机器学习复习5

2022年7月7日 星期四 16:50

Regularization & model Selection

$M=\{M_1, M_2 \dots M_d\}$ 为 sample

1. Cross validation 交叉验证

- $h_\theta(x)=\theta^T x$
- ② 选择 $J(\theta)$ min 的函数.

eg. 70% sample 为测试集，30%

(hold-out cross validation，简单交叉验证)

测试集比例一般占 $\frac{1}{4}-\frac{1}{3}$.

- k-fold cross validation

将简单交叉验证测试集改为 $\frac{1}{k}$

每个模型训练 k 次，测试 k 次，

一般 k 取 10

2. Feature selection

① 向前搜索：

1. 扫描 $i \in (1, n)$

i 和 F 放在一起 (init F=空)

用交叉验证得出错误率

2. 将 n 中 Fi 中选错误率最小的为 Fi

更新 F 为 Fi

$o(n+(n-1)+\dots+1)=o(n^2)$

② 过滤特征选择.

$$MI(x_i, y)=\sum_{x_i\in\{0,1\}}\sum_{y\in\{0,1\}} P(x_i, y)\log\frac{P(x_i, y)}{P(x_i)P(y)}$$

$$MI(x_i, y)=KL(P(x_i, y)\,||\,P(x_i)P(y))$$

③ 贝叶斯统计和规则化 //减少过拟合的发生

$$\theta_{ML}=\arg\max_\theta \prod_{i=1}^{m} P(y^{(i)}|x^{(i)};\theta)$$