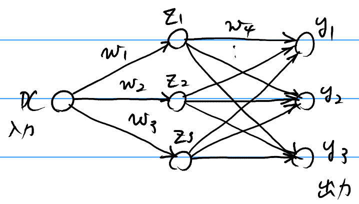
5.40

・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 $\mathbf{t}$ 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

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

目標変数 $\mathbf{t}$ (1 of K表記)の条件付分布は

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

となる。このとき $D = \{\mathbf{t}_1, \mathbf{t}_2, \cdots\}$, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots\}$, $y_{nk} = y_k(\mathbf{x}_n, \mathbf{w})$ とおくと尤度は

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

対数尤度は

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

となる。

(注) 教科書では条件付き確率の $\mathbf{x}, \mathbf{x}_n, \mathbf{X}$ 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の事後分布は

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

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

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

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

となる。

wの事前分布を等方ガウス分布(5.162)とすると

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

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

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

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

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

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

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$p(w|D,X) \simeq \delta(w - w_{MAP})$と近似して

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

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

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

となる。

これより x に対して $t=(0\cdots\underset{k\text{番目}}{1}\cdots0)$ となる確率密度は

$$p(t=(0\cdots\underset{k\text{番目}}{1}\cdots0)|x,D,\mathbb{X}) = \int y_k(x,w)\,p(w|D,\mathbb{X})\,dw$$

a_k を線形近似

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

$p(w|D,\mathbb{X})$ のラプラス近似

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

となる。

③ が w^Tb_k を変数とするシグモイド関数になっていないので、(4.145) 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

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

$$= \underbrace{-E(w_{MAP}) + \frac{\alpha}{2} w_{MAP}^T w_{MAP}}_{\text{上記 } E(w) \text{ より}} + \underbrace{\ln N(w_{MAP}|0, \alpha^{-1} I)}_{(5.162)}$$

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

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

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

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

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

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

$$\frac{\partial E(w_{MAP})}{\partial \alpha} = \frac{1}{2} w_{MAP}^T w_{MAP}$$

で与えられる。ここで

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

$$= H + \alpha I$$

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

$$\frac{\partial w^T w}{\partial w} = \frac{\partial Tr(w^T I w)}{\partial w}$$
$$= \left(\frac{\partial Tr(w^T I w)}{\partial w^T}\right)^T \overset{(C.27)}{=} \left(w^T (I + I^T)\right)^T$$
$$= 2w$$
$$\therefore \frac{\partial}{\partial w}\frac{\partial}{\partial w} w^T w = \frac{\partial}{\partial w} 2w$$
$$= 2\begin{pmatrix} \frac{\partial w_1}{\partial w_1} & \frac{\partial w_2}{\partial w_1} \\ \frac{\partial w_1}{\partial w_2} & \frac{\partial w_2}{\partial w_2} \end{pmatrix} = 2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 2I$$

である。ただし、$H = -\nabla\nabla \ln p(D|w, X)\Big|_{w=w_{MAP}}$ とする。

Hの固有値を $\lambda_i \ (i = 1 \sim W)$ とすると

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

縮退しててもOK? 固有値のおかげ?

$H u_i = \lambda_i u_i$

$(\alpha I) u_i = \alpha u_i$

$\therefore (H + \alpha I) u_i = (\lambda_i + \alpha) u_i$

よって

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

これを上の式に入れて

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

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

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

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

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)

参考までに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。

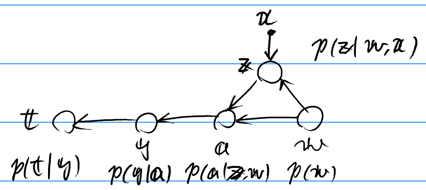
なお、ニューラルネットワークは8章の確率モデル図とは別のものである。

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。
また、各ノードに条件付分布を割り当てない。同時分布を与えない。

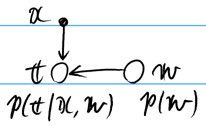
確率モデル図のノードは確率変数である。矢印はノード間の確率的依存関係を表現する。
また、各ノードに条件付分布を割り当てる。各ノードの分布の積が同時分布を与える

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

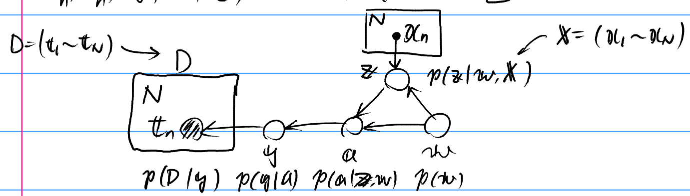


となる。① $p(t|x,w)$ のモデル図は

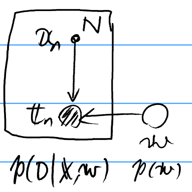


となる。

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。

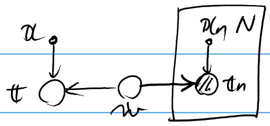
確率的依存の仕方はいくつかある (wikiより)

・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 variables

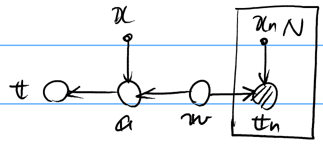
・混合(合成)分布
Compound distribution
分布のパラメータが確率変数になること
$p(y|x) = N(y|f(x), \sigma^2)$ とか
混合分布というと $p(x) = \sum_k \pi_k p_k(x)$
を指すこともある

予測分布を求めるときのモデル図は

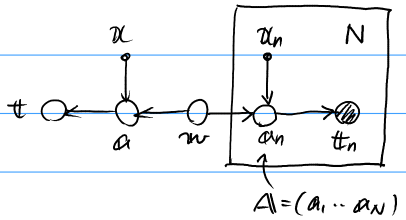


である。

失敗したが、a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anも含めたモデル図は



となる。