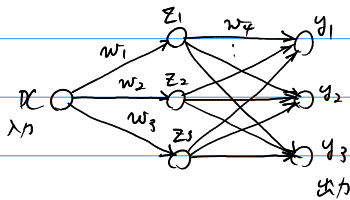


5.40

- ・ニューラルネットワークの出力ユニットの活性化関数をソフトマックス関数とする。
- ・目標変数 \mathcal{D} の条件付分布を多項分布とし、ニューラルネットの出力をこの多項分布の平均と解釈する。
というモデルを考える



出力ユニット y_k の活性を a_k とすると

$$y_k(x, w) = \frac{\exp(a_k(x, w))}{\sum_j \exp(a_j(x, w))}$$

目標変数 \mathcal{D} (1 of K 表記) の条件付分布は

$$p(\mathcal{D} | x, w) = \prod_{k=1}^K y_k^{t_k}(x, w) \quad \dots \textcircled{1}$$

と仮定。このとき $\mathcal{D} = \{t_1, t_2, \dots\}$, $\mathcal{X} = \{x_1, x_2, \dots\}$, $y_{nk} = y_k(x_n, w)$ とおくと尤度は

$$p(\mathcal{D} | w, \mathcal{X}) = \prod_{n=1}^N p(t_n | x_n, w) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}} \quad \dots \textcircled{2}$$

対数尤度は

$$\ln p(\mathcal{D} | w, \mathcal{X}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}$$

と仮定。

(注) 教科書では条件付確率の \mathcal{D}, \mathcal{X} の表記は省略されている。

(ベイズ- w の推定)

ベイズモデルにて w の推定値は 最大事後確率推定 (MAP推定) で与えられる。 ^{w の} 事後分布

$$p(w|D, X) = \frac{p(D, w|X)}{p(D|X)} = \frac{p(D|w, X) p(w)}{p(D|X)} \propto p(D|w, X) p(w)$$

よって 対数事後分布は

D, X は観測値に固定されている
よって $p(D|X)$ は定数である

$$\ln p(w|D, X) = \ln p(D|w, X) + \ln p(w) + C, \quad (C \text{ は } w \in \mathbb{R}^n \text{ の定数})$$

$$= \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} + \ln p(w) + C$$

よって

w の事前分布を 正規分布 (5.162) とすると

$$\ln p(w|D, X) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} - \frac{\alpha}{2} w^T w + C$$

よって

よって、対数事後分布 $\ln p(w|D, X)$ の最大化は、(5.182) の正則化勾配関数

$$E(w) = -\ln p(D|w, X) + \frac{\alpha}{2} w^T w$$

$$= -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} + \frac{\alpha}{2} w^T w$$

の最小化と等価であることが分かる。

$E(w)$ を最小にする w_{MAP} は $0 = \frac{\partial E}{\partial w}$ の解で与えられる。

この解は 5.2.4 節の勾配降下法で得ることができる。

勾配降下法で必要の $\frac{\partial E}{\partial w}$ の計算は 5.3 節の誤差逆伝播

を行うことができる。

(予測分布)

新しい α に対する世の予測分布は

$$p(\mathbf{t}|\alpha, D, X) = \int p(\mathbf{t}, \mathbf{w}|\alpha, D, X) d\mathbf{w} = \int p(\mathbf{t}|\alpha, \mathbf{w}) p(\mathbf{w}|D, X) d\mathbf{w}$$
$$= \int \prod_{k=1}^K y_k^{\mathbf{t}_k} g_k(\alpha, \mathbf{w}) p(\mathbf{w}|D, X) d\mathbf{w}$$

と与えられる。

この積分可算性は難しいが、 δ で近似可算。

(近似1)

(5.185)と同様に事後分布 $p(\mathbf{w}|D, X) \approx \delta(\mathbf{w} - \mathbf{w}_{\text{MAP}})$ と近似して

$$p(\mathbf{t}|\alpha, D, X) = \int \prod_{k=1}^K y_k^{\mathbf{t}_k} g_k(\alpha, \mathbf{w}) p(\mathbf{w}|D, X) d\mathbf{w}$$
$$\approx \int \prod_{k=1}^K y_k^{\mathbf{t}_k} g_k(\alpha, \mathbf{w}) \delta(\mathbf{w} - \mathbf{w}_{\text{MAP}}) d\mathbf{w} = \prod_{k=1}^K y_k^{\mathbf{t}_k} g_k(\alpha, \mathbf{w}_{\text{MAP}})$$

を得る。

(近似2)

5.7.3節の (5.190) と同様の近似は出来るがこの説明

(5.186)と同様にネットワークの出力 y_k の活性化 a_k を線形近似可算

$$a_k(\alpha, \mathbf{w}) \approx a_k^{\text{MAP}}(\alpha) + b_k^T (\mathbf{w} - \mathbf{w}_{\text{MAP}})$$

ただし $a_k^{\text{MAP}}(\alpha) = a_k(\alpha, \mathbf{w}_{\text{MAP}})$, $b_k = \nabla a_k(\alpha, \mathbf{w}) \Big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}$ とする。

このときネットワークの出力 y_k は

$$y_k(\alpha, \mathbf{w}) \approx \frac{\exp(a_k^{\text{MAP}}(\alpha) + b_k^T (\mathbf{w} - \mathbf{w}_{\text{MAP}}))}{\sum_j \exp(a_j^{\text{MAP}}(\alpha) + b_j^T (\mathbf{w} - \mathbf{w}_{\text{MAP}}))}$$

となる。

このFの予測分布は

$$p(\theta = (0, \dots, 0) | \alpha, D, X) = \int g_k(\alpha, w) p(w | D, X) dw$$

\uparrow
k番目

$$\approx \int \frac{\exp(a_k^{\text{MAP}}(\alpha) + b_k^T(w - w_{\text{MAP}}))}{\sum_j \exp(a_j^{\text{MAP}}(\alpha) + b_j^T(w - w_{\text{MAP}}))} p(w | D, X) dw$$

\swarrow a_k は線形近似

$$\approx \int \underbrace{\frac{\exp(a_k^{\text{MAP}}(\alpha) + b_k^T(w - w_{\text{MAP}}))}{\sum_j \exp(a_j^{\text{MAP}}(\alpha) + b_j^T(w - w_{\text{MAP}}))}}_{\textcircled{3}} g(w | D, X) dw$$

\swarrow $p(w | D, X)$ の近似

と近似。

③ が $w^T b_k$ を変数とする正規分布関数になっているので、(4.144)と同じ形と見做すことができる。
よって4.5.2節の結果を利用して積分を導くことができる。

5.7.3節では、③の部分から $\sigma(a^{\text{MAP}}(\alpha) + b^T(w - w_{\text{MAP}}))$ と仮定して

4.5.2節の(4.145)の積分と同じ形と見做すことができる。

(4.148)、(4.150)、(4.153) を用いて (5.190) の積分を導くことができました。

(超次元空間での推定)

α の Evidential 関数を

$$p(D|\alpha, X) = \int p(D|w, X) p(w|\alpha) dw$$

である。右辺の $f(w) = p(D|w, X) p(w|\alpha)$ は 2 次元関数である。

・ 超次元空間では $f(w)$ の値が 0 になる可能性が高い。
2 次元空間では $f(w)$ の値が 0 になる可能性は低い。

例として、(4, 135) の F'

$$p(D|\alpha, X) = \int f(w) dw \approx f(w_{MAP}) \frac{(2\pi)^{\frac{W}{2}}}{|A|^{\frac{1}{2}}}$$

である。ここで W は w の次元である。 $A = -\nabla^2 \ln f(w)|_{w=w_{MAP}}$ (4, 132) である。

したがって

$$\ln p(D|\alpha, X) = \ln f(w_{MAP}) + \frac{W}{2} \ln(2\pi) - \frac{1}{2} \ln |A|$$

ここで

$$\ln f(w_{MAP}) = \ln p(D|w_{MAP}, X) + \ln p(w_{MAP}|\alpha)$$

$$= -E(w_{MAP}) + \frac{\alpha}{2} w_{MAP}^T w_{MAP} + \ln N(w_{MAP} | 0, \alpha^{-1} I)$$

$$= -E(w_{MAP}) + \frac{\alpha}{2} w_{MAP}^T w_{MAP} + \ln \frac{1}{(2\pi)^{\frac{W}{2}}} \frac{1}{|\alpha^{-1} I|^{\frac{1}{2}}} \exp\left(-\frac{\alpha}{2} w_{MAP}^T w_{MAP}\right)$$

$$= -E(w_{MAP}) - \frac{W}{2} \ln(2\pi) + \frac{W}{2} \ln \alpha \leftarrow |\alpha^{-1} I| = \alpha^{-W}$$

したがって

$$\ln p(D|\alpha, X) = -E(w_{MAP}) - \frac{1}{2} \ln |A| + \frac{W}{2} \ln \alpha$$

を得る。

二枚を最大化する α は

$$0 = \frac{\partial}{\partial \alpha} \ln P(D|\alpha, X) = \frac{\partial}{\partial \alpha} \left(-E(w_{MAP}) - \frac{1}{2} \ln |A| + \frac{W}{2} \ln \alpha \right)$$

$$= -\frac{1}{2} w_{MAP}^T w_{MAP} - \frac{1}{2} \frac{\partial}{\partial \alpha} \ln |A| + \frac{W}{2\alpha}$$

上記 $E(w) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} + \frac{\alpha}{2} w^T w$
 $\frac{\partial E(w_{MAP})}{\partial \alpha} = \frac{1}{2} w_{MAP}^T w_{MAP}$

2" 子之" 子。二二"

$$A = -\nabla \nabla \ln f(w) \Big|_{w=w_{MAP}} = -\nabla \nabla \ln P(D|w, X) \Big|_{w=w_{MAP}} - \nabla \nabla \ln P(w|\alpha) \Big|_{w=w_{MAP}}$$

$$= H + \alpha I \quad \leftarrow \begin{aligned} -\nabla \nabla \ln P(w|\alpha) &= -\nabla \nabla \ln N(w|0, \alpha^{-1} I) \\ &= -\nabla \nabla \left(-\frac{\alpha}{2} w^T w \right) = \alpha I \end{aligned}$$

$$\frac{\partial w^T w}{\partial w} = \frac{\partial (w^T I w)}{\partial w} \quad (L.1.1)$$

$$= \left(\frac{\partial (w^T I w)}{\partial w^T} \right)^T = (w^T (I + I))^T$$

$$= 2w$$

2" 子" 子。 $T \in \mathbb{R}^L$, $H = -\nabla \nabla \ln P(D|w, X) \Big|_{w=w_{MAP}}$ 子" 子

H の固有値は λ_i ($i=1 \sim W$) 子" 子

A の固有値は $\lambda_i + \alpha$ ($i=1 \sim W$) 子" 子

特徴 (2) w_i 子" 子
 $H u_i = \lambda_i u_i$
 $(\alpha I) u_i = \alpha u_i$
 $\therefore (H + \alpha I) u_i = (\lambda_i + \alpha) u_i$

$$\therefore \frac{\partial}{\partial w} \frac{\partial}{\partial w} w^T w = \frac{\partial}{\partial w} 2w$$

$$= 2 \begin{pmatrix} \frac{\partial w_1}{\partial w_1} & \frac{\partial w_1}{\partial w_2} \\ \frac{\partial w_2}{\partial w_1} & \frac{\partial w_2}{\partial w_2} \end{pmatrix} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 2I$$

5.7

$$\frac{\partial}{\partial \alpha} \ln |A| = \frac{\partial}{\partial \alpha} \ln \prod_{i=1}^W (\lambda_i + \alpha) \quad (L.47)$$

$$= \frac{\partial}{\partial \alpha} \sum_{i=1}^W \ln (\lambda_i + \alpha) = \sum_{i=1}^W \frac{1}{\lambda_i + \alpha}$$

二枚を最大化する α は

$$0 = -\frac{1}{2} w_{MAP}^T w_{MAP} - \frac{1}{2} \sum_{i=1}^W \frac{1}{\lambda_i + \alpha} + \frac{W}{2\alpha}$$

$$\therefore 0 = -\alpha w_{MAP}^T w_{MAP} - \sum_{i=1}^W \frac{\alpha}{\lambda_i + \alpha} + \sum_{i=1}^W 1$$

$$= -\alpha w_{MAP}^T w_{MAP} + \sum_{i=1}^W \frac{\lambda_i}{\lambda_i + \alpha}$$

$$\therefore \alpha = \frac{\gamma}{w_{MAP}^T w_{MAP}} \quad \dots (5.178)$$

子" 子。 $T \in \mathbb{R}^L$

$$\gamma = \sum_{i=1}^W \frac{\lambda_i}{\lambda_i + \alpha} \quad \dots (5.179)$$

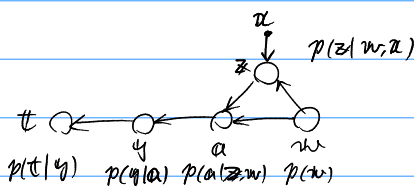
2" 子" 子。

(確率モデル)

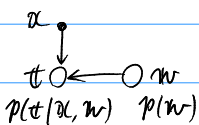
参考までに8章の確率モデルで、この問題のモデルを表現して様子。
 対応、二つのモデルは8章の確率モデルとは別のものである。

二つのモデルは確率変数と混合。矢印はト向の確率的依存関係を表現している。
 確率モデルのト向は確率変数である。矢印はト向の確率的依存関係を表現する。
 この問題モデルに記した確率変数は t, t_n, w, z, y である。 α の分布は考えない、 α は定数と見做す。

t, α, y, a, z, w の確率モデル図

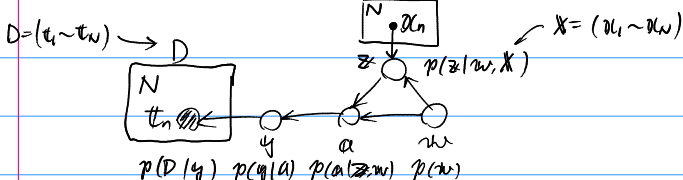


と様子。① $p(t|\alpha, w)$ のモデル図

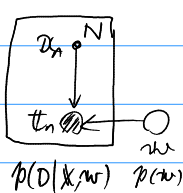


と様子。

$t_n, \alpha_n, y, a, z, w$ の確率モデル図



と様子。② $p(D|w, X)$ のモデル図



と様子。

確率的依存の仕方はいくつかあり (wikiF4)

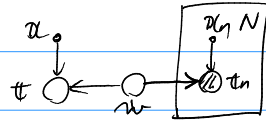
- 確率変数変換 $y = ax + b$ 等
Transform of variable
- 確率変数の関数 $z = x + y$ 等
Function of variable
- 混合(合成)分布
Compound distribution

分布の1つは x の確率変数 $z = ax + b$ 等

$$p(y|z) = N(y|f(z), \sigma^2) \text{ 等}$$

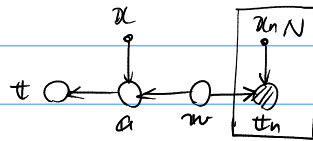
混合分布の例) $p(x) = \sum_k \pi_k p_k(x)$
 等 ($\sum \pi_k = 1$)

予測分布を求めようとするのモデル図は

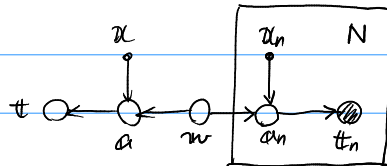


である。

失敗したが、 α と線形近似して予測分布を求めようとするのモデル図は



である。訓練時の α_n も含むモデル図は



とできる。