基本形式

what

b. L^2 参数正则化.

$$\tilde{J}(w;X,y) = \frac{\alpha}{2} w^T w + J(w;X,y)$$

$$\nabla \tilde{J}(w;X,y) = \alpha w + \nabla J_w(w;X,y)$$

$$w \leftarrow (1-\epsilon\alpha) w - \epsilon \nabla_w J(w;X,y)$$

加入权重衰减后会引起学习规则的修改，即在每步执行通常的梯度更新之前先收缩 w. 7.1

where

泰勒近似 $\hat{J}(\theta) = J(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*)$

$$\nabla_w \hat{J}(w) = H(w-w^*)$$

设最小点在 $\tilde{w}$ 上取得再加上正则项.

$$\alpha \tilde{w} + H(\tilde{w} - w^*) = 0$$

$$\tilde{w} = (H + \alpha I)^{-1} H w^*$$

因 $H = Q\Lambda Q^T$ 有.

$$\tilde{w} = Q(\Lambda + \alpha I)^{-1} \Lambda Q^T w^*$$

义: 可以看到权重衰减由 H 的特征向量所定义的轴缩放 w^*，$\frac{\lambda_i}{\lambda_i+\alpha}$.

~~意义~~

沿 H 特征值较大方向 ($\lambda_i >> \alpha$) 正则化影响小.

($\lambda_i << \alpha$) 分量将会收缩到几乎为 0.