§ 3.11　Cost Function and Regularization

PDF 版はここ↑

伊 冉 (Andre YI) - 21122200512

2025 年 9 月 4 日

Regularization

過学習を避けるために，コスト関数には追加の項を含めることがあります。パラメータが任意に大きな値を取ることを許すと，モデルは訓練データに過度に適合することがあると指摘されています。しかし，パラメータに有界な値という制約を課すと，モデルは訓練データの多くの点を通過する能力が抑えられ，過学習を防ぐことができます。したがって，パラメータ値はゼロ付近の小さな値に保たれる必要があります。小さなパラメータ値を前提にコスト関数を最小化するために，通常は L^1 または L^2 型の正則化項が用いられます。

Def. 0.1 (L^2-regularization): 重みパラメータ $w \in \mathbb{R}^n$ を考えます。その L^2-ノルム $\|w\|_2$ は $\|w\|_2^2 = \sum_{i=1}^n w_i^2$ によって与えられます。L^2 正則化を伴うコスト関数は，初期のコスト関数に L^2 ノルムを加えることで得られます。

$$L_2(w) = C(w) + \lambda \|w\|_2^2,$$

ここで λ は正のラグランジュ乗数であり，重みの大きさと $C(w)$ の最小値の間のトレードオフを制御します。大きな λ は重みを小さくし，$C(w)$ を大きくします。同様に，λ をゼロに近づけると重みは大きくなり，$C(w)$ は小さくなりますが，この場合は過学習が起きやすくなります。ハイパーパラメータ λ の値は，過学習の影響が最小化されるように選ぶべきです。

Def. 0.2 (L^1-regularization): $w \in \mathbb{R}^n$ の L^1-ノルムは $\|w\|_1 = \sum_{i=1}^n |w_i|$ と定義されます。L^1 正則化を伴うコスト関数は次のようになります。

$$L_1(w) = C(w) + \lambda \|w\|_1,$$

ここで $\lambda > 0$ はラグランジュ乗数であり，小さな重みを好む度合いの強さを制御します。なお $\|w\|_1$ はゼロで微分可能ではないため，通常の勾配法が正しく機能しない可能性があります。この欠点は L^2 正則化にはありません。

Def. 0.3 (Potential regularization): これは，上記の 2 つの正則化手続きを一般化したものです。関数 $U : \mathbb{R}^n \to \mathbb{R}_+$ が次を満たすとします。

1. $U(x) = 0$ となるのは $x = 0$ のとき，かつそのときに限る。
2. U は $x = 0$ で大域的最小値を持つ。

U が滑らかに微分可能な場合，条件 (ii) は導関数の条件 $U'(0) = 0$ および $U''(0) > 0$ によって含意されます。ポテンシャル関数 U は，前述の L^1 と L^2 ノルムの一般化です。

正則化されたコスト関数は次で定義されます。

$$G(w) = C(w) + \lambda U(w), \qquad \lambda > 0.$$

あるコスト関数に対して最良の正則化特性を持つ最適なポテンシャルを選ぶには，テスト誤差における性能を検証します。初期のコスト関数 $C(w)$ に項 $U(w)$ を加えるとテスト誤差が有意に減少しなければなりません。より明確に言えば，$\varepsilon_{\text{test},1}$，$\varepsilon_{\text{test},2}$ をそれぞれ $C(w)$ と正則化されたコスト $G(w)$ に対応するテスト誤差とすると，次を満たすように U を選ぶべきです。

$$\varepsilon_{\text{test},2} < \varepsilon_{\text{test},1}.$$

この場合，ニューラルネットワークは新しい未見データへの汎化が向上したと言います。次の節でこの種の誤差について詳しく扱います。