中山大学软件工程学院

SCHOOL OF SOFTWARE ENGINEERING

模型评估

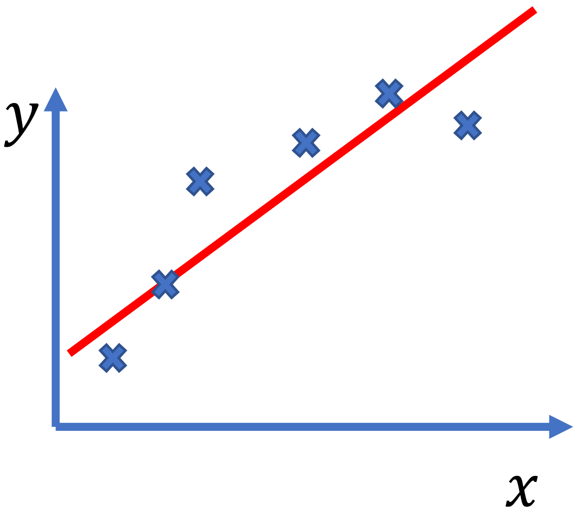
廖国成

liaogch6@mail.sysu.edu.cn

内容

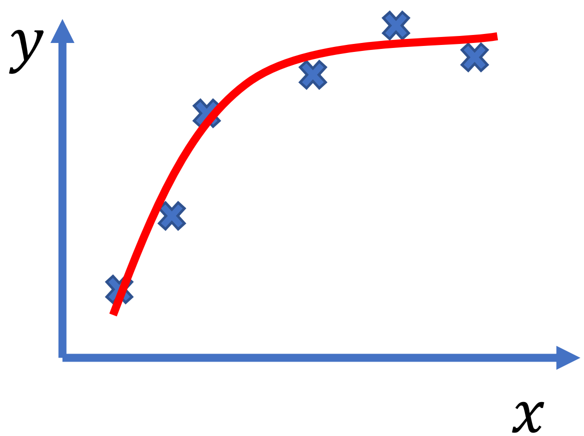
- 过拟合
- 数据集划分
- 性能度量
- 方差和偏差

过拟合与欠拟合



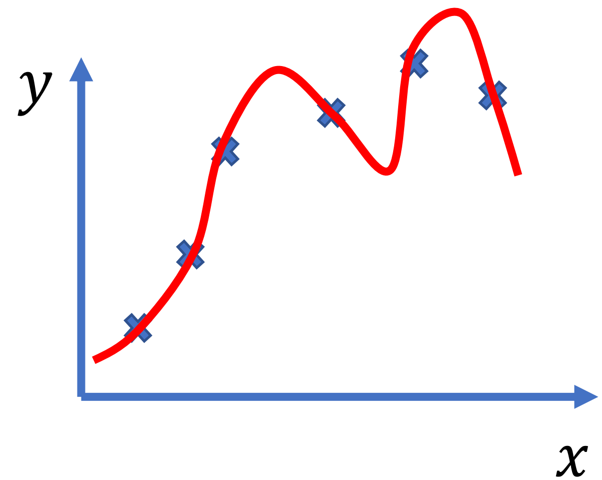
$f(x) = \omega x + b$

欠拟合，拟合不够



$f(x) = \omega_1 x + \omega_2 x^2 + b$

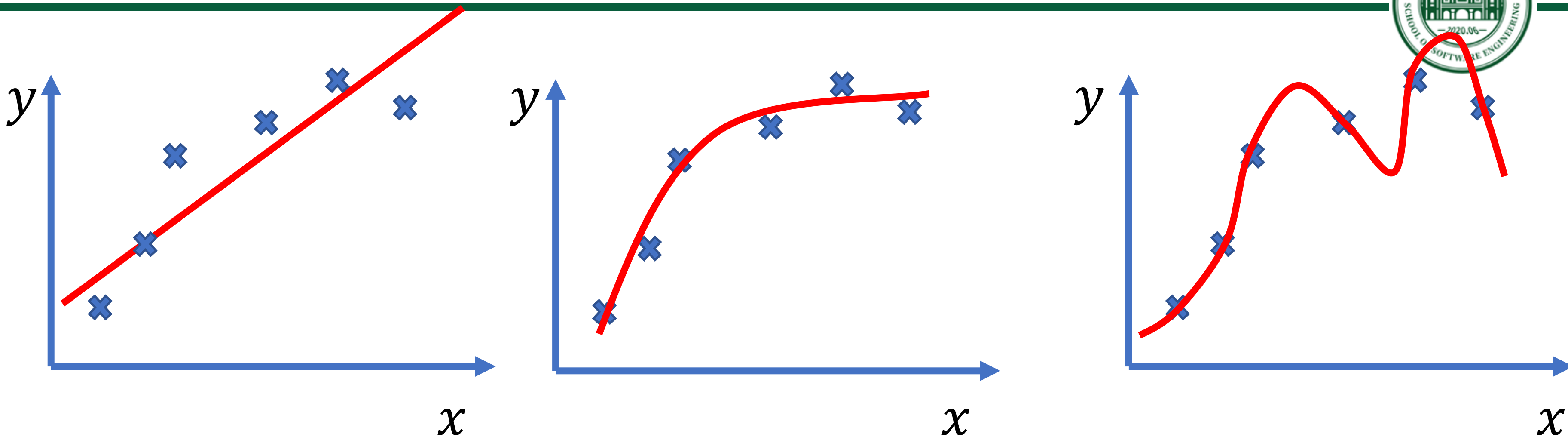
拟合较好



$f(x) = \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \omega_4 x^4 + b$

过拟合，拟合过度

过拟合



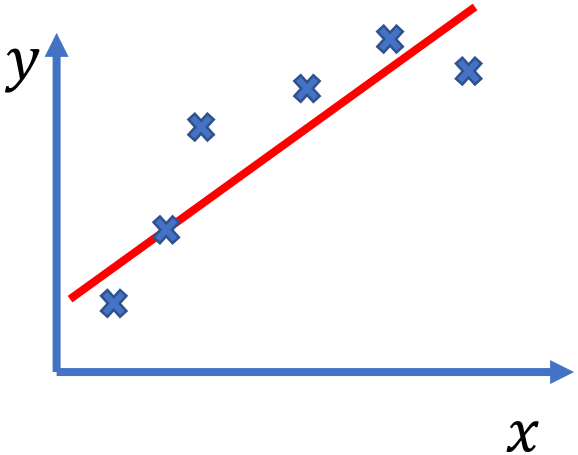
$f(x) = \omega x + b$ $\quad$ $f(x) = \omega_1 x + \omega_2 x^2 + b$ $\quad$ $f(x) = \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \omega_4 x^4 + b$

欠拟合，误差较大　　拟合较好，误差较小　　过拟合，没有误差

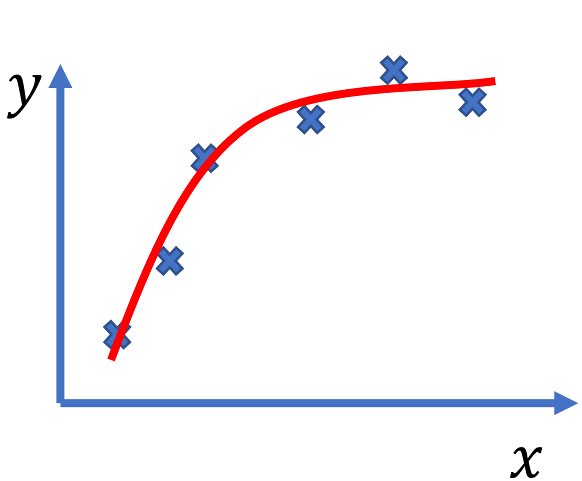
- 误差：预测输出与样本的真实输出之间的差异
- 训练误差：在训练数据上的误差
- 测试误差：在新的测试数据上的误差
- 过拟合：在训练数据上表现很好（训练误差小），在新的数据上表现较差的（测试误差大）

原因

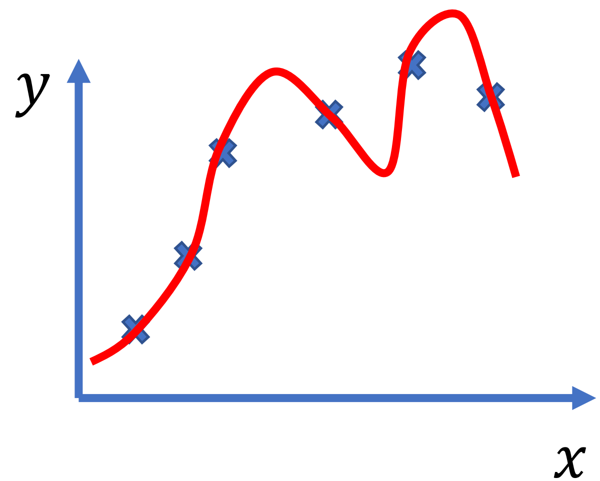
- 模型层面：模型复杂度过高，参数过多



$f(x) = \omega x + b$



$f(x) = \omega_1 x + \omega_2 x^2 + b$



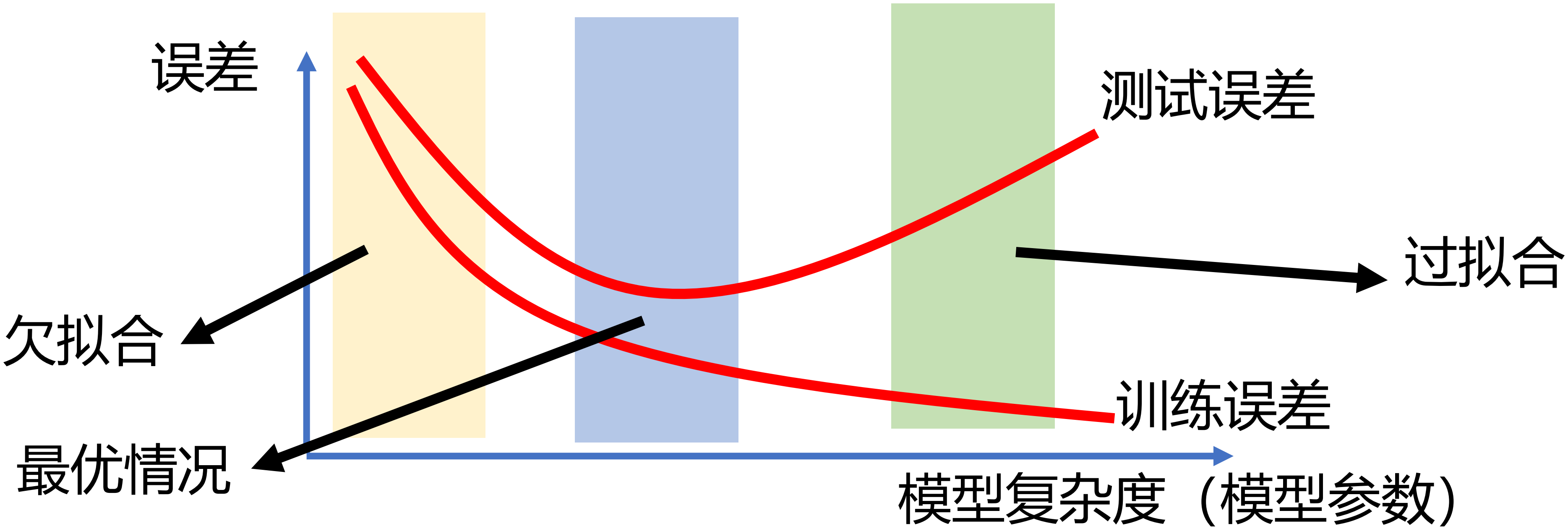
$f(x) = \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \omega_4 x^4 + b$

欠拟合：模型参数太少

过拟合：模型参数太多

原因

- 模型层面：模型复杂度过高，参数过多



- 数据层面：训练数据不足，数据存在噪声

过拟合举例

观察数列，补充括号的数字

1, 2, 4, 8，（ ）

直观的解法：

- $1 = 2^0$
- $2 = 2^1$
- $4 = 2^2$
- $8 = 2^3$
- $16 = 2^4$

过拟合解法：$f(n) = \frac{1}{6}n^3 - \frac{1}{2}n^2 + \frac{11}{6}n$

- $f(1) = \frac{1}{6} \times 1^3 - \frac{1}{2} \times 1^2 + \frac{11}{6} \times 1 = 1$
- $f(2) = \frac{1}{6} \times 2^3 - \frac{1}{2} \times 2^2 + \frac{11}{6} \times 2 = 2$
- $f(3) = \frac{1}{6} \times 3^3 - \frac{1}{2} \times 3^2 + \frac{11}{6} \times 3 = 4$
- $f(4) = \frac{1}{6} \times 4^3 - \frac{1}{2} \times 4^2 + \frac{11}{6} \times 4 = 8$
- $f(5) = \frac{1}{6} \times 5^3 - \frac{1}{2} \times 5^2 + \frac{11}{6} \times 5 = 17.5$

过拟合的解决方法

- 正则化
- 减少模型复杂度
- 增加数据量
- 早停法

正则化

正则化：往目标函数添加正则项

$$\min_{\boldsymbol{\omega}} L(\boldsymbol{\omega}) = \frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\omega_i^2$$

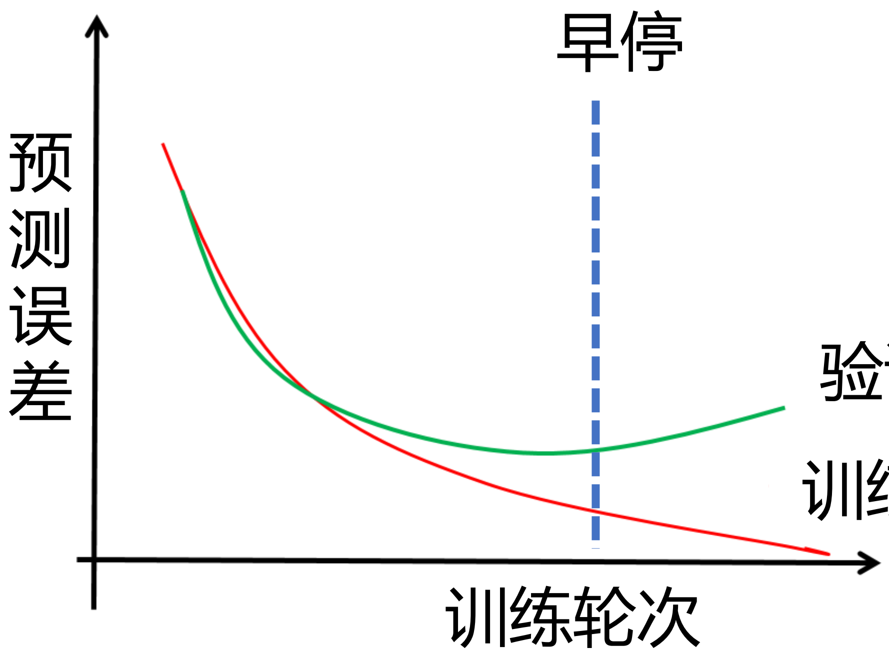
- ➢ L2正则项：$\sum_{i=1}^{d}\omega_i^2$, $\boldsymbol{\omega}$ 的$\boxed{\text{2范数}}$的平方
  - 也可用1范数作为正则项
- ➢ 通过引入惩罚项，驱使ω_i减少
- ➢ λ：超参数，用于调整正则项的影响
  - λ越大，最终的ω_i越小

p 范数：

$$\|\boldsymbol{\omega}\|_p \triangleq \left(\sum_{i=1}^{d}|\omega_i|^p\right)^{1/p}$$

过拟合的解决方法

- 正则化
- 减少模型复杂度
- 增加数据量
- 早停法



随堂小测

以下情况中，（）表明模型出现了过拟合，（）表明欠拟合

A. 模型在训练集上的准确率为98%，在测试集上的准确率为97%

B. 模型在训练集上的准确率为50%，在测试集上的准确率为49%

C. 模型在训练集上的准确率为98%，在测试集上的准确率为50%

D. 模型在训练集上的准确率为70%，在测试集上的准确率为72%

随堂小测

以下说法，错误的是（）

A. 模型过于复杂会导致过拟合，模型过于简单会导致欠拟合

B. 增加模型复杂度可以减缓过拟合

C. 正则化中的惩罚参数λ可以通过梯度下降法进行优化

D. 过拟合是指模型过度学习了训练数据的细节，导致泛化能力差

内容

- 过拟合
- 数据集划分
- 性能度量
- 方差和偏差

数据集划分

如何评估模型在新数据上的表现（即泛化能力）？

数据集划分

- 训练集：训练模型
- 验证集：调整超参数，如步长、正则项参数
- 测试集：评估模型性能

划分比例：60%-20%-20%、70% - 15% - 15%

数据集划分方法

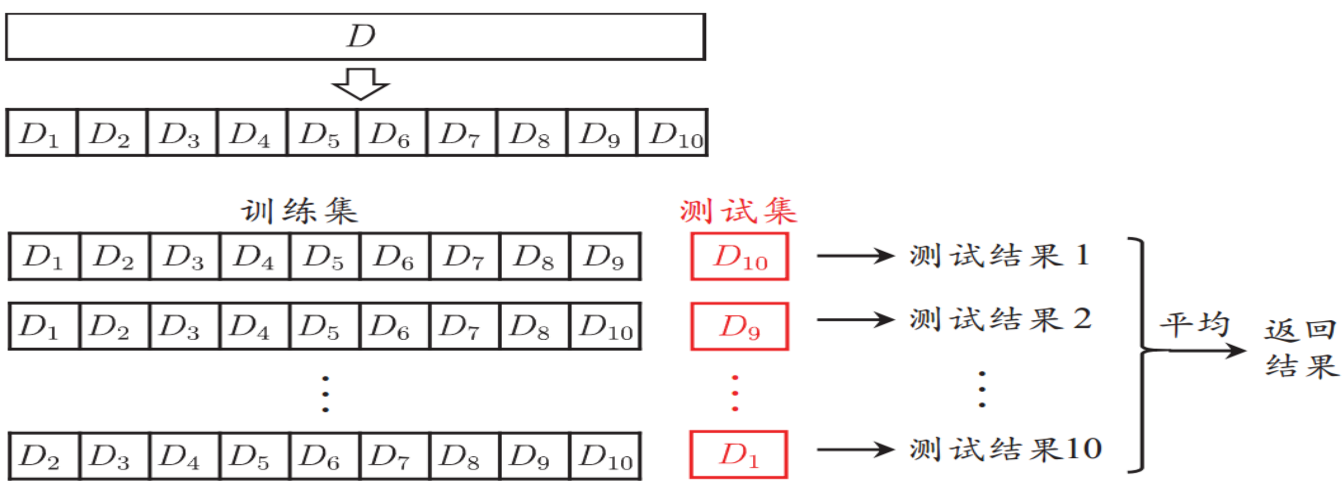
- 留出法
- 交叉验证法

留出法

- 直接将数据集划分为两个互斥集合作为训练/测试集
  - 例如，总共1000条数据，把其中700条组成训练集，剩余300条组成测试集
- 分层抽样：让训练/测试集划分要尽可能保持数据分布的一致性
  - 例如，1000条数据，正样本500条，负样本500条，训练集中正负样本各350条，测试集正负样本各150条
- 训练/测试样本比例通常为7:3、8:2

交叉验证法（k折交叉验证）

- 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



10 折交叉验证示意图

留一法：k为数据集的大小

对比

- 留出法：计算简单，评估结果不稳定，适用于数据量大的场景
- k折交叉验证法：计算复杂，评估稳定，适用于数据量小的场景，例如医疗数据

注意事项

- 训练集、验证集和测试集互斥
- 多次随机划分，然后取平均评估结果
- 分层抽样，保证子集数据分布一致

随堂小测

以下说法正确的是（）

A.划分训练集和测试集的主要目的是评估模型泛化能力

B.当数据量较小时，通常采用交叉验证法进行数据集划分

C.分层抽样的主要作用是保持类别分布在子集中一致

内容

- 过拟合
- 数据集划分
- 性能度量
- 方差和偏差

性能度量

回归任务:

- 均方误差/均方根误差：强调大误差的惩罚

$$\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right)^{2};\ \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right)^{2}}$$

- 平均绝对值误差：对异常值鲁棒

$$\frac{1}{m}\sum_{i=1}^{m}\left|f\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}\right|$$

- 平均绝对百分比误差：相对误差评估

$$\frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{f\left(\boldsymbol{x}^{(i)}\right)-y^{(i)}}{y^{(i)}}\right|$$

性能度量

分类任务：

- 错误率（分类错误样本占比）

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}^{(i)}\right)\neq y^{(i)}\right)$$

- 精度（分类正确样本占比）

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}^{(i)}\right)= y^{(i)}\right)$$

混淆矩阵

垃圾邮件检测、核酸检测等场景中经常需要衡量预测出来的正例中正确的比率或者正例被预测出来的比率

真实情况 \ 预测结果	**正例**	**负例**
正例	真正例	假负例
负例	假正例	真负例

混淆矩阵

- 真正例（True Positive，TP）：模型判断正确，预测为正例（实际为正例）的样本数
- 假正例（False Positive，FP）：模型判断错误，预测为正例（实际为负例）的样本数
- 真负例（True Negative，TN）：模型判断正确，预测为负例（实际为负例）的样本数
- 假负例（False Negative，FN）：模型判断错误，预测为负例（实际为正例）的样本数

混淆矩阵举例

垃圾邮件识别：100封邮件，其中30封垃圾邮件（正例），70封正常邮件（负例）

- 系统找出了35封垃圾邮件，其中，25封（真正例）确实是垃圾邮件，10封（假正例）实际是正常邮件
- 系统找出了65封正常邮件，其中，60封（真负例）确实是正常邮件，5封（假负例）实际是垃圾邮件

真实情况 \ 预测结果	正例（垃圾邮件）	负例（正常邮件）
正例（垃圾邮件）	真正例：25	假负例：5
负例（正常邮件）	假正例：10	真负例：60

混淆矩阵举例

急病诊断：200人，其中50名患病（正例），150名没有患病（负例）

- 系统诊断出60个患病的，其中，40人（真正例）确实是患病的，20人（假正例）实际为健康的
- 系统诊断出140个健康的，其中，130人（真负例）确实是健康的，10人（假负例）实际为患病

预测结果 / 真实情况	正例（患病）	负例（健康）
正例（患病）	真正例：40	假负例：10
负例（健康）	假正例：20	真负例：130

查准率和查全率

需要衡量预测出来的正例中正确的比率，即查准率，

以及正例被预测出来的比率，即查全率

真实情况 \ 预测结果	正例	负例
正例	真正例	假负例
负例	假正例	真负例

查准率 (precision)$=\frac{TP}{TP+FP}$,

- 分母为预测为正例的个数
- 适用场景：注重减少误报
- 例如，垃圾邮件检测，不希望把正常邮件(0)误判为垃圾邮件(1)

查全率 (recall)$=\frac{TP}{TP+FN}$,

- 分母为实际为正例的个数
- 适用场景：注重减少漏报
- 例如，癌症筛查，不希望把患病的(1)漏掉，当成健康的(0)

随堂小测

使用一个垃圾邮件检测系统对15封邮件进行检测，系统显示有10封垃圾邮件，其中有6封确实是垃圾邮件，剩余的4封实际正常邮件。同时，还有3封垃圾邮件未被检测出来。那么查准率为（），查全率为（）

F1 Score

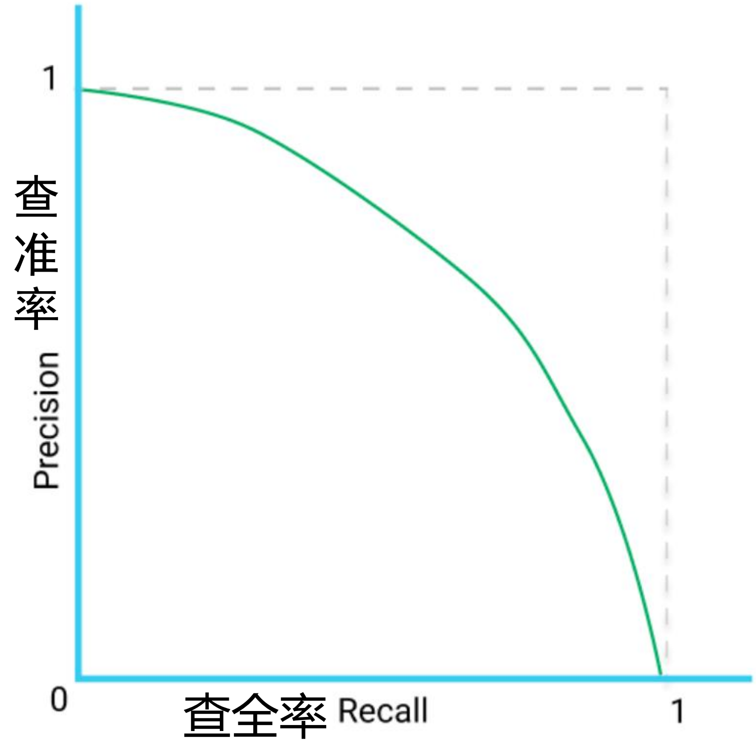
F1 score： $F1=\frac{2\times P\times R}{P+R}$ 满足：$\frac{1}{F1}=\frac{1}{2}(\frac{1}{P}+\frac{1}{R})$

- 综合了查准率（P）和查全率（R）的评估指标，适用于需要平衡这两者的场景，当P和R都较高时，F1 score才会高
- 适用场景：**数据分布不平衡**或**特定任务对误报和漏报敏感**的情况下
  - 类别不平衡：金融欺诈检测。欺诈交易占比极低，漏报（低查全）会导致损失，误报正常交易（低查准）则影响用户体验
  - 需要同时关注查准率和查全率：信息检索。用户希望返回的结果既相关（高查准）又全面（高查全）。例如，搜索“机器学习教程”时，系统需避免无关结果（高查准），返回所有优质内容（高查全）

P-R曲线

$$\hat{y} = \begin{cases} 1, & p(y=1|\boldsymbol{x}) \geq \text{阈值} \\ 0, & otherwise \end{cases}$$

根据模型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



P-R曲线

$$\hat{y} = \begin{cases} 1, & p(y=1|\boldsymbol{x}) \geq \text{阈值} \\ 0, & otherwise \end{cases}$$

查准率 =真正/(真正+假正)
分母为预测为正例的个数

查全率 =真正/(真正+假负)
分母为实际为正例的个数

样本	预测正例可能性	真实情况
1	0.95	1
2	0.90	1
3	0.85	0
4	0.80	1
5	0.70	0
6	0.60	1
7	0.55	1
8	0.50	0
9	0.40	0
10	0.35	0

判定为正，1

阈值=0.35

查准率：5/10
查全率：5/5

5个正样本，1　　5个负样本，0

P-R曲线

$$\hat{y} = \begin{cases} 1, & p(y=1|\boldsymbol{x}) \geq \text{阈值} \\ 0, & otherwise \end{cases}$$

查准率 =真正/(真正+假正)
分母为预测为正例的个数
查全率 =真正/(真正+假负)
分母为实际为正例的个数

样本	预测正例可能性	真实情况
1	0.95	1
2	0.90	1
3	0.85	0
4	0.80	1
5	0.70	0
6	0.60	1
7	0.55	1
8	0.50	0
9	0.40	0
10	0.35	0

判定为正，1

阈值=0.6

判定为负，0

查准率：4/6
查全率：4/5

5个正样本，1

5个负样本，0

P-R曲线

$$\hat{y} = \begin{cases} 1, & p(y=1|\boldsymbol{x}) \geq \text{阈值} \\ 0, & otherwise \end{cases}$$

查准率 =真正/(真正+假正)
分母为预测为正例的个数

查全率 =真正/(真正+假负)
分母为实际为正例的个数

样本	预测正例可能性	真实情况
1	0.95	1
2	0.90	1
3	0.85	0
4	0.80	1
5	0.70	0
6	0.60	1
7	0.55	1
8	0.50	0
9	0.40	0
10	0.35	0

判定为正，1

阈值=0.85

查准率：2/3
查全率：2/5

判定为负，0

5个正样本，1

5个负样本，0

P-R曲线

$$\hat{y}=\begin{cases}1, & p(y=1|\boldsymbol{x}) \ \geq \text{阈值} \\ 0, & otherwise\end{cases}$$

样本	预测正例可能性	真实情况
1	0.95	1
2	0.90	1
3	0.85	0
4	0.80	1
5	0.70	0
6	0.60	1
7	0.55	1
8	0.50	0
9	0.40	0
10	0.35	0

判定为正，1

阈值=0.95

查准率：1/1

查全率：1/5

判定为负，0

5个正样本，1

5个负样本，0

查准率 =真正/(真正+假正)

分母为预测为正例的个数

查全率 =真正/(真正+假负)

分母为实际为正例的个数

P-R曲线

$$\hat{y} = \begin{cases} 1, & p(y=1|\boldsymbol{x}) \geq \text{阈值} \\ 0, & otherwise \end{cases}$$

样本	预测正例可能性	真实情况
1	0.95	1
2	0.90	1
3	0.85	0
4	0.80	1
5	0.70	0
6	0.60	1
7	0.55	1
8	0.50	0
9	0.40	0
10	0.35	0

阈值

查准率：1/1
查全率：1/5

阈值

查准率：2/3
查全率：2/5

阈值

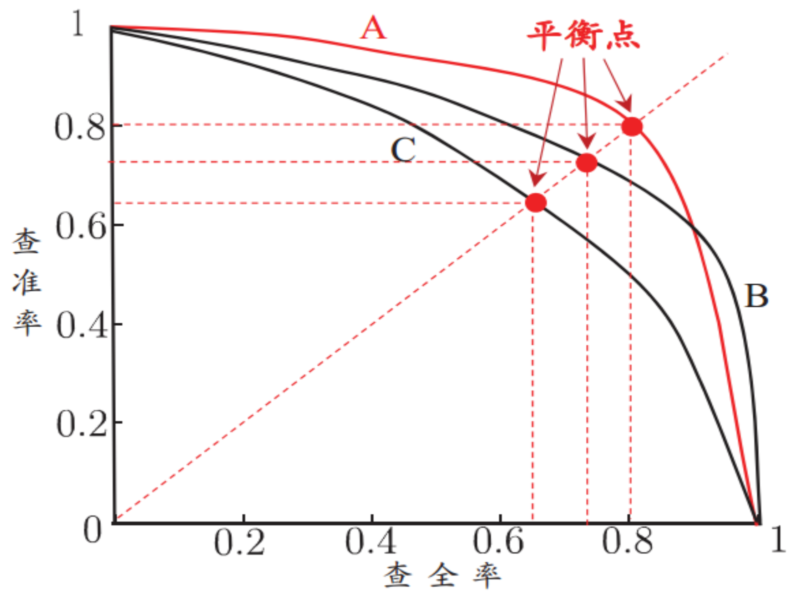
查准率：4/6
查全率：4/5

阈值

查准率：5/10
查全率：5/5

阈值越低，查全率增加，查准率降低

P-R曲线



P-R曲线与平衡点示意图

- P-R曲线越往外的模型能力越好
- 有交叉：曲线上“查准率=查全率”时的取值（平衡点）

ROC曲线

根据模型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”(Receiver Operating Characteristic)

真实情况 \ 预测结果	**正例**	**负例**
正例	真正例	假负例
负例	假正例	真负例

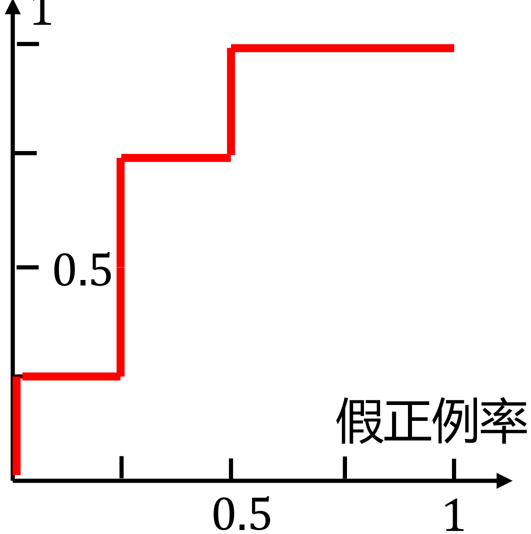
假正例率 = $\frac{FP}{TN+FP}$

分母为实际为负的个数

真正例率 = $\frac{TP}{TP+FN}$，即查全率

分母为实际为正的个数

ROC曲线



4个正样本　　4个负样本

真正率 =真正/(真正+假反)

假正率 =假正/(假正+真反)

样本	预测正例可能性	标记
1	0.91	1
2	0.85	0
3	0.77	1
4	0.72	1
5	0.61	0
6	0.48	1
7	0.42	0
8	0.33	0

阈值　$(x, y) = (0, 0)$

阈值　当前为真正例 $\left(x, y+\frac{1}{4}\right) = (0, \frac{1}{4})$

阈值　当前为假正例 $\left(x+\frac{1}{4}, y\right) = (\frac{1}{4}, \frac{1}{4})$

阈值　当前为真正例 $\left(x, y+\frac{1}{4}\right) = (\frac{1}{4}, \frac{2}{4})$

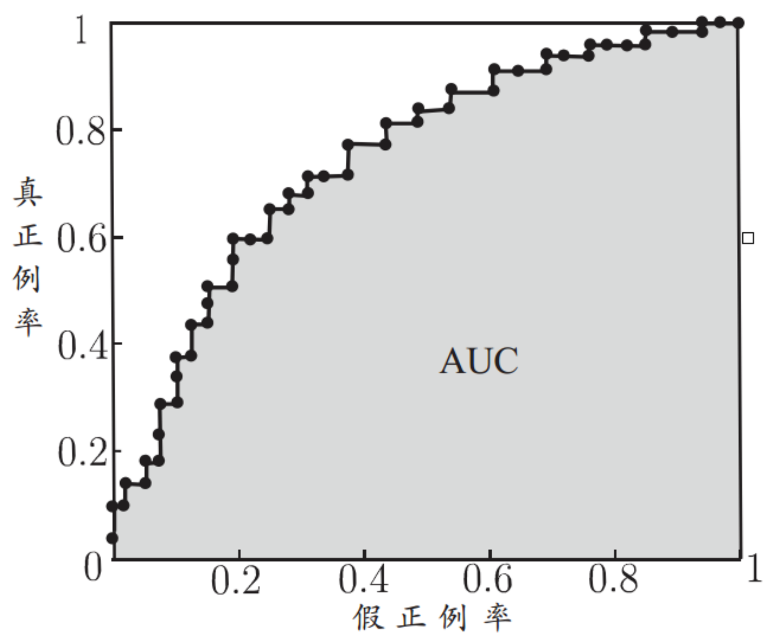
阈值　当前为真正例 $\left(x, y+\frac{1}{4}\right) = (\frac{1}{4}, \frac{3}{4})$

阈值　当前为假正例 $\left(x+\frac{1}{4}, y\right) = (\frac{2}{4}, \frac{3}{4})$

阈值　当前为真正例 $\left(x, y+\frac{1}{4}\right) = (\frac{2}{4}, \frac{4}{4})$

ROC曲线

- ROC曲线越往外的模型能力越好
- 如果曲线交叉，可以根据ROC曲线下面积 (Area under curve, AUC)大小进行比较
  - AUC越大，性能越好



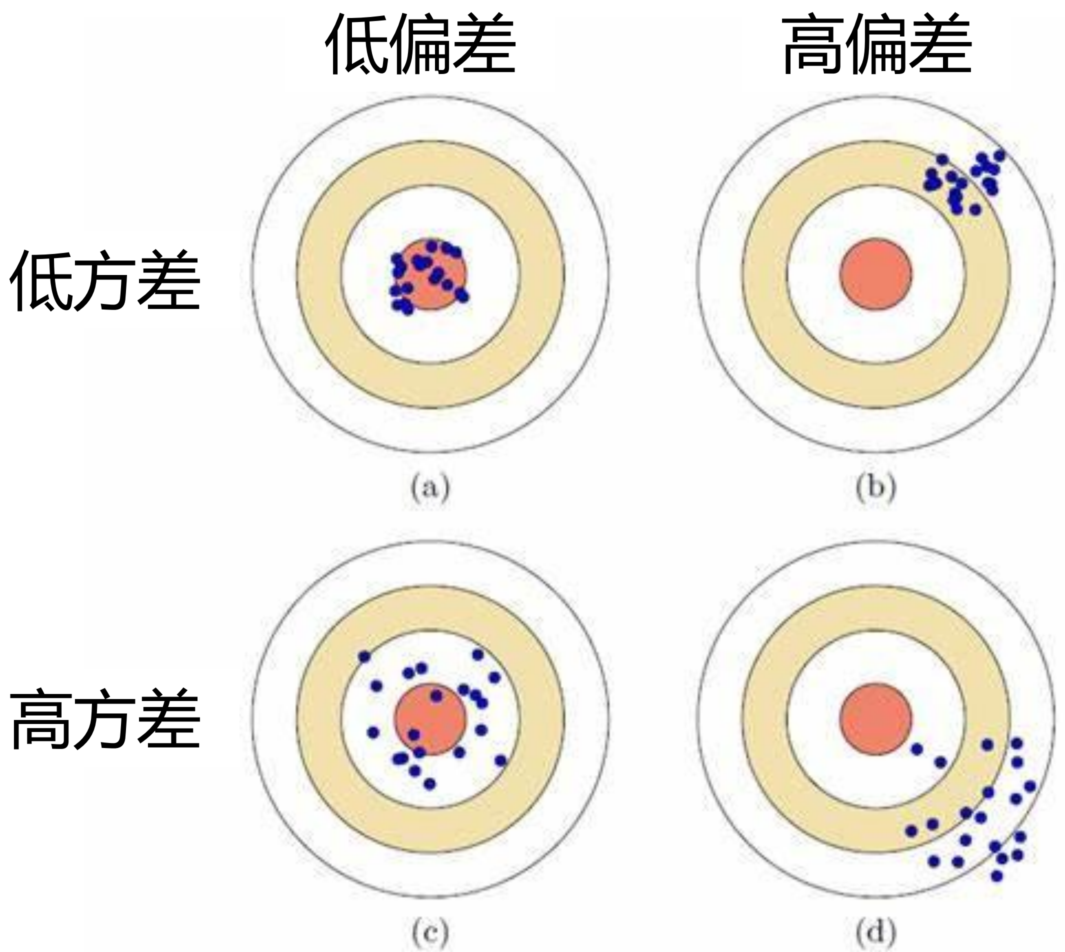
基于有限样例绘制的 ROC 曲线
与 AUC

内容

- 过拟合
- 数据集划分
- 性能度量
- 方差和偏差

偏差与方差

解释模型泛化性能：偏差和方差



偏差与方差

以回归任务为例：对测试样本$\boldsymbol{x}$, 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

- 模型预测输出的期望:

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

- 期望输出与真实标记的差别，即偏差:

$$bias(\boldsymbol{x}) = \bar{f}(\boldsymbol{x}) - y$$

- 使用样本数目相同的不同训练集产生的方差:

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

偏差与方差分解

- 算法的期望：$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$
- 方差：$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right]$
- 偏差：$bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$
- 泛化误差：$E_D[(f(x; D) - y)^2]$

偏差

$$E_D[(f(x; D) - y)^2] = E_D\left[\left(f(x; D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y\right)^2\right]$$

方差

$$= E_D\left[\left(f(x; D) - \bar{f}(\boldsymbol{x})\right)^2\right] + E_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)^2\right]$$

$$+ E_D\left[2\left(f(x; D) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - y\right)\right]$$

$$= var(\boldsymbol{x}) + bias^2(\boldsymbol{x})$$

$$\begin{aligned} & E_D[f(x; D) - \bar{f}(\boldsymbol{x})] \\ = & E_D[f(x; D)] - \bar{f}(\boldsymbol{x}) \\ = & \bar{f}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}) \\ = & 0 \end{aligned}$$

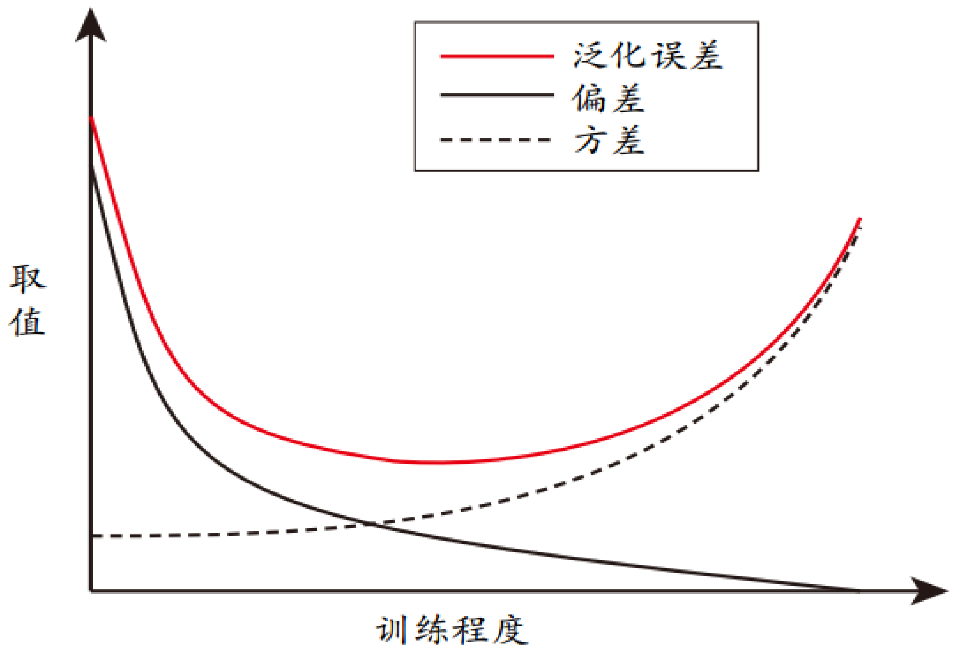
偏差与方差分解

泛化误差等于偏差的平方和方差之和:

- **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

偏差与方差

- 在训练不足时，模型拟合能力不强，偏差主导泛化误差
- 训练程度加深，模型拟合能力逐渐增强，偏差↓，方差↑
- 训练过度后，产生过拟合，训练数据的轻微扰动都会导致模型的显著变化，方差主导泛化误差



泛化误差与偏差、方差的关系示意图

总结

- 过拟合：训练误差低，测试误差高，原因：模型参数过多
- 训练集/测试集的划分：评估模型的泛化能力
- 评估度量：查准率、查全率、F1、P-R曲线、ROC曲线
- 误差来源：偏差-方差分解