Ch.3 Bayes推論の基礎

§3.1 確率推論

3.1.1 確率密度函数と確率質量函数

$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_M \end{pmatrix} \in \mathbb{R}^M$ の実数値函数 $p(\boldsymbol{x})$ が次の二つの条件をみたすとき、

$p(\boldsymbol{x})$ を確率密度函数 (probability density function : pdf) という.

(1) $p(\boldsymbol{x}) \geq 0$.

(2) $\int p(\boldsymbol{x}) d\boldsymbol{x} = 1$.

各要素が離散値の $\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_M \end{pmatrix} \in \mathbb{Z}^M$ に対して,

実数値函数 $p(\boldsymbol{x})$ が次の二つの条件をみたすとき, $p(\boldsymbol{x})$ を

確率質量函数 (probability mass function : pmf) という.

(1) $p(\boldsymbol{x}) \geq 0$.

(2) $\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$.

pdf や pmf で決められる $\boldsymbol{x}$ の分布を確率分布 (probabilistic distribution),

あるいは確率モデル (probabilistic model) という.

3.1.2 条件付き分布と周辺分布

・二つの変数 x, y に対する確率分布 $p(x, y)$ を同時分布 (joint distribution)

という. $p(y) = \int p(x, y) dx$

のようにして一方の変数を積分で除去する操作を周辺化 (marginalization)

といい, $p(y)$ を y の周辺分布 (marginal distribution) という.

・同時分布 $p(x,y)$ で y に対して特定の値が決められたときの x の確率分布を条件付き分布 (conditional distribution) といい、次で定義する：

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

・「y は x の分布 $p(x|y)$ の特性を決めるパラメータのようなもの」と解釈できる.

・$p(x|y)$ は確かに pdf になっている.

pf. $p(x|y) \geq 0$ は明らか.

$$\int p(x|y)dx = \frac{1}{p(y)}\int p(x,y)\,dx = \frac{1}{p(y)}p(y) = 1. \quad ▨$$

・同時分布が

$$p(x,y) = p(x)p(y)$$

をみたすとき，x と y は独立 (independent) という.

・ある同時分布が与えられたときに，そこから興味の対象となる条件付き分布や周辺分布を算出することを (Bayes) 推論という.

3.1.3 期待値

・分布 $p(x)$ に対して，函数 $f(x)$ の期待値 (expectation) は

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\,dx$$

と計算される.

$\mathbb{E}_p[f(x)]$ とか $\mathbb{E}[f(x)]$ とか書くこともある.

・二つの確率分布 $p(x)$ と $q(x)$ に対して，q の p に対する

Kullback-Leiblerダイバージェンス (Kullback-Leibler divergence) を

$$D_{KL}[q(x) \| p(x)] := \int q(x) \log \frac{q(x)}{p(x)} dx$$
$$= \mathbb{E}_q[\log q(x)] - \mathbb{E}_q[\log p(x)]$$

で定義する.

「x が p よりも q から出たと考えられる度合い」. より正確には「帰無仮説 $p(x)$ を棄却できるかどうかを調べるための統計量」

◦ $\dfrac{q(x)}{p(x)}$: 尤度比 (likelihood ratio).

KL-divergence は q を真の分布とみたときの対数尤度比 $\dfrac{q(x)}{p(x)}$ の平均.

→「KL-divergence が大きい」とは「平均的に分布 q の方が p よりも尤もらしい」ということ. p は q の分布から「かけ離れている」.

● もう少し深入りしてみる.

◦ $H(q) := -\mathbb{E}_q[\log q(x)]$ は分布 q のエントロピー (entropy).

分布 q の平均の「情報量」であり，q に従う事象の予測の難しさを表す.

ex) $x \in \{0,1\}$. $p(x) = \begin{cases} 0.9 & (x=1) \\ 0.1 & (x=0) \end{cases}$ ← p に従うなら だいたい $x=1$ になる.

$q(x) = 0.5$ ← $x=0$ か $x=1$ か はっきりしない.

$H(p) = -0.9 \log 0.9 - 0.1 \log 0.1 \approx 0.14$

$H(q) = -0.5 \log 0.5 - 0.5 \log 0.5 \approx 0.30$

$H(q) > H(p)$ なので q の方が予測の難しい分布!

- $H(q,p) := -\mathbb{E}_q[\log p(x)]$ は分布 p の q に対する交差エントロピー (cross entropy)

真の分布を q とみたときの負の対数尤度 $-\log p(x)$ の平均.

→ 小さいほど分布 p が尤もらしい, i.e., p が q に「似ている」.

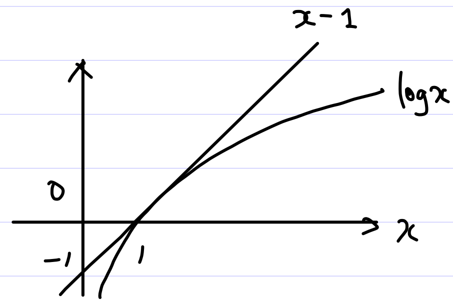
- $D_{KL}[q(x) \| p(x)] = H(q,p) - H(q)$.

KL-divergence は, 「p の q からの解離度」に「q の複雑さ」を加味した値になっている.

Th. 任意の p, q に対して $D_{KL}[q(x) \| p(x)] \geq 0$. (Gibbs の不等式)

(等号成立は $q(x) = p(x), \forall x$ のとき)

pf. 以下の不等式を利用する.



$$\log x \leq x - 1 \qquad (x > 0)$$

(等号成立は $x = 1$ のとき.)

$$
\begin{aligned}
D_{KL}[q(x) \| p(x)] &= \int q(x) \log \frac{q(x)}{p(x)} dx \\
&= -\int q(x) \log \frac{p(x)}{q(x)} dx \\
&\geq -\int q(x) \left(\frac{p(x)}{q(x)} - 1 \right) dx \\
&= -\underbrace{\int p(x) dx}_{=1} + \underbrace{\int q(x) dx}_{=1} = 0.
\end{aligned}
$$

等号成立は $\frac{p(x)}{q(x)} = 1 \Leftrightarrow p(x) = q(x)$ のとき. ▨

cf. Jensen の不等式を使っても示せる.

・KL-divergenceは非対称的：$D_{KL}[p\|q] \neq D_{KL}[q\|p]$.
（定義式から分かる）

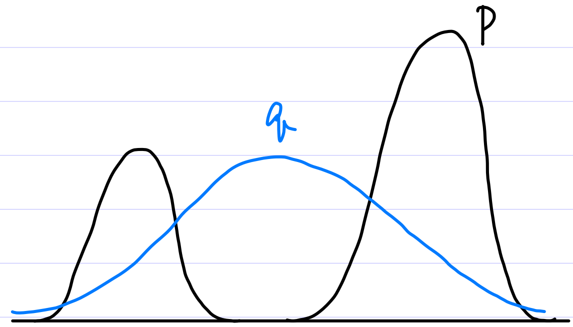
・三角不等式も満たさない.　実は「距離の2乗」のような性質をもつ.
詳しくは情報幾何の本を見ること

→ KL-divergenceは「距離」ではない.

cf. 後々に「真の分布 p に近似分布 q を近づけるために
$D_{KL}[q\|p]$ を最小化する」といった操作が出る.
上では「q を真の分布と考えて p が q からどれだけ乖離があるか」を
見ていたので違和感がある. $D_{KL}[p\|q]$ を最小化すべきでは？

・$D_{KL}[p\|q]$ を forward KL-divergence,
$D_{KL}[q\|p]$ を reverse KL-divergence という.

・forward divergence $D_{KL}[p\|q]$ では, $q(x)=0$ となる x で　「絶対連続性」
被積分関数が発散する. $D_{KL}[p\|q]$ の最小化では $q(x)=0$ となる
x に対しては $p(x)=0$ とならざるをえない（対偶をとれば, $p(x)>0$
なら必ず $q(x)>0$ となる）. p の台（正値となるところ）を覆うように q が決定.

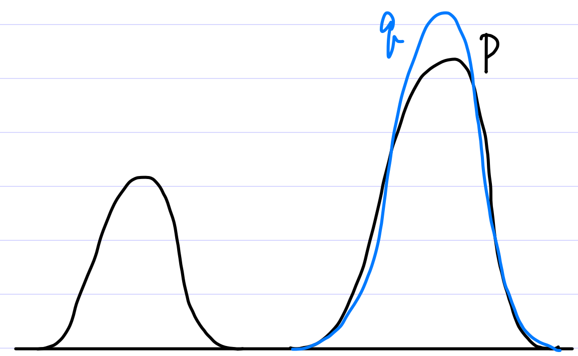


q：Gauss分布で
p を近似.

・一方，$D_{KL}[q\|p]$ では，$p(x)=0$ となる x で発散するので，

最小化すると $p(x)=0$ となる x で $q(x)=0$ となる

($q(x)>0$ なら $p(x)>0$). p の台の一部にフィットするように q が決定



q: Gauss分布で

p を近似.

・どちらを最小化するかは，やりたいことによる.

・$D_{KL}[q\|p]$ の近似計算は，q からのサンプリングが必要.

→ q は簡単な分布でないと計算ができない.

3.1.4 変数変換

・$f:\mathbb{R}^M\to\mathbb{R}^M$: bijective (逆写像 f^{-1} がある)

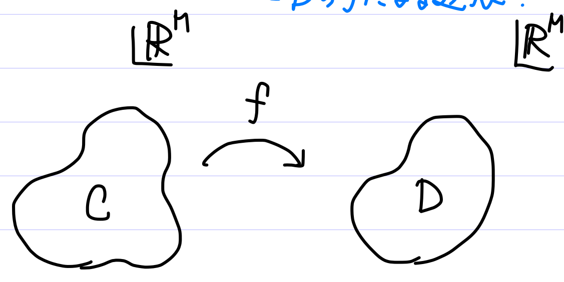
$y=f(x)$ と変数変換する.

x についての pdf を $p_x(x)$ とする. y の pdf $p_y(y)$ はどう書けるか?

$D\subseteq\mathbb{R}^M$ とする. $C:=\{x\in\mathbb{R}^M\mid f(x)\in D\}=f^{-1}(D)$ とする.

← D の f による逆像.

$$
\begin{aligned}
P(y\in D) &= P(f(x)\in D)\\
&= P(x\in f^{-1}(D))\\
&= P(x\in C)
\end{aligned}
$$



$$P(y \in D) = \int_D p_y(y)\,dy.$$

$$P(x \in C) = \int_C p_x(x)\,dx$$

重積分の変数変換.

$$= \int_D p_x(f^{-1}(y)) \left| \det \frac{\partial x}{\partial y} \right| dy. \quad -(*)$$

ここで, $\frac{\partial x}{\partial y}$ は f^{-1} の Jacobi行列で,

$$\frac{\partial x}{\partial y} := \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix} = \left(\frac{\partial x_i}{\partial y_j} \right)_{ij}.$$

$\det \frac{\partial x}{\partial y}$ は Jacobian と呼ばれる.

変換 f の「体積拡大率」を表す.

(*)の式より, 変数変換の公式

$$p_y(y) = p_x(f^{-1}(y)) \left| \det \frac{\partial x}{\partial y} \right|$$

を得る.

$p_y(y)\,dy = p_x(x)\,dx$

と覚えておくと思い出しやすいかも

ex) Gauss分布に従うr.v. x を tanh により変換して $y = \tanh(x)$ の分布の pdf を求める.

x についての pdf : $p_x(x) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$

$$\frac{dy}{dx} = \frac{d}{dx}\left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right) = \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} = 1 - \tanh^2(x).$$

$$= 1 - y^2. \quad (>0)$$

$$\therefore \frac{dx}{dy} = \left(\frac{dy}{dx}\right)^{-1} = \frac{1}{1-y^2}$$

$$\therefore p_y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\tanh^{-1}(y) - \mu)^2\right) \frac{1}{1-y^2}.$$

$$y = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \iff (e^x + e^{-x})y = e^x - e^{-x}$$
$$\iff (e^{2x} + 1)y = e^{2x} - 1$$
$$\iff e^{2x}(y-1) = -y-1$$
$$\iff e^{2x} = \frac{1+y}{1-y}$$
$$\iff 2x = \log\frac{1+y}{1-y}$$
$$\iff x = \frac{1}{2}\log\frac{1+y}{1-y}.$$

$$\therefore \tanh^{-1}(y) = \frac{1}{2}\log\frac{1+y}{1-y}.$$

$$\therefore p_y(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(\frac{1}{2}\log\frac{1+y}{1-y} - \mu\right)^2\right)\frac{1}{1-y^2}.$$

3.1.5 グラフィカルモデル.

- グラフィカルモデル(graphical model):

確率モデルに存在する複数の変数の関係性を有向グラフで表したもの.

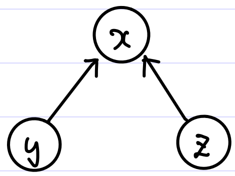
DAG(directed acyclic graph:非閉路的有向グラフ)による表現を考える.

- 頂点は変数に対応. 辺は変数間の依存関係を表す.
- ある変数 x が与えられると y の分布が定まる ($p(y|x)$: y は x によって条件付けられる) とき, 頂点 x から y へ向かう有向辺を張る.

ex) 3変数 x, y, z からなる確率モデル

$$p(x,y,z) = p(x|y,z)\,p(y)\,p(z)$$

→

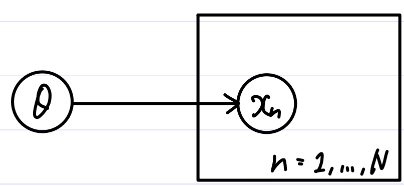


ex) パラメータ θ に依存して N 個の変数 $\mathcal{X} = \{x_1, \cdots, x_N\}$ が発生する確率モデル

$$p(\mathcal{X}, \theta) = \underbrace{p(\mathcal{X}|\theta)}_{\text{尤度関数}}\ \underbrace{p(\theta)}_{\text{事前分布}} = \left(\prod_{n=1}^{N} p(x_n|\theta)\right) p(\theta).$$

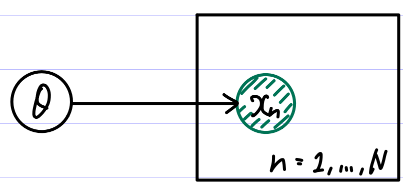
尤度関数 (likelihood function)　事前分布 (prior distribution)

→



プレート表現

$\mathcal{X}$: 観測データ　のときは上の図を



などと塗ることで観測されていることを表すことがある．

§3.2 指数型分布族

3.2.1. 確率分布の例.

- 1次元の Gauss分布.

pdf: $\mathcal{N}(x|\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{1}{2\sigma^2}(x-\mu)^2\right).$

$\mu\in\mathbb{R}$: 平均パラメータ. $\sigma^2>0$: 分散パラメータ.

・ $\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}[x] = \mu.$

pf. (ゴリ押し)

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}[x]$$

$$= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx \qquad t=\frac{x-\mu}{\sqrt{2\sigma^2}}.\quad dt=\frac{dx}{\sqrt{2\sigma^2}}.$$

$$= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}(\sqrt{2\sigma^2}\,t+\mu)e^{-t^2}dt$$

$$= \sqrt{\frac{2\sigma^2}{\pi}}\int_{-\infty}^{\infty} te^{-t^2}dt + \frac{\mu}{\sqrt{\pi}}\underbrace{\int_{-\infty}^{\infty}e^{-t^2}dt}_{\text{Gauss積分. }=\sqrt{\pi}}$$

$$= \sqrt{\frac{2\sigma^2}{\pi}}\int_{-\infty}^{\infty}\left(-\frac{1}{2}\right)\left(e^{-t^2}\right)'dt + \mu$$

$$= \sqrt{\frac{2\sigma^2}{\pi}}\cdot\left(-\frac{1}{2}\right)\underbrace{\left[e^{-t^2}\right]_{-\infty}^{\infty}}_{=0} + \mu$$

$$= \mu. \qquad \blacksquare$$

(パラメータによる微分を利用する)

$$\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}[x]$$

$$= \int_{-\infty}^{\infty} x\,\mathcal{N}(x|\mu,\sigma^2)\,dx$$

$$= \int_{-\infty}^{\infty}(\mu+x-\mu)\,\mathcal{N}(x|\mu,\sigma^2)\,dx$$

$$= \mu \underbrace{\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\,dx}_{=1\ (\because\ \text{pdfの全領域での積分})} + \int_{-\infty}^{\infty} (x-\mu)\mathcal{N}(x|\mu,\sigma^2)\,dx$$

$$= \mu + \int_{-\infty}^{\infty} (x-\mu)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx$$

$$= \mu + \sigma^2\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\left(\frac{\partial}{\partial\mu}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right)dx$$

微分と積分の順序交換.

$$= \mu + \sigma^2\frac{\partial}{\partial\mu}\underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx}_{=1\ (\because\ \text{被積分関数は}\ \mathcal{N}(x|\mu,\sigma^2))}$$

$$= \mu + \sigma^2\left(\frac{\partial}{\partial\mu}1\right)$$

$$= \mu. \qquad \blacksquare$$

・ $\mathbb{V}_{\mathcal{N}(\mu,\sigma^2)}[x] = \sigma^2.$

pf. (ゴリ押し)

$$\mathbb{V}_{\mathcal{N}(\mu,\sigma^2)}[x]$$

$$= \int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx$$

$t = \frac{x-\mu}{\sqrt{2\sigma^2}}$

$$= \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^2 e^{-t^2}dt$$

部分積分

$$= \frac{2\sigma^2}{\sqrt{\pi}}\left(\underbrace{\left[t\cdot\left(-\frac{1}{2}e^{-t^2}\right)\right]_{-\infty}^{\infty}}_{=0} + \frac{1}{2}\underbrace{\int_{-\infty}^{\infty} e^{-t^2}dt}_{=\sqrt{\pi}}\right)$$

$$= \frac{\cancel{2}\sigma^2}{\cancel{\sqrt{\pi}}}\cdot\frac{1}{\cancel{2}}\cancel{\sqrt{\pi}}$$

$$= \sigma^2. \qquad \blacksquare$$

(パラメータによる微分を用いる)

以下では 精度パラメータ $\lambda := \frac{1}{\sigma^2}$ による微分を利用する.

$$\mathcal{N}(x|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda}{2}(x-\mu)^2\right).$$

$$\mathbb{V}_{\mathcal{N}(\mu,\sigma^2)}[x]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) dx$$

$$= \sqrt{\frac{\lambda}{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) dx$$

$$= \sqrt{\frac{\lambda}{2\pi}} \cdot (-2) \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\lambda} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right)\right) dx$$

↓ 微分と積分の順序交換

$$= -2\sqrt{\frac{\lambda}{2\pi}} \frac{\partial}{\partial\lambda} \underbrace{\int_{-\infty}^{\infty} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right) dx}_{\text{Gauss積分}}$$

$$= -2\sqrt{\frac{\lambda}{\cancel{2\pi}}} \frac{\partial}{\partial\lambda} \sqrt{\frac{\cancel{2\pi}}{\lambda}}$$

$$= \cancel{-2}\sqrt{\lambda}\left(\cancel{-\frac{1}{2}} \frac{1}{(\sqrt{\lambda})^3}\right)$$

$$= \frac{1}{\lambda}$$

$$= \sigma^2. \qquad \blacksquare$$

- M次元Gauss分布.

pdf: $\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^M \det\Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

$\mu \in \mathbb{R}^M$: 平均パラメータ.

$\Sigma \in M_M(\mathbb{R})$: 正定値 ($\Sigma > 0$ と書く) 共分散行列

· $\mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[x] = \mu$.

pf. $\mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[x]$

$$= \int_{\mathbb{R}^M} x \mathcal{N}(x \mid \mu, \Sigma) dx$$

$$= \mu \underbrace{\int_{\mathbb{R}^M} \mathcal{N}(x \mid \mu, \Sigma) dx}_{=1} + \int_{\mathbb{R}^M} (x-\mu) \mathcal{N}(x \mid \mu, \Sigma) dx$$

$$= \mu + \int_{\mathbb{R}^M} (x-\mu) \frac{1}{\sqrt{(2\pi)^M \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) dx.$$

ここで, Σ: 対称 なので Σ^{-1}: 対称. ← $\Sigma = \Sigma^T$ より $(\Sigma^{-1})^T = (\Sigma^T)^{-1} = \Sigma^{-1}$.

微分公式 $\frac{\partial}{\partial x}(x^T A x) = 2Ax$ (A: 対称) を使うと,

$$\frac{\partial}{\partial \mu} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

合成函数の微分.

$$= \underbrace{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}_{\text{スカラー}} \frac{\partial}{\partial \mu}\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

合成函数の微分.

$$= \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)\left(-\cancel{\frac{1}{2}} \cdot \cancel{2} \Sigma^{-1}(x-\mu) \underline{\frac{\partial}{\partial \mu}(x-\mu)}\right)$$

ベクトルのベクトルでの微分.
→ Jacobi行列.

$$= \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)\left(-\Sigma^{-1}(x-\mu)(-I_M)\right)$$

$$= \Sigma^{-1}(x-\mu) \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

よって,

$$\mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[x]$$

$$= \mu + \Sigma \int_{\mathbb{R}^M} \frac{1}{\sqrt{(2\pi)^M \det \Sigma}} \left(\frac{\partial}{\partial \mu} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)\right) dx$$

$$= \mu + \Sigma \frac{\partial}{\partial \mu} \underbrace{\int_{\mathbb{R}^M} \frac{1}{\sqrt{(2\pi)^M \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) dx}_{=1 \ (\because \text{被積分函数は pdf})}$$

$$= \mu + \Sigma\left(\frac{\partial}{\partial \mu} 1\right)$$

$$= \mu + \Sigma 0$$

$$= \mu.$$

∎

・ $\mathbb{V}_{\mathcal{N}(\mu,\Sigma)}[x] = \Sigma$.

pf. まず, 以下をFactとして使う(Cholesky分解)

[Fact] $A > 0 \iff \exists! L$: 下三角行列, 対角項が正 s.t.

$$A = LL^T.$$

$\Sigma > 0$ なので, $\Sigma^{-1} > 0$. ← Σの固有値が $\lambda_i > 0$ $(i=1,\cdots,M)$ なら Σ^{-1}の固有値は $\lambda_i^{-1} > 0$ $(i=1,\cdots,M)$ になる.

$\Sigma^{-1} = LL^T$ とCholesky分解することで L をパラメータとして扱う.

$$\mathcal{N}(x|\mu, L) = \frac{\det L}{\sqrt{(2\pi)^M}} \exp\left(-\frac{1}{2}\underline{(x-\mu)^T LL^T(x-\mu)}\right).$$

スカラーなってトレースをとっても同じ.

$$= \frac{\det L}{\sqrt{(2\pi)^M}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((x-\mu)^T LL^T(x-\mu)\right)\right)$$

trの性質 $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$.

$$= \frac{\det L}{\sqrt{(2\pi)^M}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(L^T(x-\mu)(x-\mu)^T L\right)\right).$$

[Claim] $A \in M_M(\mathbb{R})$: 対称 として $\frac{\partial}{\partial L}\mathrm{tr}(L^TAL) = 2AL$.

pf. $L = (l_{ij})$, $A = (a_{ij})$ とする

$$\frac{\partial}{\partial l_{ij}}\mathrm{tr}(L^TAL) = \frac{\partial}{\partial l_{ij}}\sum_{m=1}^{M}(L^TAL)_{mm}$$

$$= \frac{\partial}{\partial l_{ij}}\sum_{m=1}^{M}\sum_{s=1}^{M}\sum_{t=1}^{M} l_{sm}a_{st}l_{tm}$$

$$= \sum_{m=1}^{M}\sum_{s=1}^{M}\sum_{t=1}^{M}\left(\frac{\partial l_{sm}}{\partial l_{ij}}a_{st}l_{tm} + l_{sm}a_{st}\frac{\partial l_{tm}}{\partial l_{ij}}\right)$$

$$= \sum_{t=1}^{M} a_{it}l_{tj} + \sum_{s=1}^{M} l_{sj}\underbrace{a_{si}}_{=a_{is}\ (\because A:\text{対称})}$$

$$= 2\sum_{k=1}^{M} a_{ik}l_{kj}$$

$$= 2(AL)_{ij}. \quad \blacksquare$$

以上の Claim を利用すると，

$$\frac{\partial}{\partial L}\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right) \quad \text{合成関数の微分}$$

$$= \exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)\frac{\partial}{\partial L}\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)$$

$$= -\frac{1}{2}\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)\frac{\partial}{\partial L}\left(\mathrm{tr}\left(L^T\underbrace{(x-\mu)(x-\mu)^T}_{\text{対称行列}}L\right)\right)$$

$$= -\frac{1}{\cancel{2}}\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)\cdot\cancel{2}(x-\mu)(x-\mu)^T L$$

$$= -(x-\mu)(x-\mu)^T\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)L.$$

よって．

$$\mathbb{V}_{\mathcal{N}(\mu,\Sigma)}[x]$$

$$= \int_{\mathbb{R}^M}(x-\mu)(x-\mu)^T\mathcal{N}(x|\mu,L)\,dx.$$

$$= \int_{\mathbb{R}^M}(x-\mu)(x-\mu)^T\frac{\det L}{\sqrt{(2\pi)^M}}\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)dx$$

$$= \frac{\det L}{\sqrt{(2\pi)^M}}\left(\int_{\mathbb{R}^M}-\frac{\partial}{\partial L}\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)dx\right)L^{-1} \quad \text{微分と積分の順序交換}$$

$$= -\frac{\det L}{\sqrt{(2\pi)^M}}\left(\frac{\partial}{\partial L}\int_{\mathbb{R}^M}\exp\left(-\frac{1}{2}(x-\mu)^T LL^T(x-\mu)\right)dx\right)L^{-1}$$

$$= -\frac{\det L}{\cancel{\sqrt{(2\pi)^M}}}\left(\frac{\partial}{\partial L}\frac{\cancel{\sqrt{(2\pi)^M}}}{\det L}\underbrace{\int_{\mathbb{R}}\mathcal{N}(x|\mu,L)\,dx}_{=1}\right)L^{-1}$$

$$= -\det L\left(\frac{\partial}{\partial L}(\det L)^{-1}\right)L^{-1}$$

$$= \det L\cdot\frac{1}{(\det L)^2}\left(\frac{\partial}{\partial L}\det L\right)L^{-1}$$

$$= (\det L)^{-1}\left(\frac{\partial}{\partial L}\det L\right)L^{-1}.$$

[Claim] $\frac{\partial}{\partial L}\det L = (\det L)(L^{-1})^T.$

pf. L_{ij} : L の第 i 行第 j 列を除いてできる小行列.

$\Delta_{ij} := (-1)^{i+j} \det L_{ij}$: (i,j)-余因子.　l_{ij} に依らない

第 i 行についての Laplace 展開: $\det L = \sum_{t=1}^{M} l_{it} \Delta_{it}$

を用いて

$$\frac{\partial}{\partial l_{ij}} \det L = \sum_{t=1}^{M} \frac{\partial l_{it}}{\partial l_{ij}} \Delta_{it} = \sum_{t=1}^{M} \delta_{ii} \delta_{tj} \Delta_{it} = \Delta_{ij}$$

L の余因子行列 $\mathrm{Cof}(L) := (\Delta_{ij})$ に対しては,

$$L^{-1} = \frac{1}{\det L} (\mathrm{Cof}(L))^T$$

が成立するので, $\frac{\partial}{\partial l_{ij}} \det L = (\det L)((L^{-1})^T)_{ij}$. ▨

$\therefore \mathbb{V}_{\mathcal{N}(\mu,\Sigma)}[x]$

$= \cancel{(\det L)^{-1}} \cancel{(\det L)} (L^{-1})^T L^{-1}$

$= (L^T)^{-1} L^{-1}$　$(AB)^{-1} = B^{-1}A^{-1}$

$= (L L^T)^{-1}$

$= (\Sigma^{-1})^{-1}$

$= \Sigma$. ▨

cf. 直接

$$\mathbb{V}_{\mathcal{N}(\mu,\Sigma)}[x] = \int_{\mathbb{R}^M} (x-\mu)(x-\mu)^T \mathcal{N}(x \mid \mu, \Sigma)\, dx$$

と計算するのもアリ. Σ^{-1}: 対称なので, $\exists U$: 直交行列 s.t.

$\Sigma^{-1} = U \operatorname{diag}(\lambda_1^{-1}, \dots, \lambda_M^{-1}) U^T$. これを用いて,

$\Sigma^{-1} = U \operatorname{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_M^{-\frac{1}{2}}) U^T \cdot U \operatorname{diag}(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_M^{-\frac{1}{2}}) U^T =: \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}}$

($\Sigma^{-\frac{1}{2}} > 0$ になっている) そこで、$\mathbf{z} := \Sigma^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})$ とおくと、

$$\mathbf{x}-\boldsymbol{\mu} = \Sigma^{\frac{1}{2}}\mathbf{z}, \quad (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T = \Sigma^{\frac{1}{2}}\mathbf{z}\mathbf{z}^T\Sigma^{\frac{1}{2}}$$

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \left(\Sigma^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})\right)^T\Sigma^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{z}^T\mathbf{z}.$$

変数変換のJacobianは　　det(AB) = det A · det B などを利用

$$\det\frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \det\left(\Sigma^{\frac{1}{2}}\right) = (\det\Sigma)^{\frac{1}{2}}.$$

$\therefore \mathbb{V}_{\mathcal{N}(\boldsymbol{\mu},\Sigma)}[\mathbf{x}]$

$$= \Sigma^{\frac{1}{2}}\left(\int_{\mathbb{R}^M} \mathbf{z}\mathbf{z}^T \frac{1}{\sqrt{(2\pi)^M \cancel{(\det\Sigma)}}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right)\cdot \cancel{(\det\Sigma)^{\frac{1}{2}}}\, d\mathbf{z}\right)\Sigma^{\frac{1}{2}}$$

$$= \frac{1}{\sqrt{(2\pi)^M}}\Sigma^{\frac{1}{2}}\left(\int_{\mathbb{R}^M} \mathbf{z}\mathbf{z}^T \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right) d\mathbf{z}\right)\Sigma^{\frac{1}{2}}.$$

$\int_{\mathbb{R}^M} \mathbf{z}\mathbf{z}^T \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right) d\mathbf{z} = \sqrt{(2\pi)^M}\, I_M$ となるので (*exercise*)

$\mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_M \end{pmatrix}$ として頑張って計算する.

$$\mathbb{V}_{\mathcal{N}(\boldsymbol{\mu},\Sigma)}[\mathbf{x}] = \Sigma.$$

- Bernoulli分布. 2値をとる変数 $x \in \{0,1\}$ を生成するための分布.

"成功確率"を表す.

pmf: $\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$.　　$\mu \in (0,1)$: パラメータ.

· $\mathbb{E}_{\mathrm{Bern}(\mu)}[x] = \mu$

pf. $\mathbb{E}_{\mathrm{Bern}(\mu)}[x] = \sum_{x=0}^{1} x\mu^x(1-\mu)^{1-x} = 1\cdot\mu^1(1-\mu)^{1-1} = \mu$. ▨

· $\mathbb{V}_{\mathrm{Bern}(\mu)}[x] = \mu(1-\mu)$.

pf. $\mathbb{E}_{\mathrm{Bern}(\mu)}[x(x-1)] = 0$.

$$\mathbb{E}_{\mathrm{Bern}(\mu)}[x(x-1)] = \mathbb{E}_{\mathrm{Bern}(\mu)}[x^2] - \mathbb{E}_{\mathrm{Bern}(\mu)}[x]$$

$$\mathbb{V}_{\mathrm{Bern}(\mu)}[x] = \mathbb{E}_{\mathrm{Bern}(\mu)}[x(x-1)] - \mathbb{E}_{\mathrm{Bern}(\mu)}[x]^2 + \mathbb{E}_{\mathrm{Bern}(\mu)}[x] = \mu(1-\mu).$$ ▨

- カテゴリ分布. Bernoulli分布をD値に拡張したもの.

$\mathbb{s} \in \{0,1\}^D$ かつ $\sum_{d=1}^{D} s_d = 1$ となる one-hot ベクトルを生成する.

pmf: $\mathrm{Cat}(\mathbb{s}|\pi) = \prod_{d=1}^{D} \pi_d^{s_d}$, $\pi = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_D \end{pmatrix} \in (0,1)^D$ かつ $\sum_{d=1}^{D} \pi_d = 1$.

・$\mathbb{E}_{\mathrm{Cat}(\pi)}[\mathbb{s}] = \pi$

pf. $\mathcal{S} := \{ e_d \mid d = 1, \dots, D \}$ ($e_d = (\delta_{di})_i$: 標準基底)

$$\mathbb{E}_{\mathrm{Cat}(\pi)}[\mathbb{s}] = \sum_{\mathbb{s} \in \mathcal{S}} \mathbb{s} \prod_{d=1}^{D} \pi_d^{s_d} = \sum_{d=1}^{D} \pi_d e_d = \pi. \qquad \blacksquare$$

・$\mathbb{V}_{\mathrm{Cat}(\pi)}[\mathbb{s}] = \mathrm{diag}(\pi) - \pi\pi^T$.

pf. $\mathbb{E}_{\mathrm{Cat}(\pi)}[\mathbb{s}\mathbb{s}^T] = \sum_{\mathbb{s} \in \mathcal{S}} \mathbb{s}\mathbb{s}^T \prod_{d=1}^{D} \pi_d^{s_d} = \sum_{d=1}^{D} \pi_d e_d e_d^T = \mathrm{diag}(\pi)$.

$$\therefore \mathbb{V}_{\mathrm{Cat}(\pi)}[\mathbb{s}] = \mathbb{E}_{\mathrm{Cat}(\pi)}[\mathbb{s}\mathbb{s}^T] - \mathbb{E}_{\mathrm{Cat}(\pi)}[\mathbb{s}]\mathbb{E}_{\mathrm{Cat}(\pi)}[\mathbb{s}]^T$$

$$= \mathrm{diag}(\pi) - \pi\pi^T. \qquad \blacksquare$$

- ガンマ分布. $\lambda > 0$ を生成する分布.

pdf: $\mathrm{Gam}(\lambda | a, b) = \dfrac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$. $a, b > 0$.

$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ $(x > 0)$: ガンマ関数

・$\Gamma(x+1) = x\Gamma(x)$. ← x : 整数のとき, $\Gamma(x+1) = x!$.
ガンマ関数は階乗の一般化

pf. $\Gamma(x+1)$

$$= \int_0^\infty t^x e^{-t} dt$$

$$= \left[-t^x e^{-t} \right]_0^\infty + x \int_0^\infty t^{x-1} e^{-t} dt$$

$$= x\Gamma(x). \qquad \blacksquare$$

- $\mathbb{E}_{\mathrm{Gam}(a,b)}[\lambda] = \dfrac{a}{b}$

pf. $\mathbb{E}_{\mathrm{Gam}(a,b)}[\lambda]$

$$= \int_0^\infty \lambda \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda$$

$$= \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^a e^{-b\lambda} d\lambda$$

$$= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+1)}{b^{a+1}} \underbrace{\int_0^\infty \mathrm{Gam}(\lambda | a+1, b) d\lambda}_{=1} \quad \Gamma(a+1) = a\Gamma(a)$$

$$= \frac{a}{b}. \qquad \blacksquare$$

- $\mathbb{V}_{\mathrm{Gam}(a,b)}[x] = \dfrac{a}{b^2}$

pf. $\mathbb{E}_{\mathrm{Gam}(a,b)}[\lambda^2]$

$$= \int_0^\infty \lambda^2 \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda$$

$$= \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1} e^{-b\lambda} d\lambda$$

$$= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+2)}{b^{a+2}}$$

$$= \frac{(a+1)a}{b^2}.$$

$$\therefore \mathbb{V}_{\mathrm{Gam}(a,b)}[x] = \mathbb{E}_{\mathrm{Gam}(a,b)}[\lambda^2] - \mathbb{E}_{\mathrm{Gam}(a,b)}[\lambda]^2 = \frac{a}{b^2}. \qquad \blacksquare$$

3.2.2 Gauss分布の計算例.

- $x \in \mathbb{R}^D$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $x_1 \in \mathbb{R}^{D_1}$, $x_2 \in \mathbb{R}^{D_2}$ $(D = D_1 + D_2)$ とする.

$p(x) = p(x_1, x_2) = \mathcal{N}(x | \mu, \Sigma)$ とする. (D次元Gauss分布)

Σ: 対称なので.

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \begin{matrix} \mu_1 \in \mathbb{R}^{D_1} \\ \mu_2 \in \mathbb{R}^{D_2} \end{matrix} \qquad \Sigma = \begin{array}{cc} \left(\begin{array}{c|c} \Sigma_{1,1} & \Sigma_{1,2} \\ \hline \Sigma_{2,1} & \Sigma_{2,2} \end{array}\right) & \begin{matrix} D_1 \\ D_2 \end{matrix} \\ \begin{matrix} D_1 & \quad D_2 \end{matrix} & \end{array} \qquad (\Sigma_{1,2} = \Sigma_{2,1}^\top)$$

精度行列 (precision matrix). $\Lambda := \Sigma^{-1}$.

$$\Lambda = \begin{array}{cc} \left(\begin{array}{c|c} \Lambda_{1,1} & \Lambda_{1,2} \\ \hline \Lambda_{2,1} & \Lambda_{2,2} \end{array}\right) & \begin{array}{c} D_1 \\ D_2 \end{array} \\ \begin{array}{cc} D_1 & D_2 \end{array} & \end{array} \qquad (\Lambda_{1,2} = \Lambda_{2,1}^{T})$$

<u>Th.</u> (ブロック行列の逆行列)

$A \in M_n(\mathbb{R}),\ B \in M_{n,m}(\mathbb{R}),\ C \in M_{m,n}(\mathbb{R})$,

$D \in M_m(\mathbb{R})$: regular.

$M := A - BD^{-1}C$ (Schur補行列) : regular とする.

$$\Rightarrow \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M^{-1} & -M^{-1}BD^{-1} \\ -D^{-1}CM^{-1} & D^{-1} + D^{-1}CM^{-1}BD \end{pmatrix}.$$

<u>pf.</u> 行列 $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ の基本変形を行う

以下, $\longrightarrow$ で基本変形を表すものとする.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad -BD^{-1}\times \quad (\text{行基本変形})$$

$$\longrightarrow \begin{pmatrix} \overset{M}{\overset{\shortparallel}{A - BD^{-1}C}} & O \\ C & D \end{pmatrix}$$

$$\times(-D^{-1}C) \quad (\text{列基本変形})$$

$$\longrightarrow \begin{pmatrix} M & O \\ O & D \end{pmatrix}.$$

これを基本変形行列を用いて書けば,

$$\begin{pmatrix} I_n & -BD^{-1} \\ O & I_m \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} I_n & O \\ -D^{-1}C & I_m \end{pmatrix} = \begin{pmatrix} M & O \\ O & D \end{pmatrix}.$$

$$\Leftrightarrow \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & BD^{-1} \\ O & I_m \end{pmatrix} \begin{pmatrix} M & O \\ O & D \end{pmatrix} \begin{pmatrix} I_n & O \\ D^{-1}C & I_m \end{pmatrix}$$

両辺逆行列をとって， $(XY)^{-1} = Y^{-1}X^{-1}$

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I_n & O \\ D^{-1}C & I_m \end{pmatrix}^{-1} \begin{pmatrix} M & O \\ O & D \end{pmatrix}^{-1} \begin{pmatrix} I_n & BD^{-1} \\ O & I_m \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} I_n & O \\ -D^{-1}C & I_m \end{pmatrix} \begin{pmatrix} M^{-1} & O \\ O & D^{-1} \end{pmatrix} \begin{pmatrix} I_n & -BD^{-1} \\ O & I_m \end{pmatrix}$$

$$= \begin{pmatrix} M^{-1} & O \\ -D^{-1}CM^{-1} & D^{-1} \end{pmatrix} \begin{pmatrix} I_n & -BD^{-1} \\ O & I_m \end{pmatrix}$$

$$= \begin{pmatrix} M^{-1} & -M^{-1}BD^{-1} \\ -D^{-1}CM^{-1} & D^{-1} + D^{-1}CM^{-1}BD^{-1} \end{pmatrix}. \quad \blacksquare$$

- Th. より，

$$\Lambda_{1,1} = (\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1},$$

$$\Lambda_{1,2} = -\Lambda_{1,1}\Sigma_{1,2}\Sigma_{2,2}^{-1},$$

$$\Lambda_{2,2} = \Sigma_{2,2}^{-1} + \Sigma_{2,2}^{-1}\Sigma_{1,2}\Lambda_{1,1}\Sigma_{1,2}\Sigma_{2,2}^{-1}.$$

この $\Lambda_{1,1}$ は，次の Sherman-Morrison-Woodbury 公式で求まる．

Th. (Sherman-Morrison-Woodbury)

$A \in M_n(\mathbb{R})$: regular, $D \in M_m(\mathbb{R})$: regular.

$B \in M_{n,m}(\mathbb{R})$, $C \in M_{m,n}(\mathbb{R})$, $D^{-1} + CA^{-1}B$: regular

$$\Rightarrow (A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1}.$$

pf. ブロック行列の基本変形を考える.

$$\overset{\text{行の交換}}{\begin{pmatrix} O & I_m \\ I_n & O \end{pmatrix}}\begin{pmatrix} A & -B \\ C & D^{-1} \end{pmatrix}\overset{\text{列の交換}}{\begin{pmatrix} O & I_n \\ I_m & O \end{pmatrix}} = \begin{pmatrix} D^{-1} & C \\ -B & A \end{pmatrix}.$$

両辺逆行列をとると,

$$\begin{pmatrix} O & I_n \\ I_m & O \end{pmatrix}^{-1}\begin{pmatrix} A & -B \\ C & D^{-1} \end{pmatrix}^{-1}\begin{pmatrix} O & I_m \\ I_n & O \end{pmatrix}^{-1} = \begin{pmatrix} D^{-1} & C \\ -B & A \end{pmatrix}^{-1}$$

$$\Leftrightarrow \begin{pmatrix} A & -B \\ C & D^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} O & I_n \\ I_m & O \end{pmatrix}\begin{pmatrix} D^{-1} & C \\ -B & A \end{pmatrix}^{-1}\begin{pmatrix} O & I_m \\ I_n & O \end{pmatrix} \quad -(*)$$

上のTh.より, $M = A + BDC$

$$\begin{pmatrix} A & -B \\ C & D^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} M^{-1} & * \\ * & * \end{pmatrix}.$$

$$\begin{pmatrix} D^{-1} & C \\ -B & A \end{pmatrix}^{-1} = \begin{pmatrix} * & * \\ * & A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1} \end{pmatrix}$$

$(*)$ の左上ブロックを比較して,

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1}. \qquad ▨$$

Cor. $P \in M_n(\mathbb{R}),\ R \in M_m(\mathbb{R}),\ P, R \succ O \quad B \in M_{m,n}(\mathbb{R})$

$$\Rightarrow (P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}.$$

pf. $P, R \succ O$ より $R^{-1} + BP^{-1}B^T \succ O$.

SMW公式より,

$$(P^{-1} + B^T R^{-1} B)^{-1} = P - PB^T(R + BPB^T)^{-1}BP$$

$$(P^{-1}+B^TR^{-1}B)^{-1}B^TR^{-1} = PB^TR^{-1} - PB^T(BPB^T+R)^{-1}BPB^TR^{-1}.$$

$$(\text{r.h.s.}) = PB^T(BPB^T+R)^{-1}\left((\cancel{BPB^T}+R)R^{-1} - \cancel{BPB^TR^{-1}}\right)$$

$$= PB^T(BPB^T+R)^{-1}I_m$$

$$= PB^T(BPB^T+R). \qquad \blacksquare$$

- Gauss分布のpdfの対数をとると,

$$\log \mathcal{N}(x|\mu,\Sigma)$$

$$= -\frac{1}{2}(M\log 2\pi + \log\det\Sigma) - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

$$= -\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu) + \log\det\Sigma + M\log 2\pi\right)$$

$$= -\frac{1}{2}\left(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu + \underbrace{\mu^T\Sigma^{-1}\mu + \log\det\Sigma + M\log 2\pi}_{=:\,C(\mu,\Sigma)\text{ とおく}}\right).$$

$$= -\frac{1}{2}\left(\underbrace{x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu}_{} + C(\mu,\Sigma)\right)$$

この部分を見ればGauss分布のパラメータが分かる.

- $p(x_1|x_2)$を求める.

$$\log p(x_1|x_2)$$

x_2: given のもとでは定数

$$= \log p(x_1,x_2) - \log p(x_2)$$

$$= -\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} \Lambda_{1,1} & \Lambda_{1,2} \\ \Lambda_{2,1} & \Lambda_{2,2} \end{pmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} + \text{Const.}$$

$$= -\frac{1}{2}\begin{pmatrix} (x_1-\mu_1)^T\Lambda_{1,1} + (x_2-\mu_2)^T\Lambda_{2,1} \\ (x_1-\mu_1)^T\Lambda_{1,2} + (x_2-\mu_2)^T\Lambda_{2,2} \end{pmatrix}^T \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} + \text{Const.}$$

x_1に関係しない項はConst.に含める.

$$= -\frac{1}{2}\left((x_1-\mu_1)^T\Lambda_{1,1}(x_1-\mu_1) + 2(x_2-\mu_2)^T\Lambda_{2,1}(x_1-\mu_1)\right) + \text{Const.}$$

$$= -\frac{1}{2}\left(x_1^T\Lambda_{1,1}x_1 - 2x_1^T\left(\Lambda_{1,1}\mu_1 - \Lambda_{1,2}(x_2-\mu_2)\right)\right) + \text{Const.}$$

$$= -\frac{1}{2}\left(x_1^T \underbrace{\Lambda_{1,1}}_{=:\Lambda_{1|2}} x_1 - 2x_1^T \Lambda_{1,1} \underbrace{\left(\mu_1 - \Lambda_{1,1}^{-1}\Lambda_{1,2}(x_2-\mu_2)\right)}_{=:\mu_{1|2}} \right) + \text{Const.}$$

$\therefore$ $\Lambda_{1|2} = \Lambda_{1,1}$, $\mu_{1|2} = \mu_1 - \Lambda_{1,1}^{-1}\Lambda_{1,2}(x_2-\mu_2)$ として

$p(x_1|x_2) = \mathcal{N}(x_1 | \mu_{1|2}, \Lambda_{1|2}^{-1})$ となる.

- $p(x_2)$ を求める.

$$\log p(x_2)$$

$$= \log p(x_1, x_2) - \log p(x_1|x_2)$$

$$= -\frac{1}{2}\begin{pmatrix} (x_1-\mu_1)^T\Lambda_{1,1} + (x_2-\mu_2)^T\Lambda_{2,1} \\ (x_1-\mu_1)^T\Lambda_{1,2} + (x_2-\mu_2)^T\Lambda_{2,2} \end{pmatrix}^T \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}$$

$$+\frac{1}{2}(x_1-\mu_{1|2})^T\Lambda_{1|2}(x_1-\mu_{1|2}) + \text{Const.}$$

x_2に関係ない項は Const. に含める.

$$= -\frac{1}{2}\left((x_2-\mu_2)^T\Lambda_{2,2}(x_2-\mu_2) + 2(x_2-\mu_2)^T\Lambda_{2,1}(x_1-\mu_1) + 2\mu_{1|2}^T\Lambda_{1|2}x_1 - \mu_{1|2}^T\Lambda_{1|2}\mu_{1|2} \right) + \text{Const.}$$

$$= -\frac{1}{2}\Big(x_2^T\Lambda_{2,2}x_2 - 2x_2^T\Lambda_{2,2}\mu_2 + 2x_2^T\Lambda_{2,1}(x_1-\mu_1)$$

$$+ 2\left(\mu_1 - \Lambda_{1,1}^{-1}\Lambda_{1,2}(x_2-\mu_2)\right)^T\Lambda_{1,1}x_1$$

$$- \left(\mu_1-\Lambda_{1,1}^{-1}\Lambda_{1,2}(x_2-\mu_2)\right)^T\Lambda_{1,1}\left(\mu_1-\Lambda_{1,1}^{-1}\Lambda_{1,2}(x_2-\mu_2)\right)\Big) + \text{Const.}$$

$$= -\frac{1}{2}\Big(x_2^T\left(\Lambda_{2,2} - \Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2}\right)x_2$$

$$- 2x_2^T\left(\Lambda_{2,2}\mu_2 \cancel{-\Lambda_{2,1}(x_1-\mu_1)} + \cancel{\Lambda_{2,1}x_1 - \Lambda_{2,1}\mu_1} - \Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{2,1}\mu_2\right)\Big) + \text{Const.}$$

$$= -\frac{1}{2}\left(x_2^T\left(\Lambda_{2,2}-\Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2}\right)x_2 - 2x_2^T\left(\Lambda_{2,2}-\Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2}\right)\mu_2 \right) + \text{Const.}$$

ここで, SMW公式を使うと,

$$(\Lambda_{2,2}-\Lambda_{2,1}\Lambda_{1,1}^{-1}\Lambda_{1,2})^{-1}$$ SMW公式

$$=\Lambda_{2,2}^{-1}+\Lambda_{2,2}^{-1}\Lambda_{2,1}(\Lambda_{1,1}-\Lambda_{1,2}\Lambda_{2,2}^{-1}\Lambda_{2,1})^{-1}\Lambda_{1,2}\Lambda_{2,2}^{-1}$$ ブロック行列 $\begin{pmatrix}\Lambda_{1,1} & \Lambda_{1,2}\\ \Lambda_{2,1} & \Lambda_{2,2}\end{pmatrix}^{-1}$ の右下のブロック

$\begin{pmatrix}\Sigma_{1,1} & \Sigma_{1,2}\\ \Sigma_{2,1} & \Sigma_{2,2}\end{pmatrix}$ ＝

$$=\Sigma_{2,2}.$$

$$\therefore p(x_2)=\mathcal{N}(x_2|\mu_2,\Sigma_{2,2}).$$

・対称性を考えて,

$$p(x_2|x_1)=\mathcal{N}(x_2|\mu_{2|1},\Sigma_{2|1}),$$

$$\mu_{2|1}=\mu_2-\Lambda_{2,2}^{-1}\Lambda_{2,1}(x_1-\mu_1),$$

$$\Sigma_{2|1}=\Lambda_{2,2}^{-1}.$$

$$p(x_1)=\mathcal{N}(x_1|\mu_1,\Sigma_{1,1}).$$

3.2.3 指数型分布族

3.2.3.1 定義.

<u>Def.</u> (指数型分布族)

指数型分布族 (exponential family) とは, 次のように pdf / pmf を表せる確率分布の族:

$$p(x|\eta)=h(x)\exp(\eta^T t(x)-a(\eta))$$

η : 自然パラメータ (natural parameter),

$h(x)$: 基底測度 (base measure)

$a(\eta)$: 対数分配函数 (log partition func.) という.

- $a(\eta)$ は $p(x|\eta)$ が確率分布になるように調整するための項.

$$a(\eta) = \log \int h(x) \exp(\eta^T t(x))\, dx.$$

x が不明でも, $t(x)$ の値が分かっていれば Θ の最尤推定をするのに十分になる.

- $t(x)$ は分布のパラメータ Θ の十分統計量 (sufficient statistic) になっている.

cf. $t(x)$ が Θ の十分統計量とは, $t(x)$ を与えたときの x の条件付き分布が

Θ に依存しないことをいう. 同値な表現では, x : given のときの

Θ の尤度函数 $L(\Theta|x)$ が,

$$L(\Theta|x) = g_\Theta(t(x)) \cdot h(x)$$

と表せることをいう. (g_Θ は t の函数で Θ を含む)

対数尤度 $l(\Theta|x) = \log g_\Theta(t(x)) + \log h(x)$ なので 最尤法では

$$\frac{\partial}{\partial\Theta} l(\Theta|x) = \frac{\partial}{\partial\Theta} g_\Theta(t(x)) = 0$$

となる Θ を求めることになり, x が分からなくても $t(x)$ さえ分かれば Θ の

最尤推定ができる.

3.2.3.2 分布の例.

- Bernoulli分布は指数型分布族に入る.

pf. $\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} = \exp(\log \mu^x (1-\mu)^{1-x})$

$= \exp(x \log \mu + (1-x)\log(1-\mu)) = \exp\left(x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right)$

$\therefore h(x) = 1,\ \eta = \log\frac{\mu}{1-\mu},\ t(x) = x,\ a(\eta) = \log(1+e^\eta).$

$\because e^\eta = \frac{\mu}{1-\mu} \Leftrightarrow \mu = \frac{e^\eta}{1+e^\eta}$

・ Poisson分布. pmf: $\mathrm{Poi}(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$.

Poisson分布は指数型分布族に入る.

pf. $\mathrm{Poi}(x|\lambda) = \frac{1}{x!} \exp(x \log \lambda - \lambda)$.

$$h(x) = \frac{1}{x!},\ \eta = \log \lambda,\ t(x) = x,\ a(\eta) = \lambda = e^{\eta}.$$

・ Gauss分布や多項分布も指数型分布族 (証明略)

3.2.3.3 対数分配関数と十分統計量の関係.

・ $\nabla_\eta a(\eta) = \mathbb{E}[\mathbf{t}(x)]$.

pf. $\nabla_\eta a(\eta)$

$$= \nabla_\eta \log \int h(x) \exp(\eta^T \mathbf{t}(x))\, dx$$

$$= \left(\int h(x) \exp(\eta^T \mathbf{t}(x))\, dx\right)^{-1} \nabla_\eta \left(\int h(x) \exp(\eta^T \mathbf{t}(x))\, dx\right)$$

$$= \exp(-a(\eta)) \int h(x)\, \nabla_\eta \exp(\eta^T \mathbf{t}(x))\, dx$$

↓ 微分と積分の順序交換

$$= \exp(-a(\eta)) \int \mathbf{t}(x) h(x) \exp(\eta^T \mathbf{t}(x))\, dx$$

$$= \int \mathbf{t}(x)\, \underbrace{h(x) \exp(\eta^T \mathbf{t}(x) - a(\eta))}_{p(x|\eta)}\, dx$$

$$= \mathbb{E}[\mathbf{t}(x)].$$ ▨

・ $\frac{\partial^2}{\partial \eta\, \partial \eta^T} a(\eta) = \mathbb{V}[\mathbf{t}(x)]$

pf. $\frac{\partial^2}{\partial \eta\, \partial \eta^T} a(\eta)$

$$= \frac{\partial}{\partial \eta^T} \underbrace{\int \mathbf{t}(x)\, p(x|\eta)\, dx}_{ベクトル値}$$

↓ 微分と積分の順序交換

$$= \int \mathbf{t}(x) \frac{\partial}{\partial \eta^T} p(x|\eta)\, dx$$

$$= \int \mathbf{t}(x) h(x) \exp(\eta^T \mathbf{t}(x) - a(\eta)) \left(\mathbf{t}(x)^T - \frac{\partial}{\partial \eta^T} a(\eta)\right) dx$$

$$= \int \mathbf{t}(x)\mathbf{t}(x)^T p(x|\eta) dx - \left(\int \mathbf{t}(x) p(x|\eta) dx\right) \mathbb{E}[\mathbf{t}(x)]^T$$

$$= \mathbb{E}[\mathbf{t}(x)\mathbf{t}(x)^T] - \mathbb{E}[\mathbf{t}(x)]\mathbb{E}[\mathbf{t}(x)]^T$$

$$= \mathbb{V}[\mathbf{t}(x)]. \quad \blacksquare$$

3.2.4 分布の共役性

Def. (共役事前分布)

指数型分布族 $p(x|\eta)$ に対して，次の形の η の事前分布 $p_\lambda(\eta)$ を $p(x|\eta)$ の共役事前分布 (conjugate prior) という：

$$p_\lambda(\eta) = h_c(\eta) \exp(\eta^T \lambda_1 - a(\eta)\lambda_2 - a_c(\lambda_1, \lambda_2)).$$

- $a_c(\lambda_1, \lambda_2)$ は $p_\lambda(\eta)$ が確率密度関数になるように正規化するための項

$$a_c(\lambda_1, \lambda_2) = \log \int h_c(\eta) \exp(\eta^T \lambda_1 - a(\eta)\lambda_2) d\eta.$$

3.2.4.1 事後分布の解析的計算

- 共役事前分布を使うと，指数型分布族の尤度関数に対して事後分布も事前分布と同じ形式になる．
- $\mathcal{X} = \{x_1, \dots, x_N\}$: Observed. 事後分布 $p_\lambda(\eta|\mathcal{X})$ は，

$$p_\lambda(\eta|\mathcal{X})$$

$$\propto p_\lambda(\eta) p(\mathcal{X}|\eta)$$

$$= p_\lambda(\eta) \prod_{n=1}^{N} p(x_n|\eta)$$

Bayesの定理 $\underbrace{p_\lambda(\eta|\mathcal{X})}_{\text{事後分布}} = \dfrac{\overbrace{p_\lambda(\eta)}^{\text{事前分布}} \overbrace{p(\mathcal{X}|\eta)}^{\text{尤度}}}{p(\mathcal{X})}$

$p(\mathcal{X})$ ← $\mathcal{X}$: givenなら定数．周辺尤度

ηによらない

$$= h_c(\eta)\exp\left(\eta^T\lambda_1 - a(\eta)\lambda_2 - a_c(\lambda_1,\lambda_2)\right)\prod_{n=1}^{N} h(x_n)\exp\left(\eta^T t(x_n) - a(\eta)\right)$$

$$\propto h_c(\eta)\exp\left(\eta^T\left(\lambda_1 + \sum_{n=1}^{N} t(x_n)\right) - a(\eta)(\lambda_2 + N)\right).$$

∴ 事後分布は事前分布と同じ形式. パラメータは

$$\hat{\lambda}_1 = \lambda_1 + \sum_{n=1}^{N} t(x_n), \quad \hat{\lambda}_2 = \lambda_2 + N$$

と更新される. ← 事後分布が解析的に求まる!

3.2.4.2 予測分布の解析的計算

・データ $\mathcal{X}$ が観測された後の未観測データ x_* の予測分布も解析的に求まる.

$$p(x_*|\mathcal{X}) = \mathbb{E}_{p(\eta|\mathcal{X})}\left[p(x_*|\eta)\right]$$

$$= \int p(x_*|\eta)\,p(\eta|\mathcal{X})\,d\eta$$

$$= \int h(x_*)\exp\left(\eta^T t(x_*) - a(\eta)\right)h_c(\eta)\exp\left(\eta^T\hat{\lambda}_1 - a(\eta)\hat{\lambda}_2 - a_c(\hat{\lambda}_1,\hat{\lambda}_2)\right)d\eta.$$

$$= h(x_*)\exp\left(-a_c(\hat{\lambda}_1,\hat{\lambda}_2)\right)\int \underbrace{h_c(\eta)\exp\left(\eta^T\left(\hat{\lambda}_1 + t(x_*)\right) - a(\eta)(\hat{\lambda}_2 + 1)\right)}_{\exp\left(a_c(\hat{\lambda}_1 + t(x_*),\, \hat{\lambda}_2 + 1)\right)} d\eta$$

$$= h(x_*)\exp\left(a_c(\hat{\lambda}_1 + t(x_*), \hat{\lambda}_2 + 1) - a_c(\hat{\lambda}_1,\hat{\lambda}_2)\right).$$

これは、一般には指数型分布族にならない.

3.2.4.3 Bernoulli分布のパラメータの推論

・Bernoulli分布の共役事前分布はベータ分布:

$$\text{pdf}: \mathrm{Beta}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}, \quad \alpha,\beta > 0$$

$$\mathrm{Bern}(x|\mu) = \exp\left(\underbrace{x}_{t(x)}\underbrace{\log\frac{\mu}{1-\mu}}_{\eta} - \underbrace{\log\left(1+e^{\eta}\right)}_{a(\eta)}\right).$$

自然パラメータ $\eta = \log\frac{\mu}{1-\mu}$ を用いてベータ分布を表現する.

$$e^{\eta} = \frac{\mu}{1-\mu} \Leftrightarrow (1-\mu)e^{\eta} = \mu \Leftrightarrow \mu = \frac{e^{\eta}}{1+e^{\eta}}.$$

pdf の変数変換の式より η についての pdf $\mathrm{Beta}_{\eta}(\eta|\lambda_1,\lambda_2)$ は

$$
\begin{aligned}
&\mathrm{Beta}_{\eta}(\eta|\lambda_1,\lambda_2) \quad \big\downarrow \text{変数変換} \\
&= \mathrm{Beta}(\mu|\alpha,\beta)\left|\frac{d\mu}{d\eta}\right| \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}\left|\frac{e^{\eta}(1+\cancel{e^{\eta}}) - \cancel{e^{\eta}\cdot e^{\eta}}}{(1+e^{\eta})^2}\right| \\
&= \exp\left((\alpha-1)\log\mu + (\beta-1)\log(1-\mu) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \eta - 2\underbrace{\log(1+e^{\eta})}_{a(\eta)}\right) \\
&= \exp\left((\alpha-1)(\eta - \underbrace{\log(1+e^{\eta})}_{a(\eta)}) + (\beta-1)(-\underbrace{\log(1+e^{\eta})}_{a(\eta)}) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \eta - 2a(\eta)\right) \\
&= \exp\left(\eta\alpha - a(\eta)(\alpha+\beta) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right).
\end{aligned}
$$

$$\therefore h_c(\eta) = 1,\ \lambda_1 = \alpha,\ \lambda_2 = \alpha+\beta,\ a_c(\lambda_1,\lambda_2) = \log\frac{\Gamma(\lambda_1)\Gamma(\lambda_2-\lambda_1)}{\Gamma(\lambda_2)}.$$

確かに共役事前分布!

- $\mathcal{X} = \{x_1,\ldots,x_N\}$ $(x_n \in \{0,1\})$ が観測されると,

$$\hat{\lambda}_1 = \lambda_1 + \sum_{n=1}^{N} t(x_n) = \alpha + \sum_{n=1}^{N} x_n =: \hat{\alpha}$$

$$\hat{\lambda}_2 = \lambda_2 + N = \alpha+\beta+N =: \hat{\alpha}+\hat{\beta}$$

と更新される. $\hat{\beta} = \alpha+\beta+N-\hat{\alpha} = \beta+N-\sum_{n=1}^{N} x_n.$

$\therefore$ 事後分布: $p(\mu|\mathcal{X}) = \mathrm{Beta}(\mu|\hat{\alpha},\hat{\beta}).$

予測分布は,

$$
\begin{aligned}
p(x_*|\mathcal{X}) &= \exp\left(a_c(\hat{\lambda}_1 + t(x_*), \hat{\lambda}_2+1) - a_c(\hat{\lambda}_1,\hat{\lambda}_2)\right) \\
&= \exp\left(\log\frac{\Gamma(\hat{\lambda}_1+x_*)\Gamma(\hat{\lambda}_2+1-\hat{\lambda}_1-x_*)}{\Gamma(\hat{\lambda}_2+1)}\frac{\Gamma(\hat{\lambda}_2)}{\Gamma(\hat{\lambda}_1)\Gamma(\hat{\lambda}_2-\hat{\lambda}_1)}\right) \\
&= \frac{\Gamma(\hat{\alpha}+x_*)\Gamma(\hat{\beta}+1-x_*)}{\hat{\lambda}_2\cancel{\Gamma(\hat{\lambda}_2)}}\frac{\cancel{\Gamma(\hat{\lambda}_2)}}{\Gamma(\hat{\alpha})\Gamma(\hat{\beta})}
\end{aligned}
$$

$$= \frac{1}{\hat{\alpha}+\hat{\beta}} \frac{\Gamma(\hat{\alpha}+x_*)\Gamma(\hat{\beta}+1-x_*)}{\Gamma(\hat{\alpha})\Gamma(\hat{\beta})}.$$

$\Gamma(\hat{\alpha}+1) = \hat{\alpha}\Gamma(\hat{\alpha}).$

$$p(1|\mathcal{X}) = \frac{1}{\hat{\alpha}+\hat{\beta}} \frac{\Gamma(\hat{\alpha}+1)\cancel{\Gamma(\hat{\beta}+1-1)}}{\Gamma(\hat{\alpha})\cancel{\Gamma(\hat{\beta})}} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}},$$

$$p(0|\mathcal{X}) = \frac{1}{\hat{\alpha}+\hat{\beta}} \frac{\cancel{\Gamma(\hat{\alpha})}\Gamma(\hat{\beta}+1)}{\cancel{\Gamma(\hat{\alpha})}\Gamma(\hat{\beta})} = \frac{\hat{\beta}}{\hat{\alpha}+\hat{\beta}}.$$

$$\therefore p(x_*|\mathcal{X}) = \mathrm{Bern}\left(x_* \middle| \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}\right).$$

cf. ベータ分布も指数型分布族.

$$\mathrm{Beta}(\mu|\alpha,\beta) = \exp\left((\alpha-1)\log\mu + (\beta-1)\log(1-\mu) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)$$

$$= \exp\left(\underbrace{\begin{pmatrix}\alpha-1\\ \beta-1\end{pmatrix}^T}_{\eta}\underbrace{\begin{pmatrix}\log\mu\\ \log(1-\mu)\end{pmatrix}}_{t(\mu)} - \underbrace{\log\frac{\Gamma(\alpha-1+1)\Gamma(\beta-1+1)}{\Gamma(\alpha-1+\beta-1+2)}}_{a(\eta)}\right).$$

3.2.4.4 Gauss分布の精度パラメータの推論

・1次元Gauss分布 $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$ の

平均パラメータ μ を固定したとき, 精度パラメータ $\gamma := \sigma^{-2}$ の共役事前分布は

ガンマ分布になる.

$$\mathcal{N}(x|\mu,\gamma) = \sqrt{\frac{\gamma}{2\pi}}\exp\left(-\frac{\gamma}{2}(x-\mu)^2\right)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)}\exp\left(\underbrace{\gamma}_{\eta}\underbrace{\left(-\frac{1}{2}(x-\mu)^2\right)}_{t(x)} - \underbrace{\left(-\frac{1}{2}\log\gamma\right)}_{a(\eta)}\right)$$

$C_G(a,b) := \frac{b^a}{\Gamma(a)}$ (ガンマ分布の正規化定数) として.

$$\mathrm{Gam}(\eta|a,b) = C_G(a,b)\gamma^{a-1}e^{-b\gamma}$$

$$= \exp\left((a-1)\log\gamma - b\gamma + \log C_G(a,b)\right)$$

$$= \exp\left(\eta(-b) - \underbrace{\left(-\frac{1}{2}\log\gamma\right)}_{a(\eta)}(2(a-1)) - (-\log C_G(a,b))\right)$$

$$= \exp\Big(\eta\underbrace{(-b)}_{\lambda_1} - a(\eta)\underbrace{(2(a-1))}_{\lambda_2} - \underbrace{(-\log C_G(a,b))}_{a_c(\lambda_1,\lambda_2)}\Big)$$

$\therefore h_c(\eta) = 1,\ \lambda_1 = -b,\ \lambda_2 = 2(a-1),$

$a_c(\lambda_1, \lambda_2) = -\log C_G\left(\frac{\lambda_2}{2}+1, -\lambda_1\right).$

・データ $\mathcal{X} = \{x_1, \ldots, x_N\}$ 観測後の事後分布のパラメータは,

$$\hat{\lambda}_1 := -b + \sum_{n=1}^{N} t(x_n) = -\left(b + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right) =: -\hat{b}.$$

$$\hat{\lambda}_2 := \lambda_2 + N =: 2(\hat{a}-1). \iff \hat{a} = a + \frac{N}{2}.$$

$\therefore$ 事後分布は $p(\gamma|\mathcal{X}) = \mathrm{Gam}(\gamma|\hat{a}, \hat{b}).$

予測分布は,

$$p(x_*|\mathcal{X}) = \frac{1}{\sqrt{2\pi}}\exp\left(a_c(\hat{\lambda}_1 + t(x_*), \hat{\lambda}_2+1) - a_c(\hat{\lambda}_1, \hat{\lambda}_2)\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\log C_G\left(\frac{\hat{\lambda}_2+1}{2}+1, -(\hat{\lambda}_1+t(x_*))\right) + \log C_G\left(\frac{\hat{\lambda}_2}{2}+1, -\hat{\lambda}_1\right)\right)$$

$$= \frac{1}{\sqrt{2\pi}}\frac{C_G\left(\frac{\hat{\lambda}_2}{2}+1, -\hat{\lambda}_1\right)}{C_G\left(\frac{\hat{\lambda}_2+1}{2}+1, -(\hat{\lambda}_1+t(x_*))\right)}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{C_G(\hat{a}, \hat{b})}{C_G(\hat{a}+\frac{1}{2}, \hat{b}-t(x_*))}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{\Gamma(\hat{a}+\frac{1}{2})}{\Gamma(\hat{a})}\frac{\hat{b}^{\hat{a}}}{(\hat{b}-t(x_*))^{\hat{a}+\frac{1}{2}}}$$

$$= \frac{1}{\sqrt{2\pi}}\frac{\Gamma(\hat{a}+\frac{1}{2})}{\Gamma(\hat{a})}\frac{\cancel{\hat{b}^{\hat{a}}}}{\hat{b}^{\cancel{\hat{a}}+\frac{1}{2}}\left(1-\frac{1}{\hat{b}}t(x_*)\right)^{\hat{a}+\frac{1}{2}}}$$

$$= \frac{1}{\sqrt{2\pi\hat{b}}}\frac{\Gamma(\hat{a}+\frac{1}{2})}{\Gamma(\hat{a})}\left(1+\frac{1}{2\hat{b}}(x_*-\mu)^2\right)^{-\frac{2\hat{a}+1}{2}}$$

$$= \sqrt{\frac{\hat{a}/\hat{b}}{\pi\cdot 2\hat{a}}}\frac{\Gamma(\frac{2\hat{a}+1}{2})}{\Gamma(\frac{2\hat{a}}{2})}\left(1+\frac{\hat{a}/\hat{b}}{2\hat{a}}(x_*-\mu)^2\right)^{-\frac{2\hat{a}+1}{2}}$$

$= \mathrm{St}(x_* | \mu_s, \lambda_s, \nu_s)$: Studentのt分布

$\mu_s := \mu, \ \lambda_s := \frac{\hat{a}}{\hat{b}}, \ \nu_s := 2\hat{a}.$

Studentのt分布のpdf : $\mathrm{St}(x_* | \mu_s, \lambda_s, \nu_s) = \sqrt{\frac{\lambda_s}{\pi \nu_s}} \frac{\Gamma(\frac{\nu_s + 1}{2})}{\Gamma(\frac{\nu_s}{2})} \left(1 + \frac{\lambda_s}{\nu_s}(x_* - \mu_s)^2\right)^{-\frac{\nu_s + 1}{2}}.$

- $\mathbb{E}_{\mathrm{St}(\mu_s, \lambda_s, \nu_s)}[x] = \mu_s. \quad (\nu_s > 1)$
- $\mathbb{V}_{\mathrm{St}(\mu_s, \lambda_s, \nu_s)}[x] = \frac{\nu_s}{\lambda_s(\nu_s - 2)} \quad (\nu_s > 2).$

証明はけっこう大変

§3.3 Bayes線形回帰.

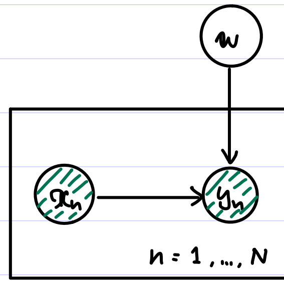
3.3.1 モデル

データ : $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ 連続値

$\mathcal{X} = \{x_1, \dots, x_N\}$, $\mathcal{Y} = \{y_1, \dots, y_N\}$.

$\mathcal{X}$ から $\mathcal{Y}$ を予測する.

Bayes線形回帰モデルの同時分布を次で定義



$$
\begin{aligned}
p(\mathcal{Y}, w \mid \mathcal{X}) &= p(w) p(\mathcal{Y} \mid \mathcal{X}, w) \\
&= p(w) \prod_{n=1}^{N} p(y_n \mid x_n, w).
\end{aligned}
$$

y_n は分散 σ_y^2 の Gauss分布に従うとする.

$$
p(y_n \mid x_n, w) = \mathcal{N}(y_n \mid w^T \phi(x_n), \sigma_y^2),
$$

$\phi : \mathbb{R}^{H_0} \to \mathbb{R}^{H_1}$: 特徴量函数.

w を学習する. Gauss事前分布を与えておく.

$$
p(w) = \mathcal{N}(w \mid 0, \sigma_w^2 I_{H_1})
$$

3.3.2 学習と予測.

3.3.2.1 事後分布の解析的計算.

事後分布は.

$$
p(w \mid \mathcal{Y}, \mathcal{X}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, w) p(w)}{p(\mathcal{Y} \mid \mathcal{X})} \propto p(\mathcal{Y} \mid \mathcal{X}, w) p(w)
$$

wに依らない

対数をとって w について整理していく.

$$\log p(\boldsymbol{w}|\mathcal{Y},\mathcal{X})$$

$$= \log p(\mathcal{Y}|\mathcal{X},\boldsymbol{w}) + \log p(\boldsymbol{w}) + \mathrm{const.}$$

$$= \log \prod_{n=1}^{N} p(y_n|x_n,\boldsymbol{w}) + \log \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\sigma_w^2 I_{H_1}) + \mathrm{const.}$$

$$= \sum_{n=1}^{N} \log \mathcal{N}(y_n|\boldsymbol{w}^T\boldsymbol{\phi}(x_n),\sigma_y^2) + \log \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\sigma_w^2 I_{H_1}) + \mathrm{const.}$$

$$= \sum_{n=1}^{N}\left(-\frac{1}{2}\left(\underbrace{\log 2\pi + \log\sigma_y^2}_{\boldsymbol{w}\text{に依らない}} + \sigma_y^{-2}\left(y_n - \boldsymbol{w}^T\boldsymbol{\phi}(x_n)\right)^2\right)\right)$$

$$-\frac{1}{2}\left(\underbrace{H_1\log 2\pi + \log\det(\sigma_w^2 I_{H_1})}_{\boldsymbol{w}\text{に依らない}} + \boldsymbol{w}^T(\sigma_w^2 I_{H_1})^{-1}\boldsymbol{w}\right) + \mathrm{const.}$$

$$= -\frac{1}{2}\sigma_y^{-2}\sum_{n=1}^{N}\left(\underbrace{y_n^2}_{\boldsymbol{w}\text{に依らない}} - 2y_n\boldsymbol{w}^T\boldsymbol{\phi}(x_n) + \boldsymbol{w}^T\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T\boldsymbol{w}\right) - \frac{1}{2}\sigma_w^{-2}\boldsymbol{w}^T I_{H_1}\boldsymbol{w} + \mathrm{const.}$$

$$= -\frac{1}{2}\left(\boldsymbol{w}^T\underbrace{\left(\sigma_y^{-2}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T + \sigma_w^{-2}I_{H_1}\right)}_{=:\hat{\Sigma}^{-1}}\boldsymbol{w} - 2\boldsymbol{w}^T\underbrace{\left(\sigma_y^{-2}\sum_{n=1}^{N}y_n\boldsymbol{\phi}(x_n)\right)}_{=:\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}}\right) + \mathrm{const.}$$

これはGauss分布の対数表示の形と一致.

→ 事後分布もGauss分布で $p(\boldsymbol{w}|\mathcal{Y},\mathcal{X}) = \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{\mu}},\hat{\Sigma})$.

$$\hat{\Sigma}^{-1} = \sigma_y^{-2}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T + \sigma_w^{-2}I_{H_1}, \quad \hat{\boldsymbol{\mu}} = \hat{\Sigma}\left(\sigma_y^{-2}\sum_{n=1}^{N}y_n\boldsymbol{\phi}(x_n)\right).$$

3.3.2.2 予測分布の解析的計算.

入力値 x_* : given のときの予測値 y_* の分布 $p(y_*|x_*,\mathcal{Y},\mathcal{X})$ を計算.

まず, Bayesの定理より

$$p(\boldsymbol{w}|y_*,x_*,\mathcal{Y},\mathcal{X}) = \frac{p(y_*|x_*,\mathcal{Y},\mathcal{X},\boldsymbol{w})\,p(\boldsymbol{w}|\mathcal{Y},\mathcal{X})}{p(y_*|x_*,\mathcal{Y},\mathcal{X})}$$

$$\therefore \log p(y_*|x_*,\mathcal{Y},\mathcal{X})$$

$$= \log p(y_*|x_*,\mathcal{Y},\mathcal{X},\boldsymbol{w}) - \log p(\boldsymbol{w}|y_*,x_*,\mathcal{Y},\mathcal{X}) + \overset{y_*\text{に関係ない項}}{\mathrm{const.}}$$

・ $\log p(w \mid y_*, x_*, \mathcal{Y}, \mathcal{X})$ については,

データ $\{(x_1, y_1), \ldots, (x_N, y_N), (x_*, y_*)\}$: given とみなして

事後分布を計算したものを考えればよい.

$$\Sigma(x_*)^{-1} := \hat{\Sigma}^{-1} + \sigma_y^{-2} \phi(x_*) \phi(x_*)^T$$

$$\mu(y_*) := \Sigma(x_*) \left(\hat{\Sigma}^{-1} \hat{\mu} + \sigma_y^{-2} y_* \phi(x_*) \right)$$

とすれば,

$$\log p(w \mid y_*, x_*, \mathcal{Y}, \mathcal{X})$$

$$= \log \mathcal{N}(w \mid \mu(y_*), \Sigma(x_*))$$

$$= -\frac{1}{2}\Big(\underbrace{H_1 \log 2\pi + \log \det \Sigma(x_*)}_{y_* \text{に依らない}} + (w - \mu(y_*))^T \Sigma(x_*)^{-1} (w - \mu(y_*)) \Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\underbrace{w^T \Sigma(x_*)^{-1} w}_{y_* \text{に依らない}} - 2 w^T \Sigma(x_*)^{-1} \mu(y_*) + \mu(y_*)^T \Sigma(x_*)^{-1} \mu(y_*) \Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\left(\hat{\mu}^T \hat{\Sigma}^{-1} + \sigma_y^{-2} y_* \phi(x_*)^T \right) \cancel{\Sigma(x_*) \Sigma(x_*)^{-1}} \Sigma(x_*) \left(\hat{\Sigma}^{-1} \hat{\mu} + \sigma_y^{-2} y_* \phi(x_*) \right)$$

$$- 2 w^T \cancel{\Sigma(x_*)^{-1} \Sigma(x_*)} \left(\hat{\Sigma}^{-1} \hat{\mu} + \sigma_y^{-2} y_* \phi(x_*) \right) \Big) + \text{const.}$$

$$= -\frac{1}{2}\Big(\big(\underbrace{\hat{\mu}^T \hat{\Sigma}^{-1} \Sigma(x_*) \hat{\Sigma}^{-1} \hat{\mu}}_{y_* \text{に依らない}} + 2 y_* \sigma_y^{-2} \phi(x_*)^T \Sigma(x_*) \hat{\Sigma}^{-1} \hat{\mu}$$

$$+ y_*^2 \sigma_y^{-4} \phi(x_*)^T \Sigma(x_*) \phi(x_*) - 2 \big(\underbrace{w^T \hat{\Sigma}^{-1} \hat{\mu}}_{y_* \text{に依らない}} + \sigma_y^{-2} y_* w^T \phi(x_*) \big) \Big) + \text{const.}$$

$$= -\frac{1}{2}\left(\sigma_y^{-4} \phi(x_*)^T \Sigma(x_*) \phi(x_*) \, y_*^2 - 2 \sigma_y^{-2} \phi(x_*)^T \left(w - \Sigma(x_*) \hat{\Sigma}^{-1} \hat{\mu} \right) y_* \right) + \text{const.}$$

・ $\log p(y_* \mid x_*, \mathcal{Y}, \mathcal{X}, w)$

↓ y_* は $\mathcal{Y}, \mathcal{X}$ に依らない.

$$= \log p(y_* \mid x_*, w)$$

$$= \log \mathcal{N}(y_* \mid w^T \phi(x_*), \sigma_y^2)$$

$$= -\frac{1}{2}\Big(\underbrace{\log 2\pi + \log \sigma_y^2}_{y_*\text{に依らない}} + \sigma_y^{-2}\left(y_* - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_*)\right)^2\Big)$$

$$= -\frac{1}{2}\sigma_y^{-2}\Big(y_*^2 - 2\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{w}\,y_* + \underbrace{\left(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_*)\right)^2}_{y_*\text{に依らない}}\Big) + \text{const.}$$

$$= -\frac{1}{2}\sigma_y^{-2}\left(y_*^2 - 2\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{w}\,y_*\right) + \text{const.}$$

・以上より、

$$\log p(y_* \mid \boldsymbol{x}_*, \mathcal{Y}, \mathcal{X})$$

$$= -\frac{1}{2}\sigma_y^{-2}\left(y_*^2 - \cancel{2\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{w}\,y_*}\right)$$

$$+\frac{1}{2}\left(\sigma_y^{-4}\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}(\boldsymbol{x}_*)\boldsymbol{\phi}(\boldsymbol{x}_*)y_*^2 - 2\sigma_y^{-2}\boldsymbol{\phi}(\boldsymbol{x}_*)^T\left(\cancel{\boldsymbol{w}} - \boldsymbol{\Sigma}(\boldsymbol{x}_*)\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\right)y_*\right) + \text{const.}$$

$$= -\frac{1}{2}\left(\left(\sigma_y^{-2} - \sigma_y^{-4}\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}(\boldsymbol{x}_*)\boldsymbol{\phi}(\boldsymbol{x}_*)\right)y_*^2 - 2\sigma_y^{-2}\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}(\boldsymbol{x}_*)\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\,y_*\right) + \text{const.}$$

ところで、

$$\log \mathcal{N}(x \mid \mu, \sigma^2)$$

$$= -\frac{1}{2}\left(\log 2\pi + \log \sigma^2 + \sigma^{-2}(x-\mu)^2\right)$$

xに関係ないところはconstにまとめる

$$= -\frac{1}{2}\left(\sigma^{-2}x^2 - 2\sigma^{-2}\mu x\right) + \text{const.}$$

なので、

$$\left(\sigma^2(\boldsymbol{x}_*)\right)^{-1} := \sigma_y^{-2} - \sigma_y^{-4}\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}(\boldsymbol{x}_*)\boldsymbol{\phi}(\boldsymbol{x}_*),$$

$$\mu(\boldsymbol{x}_*) := \sigma^2(\boldsymbol{x}_*)\sigma_y^{-2}\boldsymbol{\phi}(\boldsymbol{x}_*)^T\boldsymbol{\Sigma}(\boldsymbol{x}_*)\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}$$

とおくと、

$$p(y_* \mid \boldsymbol{x}_*, \mathcal{Y}, \mathcal{X}) = \mathcal{N}\left(y_* \mid \mu(\boldsymbol{x}_*), \sigma^2(\boldsymbol{x}_*)\right)$$

と分かる

・もう少し計算をすすめる．

$$\Sigma(x_*)$$

$$= \left(\hat{\Sigma}^{-1} + \sigma_y^{-2}\phi(x_*)\phi(x_*)^T\right)^{-1} \quad \text{Sherman-Morrison-Woodbury}$$

$$= \hat{\Sigma} - \hat{\Sigma}\phi(x_*)\underbrace{\left((\sigma_y^{-2} I_1)^{-1} + \phi(x_*)^T \hat{\Sigma}\phi(x_*)\right)^{-1}}_{\text{スカラー値}}\phi(x_*)^T\hat{\Sigma}$$

$$= \hat{\Sigma} - \left(\sigma_y^2 + \phi(x_*)^T\hat{\Sigma}\phi(x_*)\right)^{-1}\hat{\Sigma}\phi(x_*)\phi(x_*)^T\hat{\Sigma}.$$

ゆえに， $q = \phi(x_*)^T\hat{\Sigma}\phi(x_*)$ とおくと，

$$(\sigma^2(x_*))^{-1}$$

$$= \sigma_y^{-2} - \sigma_y^{-4}\left(\phi(x_*)^T\hat{\Sigma}\phi(x_*) - \left(\sigma_y^2 + \phi(x_*)^T\hat{\Sigma}\phi(x_*)\right)^{-1}\left(\phi(x_*)^T\hat{\Sigma}\phi(x_*)\right)^2\right)$$

$$= \sigma_y^{-2} - \sigma_y^{-4}\left(q - (\sigma_y^2 + q)^{-1} q^2\right)$$

$$= \sigma_y^{-2} - \sigma_y^{-4}(\sigma_y^2 + q)^{-1} q\left(\sigma_y^2 \cancel{+ q - q}\right)$$

$$= \sigma_y^{-2} - \frac{\sigma_y^{-2} q}{\sigma_y^2 + q}$$

$$= \frac{1 + \cancel{\sigma_y^{-2} q - \sigma_y^{-2} q}}{\sigma_y^2 + q}$$

$$= \left(\sigma_y^2 + \phi(x_*)^T\hat{\Sigma}\phi(x_*)\right)^{-1}$$

$$\therefore \ \sigma^2(x_*) = \sigma_y^2 + \phi(x_*)^T\hat{\Sigma}\phi(x_*).$$

更に，

$$\mu(x_*)$$

$$= \sigma^2(x_*)\,\sigma_y^{-2}\phi(x_*)^T\left(\hat{\Sigma} - (\sigma_y^2 + q)^{-1}\hat{\Sigma}\phi(x_*)\phi(x_*)^T\hat{\Sigma}\right)\hat{\Sigma}^{-1}\hat{\mu}$$

$$= \sigma^2(x_*)\,\sigma_y^{-2}\phi(x_*)^T\left(\hat{\mu} - \sigma^2(x_*)^{-1}\hat{\Sigma}\phi(x_*)\phi(x_*)^T\hat{\mu}\right)$$

$$= \sigma^2(x_*)\sigma_y^{-2}\phi(x_*)^T\hat{\mu} - \sigma_y^{-2} q\, \phi(x_*)^T\hat{\mu}$$

$$= \left((\sigma_y^2 + \cancel{q})\sigma_y^{-2} - \cancel{\sigma_y^{-2} q}\right)\phi(x_*)^T\hat{\mu}$$

$$= \hat{\mu}^T\phi(x_*).$$

結局,

$$p(y_* | x_*, \mathcal{Y}, \mathcal{X}) = \mathcal{N}(y_* | \mu(x_*), \sigma^2(x_*)),$$

$$\mu(x_*) = \hat{\mu}^T\phi(x_*), \quad \sigma^2(x_*) = \sigma_y^2 + \phi(x_*)^T\hat{\Sigma}\phi(x_*).$$

3.3.2.3 最尤推定との比較.

・最尤推定ではデータ数が変化しても大きく予測は変わらない

(外れ値などがなければ)

・Bayes推定では,データ数が増えるにつれて予測の不確実性 (i.e., 予測値の分散) が変化していく.

3.3.3 周辺尤度.

・ $\mathcal{X} = \{x_1, \dots, x_N\}$, $\mathcal{Y} = \{y_1, \dots, y_N\}$: given.

パラメータ w を同時分布 $p(\mathcal{Y}, w | \mathcal{X}) = p(\mathcal{Y} | \mathcal{X}, w)p(w)$ から積分除去した

$$p(\mathcal{Y} | \mathcal{X}) = \int p(\mathcal{Y} | \mathcal{X}, w)p(w)dw$$

を, Bayes線形回帰モデルの周辺尤度 (marginal likelihood), あるいは

エビデンス (evidence) という.

モデルが与えられたもとでの,データ $\mathcal{Y}$ の出現する尤もらしさを表す.

→ 複数個のモデルをエビデンスで定量的に比較できる.

・積分除去する代わりに，対数計算で $p(\mathbf{y}|\mathbf{X})$ を求める.

$p(\mathbf{y}, \mathbf{w}|\mathbf{X}) = p(\mathbf{w}|\mathbf{y}, \mathbf{X})p(\mathbf{y}|\mathbf{X})$ なので，

$$
\begin{aligned}
&\log p(\mathbf{y}|\mathbf{X}) \\
&= \log p(\mathbf{w}) + \log \underbrace{p(\mathbf{y}|\mathbf{X}, \mathbf{w})}_{= \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \mathbf{w})} - \log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \\
&= \log \mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma_w^2 \mathbf{I}_{H_1}) + \sum_{n=1}^{N} \log \mathcal{N}(y_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \sigma_y^2) - \log \mathcal{N}(\mathbf{w}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \\
&= -\frac{1}{2}\left(\cancel{H_1 \log 2\pi} + \log \underbrace{\det(\sigma_w^2 \mathbf{I}_{H_1})}_{=(\sigma_w^2)^{H_1}} + \mathbf{w}^T(\sigma_w^2 \mathbf{I}_{H_1})^{-1}\mathbf{w}\right) \\
&\quad - \frac{1}{2}\sum_{n=1}^{N}\left(\log 2\pi + \log \sigma_y^2 + \sigma_y^{-2}(y_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right) \\
&\quad + \frac{1}{2}\left(\cancel{H_1 \log 2\pi} + \log \det \hat{\boldsymbol{\Sigma}} + (\mathbf{w} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{w} - \hat{\boldsymbol{\mu}})\right) \\
&= -\frac{1}{2}\Big(H_1 \log \sigma_w^2 + \sigma_w^{-2}\mathbf{w}^T\mathbf{w} \\
&\qquad + N\log 2\pi + N \log \sigma_y^2 + \sigma_y^{-2}\sum_{n=1}^{N} y_n^2 \cancel{- 2\mathbf{w}^T\Big(\underbrace{\sigma_y^{-2}\sum_{n=1}^{N} y_n \boldsymbol{\phi}(\mathbf{x}_n)}_{= \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}\Big)} + \mathbf{w}^T\Big(\sigma_y^{-2}\sum_{n=1}^{N}\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T\Big)\mathbf{w} \\
&\qquad - \left(\log \det \hat{\boldsymbol{\Sigma}} - \mathbf{w}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{w} + \cancel{2\mathbf{w}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}} - \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\right)\Big) \\
&= -\frac{1}{2}\Big(H_1 \log \sigma_w^2 + N \log 2\pi + N \log \sigma_y^2 + \sigma_y^{-2}\sum_{n=1}^{N} y_n^2 - \log \det \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} \\
&\qquad + \cancel{\mathbf{w}^T\Big(\underbrace{\sigma_w^{-2}\mathbf{I}_{H_1} + \sigma_y^{-2}\sum_{n=1}^{N}\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T}_{= \hat{\boldsymbol{\Sigma}}^{-1}}\Big)\mathbf{w}} \cancel{- \mathbf{w}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{w}}\Big) \\
&= -\frac{1}{2}\left(\sigma_y^{-2}\sum_{n=1}^{N} y_n^2 + N\log \sigma_y^2 + N \log 2\pi + H_1 \log \sigma_w^2 - \log \det \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\right).
\end{aligned}
$$

$$
\begin{aligned}
&\therefore p(\mathbf{y}|\mathbf{X}) \\
&= \exp\left(-\frac{1}{2}\left(\sigma_y^{-2}\sum_{n=1}^{N} y_n^2 + N\log \sigma_y^2 + N \log 2\pi + H_1 \log \sigma_w^2 - \log \det \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}\right)\right).
\end{aligned}
$$

・Bayes線形回帰ではこのように解析的に求まる

一般にはエビデンスの計算は困難．解析的に求まったとしても，計算が

大変なことも多い．

→ サンプリングにより積分を近似．変分推論によるエビデンスの下界を

求める．などして何とかする．

3.3.4 逐次学習．

・逐次学習 (sequential learning) / オンライン学習 (online learning)：

新規に入ってくる学習データに適応的に学習すること．

・共役事前分布を使って解析的にパラメータ更新式が求まる場合は，

データの生成過程に順序の依存性を仮定しないなら，

逐次的にデータを与えるのと一度に全データを与える場合とで

最終的な事後分布はどちらも同じになる．

Bayes線形回帰の例．最初のデータ $\mathcal{D}_1 = \{x_1, y_1\}$ ：given のときの事後分布．

$$p(w \mid y_1, x_1) \propto p(y_1 \mid w, x_1)\, p(w).$$

更新後の分布を事前分布として使う

次のデータ $\mathcal{D}_2 = \{x_2, y_2\}$ ：given のとき，

$$p(w \mid y_1, x_1, y_2, x_2) \propto p(y_2 \mid w, x_2)\, p(w \mid y_1, x_1)$$

・モデルが複雑になると，逐次学習の各更新で w の事後分布が解析的に求まらなくなる．

→ 近似的に事後分布を更新していく．

3.3.5 能動学習への応用.

- 予測分布の計算

→ 予測対象の値に対する不確実性を, 分散などの指標で評価できる.

→ 効率的に学習用ラベルデータを収集する能動学習 (active learning) に応用可能.

- 能動学習:

ラベルの付いていないデータ $\mathcal{X}_{pool}$ から, 適切なデータ点 x_{query} を選び, アノテータにラベル y_{query} を要求する.

- x_{query} の選び方 (一例):

新しい入力 x_* を与えたときの予測 y_* の不確かなもの (予測に自信がないもの) を選ぶ

→ 予測分布のエントロピー (entropy) 最大の x_* を x_{query} として選ぶ:

$$x_{query} \leftarrow \underset{x_* \in \mathcal{X}_{pool}}{\operatorname{argmax}} F(x_*),$$

$$F(x_*) := -\mathbb{E}_{p(y_*|x_*, y, \mathcal{X})}[\log p(y_*|x_*, y, \mathcal{X})].$$

- Bayes線形回帰では,

$$F(x_*)$$

$$= -\int p(y_*|x_*, y, \mathcal{X}) \log p(y_*|x_*, y, \mathcal{X}) \, dy_*$$

$$= -\int \mathcal{N}(y_*|\mu(x_*), \sigma^2(x_*)) \log \mathcal{N}(y_*|\mu(x_*), \sigma^2(x_*))\, dy_*$$

$$= \frac{1}{2}\int \mathcal{N}(y_*|\mu(x_*), \sigma^2(x_*))\left(\log 2\pi + \log \sigma^2(x_*) + \sigma^{-2}(x_*)(y_*-\mu(x_*))^2\right) dy_*$$

$$= \frac{1}{2}\Bigg((\log 2\pi + \log \sigma^2(x_*)) \underbrace{\int \mathcal{N}(y_*|\mu(x_*), \sigma^2(x_*))\, dy_*}_{=1}$$

$$+ \sigma^{-2}(x_*) \underbrace{\int (y_*-\mu(x_*))^2 \mathcal{N}(y_*|\mu(x_*), \sigma^2(x_*))\, dy_*}_{=\sigma^2(x_*)} \Bigg)$$

$$= \frac{1}{2}\left(1 + \log \sigma^2(x_*) + \log 2\pi\right).$$

となるので、 $x_{query} \leftarrow \underset{x_* \in \mathcal{X}_{pool}}{\operatorname{argmax}}\, \sigma^2(x_*)$ とする。

・同様にして未知関数 $f(x)$ の最適値の探索ができる

（Bayes最適化）

・予測対象に対して弱い仮定を設定できるGauss過程を用いることが一般的。

3.3.6 Gauss過程との関係

・予測分布の分散は、

$$\sigma^2(x_*)$$

$$= \sigma_y^2 + \phi(x_*)^T \hat{\Sigma} \phi(x_*)$$

$$= \sigma_y^2 + \phi(x_*)^T \left(\sigma_y^{-2}\sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T + \sigma_w^{-2} I_{H_1}\right)^{-1}\phi(x_*)$$

ここで、 $\Lambda = \sigma_w^{-2} I_{H_1}$, $\Phi = \underbrace{\left(\phi(x_1) \cdots \phi(x_N)\right)}_{N} \Big\} H_1$ とすると、

$$
\begin{aligned}
&\sigma^2(x_*) \\
&= \sigma_y^2 + \phi(x_*)^T \left(\sigma_y^{-2} \Phi \Phi^T + \Lambda \right)^{-1} \phi(x_*).
\end{aligned}
$$

・予測分布の平均は，$y = (y_1, \dots, y_N)^T$とすると，

$$
\begin{aligned}
&\mu(x_*) \\
&= \hat{\mu}^T \phi(x_*) \\
&= \sigma_y^{-2} \left(\sum_{n=1}^{N} y_n \phi(x_n)^T \right) \left(\sigma_y^{-2} \Phi \Phi^T + \Lambda \right)^{-1} \phi(x_*) \\
&= \sigma_y^{-2} y^T \Phi^T \left(\sigma_y^{-2} \Phi \Phi^T + \Lambda \right)^{-1} \phi(x_*) \\
&= \sigma_y^{-2} \phi(x_*)^T \left(\sigma_y^{-2} \Phi \Phi^T + \Lambda \right)^{-1} \Phi y.
\end{aligned}
$$

・Sherman-Morrison-Woodburyの公式より，$\Phi^T \Lambda^{-1} \Phi =: K$とおくと，

$$
\begin{aligned}
&\left(\sigma_y^{-2} \Phi \Phi^T + \Lambda \right)^{-1} \\
&= \Lambda^{-1} - \Lambda^{-1} \Phi \left((\sigma_y^{-2} I_N)^{-1} + \underbrace{\Phi^T \Lambda^{-1} \Phi}_{=K} \right)^{-1} \Phi^T \Lambda^{-1} \\
&= \Lambda^{-1} - \Lambda^{-1} \Phi \left(\sigma_y^{2} I_N + K \right)^{-1} \Phi^T \Lambda^{-1}
\end{aligned}
$$

ゆえに，

$$
\begin{aligned}
&\sigma^2(x_*) \\
&= \sigma_y^2 + \phi(x_*)^T \left(\Lambda^{-1} - \Lambda^{-1} \Phi \left(\sigma_y^{2} I_N + K \right)^{-1} \Phi^T \Lambda^{-1} \right) \phi(x_*) \\
&= \sigma_y^2 + \phi(x_*)^T \Lambda^{-1} \phi(x_*) - \phi(x_*)^T \Lambda^{-1} \Phi \left(\sigma_y^{2} I_N + K \right)^{-1} \Phi^T \Lambda^{-1} \phi(x_*).
\end{aligned}
$$

$$
\begin{aligned}
&\mu(x_*) \\
&= \sigma_y^{-2} \phi(x_*)^T \left(\Lambda^{-1} - \Lambda^{-1} \Phi \left(\sigma_y^{2} I_N + K \right)^{-1} \Phi^T \Lambda^{-1} \right) \Phi y
\end{aligned}
$$

Λ^{-1}でくくる.

$$= \sigma_y^{-2}\phi(x_*)^T\Lambda^{-1}\left(I_{H_1} - \Phi(\sigma_y^2 I_N + K)^{-1}\Phi^T\Lambda^{-1}\right)\Phi y$$

$\downarrow$ Φをかける

$$= \sigma_y^{-2}\phi(x_*)^T\Lambda^{-1}\left(\Phi - \Phi(\sigma_y^2 I_N + K)^{-1}\underbrace{\Phi^T\Lambda^{-1}\Phi}_{=K}\right) y$$

$\downarrow$ Φをくくる

$$= \sigma_y^{-2}\phi(x_*)^T\Lambda^{-1}\Phi\left(I_N - (\sigma_y^2 I_N + K)^{-1}K\right) y$$

$$= \sigma_y^{-2}\phi(x_*)^T\Lambda^{-1}\Phi\left(I_N - (\sigma_y^2 I_N + K)^{-1}(\sigma_y^2 I_N + K - \sigma_y^2 I_N)\right) y$$

$$= \cancel{\sigma_y^{-2}}\phi(x_*)^T\Lambda^{-1}\Phi\left(\cancel{I_N - I_N} + \cancel{\sigma_y^2}(\sigma_y^2 I_N + K)^{-1}\right) y$$

$$= \phi(x_*)^T\Lambda^{-1}\Phi(\sigma_y^2 I_N + K)^{-1} y.$$

・特徴量関数 ϕ は，二つの入力データ x, x' に対して常に

$$k(x, x') := \phi(x)^T\Lambda^{-1}\phi(x') \quad (\text{カーネル関数})$$

← 二つのデータ x, x' の "類似度" のようなもの

の形で現れているので，ϕ を設計しなくても k を直接設計すればよい（カーネルトリック）：

$$K = \Phi^T\Lambda^{-1}\Phi$$

$$= \begin{pmatrix} \phi(x_1)^T\Lambda^{-1} \\ \vdots \\ \phi(x_N)^T\Lambda^{-1} \end{pmatrix} \begin{pmatrix} \phi(x_1) & \cdots & \phi(x_N) \end{pmatrix}$$

$$= \begin{pmatrix} \phi(x_1)^T\Lambda^{-1}\phi(x_1) & \cdots & \phi(x_1)^T\Lambda^{-1}\phi(x_N) \\ \vdots & \ddots & \vdots \\ \phi(x_N)^T\Lambda^{-1}\phi(x_1) & \cdots & \phi(x_N)^T\Lambda^{-1}\phi(x_N) \end{pmatrix}$$

$$= \left(k(x_i, x_j)\right)_{ij}$$

$$\phi(x_*)^T\Lambda^{-1}\Phi$$

$$= \begin{pmatrix} \phi(x_*)^T\Lambda^{-1}\phi(x_1) \\ \vdots \\ \phi(x_*)^T\Lambda^{-1}\phi(x_N) \end{pmatrix} = \begin{pmatrix} k(x_*, x_1) \\ \vdots \\ k(x_*, x_N) \end{pmatrix}$$

§3.4 最尤推定, MAP推定との関係

3.4.1 最尤推定と誤差最小化.

Th. 回帰モデルを, w をパラメータ, $\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$ とし,

$$y = f(x; w) + \varepsilon$$

とする. N個のデータ $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ が与えられたときのパラメータ w の最尤推定値 w_{ML} は, 誤差関数

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n; w))^2$$

← 最尤推定は最小2乗法に対応する.

を最小化する w と一致する.

pf. $\varepsilon = y - f(x; w) \sim \mathcal{N}(0, \sigma_y^2)$ より,

$$y_n \sim \mathcal{N}(f(x_n; w), \sigma_y^2)$$

である. $\mathcal{D}$: given のとき, モデルの尤度関数は,

$$p(\mathcal{Y}|\mathcal{X}, w) = \prod_{n=1}^{N} p(y_n|x_n, w) = \prod_{n=1}^{N} \mathcal{N}(y_n | f(x_n; w), \sigma_y^2)$$

最尤推定値は,

log をとっても大小関係は変化しない.

$$\begin{aligned} w_{ML} &= \underset{w}{\mathrm{argmax}}\ p(\mathcal{Y}|\mathcal{X}, w) = \underset{w}{\mathrm{argmax}} \log p(\mathcal{Y}|\mathcal{X}, w) \\ &= \underset{w}{\mathrm{argmax}} \left[-\frac{1}{2}\left(\underline{N \log 2\pi + N \log \sigma_y^2} + \underline{\sigma_y^{-2}} \sum_{n=1}^{N} (y_n - f(x_n; w))^2 \right) \right] \\ &= \underset{w}{\mathrm{argmin}} \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n; w))^2 \\ &= \underset{w}{\mathrm{argmin}}\ E(w). \end{aligned}$$

定数は最大化に関係ない

argmin に

∎

・ $f(x_n; w)$ が NN のように複雑でパラメータの解析解が求まらない場合，勾配降下法により最適化する．

$$\nabla_w \log p(y|X, w)$$
$$= -\sigma_y^{-2} \nabla_w \left(\frac{1}{2} \sum_{n=1}^{N} (y_n - f(x_n; w))^2 \right)$$
$$= -\sigma_y^{-2} \nabla_w E(w)$$

なので，$\log p(y|X, w)$ の最大化の更新式は α' を学習率として

$$w \leftarrow w + \alpha' \nabla_w \log p(y|X, w)$$
$$= w - \alpha' \sigma_y^{-2} \nabla_w E(w)$$

$\alpha := \alpha' \sigma_y^{-2}$ とおくと，$w \leftarrow w - \alpha \nabla_w E(w)$ となり，

E を最小化する勾配降下法の更新式と同じ式が得られる．

3.4.2 MAP推定と正則化．

・最大事後確率推定 (maximum a posteriori estimation : MAP推定)

モデル $y = f(x; w)$ のパラメータ w を，データ $\mathcal{D} = \{X, y\}$: given のときの w の事後分布を最大化する w (モード) で推定する：

$$w_{MAP} := \underset{w}{\operatorname{argmax}}\, p(w|y, X) = \underset{w}{\operatorname{argmax}} \log p(w|y, X).$$

Th. 回帰モデルを，w をパラメータ，$\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$ とし，

$$y = f(x; w) + \varepsilon$$

とする．また，w の事前分布を $p(w) = \mathcal{N}(w|0, \sigma_w^2 I_{H_1})$ とする．

N個のデータ $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ が与えられたときの

パラメータ w の MAP推定値 w_{MAP} は，コスト関数

MAP推定は
正則化つきの最小2乗法に
対応する

$$J(w) = E(w) + \lambda \Omega_{L2}(w) = \frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n; w))^2 + \frac{\lambda}{2}\|w\|^2$$

を最小化する w と一致する.

pf.

$$\log p(w|\mathcal{Y}, \mathcal{X})$$

Bayesの定理

$$= \log p(\mathcal{Y}|\mathcal{X}, w) + \log p(w) + \text{const.}$$

$$= -\frac{1}{2}\sigma_y^{-2}\sum_{n=1}^{N}(y_n - f(x_n; w))^2$$

$$-\frac{1}{2}\Big(\underbrace{H_1 \log 2\pi + \log\det(\sigma_w^2 I_{H_1})}_{w\text{に依らない}} + w^\top(\sigma_w^2 I_{H_1})^{-1} w\Big) + \text{const.}$$

$$= -\sigma_y^{-2}\left(\frac{1}{2}\sum_{n=1}^{N}(y_n - f(x_n; w))^2 + \frac{1}{2}\underbrace{\frac{\sigma_w^{-2}}{\sigma_y^{-2}}}_{=:\lambda\text{とおく}}\|w\|^2\right) + \text{const.}$$

$$= -\sigma_y^{-2}\left(E(w) + \lambda\Omega_{L2}(w)\right) + \text{const.}$$

$$\therefore w_{MAP} = \operatorname*{argmax}_{w} \log p(w|\mathcal{Y}, \mathcal{X})$$

$$= \operatorname*{argmin}_{w} J(w). \quad \blacksquare$$

・事前分布を Laplace分布 $\mathrm{Lap}(w|\mu, b) = \frac{1}{2b}\exp\left(-\frac{1}{b}|w-\mu|\right)$

を使って $p(w) = \prod_{h=1}^{H_1} \mathrm{Lap}(w_h | 0, \frac{1}{\lambda})$

とすると，正則化項は $\lambda\Omega_{L1}(w) = \lambda\sum_{h=1}^{H_1}|w_h|$ になる：

$$\log p(w)$$

$$= H_1 \log\frac{\lambda}{2} - \sum_{h=1}^{H_1}\lambda|w_h|$$

$$= -\lambda\Omega_{L1}(w) + \text{const.}$$

- 事前分布にGauss分布やLaplace分布を導入したMAP推定は最尤推定より過剰適合にロバスト（正則化の効果）

→ 利用するのは w_{MAP} のみ（点推定：point estimation）なので、予測の不確実性の定量化やエビデンスによるモデルの評価などはできない。

3.4.3 分類モデルに対する誤差関数

3.4.3.1 2値分類の場合

- ラベル：$y_n \in \{0, 1\}$ とする。

→ $y_n \sim \mathrm{Bern}(\mu_n)$ と仮定。

μ_n についても、回帰モデル $\eta_n = f(x; w) + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \sigma_\eta^2)$ を利用して $\mu_n = \mathrm{Sig}(\eta_n)$ と仮定。

つまり、

$$\mu_n = \mathrm{Sig}(\eta_n), \quad \eta_n \sim \mathcal{N}(f(x; w), \sigma_\eta^2)$$

- 特に $f(x_n; w) = w^T \phi(x_n)$ のとき ロジスティック回帰モデル。

← logit関数という

- $\eta = \mathrm{Sig}^{-1}(\mu)$. $\mathrm{Sig}(\eta) = \frac{1}{1+e^{-\eta}} = \mu$ なので、

$$\eta = \log \frac{\mu}{1-\mu}.$$

← Bernoulli分布の自然パラメータ表現

シグモイド関数でBernoulli分布のパラメータを自然パラメータに変換。

- データ $\mathcal{D} = \{X, Y\}$: given とする。対数尤度関数は、

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w})$$
$$= \log \prod_{n=1}^{N} p(y_n | \mathbf{x}_n, \mathbf{w})$$
$$= \sum_{n=1}^{N} \log\left(\mu_n^{y_n} (1-\mu_n)^{1-y_n}\right)$$
$$= \sum_{n=1}^{N} \left(y_n \log \mu_n + (1-y_n) \log (1-\mu_n) \right)$$

→ このモデルに対する最尤法は，クロスエントロピー誤差

$$-\sum_{n=1}^{N} \left(y_n \log \mu_n + (1-y_n) \log (1-\mu_n) \right)$$

の最小化になる．

・ところで．NNのところでは，$\eta_n = f(\mathbf{x}_n ; \mathbf{w})$ と

決定的に求めていた．このときのクロスエントロピー誤差は．$\mathbf{w}$の函数として

$$E(\mathbf{w})$$
$$= -\sum_{n=1}^{N} \left(y_n \log Sig(f(\mathbf{x}_n; \mathbf{w})) + (1-y_n) \log (1 - Sig(f(\mathbf{x}_n; \mathbf{w}))) \right)$$

と書ける．一方で，ここでのモデル化では η_n は確率変数なので，

η_n は $\mathbf{w}$の函数になっておらず$\mathbf{w}$で微分できない．最尤法を使うためには，

誤差が $\mathbf{w}$の函数になっていないといけないのではないか…？

→ 実は．上のクロスエントロピー誤差も$\mathbf{w}$の函数 $\tilde{E}(\mathbf{w})$ とみなせる．

$\eta_n = f(\mathbf{x}_n ; \mathbf{w}) + \varepsilon_n$ ，$\varepsilon_n \sim \mathcal{N}(0, \sigma_\eta^2)$ なので，

$$\tilde{E}(\mathbf{w})$$
$$= -\sum_{n=1}^{N} \left(y_n \log Sig(f(\mathbf{x}_n; \mathbf{w}) + \varepsilon_n) + (1-y_n) \log (1 - Sig(f(\mathbf{x}_n; \mathbf{w}) + \varepsilon_n)) \right).$$

これは w について微分ができる。(reparametrization trick)

とはいえ、ε_n が確率変数であることに変わりはないので、$\tilde{E}(w)$ も

確率変数になる。そこで、平均的に loss が小さくなるように

$\tilde{E}(w)$ の期待値 $\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\tilde{E}(w)]$ を最小化するようにする。

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\tilde{E}(w)] \\
&= -\sum_{n=1}^{N}\Big(y_n\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\log \mathrm{Sig}(f(x_n;w)+\varepsilon_n)] \\
&\qquad + (1-y_n)\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\log(1-\mathrm{Sig}(f(x_n;w)+\varepsilon_n))]\Big)
\end{aligned}
$$

この最小化のためには $\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\tilde{E}(w)]$ の w に関する勾配を考えればよい

$$
\begin{aligned}
&\nabla_w \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\log \mathrm{Sig}(f(x_n;w)+\varepsilon_n)] \quad \text{微分と積分の順序交換.}\\
&= \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\nabla_w \log \mathrm{Sig}(f(x_n;w)+\varepsilon_n)] \quad \text{合成関数の微分.}\ \mathrm{Sig}'(x) = \mathrm{Sig}(x)(1-\mathrm{Sig}(x)).\\
&= \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[\frac{1}{\cancel{\mathrm{Sig}(f(x_n;w)+\varepsilon_n)}}\cancel{\mathrm{Sig}(f(x_n;w)+\varepsilon_n)}(1-\mathrm{Sig}(f(x_n;w)+\varepsilon_n))\nabla_w f(x_n;w)\right]\\
&= \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[1-\mathrm{Sig}(f(x_n;w)+\varepsilon_n)]\nabla_w f(x_n;w),
\end{aligned}
$$

$$
\begin{aligned}
&\nabla_w \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\log(1-\mathrm{Sig}(f(x_n;w)+\varepsilon_n))] \quad \text{微分と積分の順序交換.}\\
&= \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[\frac{-1}{\cancel{1-\mathrm{Sig}(f(x_n;w)+\varepsilon_n)}}\mathrm{Sig}(f(x_n;w)+\varepsilon_n)\cancel{(1-\mathrm{Sig}(f(x_n;w)+\varepsilon_n))}\nabla_w f(x_n;w)\right]\\
&= -\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\mathrm{Sig}(f(x_n;w)+\varepsilon_n)]\nabla_w f(x_n;w)
\end{aligned}
$$

よって、

$$
\nabla_w \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\tilde{E}(w)]
$$

$$= -\sum_{n=1}^{N}\Big(y_n \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[1 - Sig(f(x_n;w)+\varepsilon_n)\right]\nabla_w f(x_n;w)$$
$$- (1-y_n)\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[Sig(f(x_n;w)+\varepsilon_n)\right]\nabla_w f(x_n;w)\Big)$$
$$= -\sum_{n=1}^{N}\left(y_n - \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[Sig(f(x_n;w)+\varepsilon_n)\right]\right)\nabla_w f(x_n;w).$$

ここで,

$$\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[Sig(f(x_n;w)+\varepsilon_n)\right]$$

Taylor展開.してもいいかは不明….

$$= \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[\sum_{i=0}^{\infty}\frac{1}{i!}Sig^{(i)}(f(x_n;w))\,\varepsilon_n^i\right]$$
$$= \sum_{i=0}^{\infty}\frac{1}{i!}Sig^{(i)}(f(x_n;w))\underbrace{\mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}\left[\varepsilon_n^i\right]}_{\text{正規分布のモーメント. } i:\text{odd なら } 0,\ i:\text{even なら } (i-1)!!\sigma_y^i}$$
$$= Sig(f(x_n;w)) + \sum_{i=1}^{\infty}\frac{1}{(2i)!}Sig^{(2i)}(f(x_n;w))(2i-1)!!\sigma_y^{2i}$$
$$\approx Sig(f(x_n;w))$$

なので, $\nabla_w \mathbb{E}_{\mathcal{N}(0,\sigma_y^2)}[\tilde{E}(w)] \approx -\sum_{n=1}^{N}\left(y_n - Sig(f(x_n;w))\right)\nabla_w f(x_n;w).$

一方で,

$$\nabla_w E(w)$$
$$= -\sum_{n=1}^{N}\Big(y_n \frac{\cancel{Sig(f(x_n;w))}(1-Sig(f(x_n;w)))}{\cancel{Sig(f(x_n;w))}}\nabla_w f(x_n;w)$$
$$- (1-y_n)\frac{Sig(f(x_n;w))\cancel{(1-Sig(f(x_n;w)))}}{\cancel{1-Sig(f(x_n;w))}}\nabla_w f(x_n;w)\Big)$$
$$= -\sum_{n=1}^{N}\left(y_n - Sig(f(x_n;w))\right)\nabla_w f(x_n;w)$$

となるので,このモデルに対する最尤推定のパラメータは,

近似的には決定的に求める場合と同じになる.

3.4.3.2 多値分類の場合.

多値分類でも同様. D値分類とする.

$y_n \in \{0,1\}^D$, $\sum_{d=1}^{D} y_{n,d} = 1$ とする (one-hot)

$y_n \sim \mathrm{Cat}(\pi_n)$ と仮定. さらに,

$\pi_n = \mathrm{Softmax}(\eta_n)$,

$\eta_n = f(x_n; W) + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \sigma_\eta^2 I_D)$

と仮定. 対数尤度関数は,

$$
\begin{aligned}
&\log p(Y \mid X, W) \\
&= \log \prod_{n=1}^{N} p(y_n \mid x_n, W) \\
&= \sum_{n=1}^{N} \log p(y_n \mid x_n, W) \\
&= \sum_{n=1}^{N} \log \prod_{d=1}^{D} \pi_{n,d}^{\,y_{n,d}} \\
&= \sum_{n=1}^{N} \sum_{d=1}^{D} y_{n,d} \log \pi_{n,d}
\end{aligned}
$$

→ 最尤法は, クロスエントロピー誤差

$$
-\sum_{n=1}^{N} \sum_{d=1}^{N} y_{n,d} \log \pi_{n,d}
$$

の最小化になる.

・3.4.3.1 と同様に, η_n を決定的に求める場合と, このモデル化のように確率的に決める場合は, 近似的には同じ (はず).