5.5.6 たたみ込みニューラルネットワーク

不変性を持つモデルを構築するもう1つのアプローチ.

⇒ NNの構造そのものに不変性を持たせる.

⇒ 特徴抽出処理自体もNNに組み込む.

5.5.7 ソフト重み共有

- 一定のグループに属する重みを等しくする制限で不変性を作っていた.

  ⇒ ただこれは先に制限の形が分かっていないとダメ

- 正則化項の導入に置きかえて,同じグループに属する重みが似た値を取りやすい状態にしてあげる. ⇒ ソフト重み共有 (soft weight sharing)

- グループ分け,グループの重みの平均,分散をすべて学習過程の一部として決定.

$$P(W)=\prod_i P(w_i)$$: 複数グループには混合ガウス分布を事前分布とすればよい

$$P(w_i)=\sum_{j=1}^{M}\pi_j N(w_i|\mu_i,\sigma_j^2)$$: π_j 混合係数

負の対数を取ると

$$\Omega(w) = -\sum_i \ln\left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)\right)$$

したがって誤差関数全体は

$$\widetilde{E}(w) = E(w) + \lambda\Omega(w)$$

学習では重みだけでなく混合モデルのパラメータ $\{\pi_j, \mu_j, \sigma_j\}$ についても最小化させる.

$\{\pi_j\}$ を事前分布とみなし、ベイズの定理より事後分布は (2.192)

$$\gamma_j(w) = \frac{\pi_j \mathcal{N}(w|\mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w|\mu_k, \sigma_k^2)}$$

ⓐ すると全体の誤差関数の重みに関する微分は

$$\frac{\partial}{\partial w_i}\Omega(w) = + \frac{\sum_j \pi_j \mathcal{N}(w_i|k)\frac{(w_i - \mu_j)}{\sigma_j^2}}{\sum_k \pi_k \mathcal{N}(w_i|k)}$$

$$\Rightarrow \frac{\partial \widetilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \sum_j \gamma_j(w_i)\frac{(w_i - \mu_j)}{\sigma_j^2} \qquad (5.141)$$

正則化項の影響は各重みを j 番目のガウス分布の中心へと引き寄せる.

その強さは事後分布 $\gamma_j(w_\cdot)$ に比例 ⇒ そのグループに属する確率

が高いほど中心に近づきやすい.

ⓑ ガウス分布の中心に関する微分

$$\frac{\partial \widetilde{E}}{\partial \mu_j} = \sum_i \gamma_j(w_i)\frac{(\mu_j - w_i)}{\sigma_j^2} \qquad (5.142)$$

平均 μ_j を「重み w_i の平均」に引っ張る.

($\gamma_j(w_i)$ での)

④ 分散による微分

$$\frac{\partial E}{\partial \sigma_j} = \sum_i \gamma_j(w_i)\left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3}\right) \quad (5.143)$$

分散 σ_j が $(w_i - \mu_j)$ に近づく。

実装では $\sigma_j^2 = \exp(\xi_j)$ という ξ_j を導入して ξ_j について最小化する

(σ_j の正値性といくつかが $\sigma_j \to 0$ になってしまう事をふせぐため.)

⑤ 混合係数に関する微分

制約 $\sum_j \pi_j = 1 \quad 0 \leq \pi_j \leq 1 \quad (5.145)$

を考慮する.

π_j をソフトマックスで表す.

$$\pi_j = \frac{\exp(\eta_j)}{\sum_{k=1}^{M} \exp(\eta_k)}$$

すると
$$\frac{\partial \widetilde{E}}{\partial \eta_j} = \sum_i \{\pi_j - \gamma_j(w_i)\} \quad (5.147)$$

π_j は事後分布 γ_j に引きよせられる.