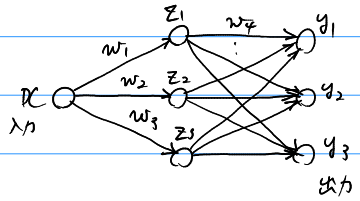
5.40

・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 t 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 t (1 of K表記)の条件付分布は

$$p(t|x,w) = \prod_{k=1}^{K} y_k^{t_k}(x,w) \quad \cdots ①$$

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

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

対数尤度は

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

となる。

(注)教科書では条件付き確率の x, x_n, X の表記は省略されている。

(パラメータ w の推定)

ベイズモデルなので 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)

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

(D, X は観測値に固定されているので $p(D|X)$ は定数である)

なので対数事後分布は

$$\ln p(w|D,X) = \ln p(D|w,X) + \ln p(w) + C, \quad (C \text{ は } w \text{ によらない定数})$$

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

となる。

w の事前分布を等方ガウス分布(5.162)とすると

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

となる。

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

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

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

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。

勾配降下法で必要な $\frac{\partial E}{\partial w}$ の計算は5.3節の誤差逆伝播で行うことができる。

（予測分布）

新しい x に対する t の予測分布は

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

$$= \int \prod_{k=1}^{K} y_k^{t_k}(x,w)\,p(w|D,\alpha)\,dw$$

で与えられる。

このまま積分するのは難しいので近似する。

（近似 1）

(5.185) と同様に事後分布を $p(w|D,\alpha) \simeq \delta(w - w_{MAP})$ と近似して

$$p(t|x,D,\alpha) = \int \prod_{k=1}^{K} y_k^{t_k}(x,w)\,p(w|D,\alpha)\,dw$$

$$\simeq \int \prod_{k=1}^{K} y_k^{t_k}(x,w)\,\delta(w - w_{MAP})\,dw = \prod_{k=1}^{K} y_k^{t_k}(x,w_{MAP})$$

を得る。

（近似 2）

5.7.3節の (5.190) と同様の近似は出来ないことの説明

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

$$a_k(x,w) \simeq a_k^{MAP}(x) + b_k^{T}(w - w_{MAP})$$

ただし $a_k^{MAP}(x) = a_k(x,w_{MAP})$, $b_k = \nabla a_k(x,w)\big|_{w=w_{MAP}}$ とする。

このときネットワークの出力 y_k は

$$y_k(x,w) \simeq \frac{\exp\left(a_k^{MAP}(x) + b_k^{T}(w - w_{MAP})\right)}{\sum_j \exp\left(a_j^{MAP}(x) + b_j^{T}(w - w_{MAP})\right)}$$

となる。

これより予測分布は

$$p(t=(0\cdots\underset{\uparrow}{1}\cdots0)|x,D,\alpha) = \int y_k(x,w)\, p(w|D,\alpha)\, dw$$

（$\uparrow$ k番目）

a_kを線形近似

$$\simeq \int \frac{\exp\left(a_k^{MAP}(x) + b_k^T(w-w_{MAP})\right)}{\sum_j \exp\left(a_j^{MAP}(x) + b_j^T(w-w_{MAP})\right)}\, p(w|D,\alpha)\, dw$$

$p(w|D,\alpha)$のうつうう近似

$$\simeq \int \underbrace{\frac{\exp\left(a_k^{MAP}(x) + b_k^T(w-w_{MAP})\right)}{\sum_j \exp\left(a_j^{MAP}(x) + b_j^T(w-w_{MAP})\right)}}_{③}\, q(w|D,\alpha)\, dw$$

となる。

③が$w^T b_k$を変数とするシグモイド関数になっていないので、(4.144)と同じ形と見なせない。

よって4.5.2節の結果を利用して積分を得ることができない。

5.7.3節では、③の部分が$\sigma(a^{MAP}(x) + b^T(w-w_{MAP}))$となっていたので

4.5.2節の(4.145)の積分と同じ形と見なすことができ、

(4.148)、(4.150)、(4.153)を用いて(5.190)の積分を得ることができた。

(超パラメータ α の推定)

α のエビデンス関数は

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

である。右辺の $f(w) = p(D|w, X)\, p(w|\alpha)$ をラプラス近似する。← ・根拠はない ラプラス近似がよい近似になると仮定しているだけ。

このとき、(4.135) より

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

となる。ただし W は w の次元で、$A = -\nabla\nabla \ln f(w)\big|_{w=w_{MAP}}$ (4.132) である。

よって

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

ここで

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

上記 $E(w)$ より　(5.162)

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

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

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

なので

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

を得る。

これを最大化するαは

$$0 = \frac{\partial}{\partial \alpha} \ln p(\mathbf{t}|\alpha, \beta) = \frac{\partial}{\partial \alpha}\left(-E(\mathbf{m}_{MAP}) - \frac{1}{2}\ln|\mathbf{A}| + \frac{W}{2}\ln\alpha\right)$$

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

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

で与えられる。ここで

$$\mathbf{A} = -\nabla\nabla\ln f(\mathbf{w})\big|_{\mathbf{w}=\mathbf{m}_{MAP}} = -\nabla\nabla\ln p(\mathbf{t}|\mathbf{w},\beta)\big|_{\mathbf{w}=\mathbf{m}_{MAP}} - \nabla\nabla\ln p(\mathbf{w}|\alpha)\big|_{\mathbf{w}=\mathbf{m}_{MAP}}$$

$$= \mathbf{H} + \alpha\mathbf{I}$$

$-\nabla\nabla\ln p(\mathbf{w}|\alpha) = -\nabla\nabla\ln\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = -\nabla\nabla\left(-\frac{\alpha}{2}\mathbf{w}^{T}\mathbf{w}\right) = \alpha\mathbf{I}$

$\frac{\partial \mathbf{w}^{T}\mathbf{w}}{\partial \mathbf{w}} = \frac{\partial \mathrm{Tr}(\mathbf{w}^{T}\mathbf{I}\mathbf{w})}{\partial \mathbf{w}}$ (1.117)

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

$= 2\mathbf{w}$

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

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

である。ただし、$\mathbf{H} = -\nabla\nabla\ln p(\mathbf{t}|\mathbf{w},\beta)\big|_{\mathbf{w}=\mathbf{m}_{MAP}}$ とする。

$\mathbf{H}$ の固有値を λ_i $(i=1\sim W)$ とすると

$\mathbf{A}$ の固有値は $\lambda_i+\alpha$ $(i=1\sim W)$ となる。

固有値は行列と定まる

$\mathbf{H}\mathbf{u}_i = \lambda_i\mathbf{u}_i$

$(\alpha\mathbf{I})\mathbf{u}_i = \alpha\mathbf{u}_i$

$\therefore (\mathbf{H}+\alpha\mathbf{I})\mathbf{u}_i = (\lambda_i+\alpha)\mathbf{u}_i$

よって

$$\frac{\partial}{\partial\alpha}\ln|\mathbf{A}| = \frac{\partial}{\partial\alpha}\ln\prod_{i=1}^{W}(\lambda_i+\alpha) \overset{(1.47)}{=} \frac{\partial}{\partial\alpha}\sum_{i=1}^{W}\ln(\lambda_i+\alpha) = \sum_{i=1}^{W}\frac{1}{\lambda_i+\alpha}$$

これを上の式に入れて

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

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

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

$$\therefore\ \alpha = \frac{\gamma}{\mathbf{m}_{MAP}^{T}\mathbf{m}_{MAP}} \quad \cdots (5.178)$$

を得る。ただし

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

である。

(確率モデル図)

今までに8章の確率モデル図で、この問題のモデル図を書いてみる。

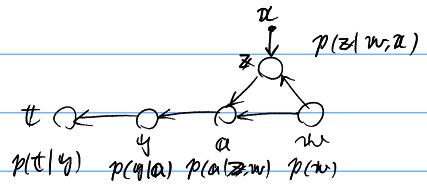
なお、ニューラルネットワークは8章の確率モデル図とは別のものである。

ニューラルネットワークのノードは確率変数とは限らない。矢印はノード間の確率的依存関係を表現しない。

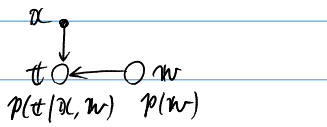
確率モデル図のノードは確率変数である。矢印はノード間の確率的依存関係を表現する。

この問題のモデルにおいて確率変数は t, t_n, w, z, y である。x の分布は考えない、x は定数としている。

t, x, y, a, z, w の確率モデル図は

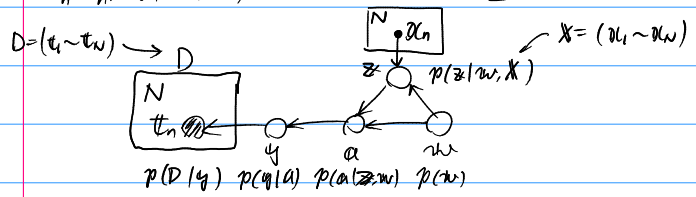


となる。① $p(t|x,w)$ のモデル図は

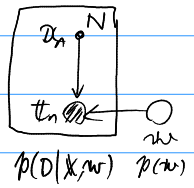


となる。

t_n, x_n, y, a, z, w の確率モデル図は



となる。② $p(D|w,X)$ のモデル図は

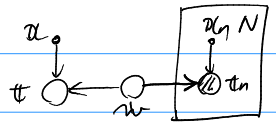


となる。

確率的依存の仕方はいくつかある (wikiFII)

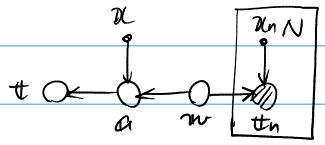
- 確率変数変換 $y=ax+b$ とか
  Transform of variable
- 確率変数の関数 $z=x+y$ とか
  Function of variable
- 混合(合成)分布
  Compound distribution
  分布のパラメータが確率変数になること
  $p(y|x)=N(y|f(x),\sigma^2)$ とか
  混合分布というと $p(x)=\sum_k \pi_k p_k(x)$
  を指すことである。

予測分布を求めるときのモデル図は

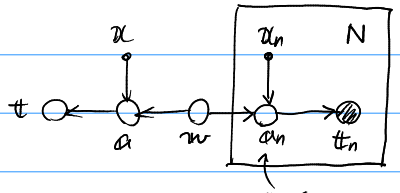


である。

失敗したが、αを線形近似して予測分布を求めるときのモデル図は



である。訓練時の a_n も含めたモデル図は



となる。