what

8. 提前终止.

a. 在每次记录误差有所改善后，我们存储模型参数的副本，当训练算法终止时，我们返回这些参数而不是最新的参数。

算法7.1 含n为评估间隔和p为耐心和θ_0为初始参数

$\theta \leftarrow \theta_0$

$i \leftarrow 0$

$j \leftarrow 0$

$v \leftarrow \infty$

$\theta^* \leftarrow \theta$

$i^* \leftarrow i$

while $j < p$ do

运行n步，更新θ.

$i \leftarrow i + n$

$v' \leftarrow ValidErr(\theta)$

if $v' < v$

$j \leftarrow 0$

$\theta^* \leftarrow \theta$

$i^* \leftarrow i$

$v \leftarrow v'$

else

$j \leftarrow j + 1$

最佳θ^*与i^*.

算法7.2 含$X^{(train)}$和$y^{(train)}$为训练集.

将$X^{(train)}$和$y^{(train)}$分为

$X^{(subtrain)}$, $y^{(subtrain)}$与$X^{(valid)}$, $y^{(valid)}$

从θ和i, 调练 7.1 到到 i^*,

将θ再次设为随机值.

在$X^{(train)}$, $y^{(train)}$上训练i^*.

算法7.3.

⋮

运行7.1到θ

$\varepsilon \leftarrow J(\theta, X^{(subtrain)}, y^{(subtrain)})$

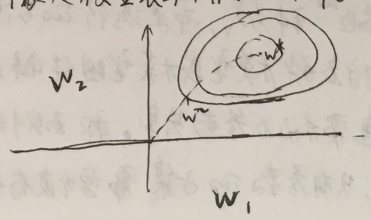
while $J(\theta, X^{(valid)}, y^{(valid)}) > \varepsilon$ do

在$X^{(train)}$, $y^{(train)}$上训练n步.

wly 有改

b. 提前终止为何具有正则化效果：

我们可以将 τ 作为有效容量的度量。假设梯度有界，限制迭代的次数和学习速率能够限制从 θ_0 到达的参数空间的大小，在这个意义上，τ 的效果就好像是权重衰减系数的倒数。



实线表示负对数似然；
虚线为SGD所经过的步九迹，
提前终止在 $\tilde{w}$ 处停止。

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

$$w^{(0)} = 0, \quad w^{(\tau)} = w^{(\tau-1)} - \varepsilon \nabla_w \hat{J}(w^{(\tau-1)})$$

$$w^{(\tau)} - w^* = (I - \varepsilon H)(w^{(\tau-1)} - w^*)$$

$$w^{(\tau)} - w^* = (I - \varepsilon Q \Lambda Q^T)(w^{(\tau-1)} - w^*)$$

假定 $w^{(0)}=0$，并且足够小使 $|1-\varepsilon\lambda_i| < 1$，经过 τ 次后：

$$Q^T w^{(\tau)} = [I - (I - \varepsilon\Lambda)^\tau] Q^T w^*$$

而L2中有 $Q^T \tilde{w} = [I - (\Lambda + \alpha I)^{-1}\alpha] Q^T w^*$

比较可知有当 $\alpha, \varepsilon, \tau$ 满足 $(I - \varepsilon\Lambda)^\tau = (\Lambda + \alpha I)^{-1}\alpha$ 时

L2正则化与提前终止等价。

$$\tau \approx \frac{1}{\varepsilon\alpha}, \quad \alpha \approx \frac{1}{\tau\varepsilon}$$

也就是说，在这些假设下，训练次数 τ 与 L^2 参数或在此作用

$\tau\varepsilon$ 的倒数与权重的衰减系数作用类似。