中山大學
SUN YAT-SEN UNIVERSITY

第2章 模型评估与选择

1. 训练误差与测试误差
2. 过拟合与模型选择
3. 性能度量
4. 偏差与方差

中山大學
SUN YAT-SEN UNIVERSITY

第2章 模型评估与选择

1. 训练误差与测试误差
2. 过拟合与模型选择
3. 性能度量
4. 偏差与方差

训练误差与测试误差

机器学习的目的是使学得的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

当损失函数给定时，基于损失函数的模型的训练/经验误差（training/empirical error）和模型的测试误差（testing error）可用于评价学习方法。（经验损失或经验风险）

训练误差，训练集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$：

$$E\left(\hat{f}; D\right) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(y_i, \hat{f}(\boldsymbol{x}_i)\right)$$

（1）0-1损失函数，$\mathcal{L}\left(y_i, f(\boldsymbol{x}_i)\right) = \begin{cases} 1, y_i \neq f(\boldsymbol{x}_i) \\ 0, y_i = f(\boldsymbol{x}_i) \end{cases}$

（2）平方损失函数，$\mathcal{L}\left(y_i, f(\boldsymbol{x}_i)\right) = \left(y_i - f(\boldsymbol{x}_i)\right)^2$

训练误差与测试误差

机器学习的目的是使学得的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

当损失函数给定时，基于损失函数的模型的训练/经验误差（training/empirical error）和模型的测试误差（testing error）可用于评价学习方法。（经验损失或经验风险）

测试误差，测试集 $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^{n}$：

$$E(\hat{f}; D) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}\left(y_j, \hat{f}(\boldsymbol{x}_j)\right)$$

（1）0-1损失函数，$\mathcal{L}(y_i, f(\boldsymbol{x}_i)) = \begin{cases} 1, y_i \neq f(\boldsymbol{x}_i) \\ 0, y_i = f(\boldsymbol{x}_i) \end{cases}$

（2）平方损失函数，$\mathcal{L}(y_i, f(\boldsymbol{x}_i)) = (y_i - f(\boldsymbol{x}_i))^2$

训练误差与测试误差

机器学习的目的是使学得的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

- 训练误差：训练集上
- 测试误差：测试集上

表 4.2 西瓜数据集 2.0 划分出的训练集(双线上部)与验证集(双线下部)

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否
编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否

训练误差与测试误差

机器学习的目的是使学得的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

训练误差的大小，对判定给定的问题是不是一个容易学习的问题具有重要意义。如果训练误差过大，则说明存在“欠拟合”。

测试误差的大小，反映了学习方法对未知的测试数据集的预测能力。给定两种学习方法，测试误差小的方法就有更好的预测能力（泛化能力，generalizationability）。

- 训练误差：训练集上
- 测试误差：测试集上
- 泛化误差：除训练集外所有样本

常用测试误差作为泛化误差的近似

中山大學
SUN YAT-SEN UNIVERSITY

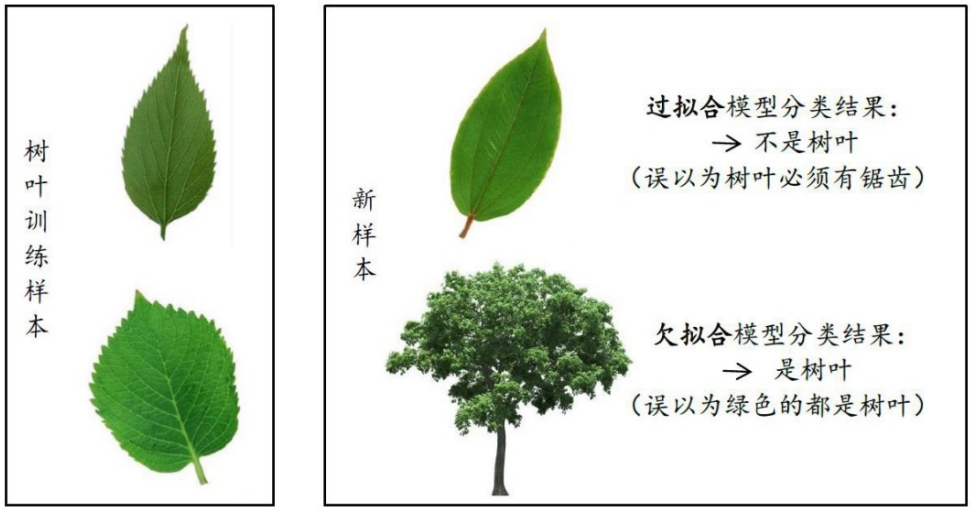
第2章 模型评估与选择

1. 训练误差与测试误差
2. 过拟合与模型选择
3. 性能度量
4. 偏差与方差

过拟合与模型选择

机器学习的目的是使学得的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

由于事先并不知道新样本的特征，我们只能努力使经验误差最小化；但是当学习器把训练样本学得“太好”，将训练样本本身的特点当做所有样本的一般性质，因此而导致泛化性能下降。这种现象叫做“过拟合”（overfitting）。



过拟合、欠拟合的直观类比

过拟合与模型选择

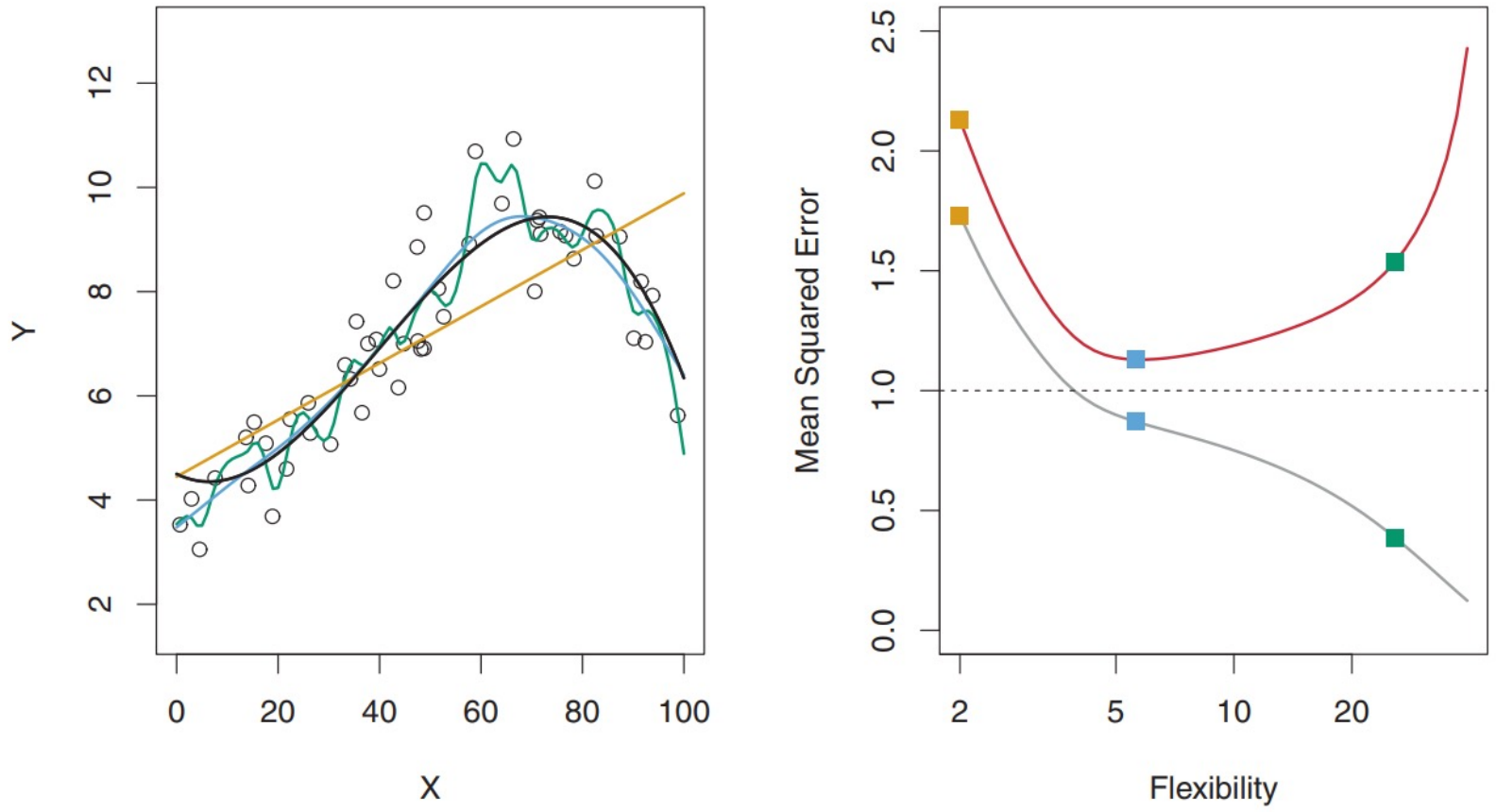


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

过拟合与模型选择

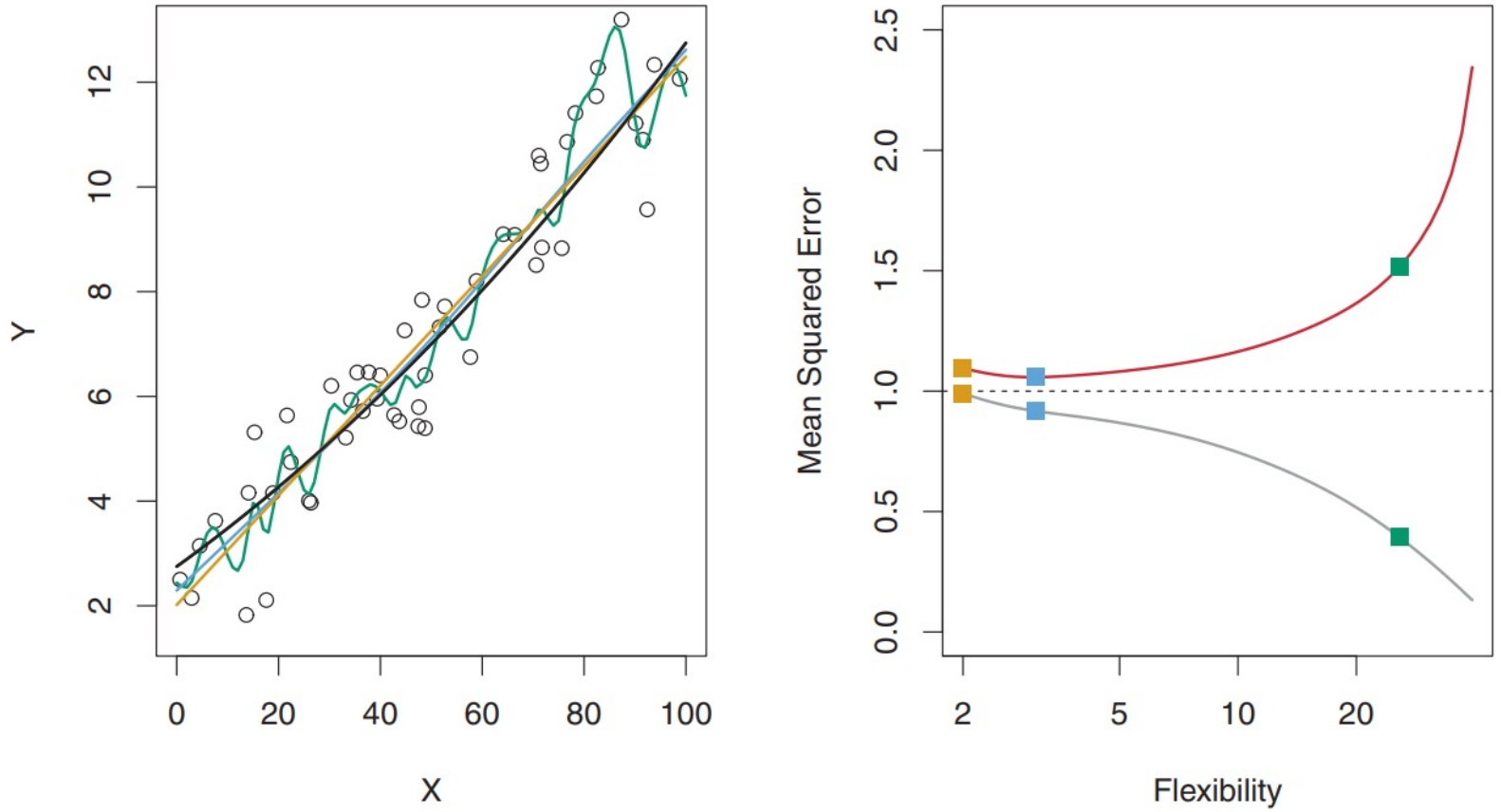


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

过拟合与模型选择

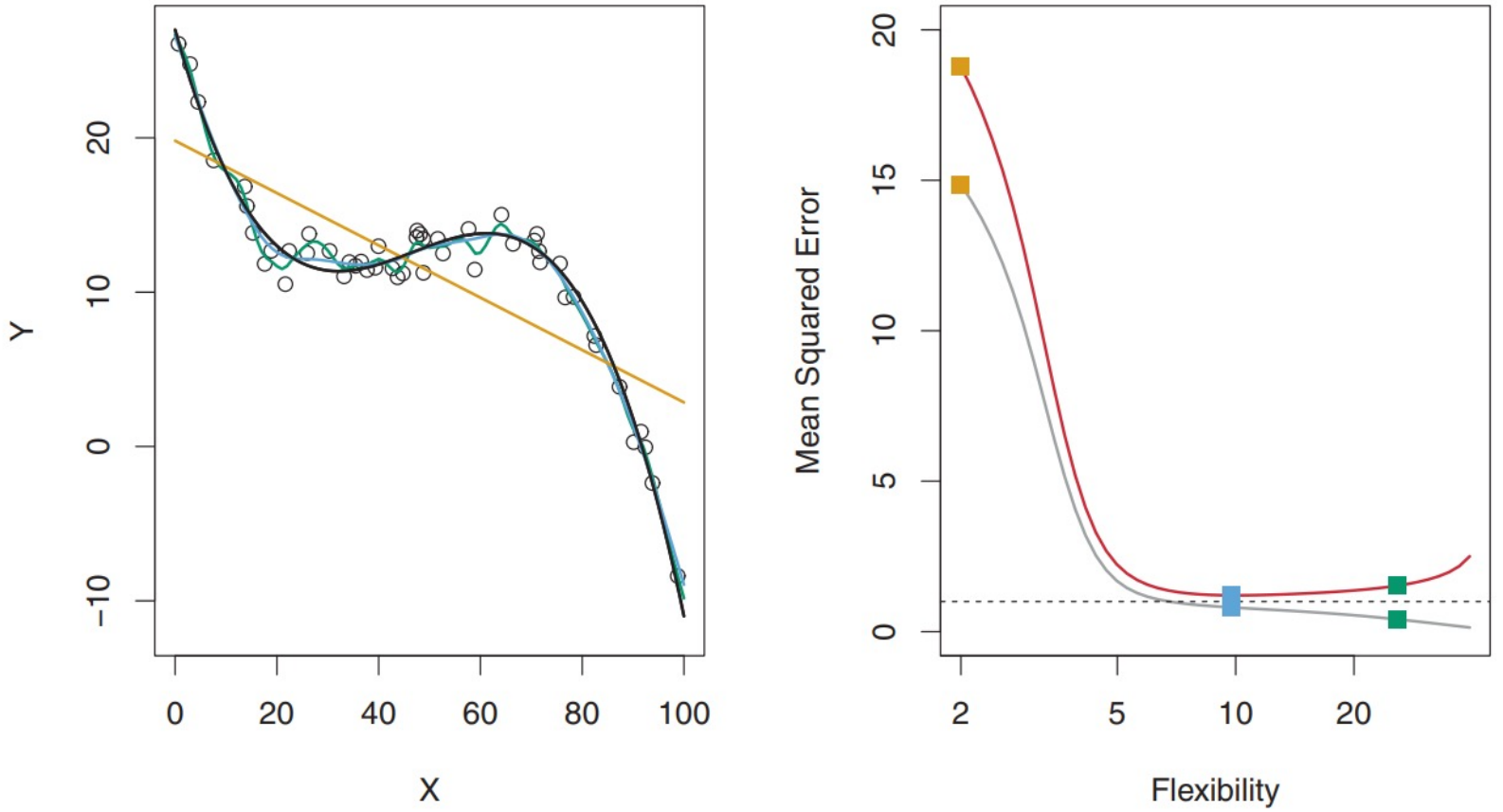


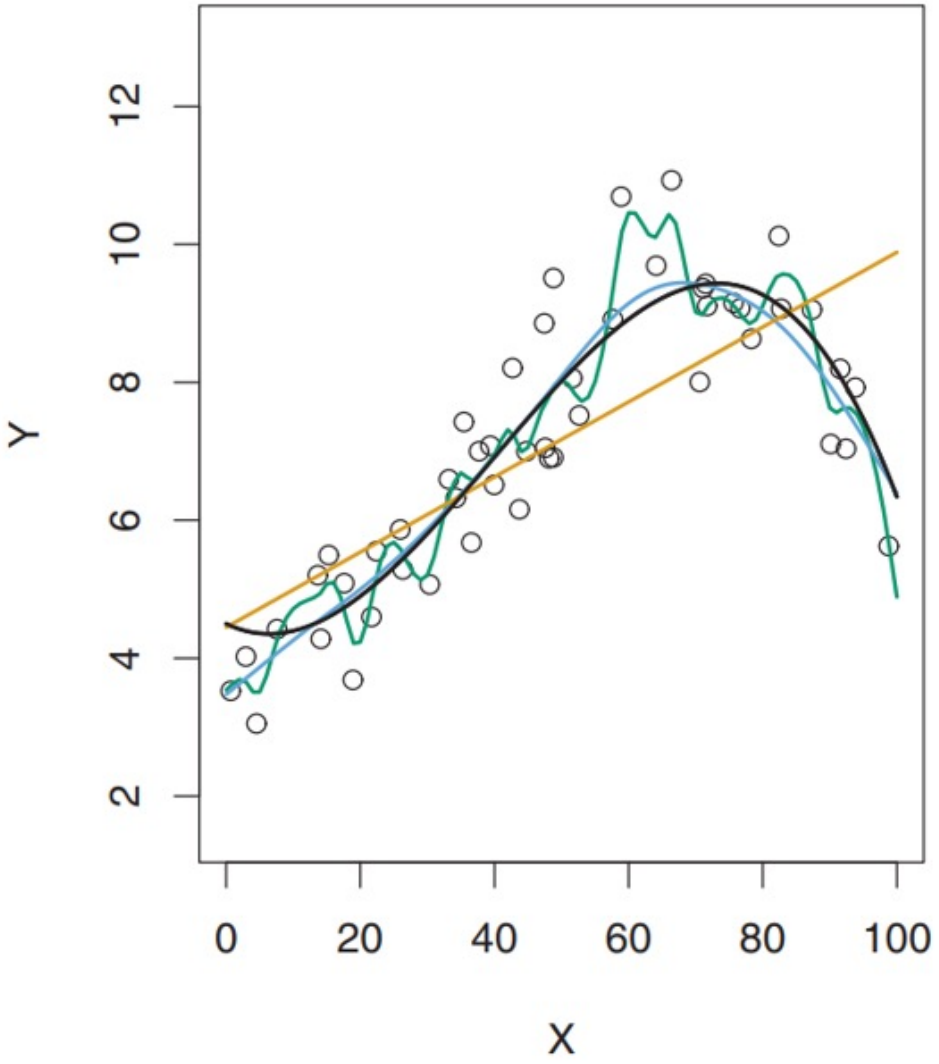
FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

过拟合与模型选择

机器学习的目的是使学得的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

“欠拟合”（underfitting）：学习能力“太弱”；克服方法：增加模型复杂度、增加训练轮数等。

“过拟合”（overfitting）：学习能力“太强”；克服方法：优化目标中增加正则项、进行模型选择等。



过拟合与模型选择

模型选择的典型方法是正则化（regularization）。正则化是结构风险最小策略的实现，是在经验风险上加一个正则化项（regularizer）或罚项（penalty term）。

正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大。正则化一般具有如下形式：

$$\min_{f} \sum_{i=1}^{m} \mathcal{L}\left(y_i, f(\boldsymbol{x}_i)\right) + \lambda \|f\|_l$$

- ERM
- SRM

$\|f\|_l$为f的L_l范数。以线性模型$f(\boldsymbol{x}_i) = \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}_i + b$为例，定义$\boldsymbol{\omega} = (\boldsymbol{b}, \boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_d}) = (\boldsymbol{w_0}, \boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_d})$，那么，$\mathrm{L}_1$范数为$\sum_{j=0}^{d} |w_j|$，$\mathrm{L}_2$范数为$\sum_{j=0}^{d} w_j^2$。

过拟合与模型选择

当给定的样本数据充足，进行模型选择的一种简单方法是随机地将数据集切分成三部分：

训练集，用来训练模型，对应训练误差

验证集，用来选择模型，对应测试误差

测试集，用来最终对学习方法进行评估，对应泛化误差的近似

我们假设测试集是从样本真实分布中独立采样获得，将测试集上的“测试误差”作为泛化误差的近似，所以测试集要和训练集中的样本尽量互斥。

过拟合与模型选择

在实际应用中数据是不充足的，为了选择好的模型，可以采用留出法、交叉验证法和自助法。

留出法：

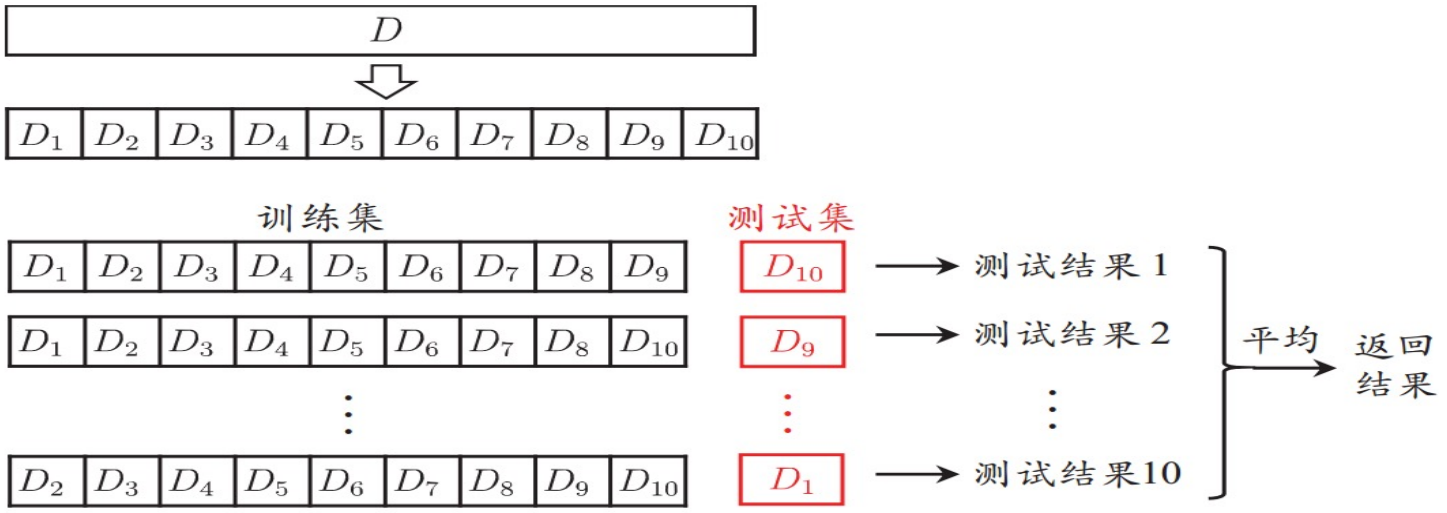
- 直接将数据集划分为两个互斥集合
- 训练/测试集划分要尽可能保持数据分布的一致性
- 一般若干次随机划分，重复实验取平均值
- 训练/测试样本比例通常为2:1~4:1

过拟合与模型选择

在实际应用中数据是不充足的，为了选择好的模型，可以采用留出法、交叉验证法和自助法。

交叉验证法：

将数据集分层采样划分为k个大小相似的互斥子集，每次用$k-1$个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值，k最常用的取值是10。



10 折交叉验证示意图

过拟合与模型选择

交叉验证法：与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值，例如常见的“10次10折交叉验证”。

假设数据集D包含m个样本，若令$k = m$，则得到留一法：

- 不受随机样本划分方式的影响
- 结果往往比较准确
- 当数据集比较大时，计算开销难以忍受

过拟合与模型选择

在实际应用中数据是不充足的，为了选择好的模型，可以采用留出法、交叉验证法和自助法。

自助法：以自助采样法为基础，对数据集D有放回采样m次得到训练集D'，用D/D'做测试集。

- 实际模型与预期模型都使用m个训练样本
- 约有1/3的样本没在训练集中出现 $\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^m \mapsto \frac{1}{e}\approx 0.368$
- 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处

自助法在数据集较小、难以有效划分训练/测试集时很有用；由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用。

中山大學
SUN YAT-SEN UNIVERSITY

第2章 模型评估与选择

1. 训练误差与测试误差
2. 过拟合与模型选择
3. 性能度量
4. 偏差与方差

性能度量：均方误差

性能度量是衡量模型泛化能力的评价标准，反映了任务需求；
使用不同的性能度量往往会导致不同的评判结果。

在预测任务中，给定样例集 $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^{n}$ ，评估学习器 $\hat{f}$ 的性能，即把预测结果 $\hat{f}(\boldsymbol{x}_j)$ 和真实标记比较。

回归任务最常用的性能度量是“均方误差”：

$$E(\hat{f}; D) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}\left(y_j, \hat{f}(\boldsymbol{x}_j)\right)$$

性能度量：错误率与精度

性能度量是衡量模型泛化能力的评价标准，反映了任务需求；
使用不同的性能度量往往会导致不同的评判结果。

在预测任务中，给定样例集 $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^{n}$ ，评估学习器 $\hat{f}$ 的性能，即把预测结果 $\hat{f}(\boldsymbol{x}_j)$ 和真实标记比较。

对于分类任务，错误率和精度是最常用的两种性能度量：

错误率

$$E(\hat{f};D)=\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(\hat{f}(\boldsymbol{x}_j)\neq y_j)$$

精度

$$acc(\hat{f};D)=\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(\hat{f}(\boldsymbol{x}_j)= y_j)$$

性能度量：查准率与查全率

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比率，此时查准率和查全率比错误率和精度更适合。

统计真实标记和预测结果的组合可以得到“混淆矩阵”。

分类结果混淆矩阵

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)

- True / False
  - True: prediction = label
  - False: prediction ≠ label
- Positive / Negative
  - Positive: predict y = 1
  - Negative: predict y = 0

性能度量：查准率与查全率

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比率，此时查准率和查全率比错误率和精度更适合。

统计真实标记和预测结果的组合可以得到“混淆矩阵”。

Prediction

Label	1	0
1	True Positive	False Negative
0	False Positive	True Negative

Prediction

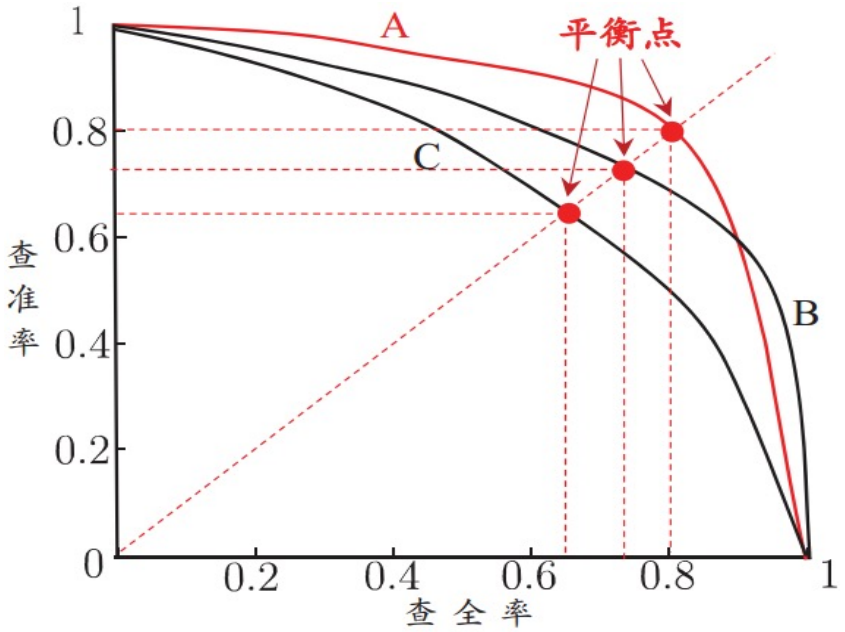
Label	1	0
1	True Positive	False Negative
0	False Positive	True Negative

查准率 $P = \frac{TP}{TP + FP}$

查全率 $R = \frac{TP}{TP + FN}$

性能度量：P-R曲线

根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”。

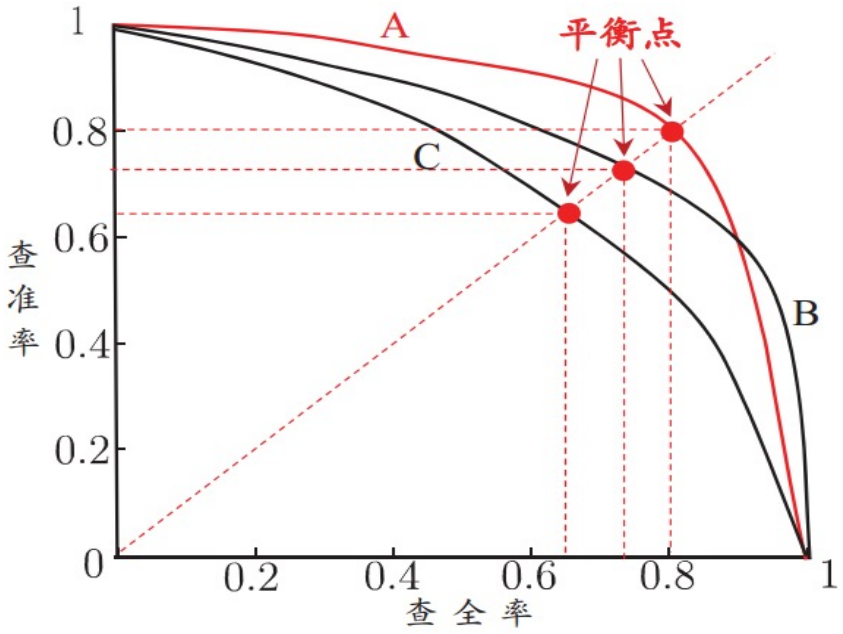


P-R曲线与平衡点示意图

Prediction	Label
0.91	1
0.85	0
0.77	1
0.72	1
0.61	0
0.48	1
0.42	0
0.33	0

性能度量：P-R曲线

根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”。



P-R曲线与平衡点示意图

平衡点是曲线上“查准率=查全率”时的取值，可用来用于度量P-R曲线有交叉的分类器性能高低

性能度量：F1度量

比P-R曲线平衡点更用常用的是$F1$度量：

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{样例综述 + TP - TN}$$

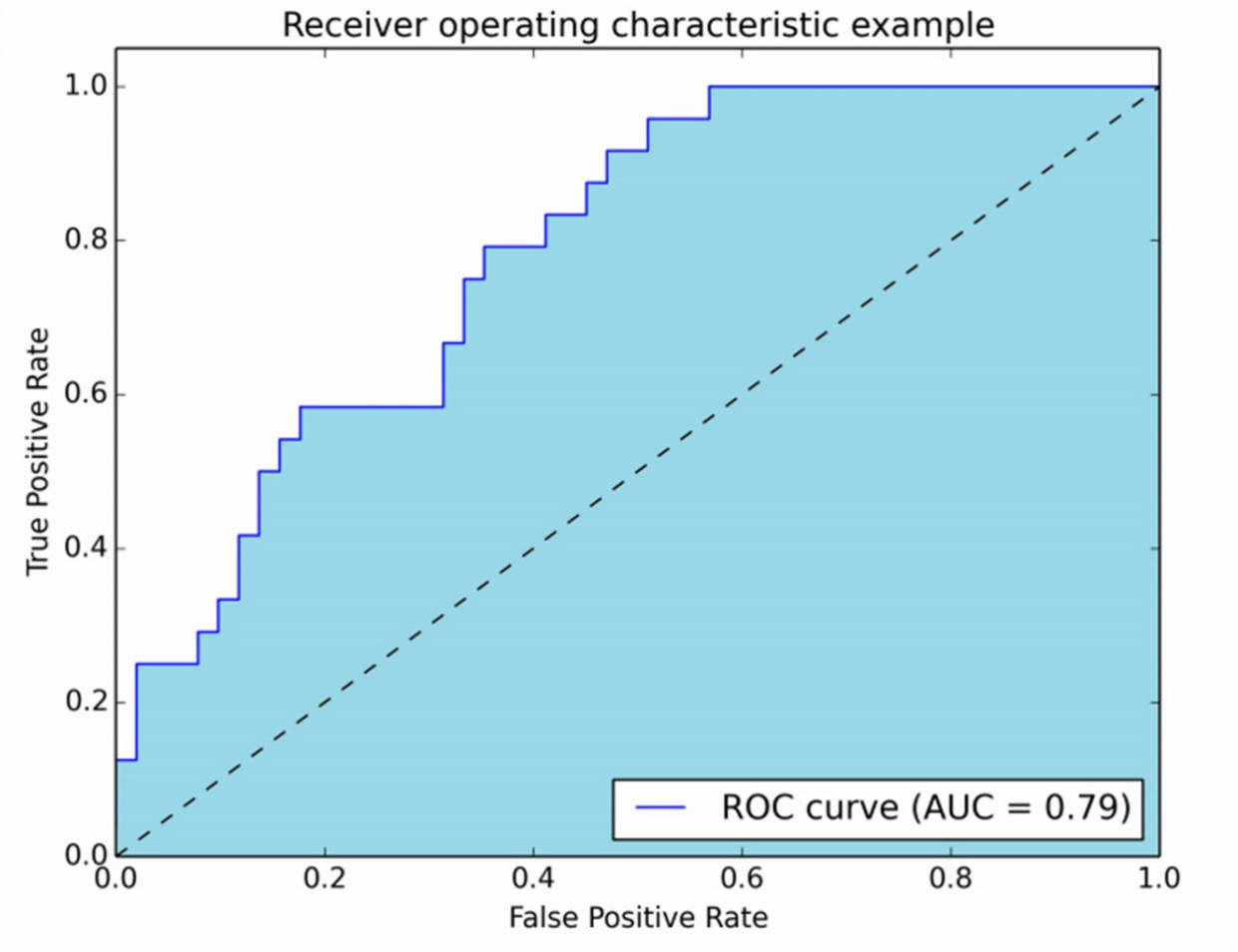
比$F1$度量更一般的形式F_β：

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

$\beta = 1$：标准$F1$
$\beta > 1$：偏重查全率（逃犯信息检索）
$\beta < 1$：偏重查准率（商品推荐系统）

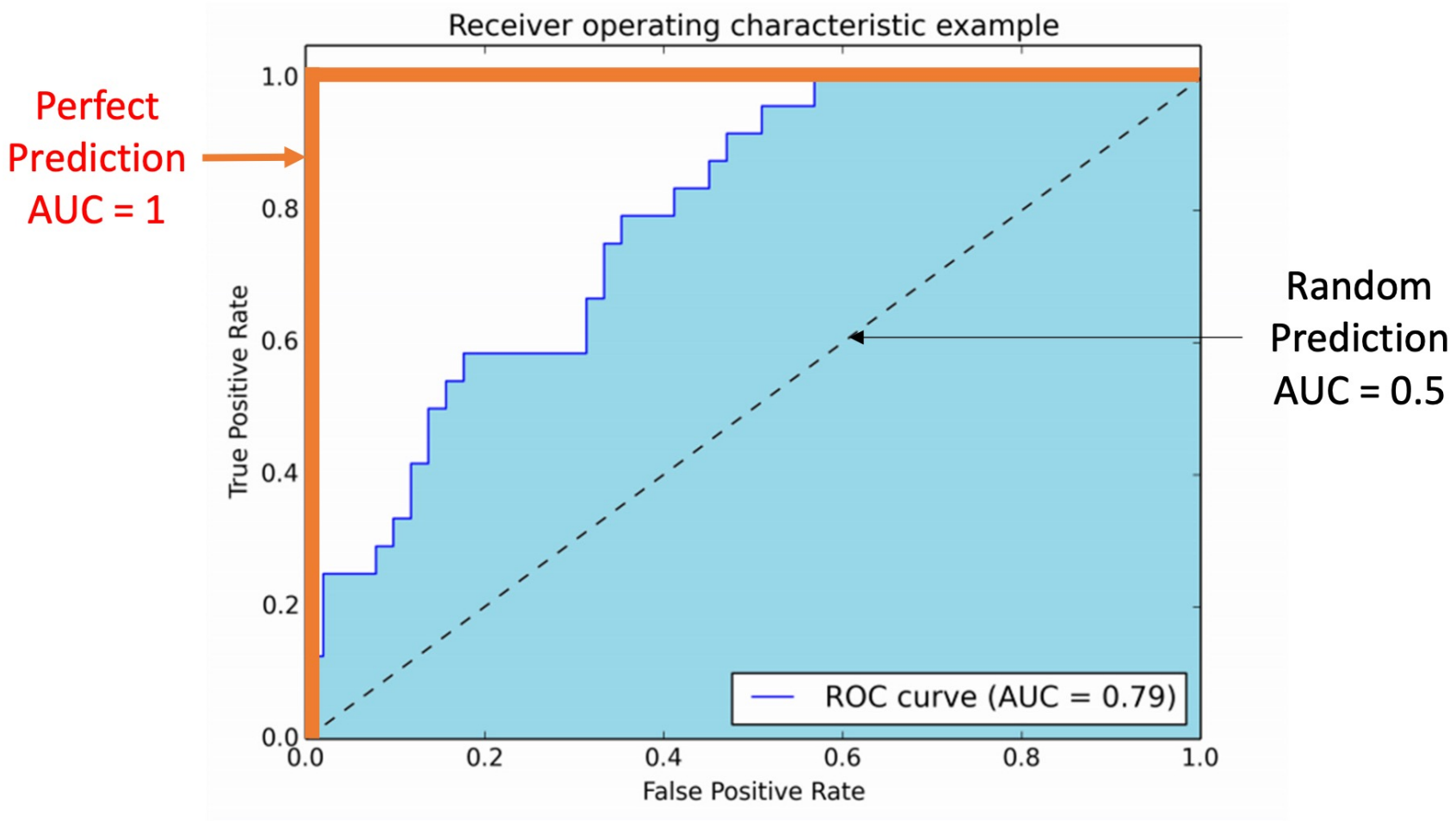
性能度量：ROC曲线与AUC值

类似P-R曲线，根据学习器的预测结果对样例排序，并逐个作为正例进行预测，以“假正例率”为横轴，“真正例率”为纵轴可得到ROC曲线，全称“受试者工作特征”。



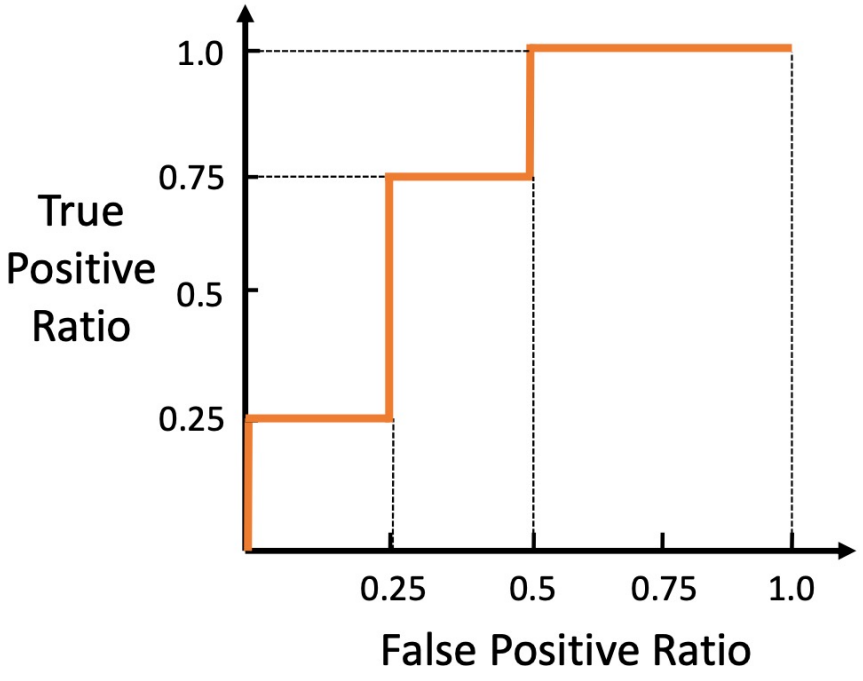
性能度量：ROC曲线与AUC值

若某个学习器的ROC曲线被另一个学习器的曲线“包住”，则后者性能优于前者；否则如果曲线交叉，可以根据ROC曲线下面积大小进行比较，也即AUC值。



性能度量：ROC曲线与AUC值

类似P-R曲线，根据学习器的预测结果对样例排序，并逐个作为正例进行预测，以“假正例率”为横轴，“真正例率”为纵轴可得到ROC曲线，全称“受试者工作特征”。



AUC = 0.75

Prediction	Label
0.91	1
0.85	0
0.77	1
0.72	1
0.61	0
0.48	1
0.42	0
0.33	0

性能度量：ROC曲线与AUC值

类似P-R曲线，根据学习器的预测结果对样例排序，并逐个作为正例进行预测，以“假正例率”为横轴，“真正例率”为纵轴可得到ROC曲线，全称“受试者工作特征”。

ROC曲线的绘制：

- 给定m^+个正例和m^-个负例，根据学习器预测结果对样例进行排序
- 然后把分类阈值设为最大，即把所有样例均预测为负例，此时真正例率和假正例率均为0，在坐标$(0,0)$处标记一个点
- 将分类阈值依次设为每个样例的预测值，即依次将每个样例划分为正例，设前一个标记点坐标为(x,y)，当前若为真正例，则对应标记点的坐标为$(x,y+1/m^+)$；当前若为假正例，则对应标记点的坐标为$(x+1/m^-,y)$
- 最后用线段连接相邻点

中山大學
SUN YAT-SEN UNIVERSITY

第2章 模型评估与选择

1. 训练误差与测试误差
2. 过拟合与模型选择
3. 性能度量
4. 偏差与方差

偏差与方差

通过实验可以估计学习算法的泛化性能，而“偏差-方差分解”可以用来帮助解释泛化性能。偏差-方差分解试图对学习算法期望的泛化性能进行拆解。

对测试样本$\boldsymbol{x}$，令y_D为$\boldsymbol{x}$在数据集中的标记，y为$\boldsymbol{x}$的真实标记，$f(\boldsymbol{x};D)$为训练集D上学得模型f在$\boldsymbol{x}$上的预测输出。以回归为例，学习算法的期望预测为：

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x};D)]$$

使用样本数目相同的不同训练集产生的方差为：

$$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right]$$

噪声为：

$$\varepsilon^2 = \mathbb{E}_D[(y_D - y)^2]$$

偏差与方差

期望输出与真实标记的差别称为偏差，$bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$

假定噪声期望为0，也即$\mathbb{E}_D[y_D - y] = 0$，对泛化误差分解。

$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right) + \bar{f}\left(\boldsymbol{x}\right) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right) - y_D\right)^2\right] \\
&\quad + \mathbb{E}_D\left[2\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)\left(\bar{f}\left(\boldsymbol{x}\right) - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right) - y + y - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right) - y\right)^2\right] + \mathbb{E}_D\left[\left(y - y_D\right)^2\right] \\
&\quad + 2\mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right) - y\right)\left(y - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \left(\bar{f}\left(\boldsymbol{x}\right) - y\right)^2 + \mathbb{E}_D\left[\left(y_D - y\right)^2\right] ,
\end{aligned}
$$

偏差与方差

泛化误差可分解为偏差、方差与噪声之和：

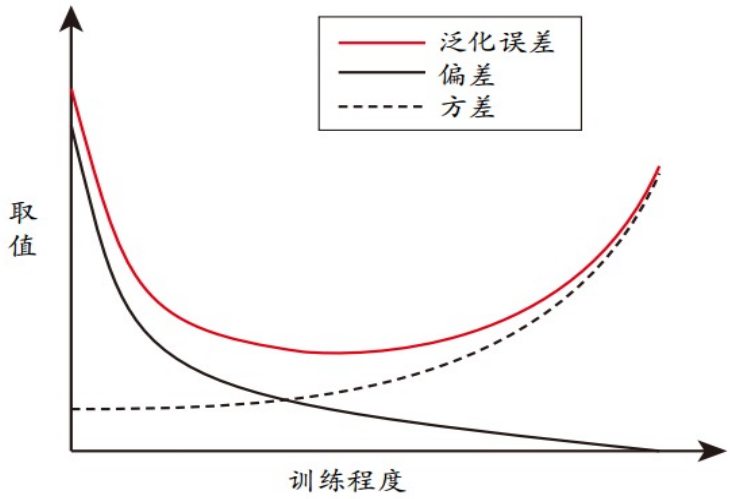
- 偏差度量了学习算法期望预测与真实结果的偏离程度；即刻画了学习算法本身的拟合能力；
- 方差度量了同样大小训练集的变动所导致的学习性能的变化；即刻画了数据扰动所造成的影响；
- 噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界；即刻画了学习问题本身的难度。

泛化性能是由学习方法的能力、数据的充分性以及学习任务本身的难度所共同决定的。给定学习任务为了取得好的泛化性能，需要使偏差小(充分拟合数据)而且方差较小(减少数据扰动产生的影响)。

偏差与方差

一般来说，偏差与方差是有冲突的，称为偏差-方差窘境。如右图所示，假如我们能控制算法的训练程度：

- 在训练不足时，学习器拟合能力不强，训练数据的扰动不足以使学习器的拟合能力产生显著变化，此时偏差主导泛化错误率；
- 随着训练程度加深，学习器拟合能力逐渐增强，方差逐渐主导泛化错误率；
- 训练充足后，学习器的拟合能力非常强，训练数据的轻微扰动都会导致学习器的显著变化，若训练数据自身非全局特性被学到则会发生过拟合。



泛化误差与偏差、方差的关系示意图

中山大學
SUN YAT-SEN UNIVERSITY

第2章 模型评估与选择

1. 训练误差与测试误差

（概念、计算方法、关系）

2. 过拟合与模型选择

（欠拟合、过拟合、正则化、三种方法）

3. 性能度量

（错误率、查准率、查全率、P-R曲线、ROC曲线）

4. 偏差与方差

（分解推导、内涵解读）

谢谢！