数据集：西瓜问

样本：单个西瓜

特征向量：

属性.

得到模型：分类 { 二分类 / 多分类

回归：价格走势.

聚类

进行预测：

奥卡姆剃刀：选最简单的那个.

模型评估与选择：

评估方法 { 泛化能力.
{ 训练集 / 测试集 / 验证集 } { 留出法 / 交叉验证法 / 自助法.

性能度量.

均方误差 $E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$

引入概率密度 $p(x)$ $E(f;D)=\int_{x\in D}(f(x)-y)^2p(x)\,dx$

错误率：$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)\neq y_i)$

精度：$acc(f;D)=1-E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)=y_i)$

引入密度函数：$E(f;D)=\int_{x\in D}\mathbb{I}(f(x)\neq y)p(x)\,dx$.

$acc(f;D)=\int_{x\in D}\mathbb{I}(f(x)=y)p(x)\,dx$

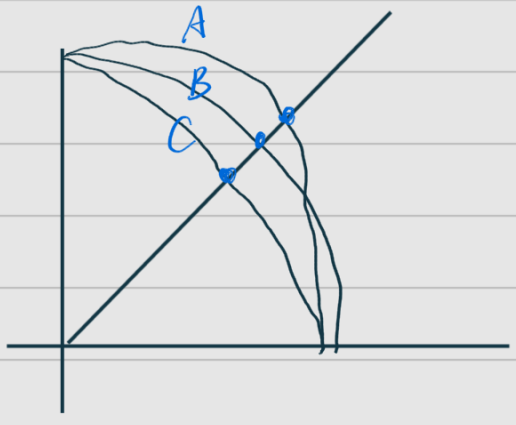
混淆矩阵：

真实 \ 预测情况	正例.	反例.
正例	TP ①	FN ②
反例	FP ③	TN ④

查准率：$\frac{TP}{TP+FP}$

查全率：$\frac{TP}{TP+FN}$

P-R图.



A > C

B > C

比较A.B.

F1度量.

查准：$\frac{①}{①+③}$

查全：$\frac{①}{①+②}$

调和平均数.

$$\frac{①+③}{2①} + \frac{①+②}{2①} = \frac{2①+②+③}{2①}$$

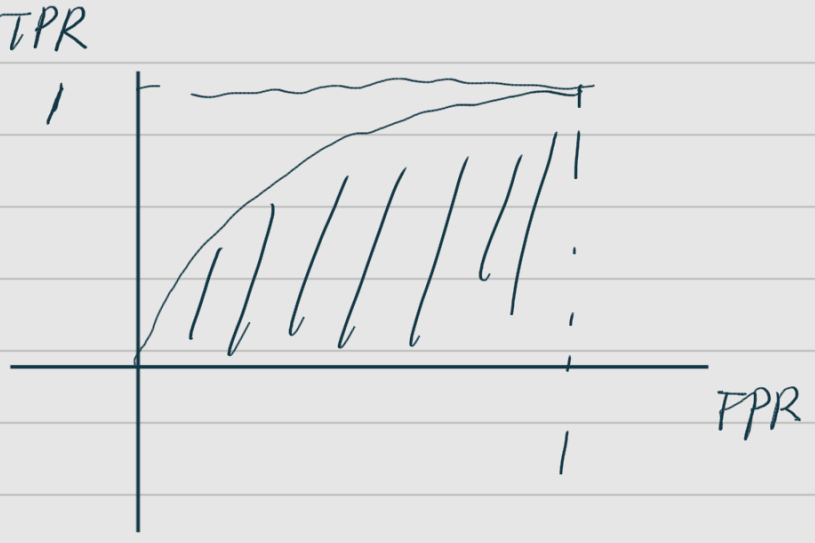
$$\frac{2①}{2①+②+③} = \frac{2\times TP}{总-TN+TP}$$

真实 \ 预测情况	正例.	反例.
正例	TP ①	FN ②
反例	FP ③	TN ④

ROC & AUC

$$TPR = \frac{TP}{TP+FN} \quad (True\ Positive\ Rate)$$

$$FPR = \frac{FP}{TP+FP} \quad (False\ Positive\ Rate)$$



AUC = 1 - Loss

代价敏感错误率、代价曲线：

真实 \ 预测	0	1
0	0	cost01
1	cost10	0

代价敏感错误率：

$$E(f;D;cost) = \frac{1}{m}\left(\sum_{x_i \in D^+} \mathbb{I}(f(x_i) \neq y_i) \cdot cost_{01} + \sum_{x \in D^-} \mathbb{I}(f(x_i) \neq y_i) \cdot cost_{10}\right)$$