パターン認識と機械学習
第5章 ニューラルネットワーク

橋本龍二

February 5, 2021

1　フィードフォワード

省略。

2　ネットワーク訓練

入力ベクトル $\boldsymbol{x}$ に対し、一次元目標ベクトル t の従う分布を

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

と置く。入力ベクトルの集合 $\{\boldsymbol{x}_n\}$ $(n = 1, ..., N)\}$ と対応する目標ベクトルの集合 $\{t_n\}$ $(n = 1, ..., N)$ が与えられたときの対数尤度関数は、

$$\begin{aligned} p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) &= \Pi_{n=1}^{N} p(t_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) \\ &= \Pi_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\{-\frac{\beta(t_n - y(\boldsymbol{x}_n, \boldsymbol{w}))^2}{2}\} \\ &= \beta^{N/2}(2\pi)^{-N/2} \exp\{-\frac{\beta}{2}\sum_{n=1}^{N}(y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n)^2\} \\ \therefore -\log p &= \frac{\beta}{2}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2 - \frac{N}{2}\log\beta + \frac{N}{2}\log(2\pi) \quad (1) \end{aligned}$$

より、尤度関数を最大化する $\boldsymbol{w}_{ML}$ は、誤差関数：

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n) - t_n\}^2$$

を最小化することによって得られる。$\boldsymbol{w}_{ML}$ が求まれば、β_{ML} は (1) より

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n, \boldsymbol{w}_{ML}) - t_n\}^2$$

と求まる。同様に、目標ベクトルを K 次元としてその分布を

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

と置くと、データ N 個の対数尤度関数は

$$\begin{aligned} p(\boldsymbol{t}|X, \boldsymbol{w}, \beta) &= \Pi_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) \\ &= \Pi_{n=1}^{N} \frac{\beta^{K/2}}{(2\pi)^{K/2}} \exp\{-\frac{\beta}{2}\|y(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}_n\|^2\} \\ &= \beta^{NK/2}(2\pi)^{-K/2} \exp\{-\frac{\beta}{2}\sum_{n=1}^{N}\|y(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}_n\|^2\} \\ \therefore -\log p &= \frac{\beta}{2}\sum_{n=1}^{N}\|y(\boldsymbol{x}_n - \boldsymbol{w}) - \boldsymbol{t}_n\|^2 - \frac{NK}{2}\log\beta - \frac{K}{2}\log(2\pi) \quad (2) \end{aligned}$$

より、尤度関数を最大化する $\boldsymbol{w}_{ML}$ は、誤差関数：

$$E(\boldsymbol{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\|y(\boldsymbol{x}_n - \boldsymbol{w}) - \boldsymbol{t}_n\|^2$$

を最小化することによって得られる。$\boldsymbol{w}_{ML}$ が求めれば、β_{ML} は (2) より

$$\frac{1}{\beta_{ML}} = \frac{1}{NK}\sum_{n=1}^{N}\|y(\boldsymbol{x}_n - \boldsymbol{w}) - \boldsymbol{t}_n\|^2$$

と求まる。
次に、二クラス分類を考える。目標変数 t は $t = 1$ でクラス $\mathcal{C}_1$、$t = 0$ でクラス $\mathcal{C}_2$ を表す。活性 a に対しネットワークの出力層は

$$y = \sigma(a) \equiv \frac{1}{1 + \exp(-a)}$$

なる活性化関数を持つ。今、入力 $\boldsymbol{x}$ が与えられたときの条件付き分布を

$$\begin{aligned} p(t|\boldsymbol{x}, \boldsymbol{w}) &= Ber(y) \\ &= y(\boldsymbol{x}, \boldsymbol{w})^t\{1 - y(\boldsymbol{x}, \boldsymbol{w})\}^{1-t} \end{aligned}$$

とすると、データ N 個の対数尤度関数は

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w}) &= \Pi_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{x}_n,\boldsymbol{w}) \\ &= \Pi_{n=1}^{N} (y_n)^{t_n}(1-y_n)^{1-t_n} \\ &= (y_n)^{\sum_{n=1}^{N} t_n}(1-y_n)^{\sum_{n=1}^{N} 1-t_n} \\ \therefore -\log p &= -\sum_{n=1}^{N}\{t_n \log y_n + (1-t_n)\log(1-y_n)\} \end{aligned}$$

ただし、$y_n = y(\boldsymbol{x}_n,\boldsymbol{w})$ と置いた。したがって尤度関数を最大化する $\boldsymbol{w}_{ML}$ は、交差エントロピー誤差：

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\{t_n \log y_n + (1-t_n)\log(1-y_n)\}$$

によって与えられる。このように、解くべき問題に応じてネットワークの出力層の活性化関数および誤差関数は自然に選択される。

2.1 パラメータ最適化

重み空間内で重みベクトル $\boldsymbol{w}$ を

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} + \Delta\boldsymbol{w}^{(\tau)}$$

の形で移動させていくことで、停留点：

$$\nabla E(\boldsymbol{w}) = 0$$

なる点を見つける。ただし、一般に誤差関数の重みパラメータへの依存性には高い非線形性があるため、重み空間内には停留点が数多く存在する。

2.2 局所二次近似

$E(\boldsymbol{w})$ を重み空間内のある点 $\tilde{\boldsymbol{w}}$ 周りでテイラー展開する。

$$E(\boldsymbol{w}) \simeq E(\tilde{\boldsymbol{w}}) + {}^{T}(\boldsymbol{w}-\tilde{\boldsymbol{w}})\frac{\partial E(\tilde{\boldsymbol{w}})}{\partial \boldsymbol{w}} + \frac{1}{2}\,{}^{T}(\boldsymbol{w}-\tilde{\boldsymbol{w}})H(\boldsymbol{w}-\tilde{\boldsymbol{w}})$$

ただし、

$$(H)_{ij} = \frac{\partial E(\tilde{\boldsymbol{w}})}{\partial w_i \partial w_j}$$

なるヘッセ行列。ゆえに、勾配の局所近似は

$$\nabla E = \frac{\partial E}{\partial \boldsymbol{w}} \simeq \frac{\partial E(\tilde{\boldsymbol{w}})}{\partial \boldsymbol{w}} + H(\boldsymbol{w} - \tilde{\boldsymbol{w}})$$

誤差関数の極小点 $\boldsymbol{w}^*$ 周りでの局所近似を考える。

$$\frac{\partial E(\boldsymbol{w}^*)}{\partial \boldsymbol{w}} = \boldsymbol{0}$$

より、

$$E(\boldsymbol{w}) \simeq E(\boldsymbol{w}^*) + \frac{1}{2}\,{}^T(\boldsymbol{w} - \boldsymbol{w}^*)H(\boldsymbol{w} - \boldsymbol{w}^*)$$

ヘッセ行列は対称より、その i 番目の固有値を λ_i、固有ベクトルを $\boldsymbol{u}_i$ と置くと $\boldsymbol{u}_i, \boldsymbol{u}_j$ $(i \neq j)$ は互いに直行する。$\boldsymbol{u}_i$ らは重み空間上の正規直交基底より、

$$\boldsymbol{w} - \boldsymbol{w}^* = \sum_i \alpha_i \boldsymbol{u}_i = U\boldsymbol{\alpha}$$

の形でかける。ゆえ、

$$\begin{aligned} {}^T(\boldsymbol{w} - \boldsymbol{w}^*)H(\boldsymbol{w} - \boldsymbol{w}^*) &= {}^T\boldsymbol{\alpha}\,{}^TUHU\boldsymbol{\alpha} \\ &= {}^T\boldsymbol{\alpha}\Lambda\boldsymbol{\alpha} = \sum_i \lambda_i \alpha_i^2 \end{aligned}$$

なので、誤差関数は

$$E(\boldsymbol{w}) = E(\boldsymbol{w}^*) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2 \tag{1}$$

という形で書ける。新しい座標系 $\{\boldsymbol{u}_i\}$ における α_i の二次関数→誤差関数の等高線は各軸がヘッセ行列の固有ベクトル $\boldsymbol{u}_i$ にそろった楕円になる。
正方行列 A について、全ての $\boldsymbol{v} \neq \boldsymbol{0}$ に対し ${}^T\boldsymbol{v}A\boldsymbol{v} > 0$ を満たす時、A は正定値であるという。A の固有ベクトルの集合 $\{\boldsymbol{u}_i\}$ は完全基底をなすので、

$$\boldsymbol{v} = \sum_i c_i \boldsymbol{u}_i = U\boldsymbol{c}$$

という形で書ける。したがって、

$$\begin{aligned} {}^T\boldsymbol{v}A\boldsymbol{v} &= {}^T\boldsymbol{c}\,{}^TUAU\boldsymbol{c} \\ &= {}^T\boldsymbol{c}\Lambda\boldsymbol{c} \\ &= \sum_i \lambda_i c_i^2 \end{aligned}$$

が得られる。したがって、全ての固有値 λ_i が正の時、またその時に限り行列 A は正定値になる。
(1) より、点 $\boldsymbol{w}^*$ が極小点である条件は

$$\sum_i \lambda_i \alpha_i^2 > 0$$

すなわち $\boldsymbol{w}^*$ で評価したヘッセ行列 H が正定値であることである。(逆なら極大)

2.3 勾配情報の利用

誤差曲面は $\frac{\partial E(\tilde{\boldsymbol{w}})}{\partial \boldsymbol{w}}$ と H で特定される。$\boldsymbol{w}$ の次元を W と置くと、パラメータの数は

$$W + W + (W-1) + ... + 1 = \frac{W(W+3)}{2}$$

より、二次近似の極小点の位置は $O(W^3)$ 個のパラメータに依存する。勾配情報 (W 次元) を利用すれば各勾配の評価は $O(W)$ ステップで実行可能。各表はで $O(W)$ 使うので、極小点は $O(W^2)$ で見つかる。

3 誤差逆伝播

3.1 誤差関数微分の評価

N 個のデータについての誤差関数は、各データに対応する誤差項の和：

$$E(\boldsymbol{w}) = \sum_{n=1}^{N} E_n(\boldsymbol{w})$$

として表される。一つの項に対する勾配 $\nabla E_n(\boldsymbol{w})$ を考えよう。ネットワークの出力 y_k が入力変数 x_i の線形和

$$y_k = \sum_i w_{ki} x_i$$

で、入力パターン n に対する誤差関数が

$$\begin{aligned} E_n &= \frac{1}{2}\sum_k (y_{nk} - t_{nk})^2 \\ &= \frac{1}{2}\sum_k \{y_k(\boldsymbol{x}_n, \boldsymbol{w}) - t_{nk}\}^2 \\ &= \frac{1}{2}\sum_k (\sum_i w_{ki} x_i - t_{nk})^2 \end{aligned}$$

という形を取る場合を考える。ただし、入力ベクトル $\boldsymbol{x}_n$ の i 番目の要素 $(\boldsymbol{x}_n)_i = x_i$ と置いた。この勾配は

$$\frac{\partial E_n}{\partial w_{ji}} = (y_{nj} - t_{nj})x_i$$

である。これをより多層なニューラルネットワークへと拡張する。ユニット j の活性 a_j は

$$a_j = \sum_i w_{ji} z_i$$

で書ける。ただし、z_i はユニット j と結合がある前層のユニット i からの出力で、w_{ji} はユニット i からユニット j への重みを表す。
ユニット j の出力 z_j は非線形活性化関数 h を用いて

$$z_j = h(a_j)$$

で表せる。E_n を w_{ji} で微分する。

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = \delta_j z_i \tag{1}$$

ただし、

$$\delta_j = \frac{\partial E_n}{\partial a_j}$$

と置いた。δ_j をユニット j における誤差という。(1) より、E_n の w_{ji} での微分は「ユニット j における誤差」×「ユニット i からの出力」の形で書けることがわかった。
δ について考える。多くの出力ユニット k で、

$$\delta_k = y_k - t_k$$

と書ける。例えば、出力層の活性化関数を恒等関数とする場合 (回帰)、$y_k = a_k$ であるから

$$\begin{aligned} E_n &= \frac{1}{2}(y_k - t_k)^2 \\ \therefore \frac{\partial E_n}{\partial a_k} &= y_k - t_k \end{aligned}$$

またロジスティックシグモイド関数を選ぶ場合 (二クラス分類)、

$$y_k = \frac{1}{1 + \exp(-a)}$$

であるから

$$\begin{aligned} E_n &= -\{t_k \log y_k + (1-t_k)\log(1-y_k)\} \\ \therefore \frac{\partial E_n}{\partial a_k} &= -(\frac{t_k}{y_k} - \frac{1-t_k}{1-y_k}) y_k(1-y_k) = y_k - t_k \end{aligned}$$

一層後の δ_k が分かっている時、隠れユニット j の δ_j は

$$\begin{aligned} \delta_j \equiv \frac{\partial E_n}{\partial a_j} &= \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \\ &= \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \\ &= \sum_k \delta_k w_{kj}\frac{\partial z_j}{\partial a_j} \\ &= h'(a_j)\sum_k \delta_k w_{kj} \end{aligned}$$

まとめ

- 入力データ n が得られた時、E_n の重み w_{ji} による微分は

  $$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = \delta_j z_j$$

  と書ける。

- 出力ユニット k では

  $$\delta_k = y_k - t_k$$

  が成り立つ。

- 一層前の δ_k が分かっていれば、δ_j は

  $$\delta_j = h'(a_j)\sum_k \delta_k w_{kj}$$

  で得られる。したがって、全ユニットの w_{ji} による E_n の重み w_{ji} が求まるため、それを並べたベクトルが勾配 ∇E_n である。