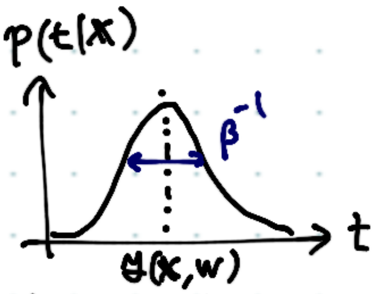
5.7 ベイズニューラルネットワーク

5.7.1 パラメータの事後分布

NNによる1次元の回帰をベイズ化

1次元・連続な目標変数：t $\xleftarrow{\text{予測}}$ 入力ベクトル$\mathbf{x}$

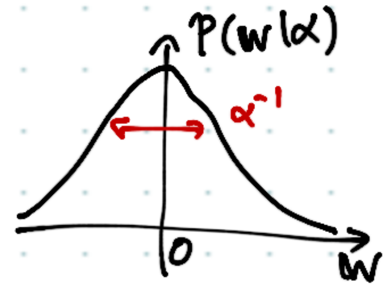


tの条件付き確率 $p(t|\mathbf{x})$は

$$p(t|\mathbf{x},\mathbf{w},\beta)=\mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \quad (5.161)$$

$$\exp\left(-\frac{\beta}{2}\sum_n\{y_n-t_n\}^2\right)$$

$y(\mathbf{x},\mathbf{w})$：NNの出力



重みに関する事前分布を

$$p(\mathbf{w}|\alpha)=\mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) \quad (5.162)$$

とする.

$$\exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right)$$

N回の観測値 $\mathbf{x}_1,\mathbf{x}_2,\cdots\mathbf{x}_N$ | 目標集合 $\mathcal{D}=\{t_1,\ldots,t_N\}$

尤度関数

$$p(\mathcal{D}|\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n|y(\mathbf{x}_n,\mathbf{w}),\beta^{-1}) \quad (5.163)$$

事後分布は

$$p(\mathbf{w}|\mathcal{D},\alpha,\beta)\propto p(\mathbf{w}|\alpha)p(\mathcal{D}|\mathbf{w},\beta) \quad (5.164)$$

$y(\mathbf{x},\mathbf{w})$は $\mathbf{w}$に非線形に依存 $\Rightarrow$ (5.164)はガウス分布ではない.

なのでラプラス近似でガウス分布に近似しよう.

事後分布 $P(w|D,\alpha,\beta)$ (5.164)を最大化しよう。

いつも通り、事後分布の対数をとって

$$\ln P(w|D) = -\frac{\alpha}{2}w^T w - \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2 + 定数 \quad (5.165)$$

これを(局所)最大化する。これは、今まで行なってきた最適化アルゴリズムにより、事後分布の最大値 W_{MAP} を見つけることができる。

を見つけたら事後分布の負の対数尤度の2階微分の行列を評価して、局所的にガウス分布で近似できる。(5.165)より、

これのこと

$$A = -\nabla\nabla \ln p(w|D,\alpha,\beta) = \alpha I + \beta H \quad (5.166)$$

ここで H は二乗和誤差関数 $E=\frac{1}{2}\sum\{y_n-t_n\}^2$ の w の2階微分からなるヘッセ行列である。⇐ これを今までやった近似で求める。

この A を利用して事後分布を近似したガウス分布は(4.134)より

$$q(w|D) = \mathcal{N}(w|w_{MAP}, A^{-1}) \quad (5.167)$$

$$A = -\nabla\nabla \ln f(z)\big|_{z=z_0} \quad \text{P.215}$$

$$q(z) = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}}\exp\left[-\frac{1}{2}(z-z_0)^T A(z-z_0)\right]$$

$$= \mathcal{N}(z|z_0, A^{-1}) \cdots (4.134)$$

この事後分布について周辺化することで予測分布が得られ

$$p(t|x,D) = \int p(t|x,w)q(w|D)dw \cdots (5.168)$$

となる。

しかし $y(x,w)$ が w に対して非線形なのでまだこの積分はあつかいづらい。

なので事後分布の分散が $y(x,w)$ が変化する w の特性スケールに比べて小さいと仮定する。

ネットワーク関数 $y(\mathbf{x},\mathbf{w})$ の $\mathbf{w}_{MAP}$ まわりでのテイラー展開で1次までみる

$$y(\mathbf{x},\mathbf{w}) \simeq y(\mathbf{x},\mathbf{w}_{MAP}) + \mathbf{g}^T(\mathbf{w}-\mathbf{w}_{MAP}) \quad (5.196)$$

$$\mathbf{g} = \nabla_{\mathbf{w}} y(\mathbf{x},\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}_{MAP}}$$

よって条件付き確率 $p(t|\mathbf{x},\mathbf{w},\beta)$ (5.161) は

$$p(t|\mathbf{x},\mathbf{w},\beta) \simeq \mathcal{N}(t|y(\mathbf{x},\mathbf{w}_{MAP})+\mathbf{g}^T(\mathbf{w}-\mathbf{w}_{MAP}),\beta^{-1}) \quad (5.171)$$

で表される。

したがって一般に成り立つ (2.115) p.90 を周辺分布 $p(t)$ に利用して

$$q(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\underbrace{\mathbf{w}_{MAP}}_{\mu},\underbrace{\mathbf{A}^{-1}}_{\Lambda^{-1}}) \quad (5.167)$$

$$p(t|\mathbf{x},\mathbf{w},\beta) \simeq \mathcal{N}(t|\underbrace{\mathbf{g}^T}_{A}\underbrace{\mathbf{w}}_{x}\underbrace{-\mathbf{g}^T\mathbf{w}_{MAP}+y(\mathbf{x},\mathbf{w}_{MAP})}_{+b},\underbrace{\beta^{-1}}_{L^{-1}})$$

$$p(x) = \mathcal{N}(x|\mu,\Lambda^{-1})$$

$$p(y|x) = \mathcal{N}(y|Ax+b, L^{-1})$$

周辺分布

$$p(y) = \mathcal{N}(y|A\mu+b, L^{-1}+A\Lambda^{-1}A^T) \quad \cdots (2.115)$$

右の (2.115) から周辺分布の平均 $A\mu+b$ は

$$A\mu+b = \mathbf{g}^T\cancel{\mathbf{w}_{MAP}} - \cancel{\mathbf{g}^T\mathbf{w}_{MAP}} + y(\mathbf{x},\mathbf{w}_{MAP})$$

分散 $L^{-1}+A\Lambda^{-1}A^T$ は

$$L^{-1}+A\Lambda^{-1}A^T = \beta^{-1} + \mathbf{g}^T\mathbf{A}^{-1}\mathbf{g} =: \sigma^2(\mathbf{x}) \quad (5.173)$$

よって周辺分布（予測分布）$p(t|\mathbf{x},\mathcal{D})$ は

$$p(t|\mathbf{x},\mathcal{D},\alpha,\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}_{MAP}),\sigma^2(\mathbf{x})) \quad (5.172)$$

$$\text{分散} \quad \sigma^2(\mathbf{x}) = \underbrace{\beta^{-1}}_{\text{目標変数の内在的なノイズ}} + \underbrace{\mathbf{g}^T\mathbf{A}^{-1}\mathbf{g}}_{\mathbf{x}\text{に依存，モデルパラメータ}\mathbf{w}\text{による不確定性に起因}} \quad (5.173)$$

5.7.2 超パラメータ最適化

ハイパーパラメータ α, β は固定して既知であるとしていた。

$p(t|x)$ の精度　　事前分布 $p(w|\alpha)$ の精度

今やった(5.7.1)近似で3.5節のエビデンス理論を使おう!!

ハイパーパラメータの周辺尤度あるいはエビデンスは

$$p(D|\alpha,\beta) = \int p(D|w,\beta)\,p(w,\alpha)\,dw \quad (5.174)$$

で得られる。

$$Z \cdot p(w|D,\alpha,\beta)$$

Z：正規化係数　　$p(w|D,\alpha,\beta)$ → (5.165)を使う。

↓

P.216のラプラス近似の結果(4.135)をつかう。

$$\frac{1}{Z} p(D|w,\beta)\,p(w,\alpha) = p(w|D,\alpha,\beta)$$

$$\Rightarrow Z = \int \underbrace{p(D|w,\beta)\,p(w,\alpha)}_{f(z)\text{に対応}}\,dw$$

$$Z \simeq f(z_0)\frac{(2\pi)^{M/2}}{|A|^{\frac{1}{2}}} \quad (4.135)$$

$p(w|D,\alpha,\beta)$ はラプラス近似で

$$q(w|D) = \mathcal{N}(w|w_{MAP}, A^{-1}) \cdots (5.167)$$

なので平均は w_{MAP} を使う。

$$p(D|\alpha,\beta) = Z\underbrace{\int p(w|D,\alpha,\beta)\,dw}_{1}$$

$$= Z$$

$$= p(D|w_{MAP},\beta)\,p(w_{MAP},\alpha)\frac{(2\pi)^{\frac{W}{2}}}{|A|^{\frac{1}{2}}}$$

$$p(D|\mathbf{w}_{MAP},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n|y(\mathbf{x}_n,\mathbf{w}_{MAP}),\beta^{-1}) \quad (5.163)\text{より}$$

$$=\prod_{n=1}^{N}\frac{1}{(2\pi)^{\frac{1}{2}}}\frac{1}{(\beta^{-1})^{\frac{1}{2}}}\exp(\sim)$$

$$p(\mathbf{w}_{MAP},\alpha)=\mathcal{N}(\mathbf{w}|0,\alpha^{-1}I) \quad (5.162)\text{より}$$

$$=\frac{1}{(2\pi)^{\frac{W}{2}}}\frac{1}{|\alpha^{-1}I|^{\frac{1}{2}}}\exp(\sim)$$

$$\downarrow$$

$$\alpha^{-\frac{W}{2}}$$

よって

$$\ln p(D|\alpha,\beta)=-E(\mathbf{w}_{MAP})-\frac{1}{2}\ln|A|+\frac{W}{2}\ln\alpha+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi) \quad (5.175)$$

$$E(\mathbf{w}_{MAP})=\frac{\beta}{2}\sum_{n=1}^{N}\{y(\mathbf{x}_n,\mathbf{w}_{MAP})-t_n\}^2+\frac{\alpha}{2}\mathbf{w}_{MAP}^T\mathbf{w}_{MAP} \quad (5.176)$$

エビデンス理論では、$\ln p(D|\alpha,\beta)$を最大化してαとβで点推定を行う.

αについては 3.5.2節の 線形回帰 モデルの場合と同様にして行う.

$$A=-\nabla\nabla\ln p(\mathbf{w}|D,\alpha,\beta)=\alpha I+\beta \mathbf{H} \quad \cdots (5.166)$$

の固有値 $(\mathbf{w}=\mathbf{w}_{MAP})$ を考えるために固有値方程式

$$\beta\mathbf{H}|_{\mathbf{w}=\mathbf{w}_{MAP}}\mathbf{u}_i=\lambda_i\mathbf{u}_i \quad \cdots (5.177)$$

を考える.

これを$\ln p(D|\alpha,\beta)$の$\ln|A|$の評価で使う. $|A|=\prod_i(\lambda_i+\alpha)$

すると $\ln p(D|\alpha,\beta)$ を最大化する α は (3.92) と同様にして

$$\alpha = \frac{\gamma}{w_{MAP}^T w_{MAP}} \qquad (5.178)$$

有効パラメータ数 (3.5.3節) → $\gamma = \sum_{i=1}^{W} \frac{\lambda_i}{\alpha + \lambda_i}$ (5.179)

ただし今回の場合はこれは厳密でない。(線形回帰では厳密.)

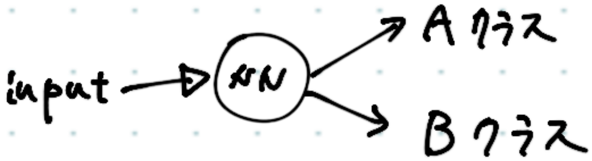
[H が α によって変わる → λ_i が α に依存するので]
計算の途中で
λ_i の α 微分をムシしている.

同様に (5.177) から $\ln|A|$ を評価して、(3.95) より β の再推定の式は

$$\frac{1}{\beta} = \frac{1}{N-\gamma} \sum_{n=1}^{N} \left[y(x_n, w_{MAP}) - t_n \right]^2 \quad \cdots \quad (5.180)$$

5.7.3 クラス分類のためのベイズNN

1つのロジスティックシグモイド出力を持つ2クラス分類を考える.



このモデルの対数尤度関数は

$$\ln p(D|w) = \sum_{n=1}^{N} \{ t_n \ln y_n + (1-t_n) \ln (1-y_n) \} \quad (5.181)$$

$t_n = \begin{cases} 0 \\ 1 \end{cases}$, $y_n = y(x_n, w)$ ← NNが出力する分類の確率

$$\left(p(D|w) = \prod_{n=1}^{N} y_n^{t_n} \{1-y_n\}^{1-t_n} \quad \cdots (4.89) \right)$$

データは正しくラベル付けされているとしてハイパーパラメータβは現れない.
事前分布を $p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) \cdots (5.162)$ とする.

1. まずαを初期化する.
2. そして対数事後分布 $p(w|D\alpha) \propto p(w|\alpha)\, p(D|w)$
   を最大化してパラメータベクトルwを決める.
   これは正則化誤差関数

$$E(w) = -\ln p(D|w) + \frac{\alpha}{2} w^T w \quad \cdots (5.182)$$

を最小化することと等価でありBackpropergationなどでできる.

3. w_{MAP}をみつけたら 負の対数尤度関数の2階微分からなるヘッセ行列Hを評価する.

4. 事後分布のガウス近似は $q(w|D)=\mathcal{N}(w|w_{MAP},A^{-1})$ (5.167) で与えられる.

5. ハイパーパラメータαを最適化するには、やはり周辺尤度を最大化する.
(5.175)を導出したときと同様にして

$$\ln p(D|\alpha)\simeq -E(w_{MAP})-\frac{1}{2}\ln|A|+\frac{W}{2}\ln\alpha \quad \cdots (5.183)$$

$$E(w_{MAP})=-\sum_{n=1}^{N}\{t_n\ln y_n+(1-t_n)\ln(1-y_n)\}+\frac{\alpha}{2}w_{MAP}^{T}w_{MAP} \cdots (5.184)$$

$$y_n\equiv y_n(x_n,w_{MAP})$$

(5.183)をαについて最大化すると再び $\alpha=\dfrac{\gamma}{w_{MAP}^{T}w_{MAP}}$ (5.178)で与えられた再推定方程式が導かれる.

6. 最後に(5.168)で定義される予測分布が必要である.
この積分もネットワークの非線形性からムズかしい.

事後分布の幅がせまいとして

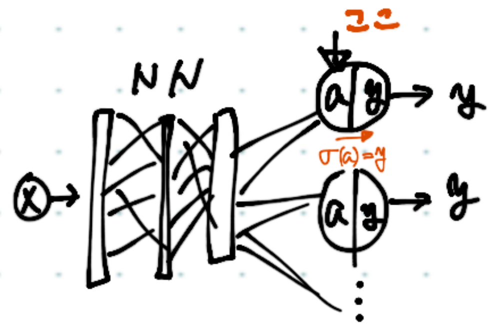
$$p(t|x,D)\simeq p(t|x,w_{MAP})$$

とするか、分散をさらに考慮するかで近似する.
しかし出力がロジスティックシグモイド活性化関数であることから出力は(0,1)に制約されるため、回帰のときのように(5.169)みたいな線形近似はよくない.
なので$\sigma(x)$の1歩手前の出力ユニット活性を線形近似する.

$$a(x, w) \simeq a_{MAP}(x) + b^T(w - w_{MAP})$$

$a_{MAP}(x) = a(x, w_{MAP})$, $b \equiv \nabla a(x, w_{MAP})$

は逆伝播により求められる.



予測分布の計算の節

7. 今までの結果を 4.5.2 節の結果にあてはめることができる.

出力ユニットの活性の値の分布は

$$p(a|x, D) = \int \delta(a - a_{MAP}(x) - b^T(x)(w - w_{MAP}))\, q(w|D)\, dw \quad \cdots (5.189)$$

$$q(w|D) \propto \exp\left[-\frac{1}{2}(w - w_{MAP})^T A (w - w_{MAP})\right]$$

▷ 平均 a_{MAP}, 分散 $\sigma_a^2(x) = b^T A^{-1} b$

のガウス分布になる.

予測分布を得るために a に関して周辺化する.

$$p(t=1|x, D) = \int \sigma(a)\, p(a|x, D)\, da \quad (5.189)$$

けど σ をもつガウス分布のたたみ込みはムズかしい.

(4.153) の近似を (5.189) に使って

$$p(t=1|x, D) = \sigma(K(\sigma_a^2)\, a_{MAP}) \quad (5.190)$$

$$K(\sigma^2) = (1 + \pi\sigma^2/8)^{-\frac{1}{2}} \quad (4.153)$$