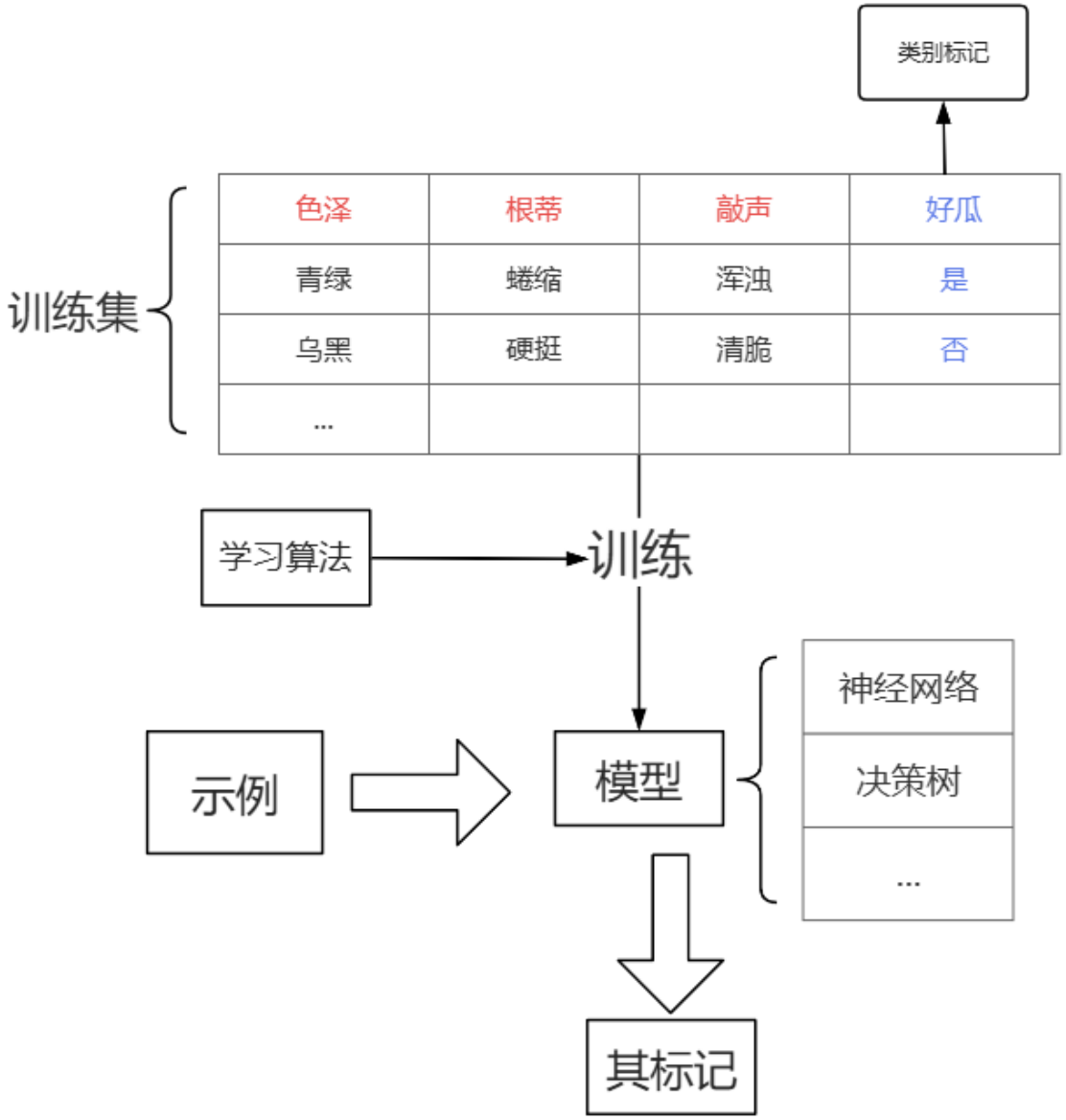
学习周报

汇报人：李俊泽

01

绪论

机器学习流程

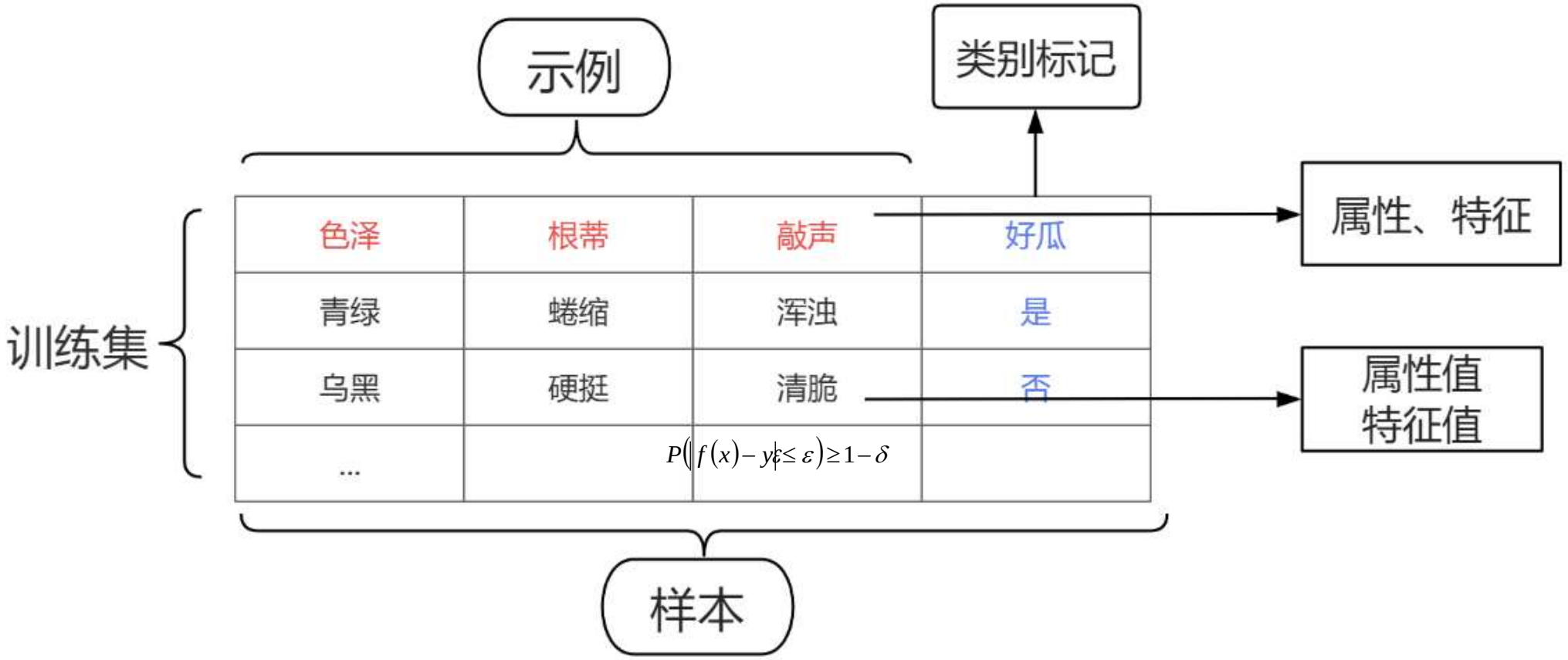


训练模型

根据处理好的数据集，划分训练集、验证集和测试集，然后根据训练集和验证集确定好学习算法和参数，训练出模型，然后用测试集验证该模型的泛化能力。

模型泛化

第一步得到了训练后的模型，然后输入示例（也就是没有标记的数据集，在本图中是还未判断为好瓜或者坏瓜的西瓜特征），最后通过模型得到标记。

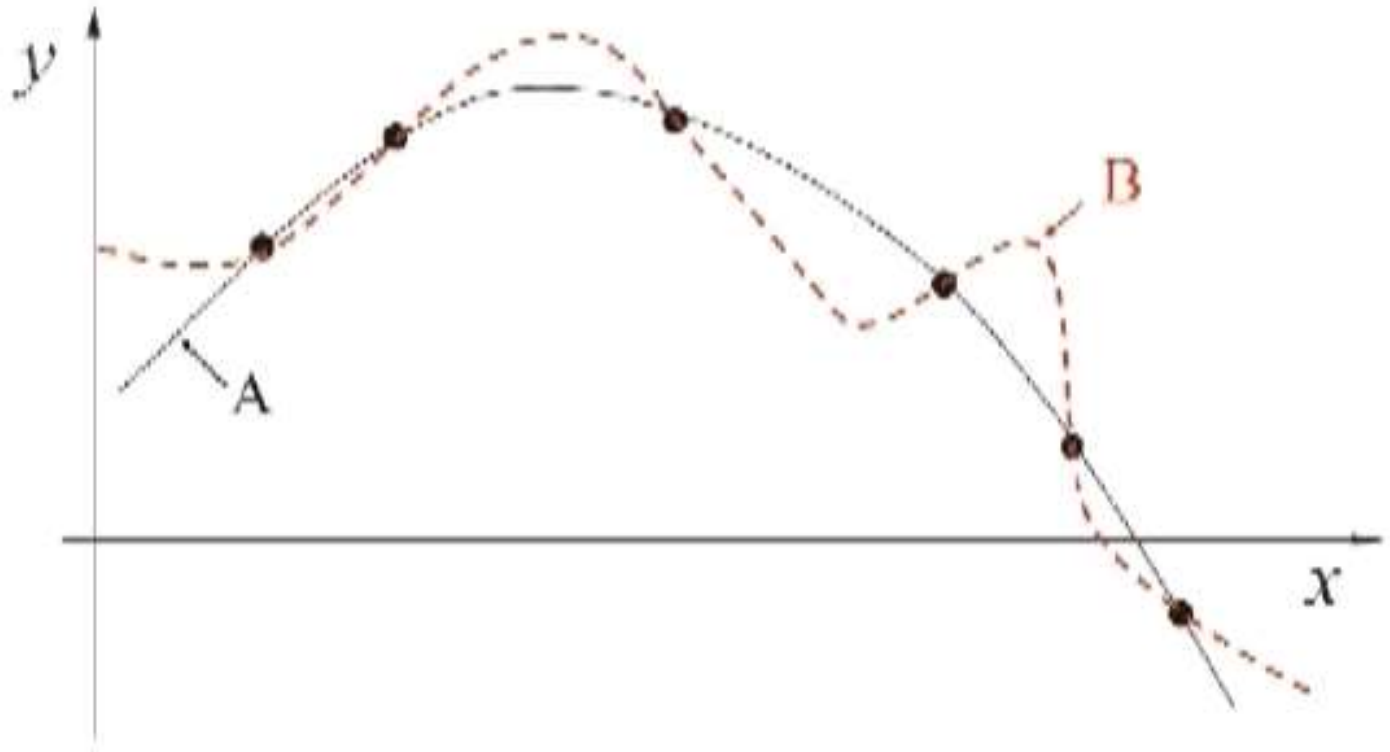


属性空间、样本空间：由n个属性(特征)张成的n维空间。
特征向量：在样本空间中，具体属性值(特征值)生成的向量。

假设：f(x)　　真相：y　　$P\left(\left|f\left(x\right)-y\right|\leq\varepsilon\right)\geq 1-\delta$

目标：训练出来的模型(假设)与一般规律(真相)的误差在一定范围内的概率比较大

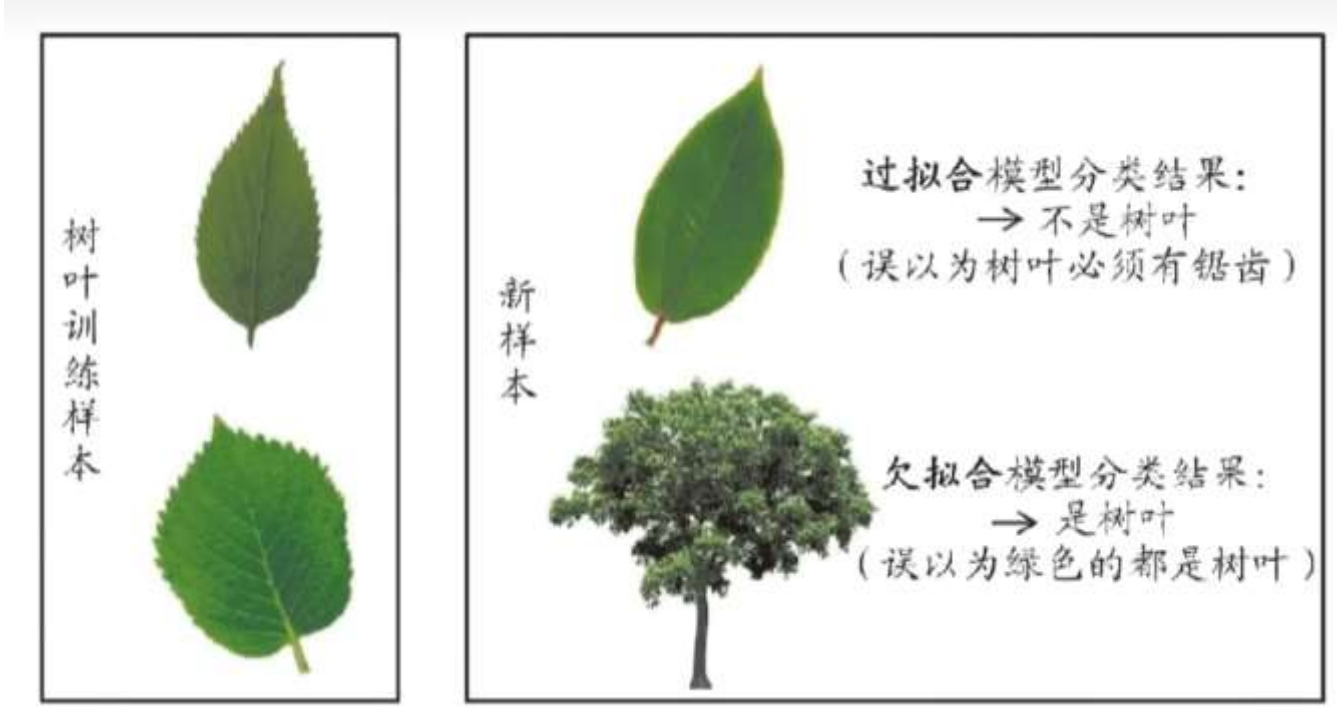
归纳偏好

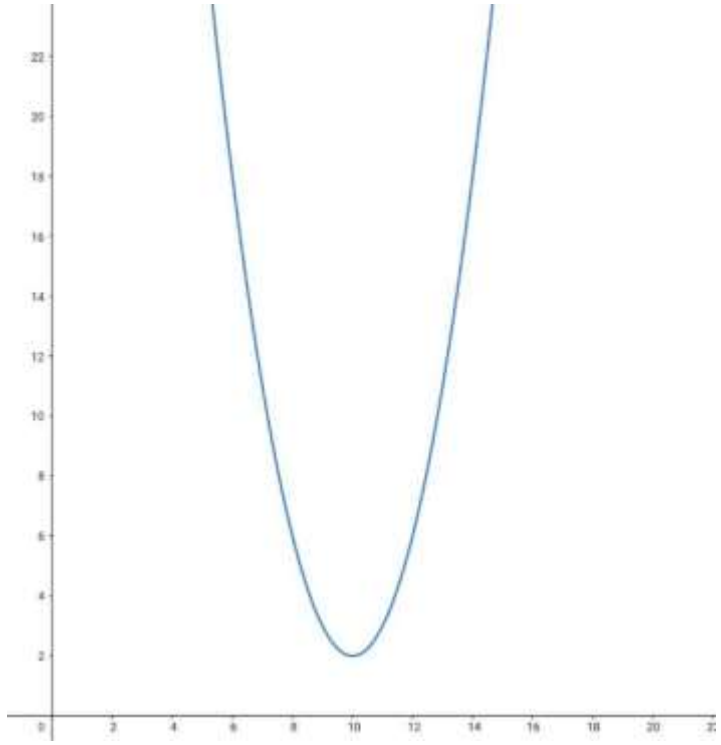


对于同样的5个样本点，不同的假设(模型)都符合，但是不同假设(不同)的泛化能力，对于要预测的真相(一般规律)，存在欠拟合或者过拟合的问题，一般来说采用奥卡姆剃刀原则，选最”简单”的那个。

NFL定理：在这个问题上，A学习算法比B学习算法表现好，则比存在一个问题，B比A好（但是一般而言，我们只关注这个问题，只是在谈论算法好坏的时候，不能脱离具体某一个任务和问题）

过拟合和欠拟合





对于一个模型，如果训练数据训练出来，经验误差很小，精确度很高，泛化能力很可能不好，因为此时有可能吧训练集的某一些特性也总结出来，使得模型过拟合。

而对于泛化误差，则呈U型曲线，在某一个地方泛化能力最强，所以我们对于某一个具体问题就是要训练一个模型尽可能使得泛化误差最小。

02

模型评估

训练集和测试集划分

留出法

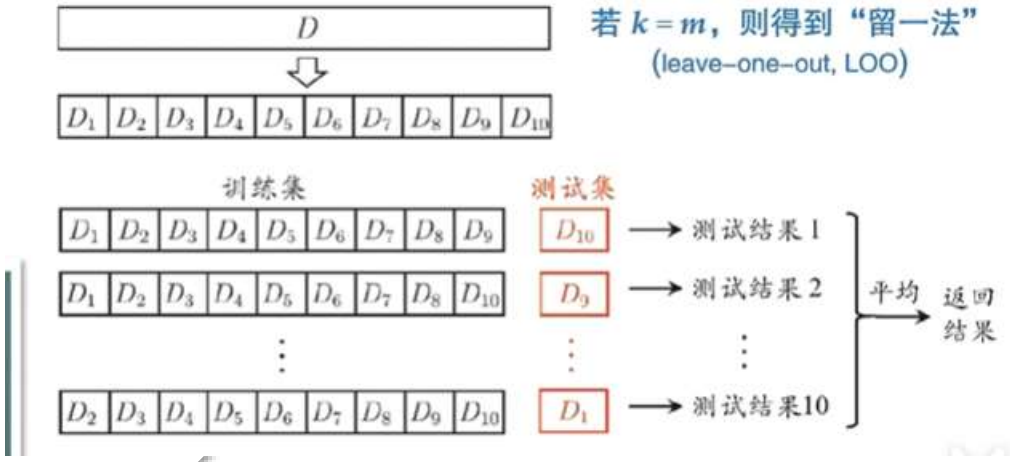
将数据集随机划分为训练集和测试集，训练集占数据集的2/3或者4/5左右，必须保证划分出来的训练集能够完成模型训练的工作。
注：训练集和测试集必须互斥，划分的时候必须采用分层抽样，无论是训练集和测试集，数据类别比例相似。

优点：简单、随机划分
缺点：随机划分的时候，有概率有些数据集并没有作为测试集，影响模型的泛化能力。

交叉验证法

把数据集划分为n份，K1-Kn，第一次K1作为测试集，剩下作为训练集，第二次K2作为测试集，以此类推，测试n次，取平均，返回测试结果。
注：训练集和测试集必须互斥

优点：解决了留出法中，有概率数据并未做测试集的弊端



训练集和测试集划分

数据集不够？

自助法

将数据集随机划分为训练集和测试集，因为数据集数量不足，所以采用自助法，对于数据集，有放回地抽取样本，然后抽取到足够多的训练集后，把互斥的那部分数据集划分为测试集。给定m个样本的数据集，然后m次采样后，始终不被采样到的数据占36.8%左右。

优点：解决数据不足的问题

缺点：重复采样的时候会改变数据集分布，有些数据重复采样。

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^{m}=\frac{1}{\mathrm{e}}\approx 0.368$$

性能度量

回归任务

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

采用二阶中心矩来计算回归任务中，模型与真相的误差

分类任务

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

精度：1-错误率

通过判断结果是否符合真相来计算错误率

性能度量

分类任务(二分类)

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

分类结果混淆矩阵

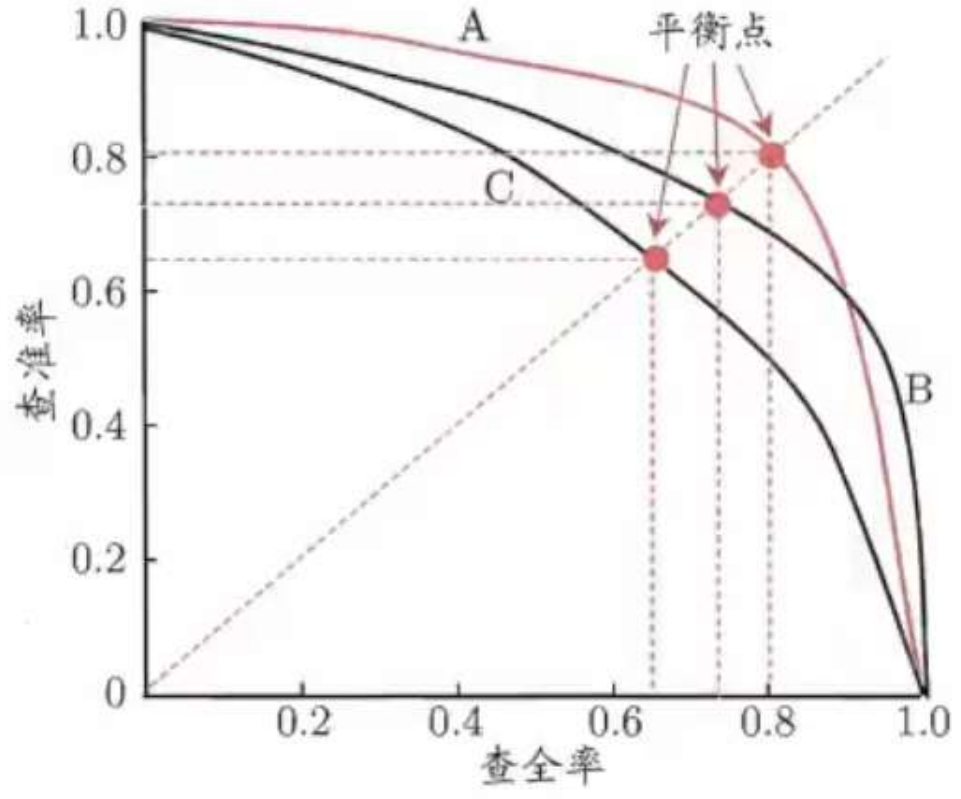
查准率P: $$\frac{TP}{TP+FP}$$

查全率R: $$\frac{TP}{TP+FN}$$

一般来说，查准率和查全率是一对矛盾的度量，只有在一些简单任务，两者才会比较高

性能度量

分类任务(二分类)



如果粗略的令P=R，则绘制这样一个P-R图，用P=R去哼两一个模型，此时能得出A是优于B和C的，而对于A“包着”C曲线，无论P与R什么关系，都是A优于C，但是对B则需要看情况，如果需要更复杂的衡量和评估，则需要F1度量。

$$\frac{1}{F1}=\frac{1}{2}\cdot\left(\frac{1}{P}+\frac{1}{R}\right)$$

$$\frac{1}{F_\beta}=\frac{1}{1+\beta^2}\cdot\left(\frac{1}{P}+\frac{\beta^2}{R}\right)$$

后者当 β >1的时候，更看重R，R由更大影响，<1的时候更看重P，=1P与R影响相同，相当于加权。