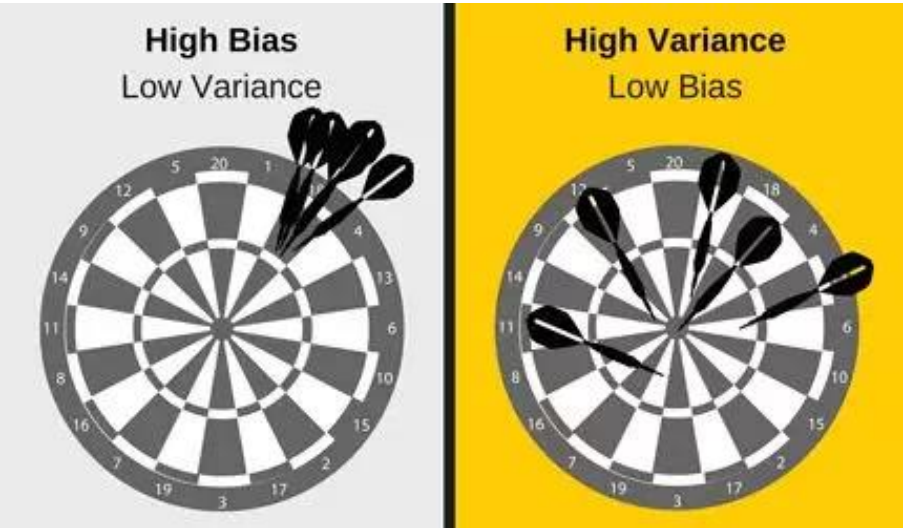
《机器学习及其在化学中的应用》2025年课程

Sec. 4

模型选择与评估

Model Selection and Assessment



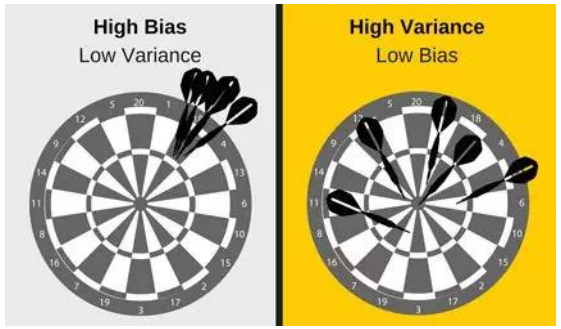
刘志荣 (LiuZhiRong@pku.edu.cn)

北京大学化学学院

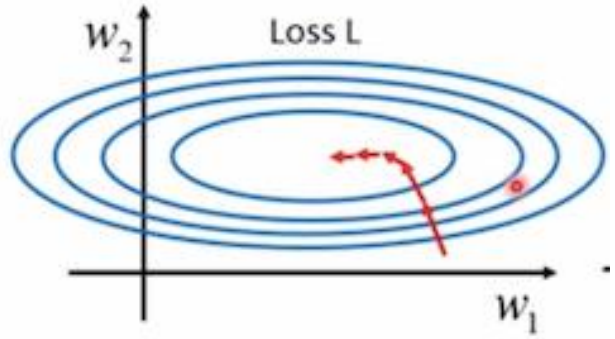
2025.10.13

内容提要

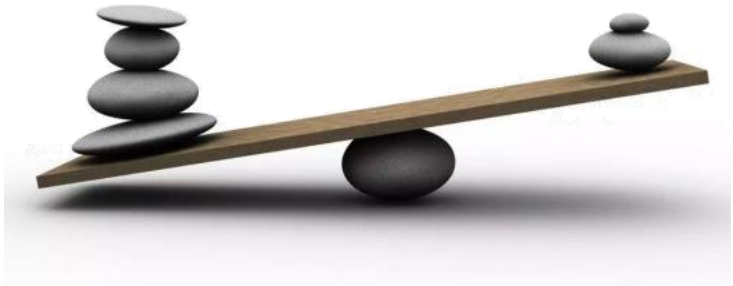
- 模型选择与交叉验证

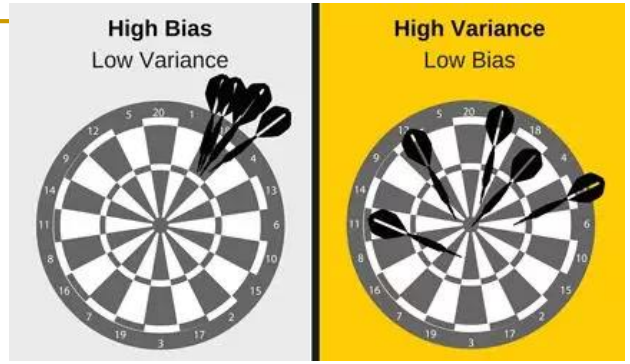


- 数据预处理与模型保存



- 数据不平衡及分类模型的评价指标





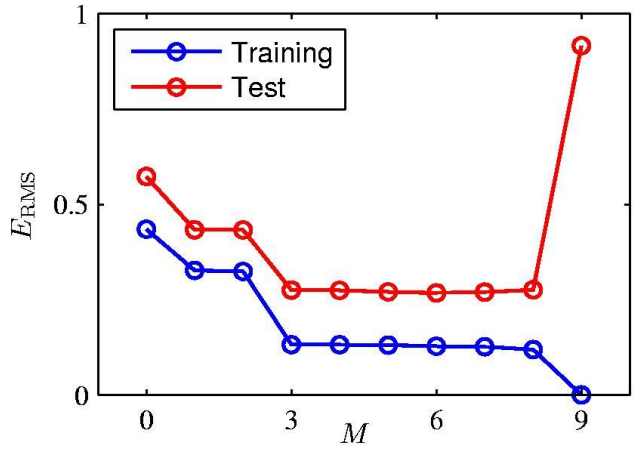
1. 模型选择与交叉检验…

- 训练集、验证集、测试集
- 偏差与方差，欠拟合与过拟合
- 判断方法：
  - 训练集与验证集（测试集）误差随模型复杂性的变化曲线；
  - 学习曲线（误差随训练集数据量的变化曲线）。
- K折交叉验证，留一法

超参数与验证集

- 模型中有超参数，如正则化强度λ、基函数的数目、神经网络的隐藏层节点数目，等等。甚至是不同模型。如何选择？
- 严格地讲，不能用测试集误差来选择超参数与模型，否则测试集的信息会“泄露”到模型选择中，从而导致测试集误差不能用来独立衡量模型的泛化能力。
- 因此，较好的做法是把数据分成训练集(training set)、验证集(validation set)与测试集(test set)。
  - 分别用于模型的训练（training）、选择（selection）与最终选中的最优模型的测试评估（assessment）。
  - （Elements）建议：50% : 25% : 25%

误差的偏差-方差分解



为什么误差会两头翘？

- 假设观察值 t 可表示成$t(x) = f(x) + \epsilon$，基于训练集$\mathcal{D}$训练得到的函数为 $\hat{f}(x|\mathcal{D})$。则在某一个测试（验证）数据（$\{x_0, t(x_0)\}$）下的误差为

$$\begin{aligned}
&\left[f(x_0) + \varepsilon - \hat{f}(x_0|\mathcal{D})\right]^2 \\
&= \left[f(x_0) + \varepsilon - \hat{f}(x_0|\mathcal{D}) + \left\langle\hat{f}(x_0|\mathcal{D})\right\rangle - \left\langle\hat{f}(x_0|\mathcal{D})\right\rangle\right]^2 \\
&= \left\{\left[f(x_0) - \left\langle\hat{f}(x_0|\mathcal{D})\right\rangle\right] - \left[\hat{f}(x_0|\mathcal{D}) - \left\langle\hat{f}(x_0|\mathcal{D})\right\rangle\right] + \varepsilon\right\}^2
\end{aligned}$$

  - 其中$\left\langle\hat{f}(x_0|\mathcal{D})\right\rangle$代表不同训练集（例如，由于$x$或$\epsilon$的不同取值而得到的数据）下得到的不同拟合结果的期望值（平均值）。

- $f(x_0) - \langle \hat{f}(x_0|\mathcal{D}) \rangle$是常数，因此误差的期望值为

$$\left\langle \left[f(x_0) + \varepsilon - \hat{f}(x_0|D) \right]^2 \right\rangle$$

$$= \langle \varepsilon^2 \rangle + \left[f(x_0) - \langle \hat{f}(x_0|D) \rangle \right]^2 + \left\langle \left[\hat{f}(x_0|D) - \langle \hat{f}(x_0|D) \rangle \right]^2 \right\rangle$$

$$-2\left\langle \left[\hat{f}(x_0|D) - \langle \hat{f}(x_0|D) \rangle \right] \varepsilon \right\rangle$$

$$= \langle \varepsilon^2 \rangle + \mathrm{Bias}^2[f(x_0)] + \mathrm{Var}[f(x_0)]$$

两部分的随机性分别来源于训练集与测试集，是独立的，因此结果为0。

偏差

与模型容量有关

（预测平均值与真实值之间的偏差）

方差

与数据多少有关

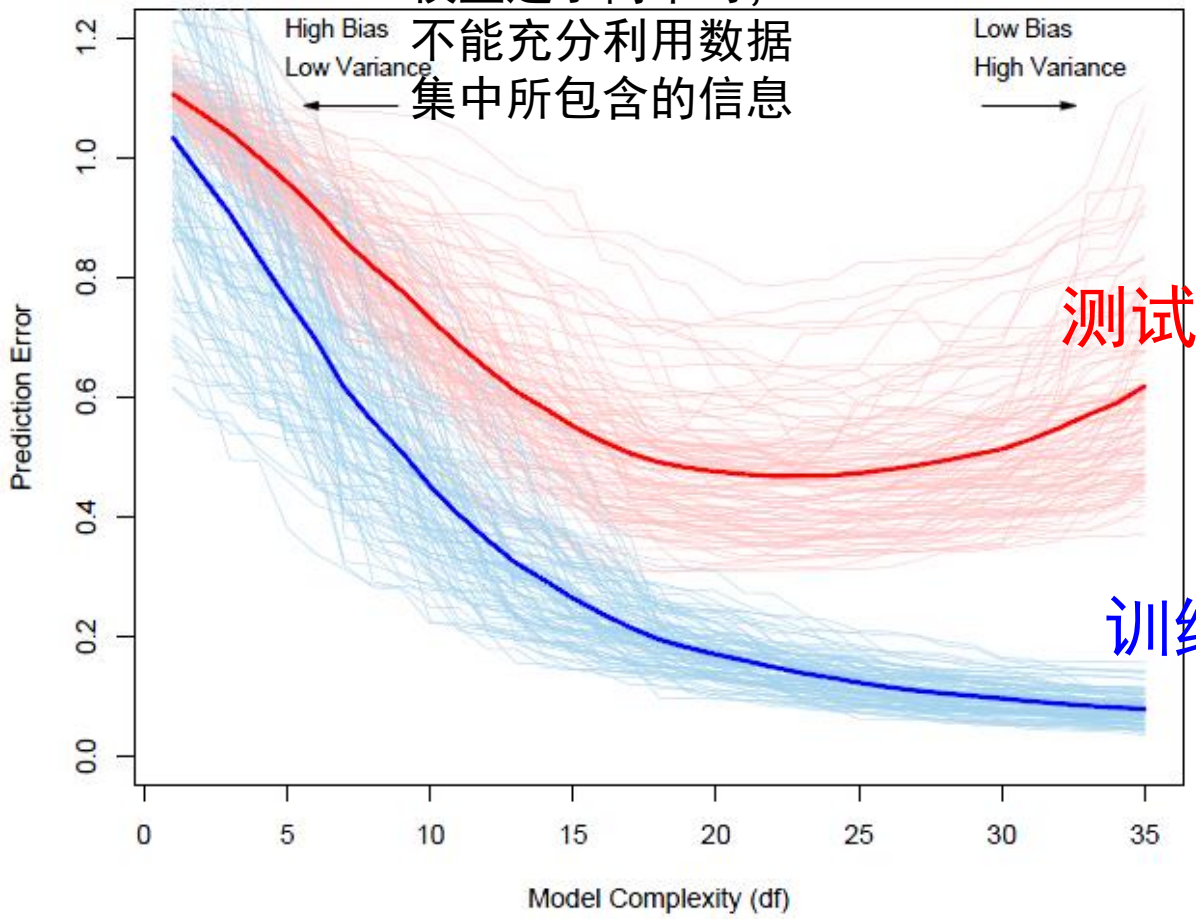
（每次预测的值与预测平均值之间的涨落）

偏差–方差权衡

欠拟合

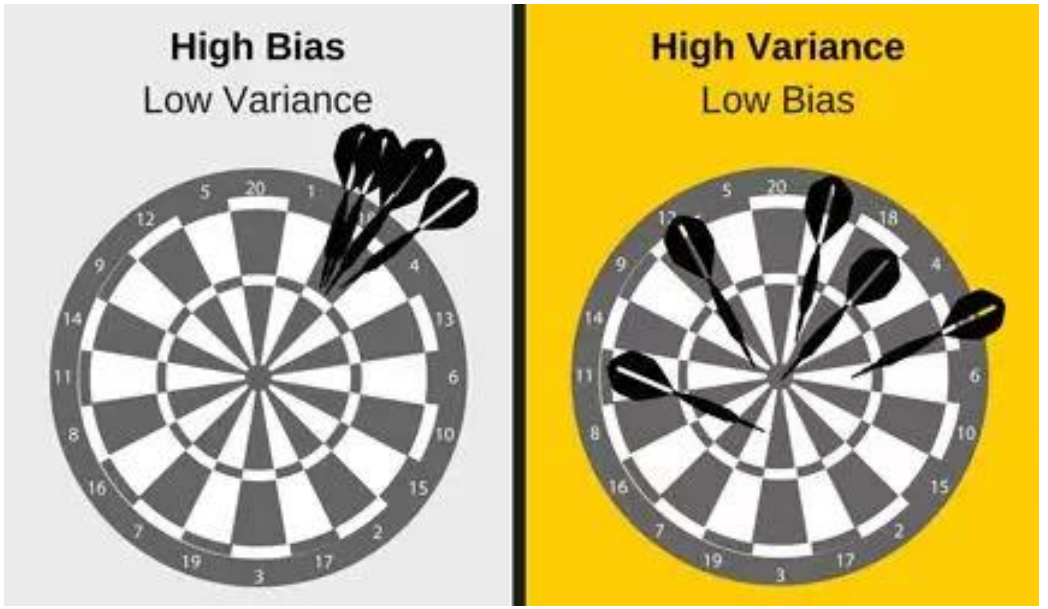
模型过于简单时，
不能充分利用数据
集中所包含的信息

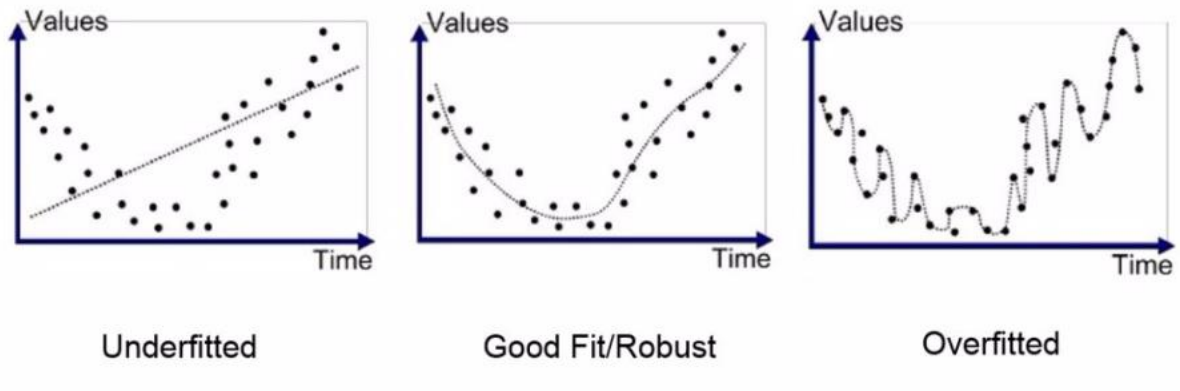
过拟合



测试误差

训练误差

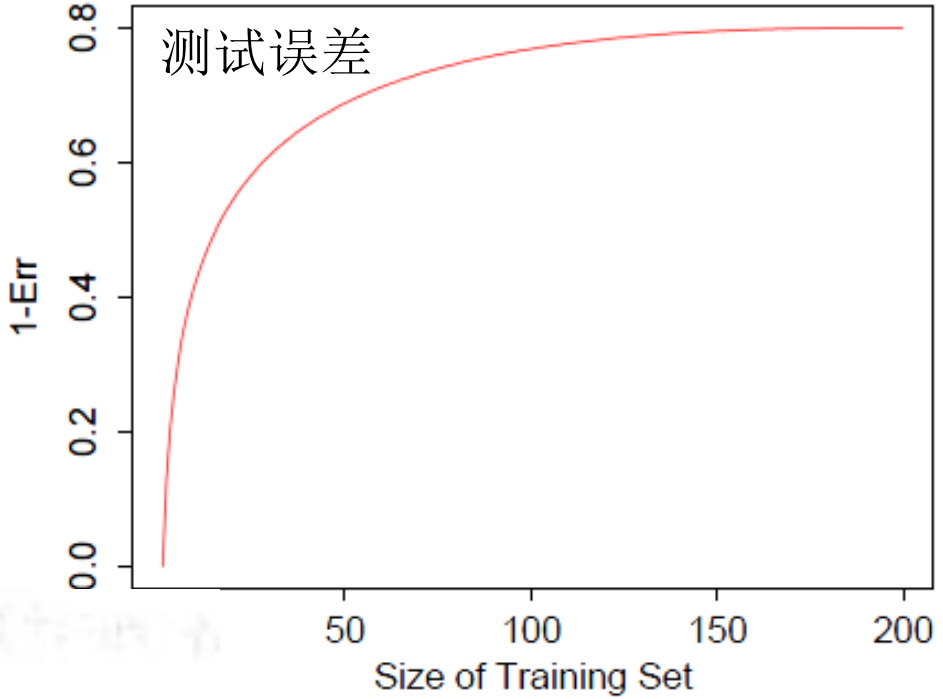


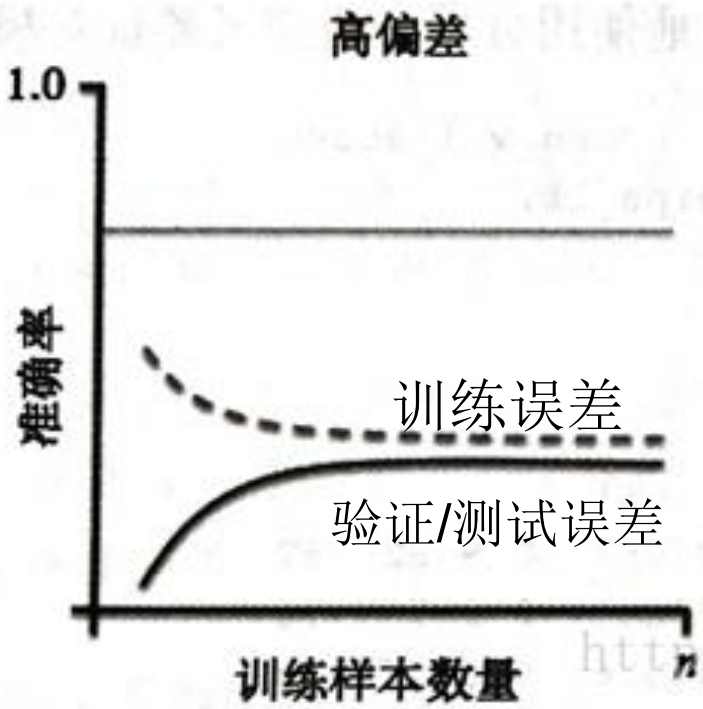


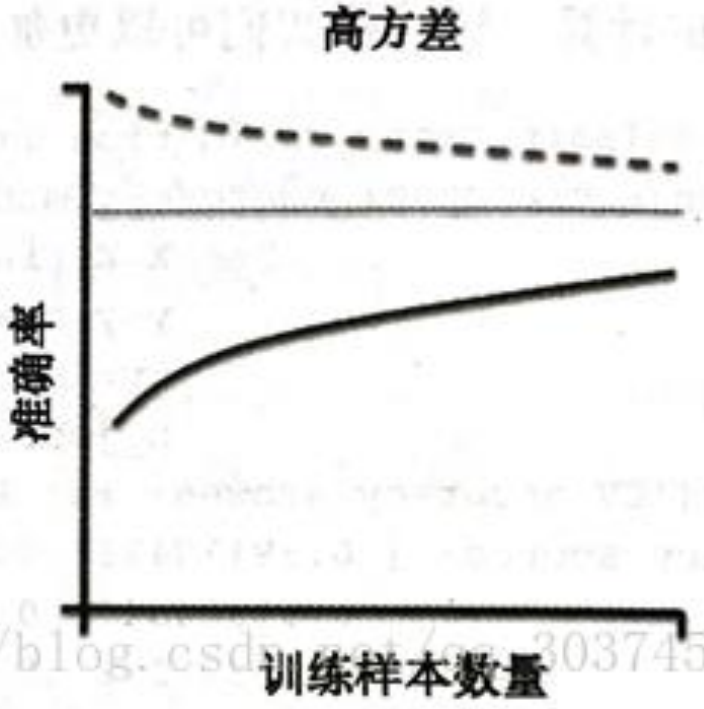
The light blue curves show the training error, while the light red curves show the conditional test error for 100 training sets of size 50 each, as the model complexity is increased.

学习曲线…

- 精度/误差随训练集数据量的变化曲线。
- 也可用于判断欠拟合/过拟合。
  - 欠拟合时过早饱和且两者差别小（方差小）；
    过拟合时未饱和且差别大。

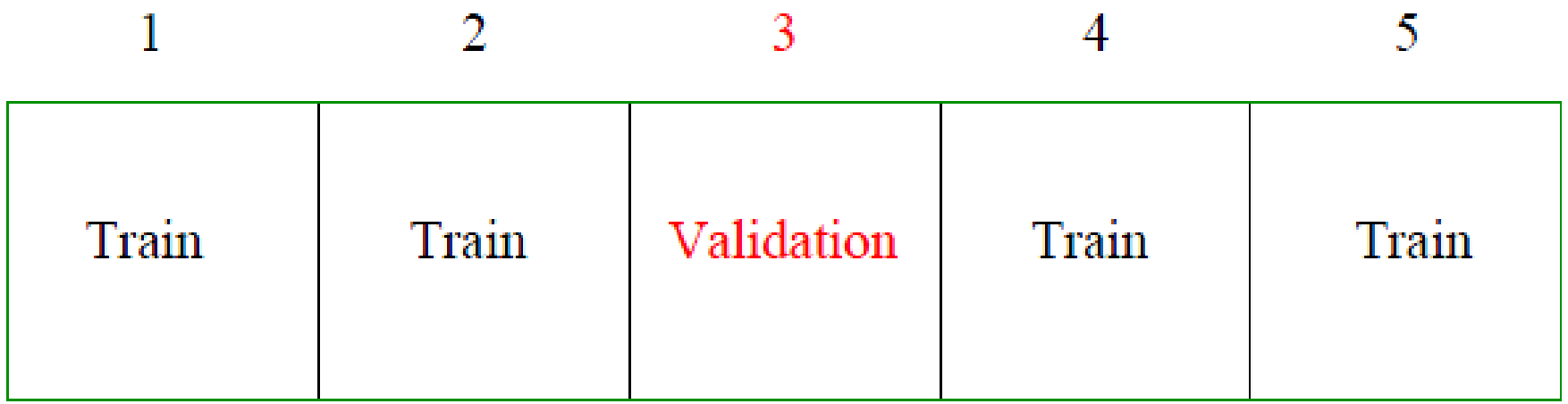


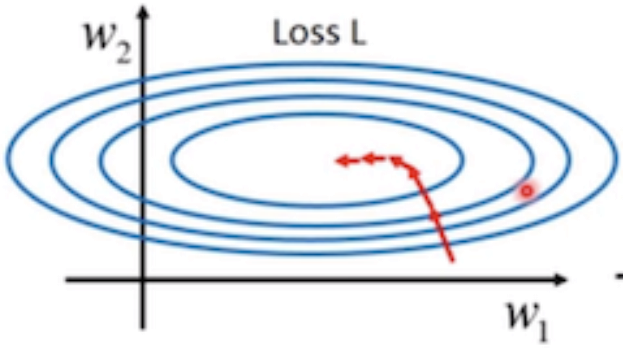




K折交叉验证（K-Fold Cross-Validation）

- 为了减少数据的浪费，可采用交叉验证。
- 将数据划分为K份，每次拿出1份作为验证集而其余$K-1$份作为训练集，最后将得到的性能/误差进行平均作为模型选择的指标。（选定模型后再用所有数据对其重新训练）
- 虽然计算代价很高，但是它不会浪费太多的数据。
- 当 $K=N$ （样本总数）时，叫做留一法（Leave one out cross validation）
- $K=5,10$



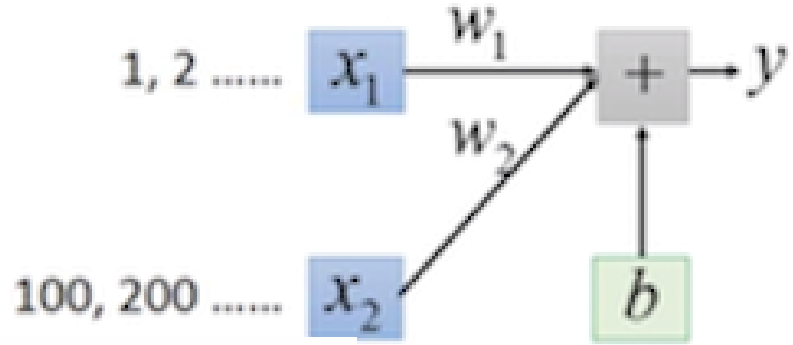


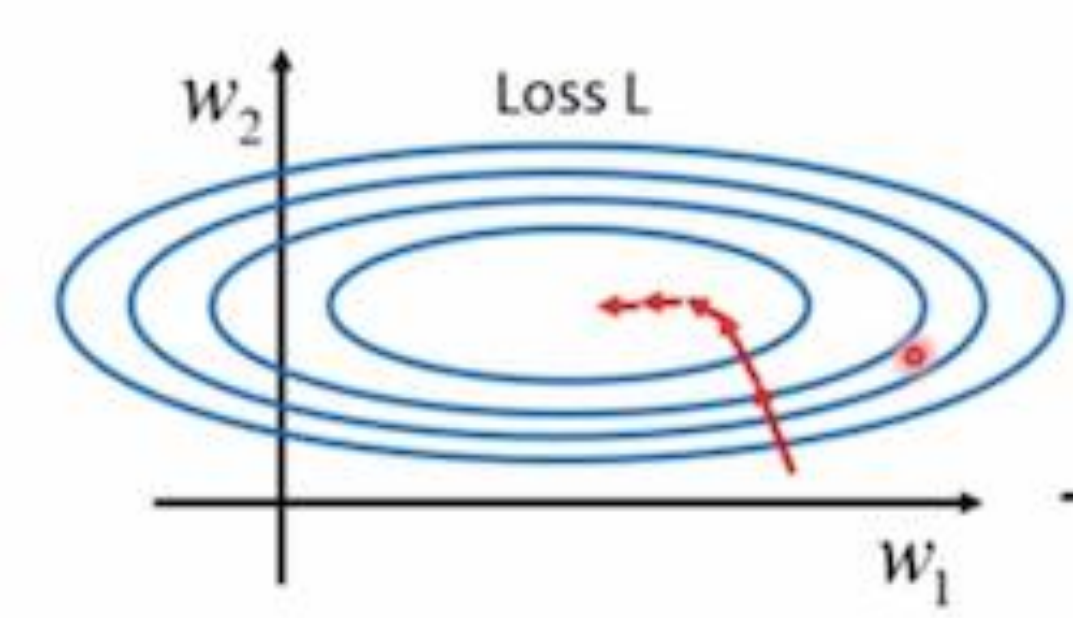
2. 数据预处理与模型保存

- 特征归一化
- 非数值型数据的处理
- 异常值检测
- 训练模型的保存与载入

预处理：特征归一化

- 当不同特征的数值相差过大时，代价函数的轮廓是扁长的，梯度下降的过程曲折且耗时。
  - 例如，$y = b + w_1 x_1 + w_2 x_2$





- 因此，往往需要对数据进行标准化处理。

- 方法1：Min-Max scaling (normalization)
  - 调节至[0,1]区间。

$$x_i' = \frac{x_i - \min(\{x_i\})}{\max(\{x_i\}) - \min(\{x_i\})}$$

- 方法2：Z-score normalization (standardization)
  - 调节后均值为0，方差为1。

$$x_i' = \frac{x_i - \mu(\{x_i\})}{\sigma(\{x_i\})}$$

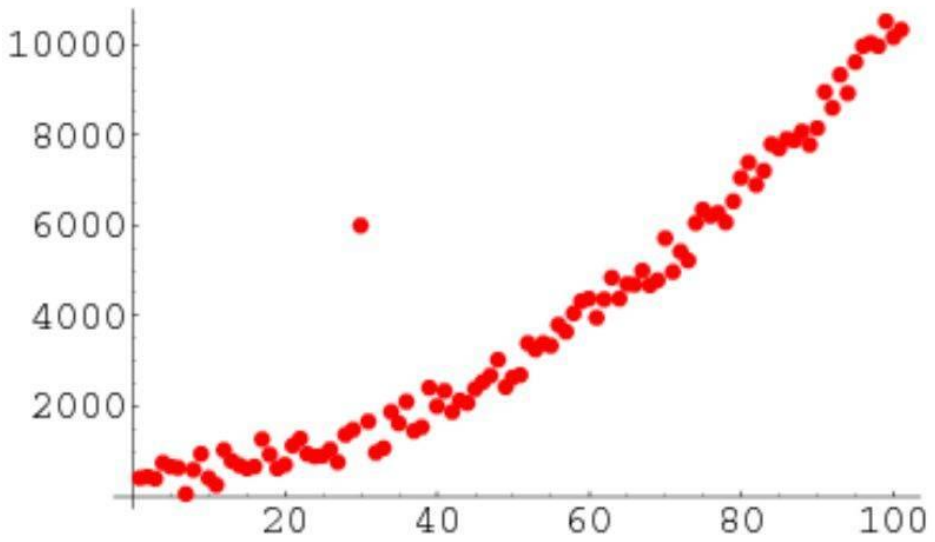
- 线性回归、逻辑回归 、神经网络、支持向量机等最优化问题一般需要归一化。决策树方法不需要归一化。

预处理：非数值型特征的处理

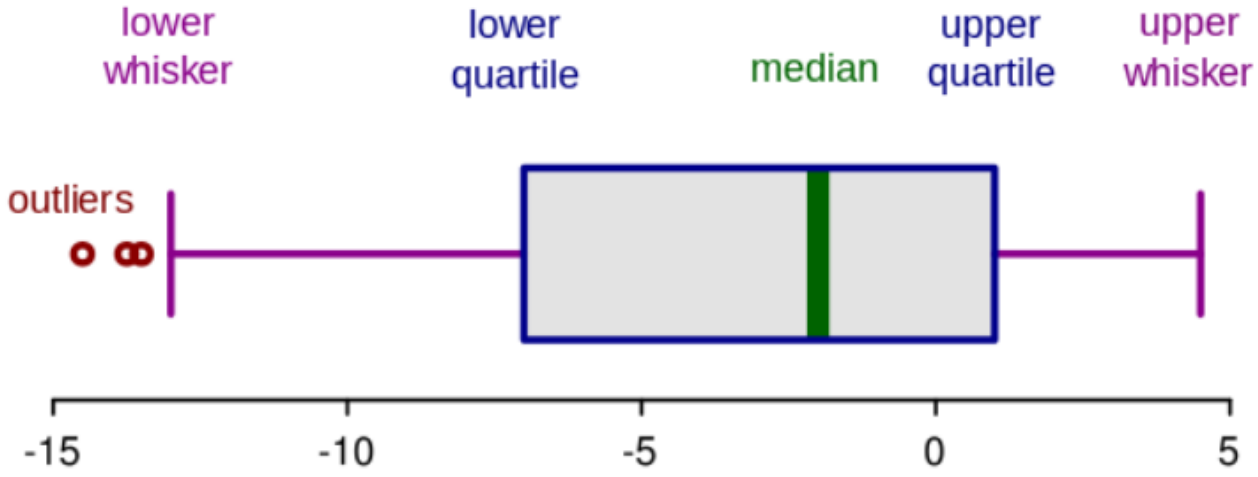
- 非数值型数据例子：
  - 薪资水平 {“low”, “medium”, “high”}
  - 岗位
    {“technical”, “sales”, “support”, “management”, “marketing”, …}
- 方法1：普通数值转换（数据具有顺序意义）
  - 例子： {“low”: 0, “medium”: 1, “high”: 2}
  - sklearn 的LabelEncoder()
- 方法2：独热编码（One-Hot Encoding）
  - 转换成多个0-1分量，且只能有一个分量为1：
  - sklean 的OneHotEncoder()和LabelBinarizer()

$$\begin{bmatrix}1\\0\\0\\0\\\dots\end{bmatrix},\begin{bmatrix}0\\1\\0\\0\\\dots\end{bmatrix},\begin{bmatrix}0\\0\\1\\0\\\dots\end{bmatrix},\begin{bmatrix}0\\0\\0\\1\\\dots\end{bmatrix},\dots$$

预处理：异常值检测



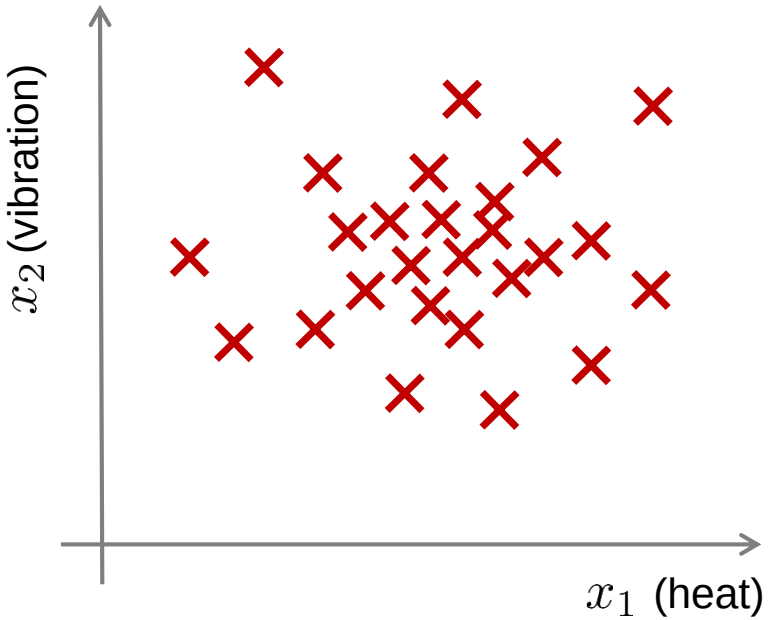
- 方法1：均方差（Z-score方法）
  - $z_i = \frac{x_i - \mu(\{x_i\})}{\sigma(\{x_i\})}$
  - 常用临界值：2.5, 3.0, 3.5
- 方法2：箱形图



- 其它方法：DBScan 聚类、孤立森林、…

异常值检测的其它应用…

事出反常必有妖？



- 产品质量检测
  - 飞机发动机：热量、振动、…
- 账户欺骗识别
  - 用户多久登录一次，访问过的页面，在论坛发布的帖子数量，甚至是打字速度等。
- 数据中心服务器故障
  - 内存使用情况，被访问的磁盘数量，CPU的负载，网络的通信量等

训练模型的保存与载入

- Model persistence
- Python（Scikit-learn）内部：
  - Python的pickle模块
  - Python的joblib模块(dump & load)

```
>>> import pickle
>>> s = pickle.dumps(clf)
>>> clf2 = pickle.loads(s)
```

```
>>> from joblib import dump, load
>>> dump(clf, 'filename.joblib')
>>> clf = load('filename.joblib')
```

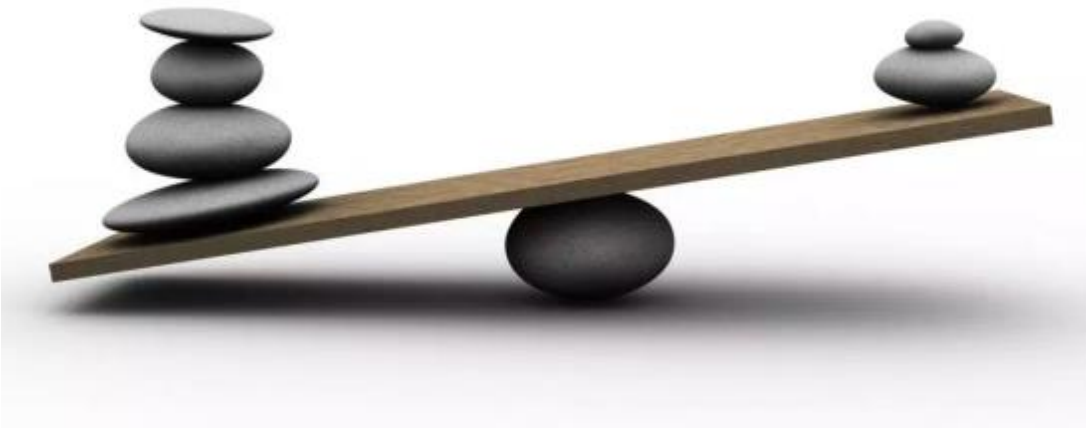
  - 参考： https://yq.aliyun.com/articles/228637

- 保持成通用格式，以便其它软件使用
  - 例子：ONNX（开放神经网络交换格式）
  - 例子：PMML（XML支持）

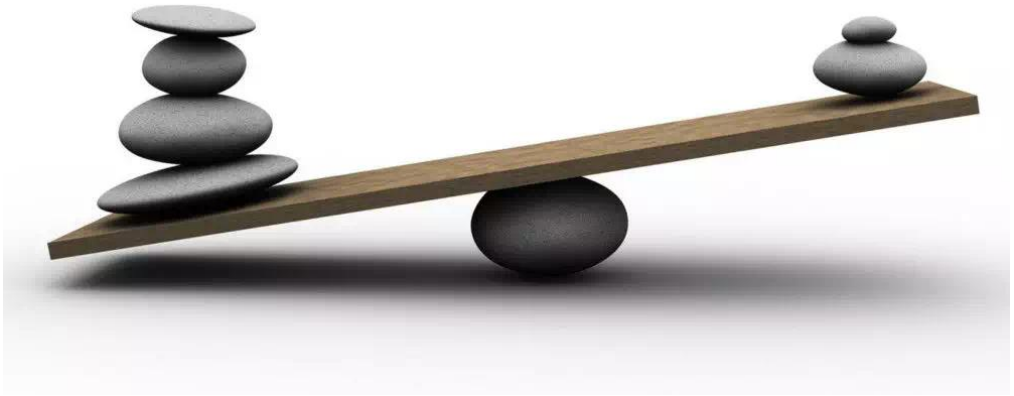
ONNX

3. 数据不平衡及分类模型的评价指标

- 不平衡数据
- 评价指标
- 训练方法



不平衡数据



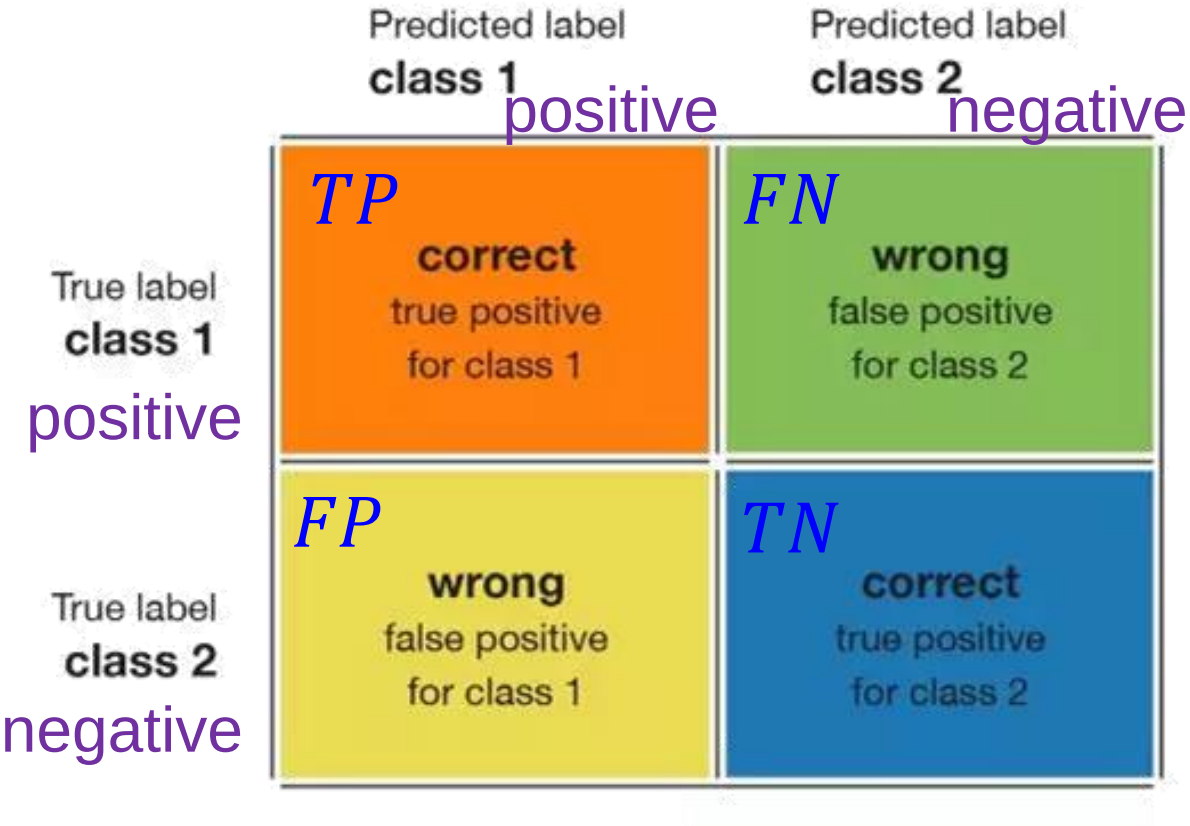
- 点击率预测；
- 信用卡欺诈检测；
- 网络攻击识别；
- 疾病监测；
- 恐怖分子识别…
- 假想例子：产品缺陷预测
  - 准确率高达 96.3%
  - 有用吗？

数据不平衡：

分类问题的正样本与负样本数量差别巨大，增加了模型训练与评价的难度。

评价指标：混淆矩阵

- 在分类问题中，混淆矩阵（confusion matrix）很好地概述了模型的运行情况，可作为任何分类模型评估的基础/起点。
  - 真实阳性/阴性而被预测为阳性/阴性的样本数目



TP: 真阳性，True Positive
FP: 假阳性，False Positive
TN: 真阴性，True Negative
FN: 假阴性，False Negative

T/F表示预测是否正确，
P/N表示预测的label

各种评价指标…

- 准确率（accuracy）：正确预测的比例。$\frac{TP+TN}{TP+FP+TN+FN}$
- （类别）精度（precision，又称查准率）：$\frac{TP}{TP+FP}$
  - 表示当模型判断一个点属于该类的情况下，判断结果的可信程度。
  - 常常被应用于推荐系统中。
- （类别）召回率（recall，又称查全率）：$\frac{TP}{TP+FN}$
  - 表示模型能够检测到该类的比率。
  - 在医学上常常被称作敏感度(Sensitivity)：不要漏诊！
- Specificity（特异性）：$\frac{TN}{FP+TN}$
  - 即负样本的召回率。医疗中的重要指标，与误诊率相关。

- 类别的 F_1 分数：精度和召回率的调和平均值，

$$F_1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- Matthews correlation coefficient (MCC)，

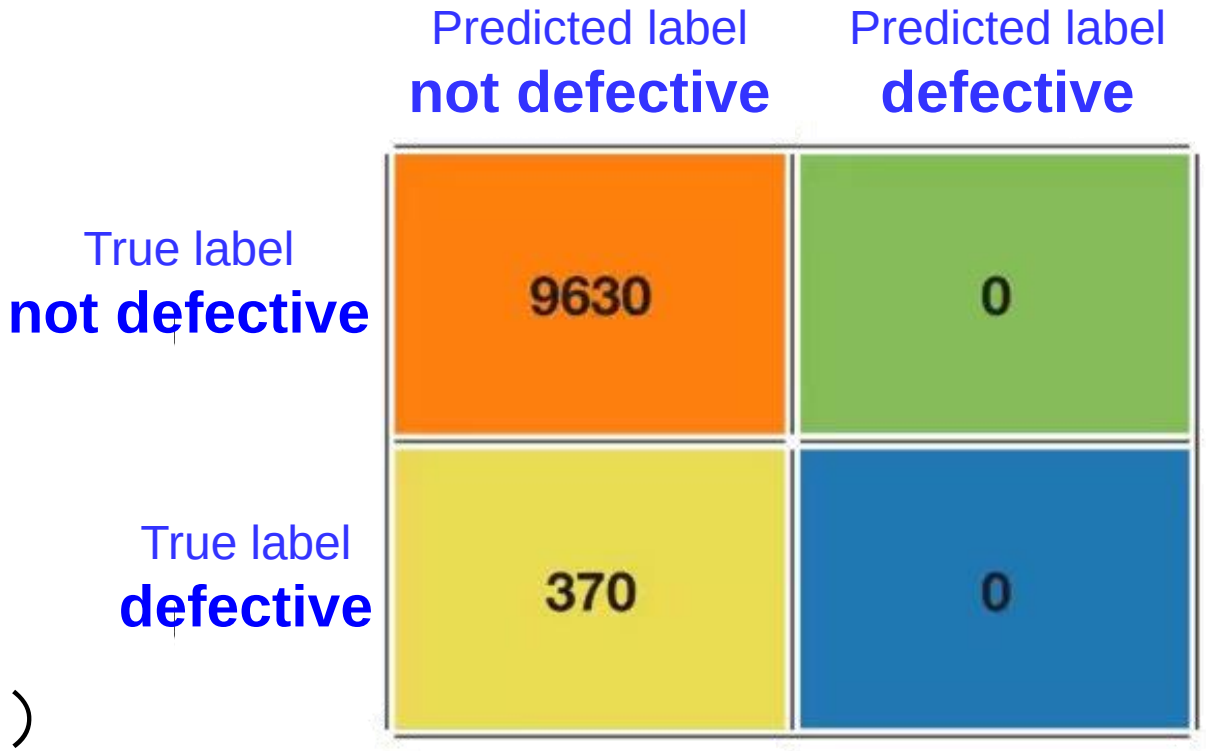
$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- 对于一个给定类，精度和召回率的不同组合：
  - 高精度+高召回率：模型能够很好地检测该类；
  - 高精度+低召回率：模型不能很好地检测该类，但是在它检测到这个类时，判断结果是高度可信的；
  - 低精度+高召回率：模型能够很好地检测该类，但检测结果中也包含其他类的点；
  - 低精度+低召回率：模型不能很好地检测该类。

假想例子…

- 准确率为 96.3%
- 无缺陷类的精度为 96.3%
- 有缺陷类的精度不可计算；
- 无缺陷类的召回率为 1.0（很好）
- 有缺陷类的召回率是 0（很糟糕！）

- 结论：这个模型对有缺陷类是无用（不友好）的。



$$\text{accuracy} = \frac{9630 + 0}{9630 + 370 + 0 + 0}$$

$$\text{defective precision} = \frac{9630}{9630 + 370}$$

$$\text{not defective recall} = \frac{0}{370 + 0}$$

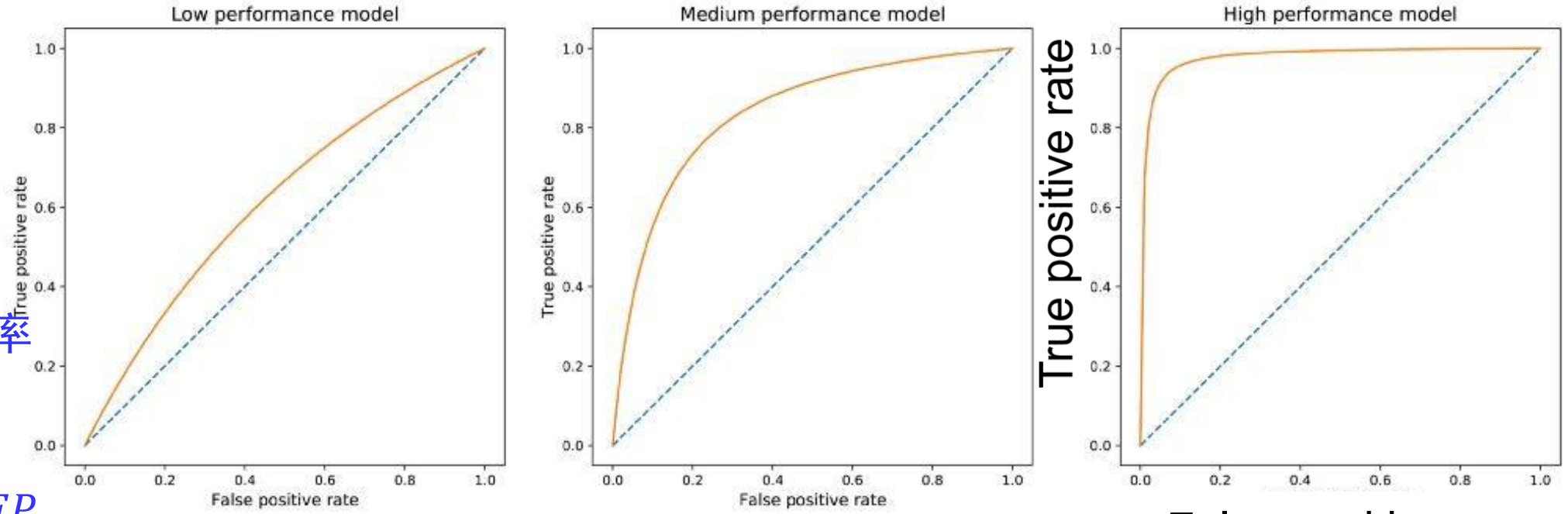
$$\text{defective recall} = \frac{9630}{9630 + 0}$$

指标：ROC曲线与AUC指标

虚线：随机猜测

$$\frac{TP}{TP+FN}$$

真阳性率
召回率
敏感度
$1-$漏诊率
击中概率



$$\frac{FP}{FP+TN}$$

假阳性率、$1-$特异性、误诊率、虚报概率

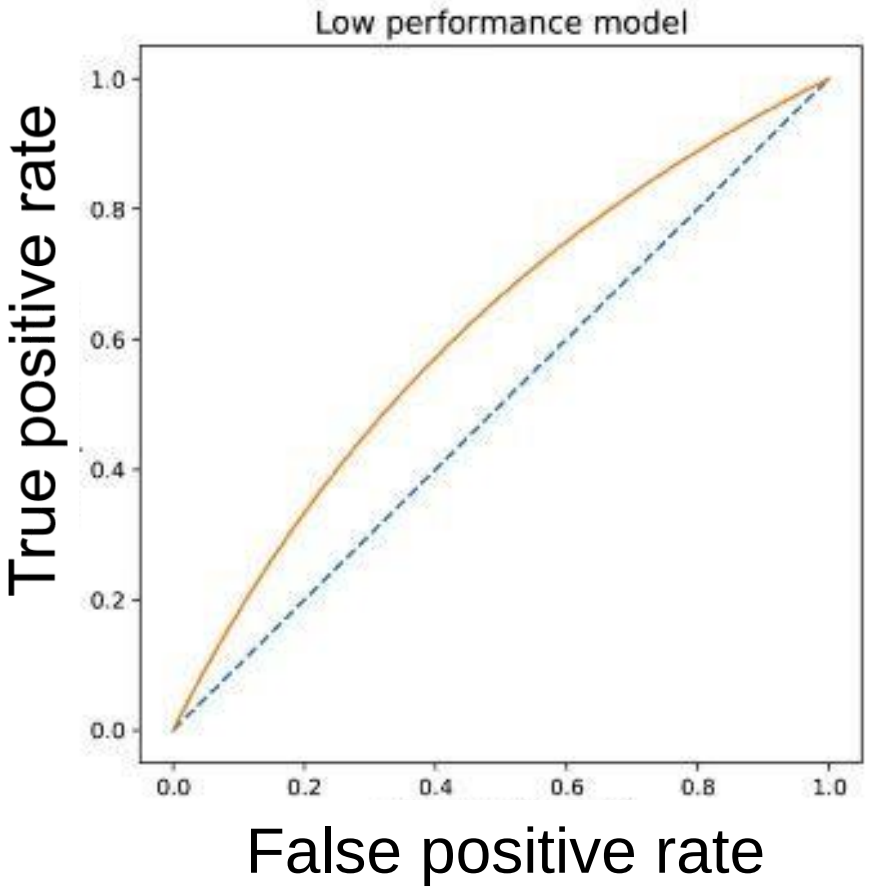
- 基于模型输出的类别概率$P(C|x)$，我们可以调整决策阈值P_{cri}（不一定为0.5）来获得不同的性能结果。ROC 曲线（Receiver Operating Characteristic curve受试者特征曲线）：正样本的真阳性率 ~ 负样本的假阳性率。

ROC曲线…

https://mp.weixin.qq.com/s/x48Ctb0_Eu1kcSGTYLt5BQ

- 假设对于给定点x，我们的模型输出该点属于类别C的概率为：$P(C|x)$。基于这个概率，我们定义一个决策规则，即当且仅当 $P(C|x) \geq P_{\text{cri}}$（其中$P_{\text{cri}}$是决策阈值）时，$x$属于类别 C。如果$P_{\text{cri}} = 1$，则仅当模型100%可信时，才将该点标注为类别C。如果 $P_{\text{cri}} = 0$，则每个点不管其x值如何都标注为类别C 。

- 阈值P_{cri}从 0 到 1 之间的每个值都会生成一个点 (false positive, true positive)，ROC 曲线就是当 P_{cri}从 1 变化到 0 所产生点的集合所描述的曲线。



- ROC曲线从点 (0,0) 开始，在点 (1,1) 处结束，且单调增加。
- 好模型的 ROC 曲线会快速从 0 增加到 1
  - 这意味着必须牺牲一点特异性就能获得高召回率（敏感性）。
- 基于 ROC 曲线，可以构建另一个指标：AUC（Area Under the Curve），即 ROC 曲线下的面积。
  - 与基尼系数相关。
- AUC 在最佳情况下将趋近于 1.0，而在最坏（随机猜测）情况下将趋向于 0.5。

- 注意：数据极端不平衡时，AUC可能不是好的评价指标。

题外…

- https://www.jiqizhixin.com/articles/amazon-aclu-facial-recognition-controversy

  美国参众两院议员中有28名罪犯？

  亚马逊AI人脸识别系统遭质疑

25,000 张罪犯照片
× 535名国会议员

- 亚马逊回应：

“Rekognition 系统中人脸识别 API 的默认置信阈值为 80%，这对很多通用案例来说很实用（如在社交媒体上识别名人或者在照片应用中识别长相相似的家庭成员），但这并不适用于公共安全领域。在需要高度精确的面部相似性匹配案例中，我们建议使用 99% 的置信阈值。”

题外…

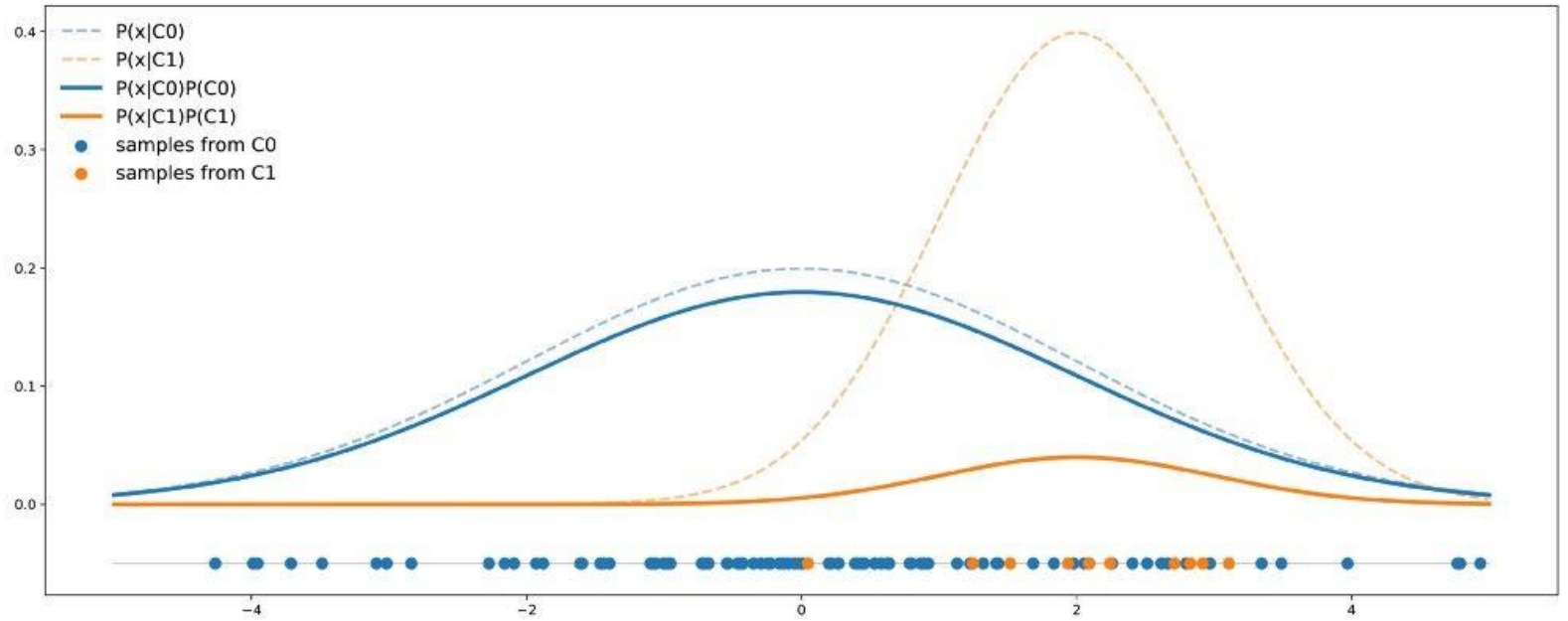
- https://news.sina.com.cn/c/2020-09-01/doc-iivhvpwy4339323.shtml

新冠病毒核酸检测：瑞典政府太阴了

- 多家西方媒体炒作说中国公司的新冠病毒核酸检测试剂“不准确”，导致瑞典9个地区的受检者中出现了3700个“假阳性”结果。
- **真相**：检测试剂非常灵敏，不会错过人群中的任何病例；在疫情中，需要抓住的是那些真正的“阳性”病例，所以就得调低阈值。

不平衡例子

$$P(x|C_0) = \mathcal{N}(x|0, 2^2)$$
$$P(x|C_1) = \mathcal{N}(x|2, 1^2)$$



- 数据集中 90% 的点来自 C_0，其余 10% 来自 C_1。
- 可以看到C_0的曲线总是在 C_1 曲线之上，因此对于任意给定点，它出自 C_0 类的概率总是大于出自 C_1 类的概率。用贝叶斯公式来表示，即：

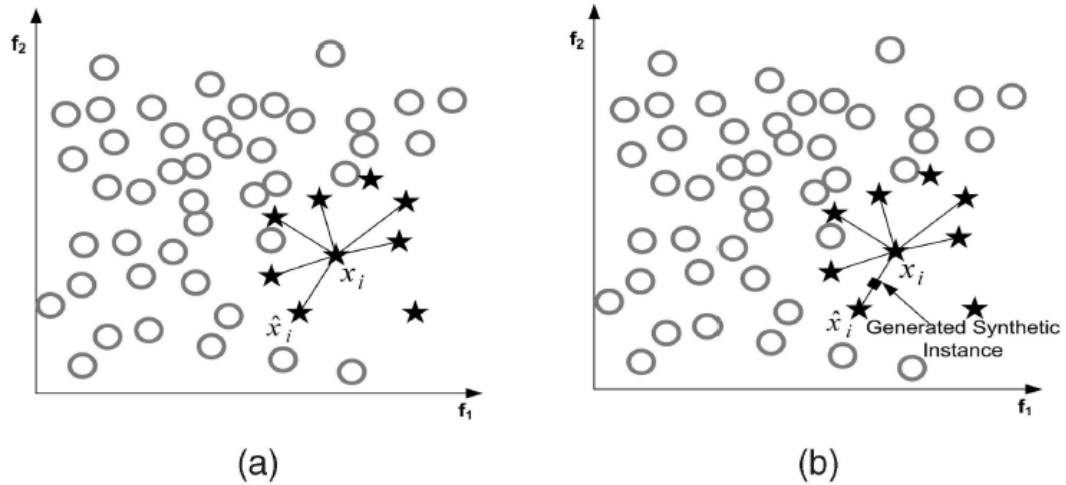
$$P(C_0|x) = \frac{P(x|C_0)P(C_0)}{P(x)} > \frac{P(x|C_1)P(C_1)}{P(x)} = P(C_1|x)$$

- 在通常的追求准确率的算法下，分类器每次都判断为 C_0。

解决方法1：重新处理数据集

- 目的：扩大少数类样本比例。
- 重新采样：
  - 过抽样（Oversampling，又称上采样）：通过复制少数类样本
    - 可以在每次生成新数据点时加入轻微的随机扰动，以减少过拟合。
  - 欠抽样（Undersampling，又称下采样）：删除部分多数类样本
    - 为减少信息的损失，可采用模型融合的方法（如EasyEnsemble），或利用组合模型中增量训练的思想（Boosting）。
- 生成合成数据（人工样本）
  - 例如：SMOTE

（Syntheic Minority Over-sampling）



解决方法2：重新考虑训练目标

- 不再以最佳准确率为目标，而是寻找较低的预测成本。
  - 例如，医疗诊断的问题。
- 权重设置（类重新加权）
  - 直接在训练时考虑误差成本

	cancer	normal
cancer	0	1000
normal	1	0

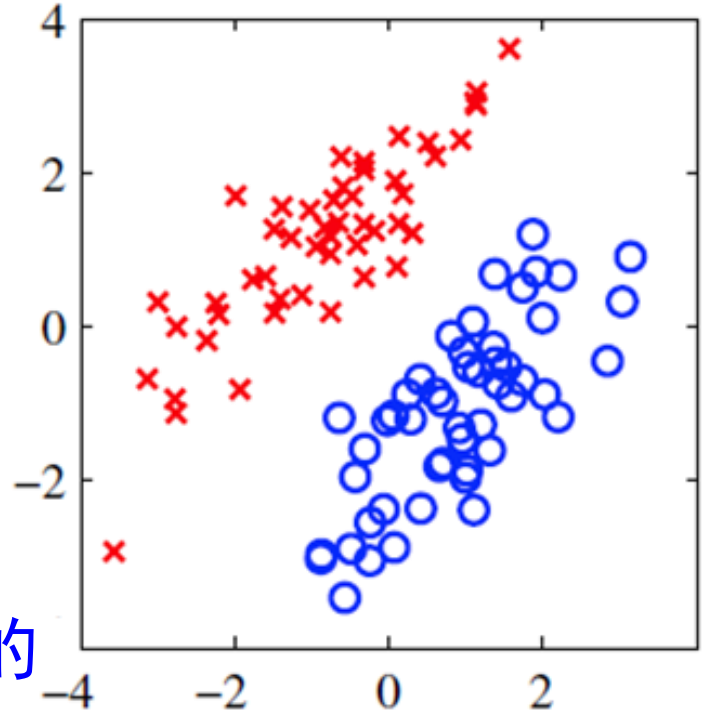
- 概率阈值的调整
  - 先按照基本的方法训练分类器，输出概率。然后，根据误差成本来调整分类预测的阈值。
  - 如果满足下述条件，则预测类为C_0：

$$\mathbb{P}(true\_C1|x) \times P01 < \mathbb{P}(true\_C0|x) \times P10$$

其它方法

- 改变思路。
- 例如，不再当做分类问题，而是一分类（One Class Learning）或异常检测（Novelty Detection）问题。
  - 例如：One-class SVM。
  - 例子：需要从高压线的航拍图片中，将松动的螺丝/零件判断为待检测站点，即负样本，其他作为正样本，这样来看，数据倾斜是非常严重的，而且在图像质量一般的情况下小物体检测的难度较大，所以不如将其转换为无监督的异常检测算法，不用过多的去考虑将数据转换为平衡问题来解决。

题外：天下没有免费午餐定理



- D. H. Wolpert and W. G. Macready.
  No free lunch theorems for optimization.
  IEEE Trans. Evol. Comput. **1**, 67 (1997).
- 对所有可能的目标函数求平均，得到的所有学习算法的“非训练集误差”的期望值都是相同的。
  - 例如，对于有N个数据点的分类问题，共有2^N 种可能的目标函数。对任一$\mathbf{x}_n$，总有一半函数认为其值为真，一半为假。不管你的算法预测这个变量为真或为假，平均总是50%正确。
- 因此，在脱离实际意义情况下，空泛地谈论哪种算法好是毫无意义的，要谈论算法优劣必须针对具体学习问题。

Scikit-Learn相关内容

https://scikit-learn.org/
https://sklearn.apachecn.org/

- **3. Model selection and evaluation**
- **3.1. Cross-validation: evaluating estimator performance**
  - **model_selection.train_test_split**
  - **model_selection.cross_val_score**
  - **model_selection.cross_validate**
- **3.2. Tuning the hyper-parameters of an estimator**
  - **model_selection.GridSearchCV**
  - **model_selection.RandomizedSearchCV**

- 3.3. Model evaluation: quantifying the quality of predictions
  - 3.3.2. Classification metrics
    - 3.3.2.5. Confusion matrix
    - 3.3.2.8. Precision, recall and F-measures
    - 3.3.2.13. Multi-label confusion matrix
- 3.4. Model persistence
  - joblib.dump，joblib.load
  - neural_network.MLPRegressor
- 2.7. Novelty and Outlier Detection
- 5. Dataset transformations
- 5.3. Preprocessing data

小结

- 把数据分成训练集、验证集与测试集，分别用于模型的训练（training，参数的确定）、选择（selection，超参数与模型的确定）与最优模型的最终评估（assessment）。
- 误差的偏差-方差分解。
- 学习曲线：精度/误差随训练集数据量的变化曲线，可用于判断欠拟合/过拟合。
- K折交叉验证：可减少数据的浪费。
- 数据预处理：
  - 特征归一化
  - 非数值型数据的处理：普通数值转换；独热编码；
  - 异常值检测：均方差；箱形图

- 训练模型的保存与载入
- 不平衡数据：正样本与负样本数量差别巨大，增加了模型训练与评价的难度
- 分类问题的评价指标：混淆矩阵
  - 准确率、精度、召回率、特异性、F_1分数
  - ROC曲线与AUC指标
- 不平衡数据的解决方法1：重新处理数据集
  - 重新采样：过抽样、欠抽样
  - 生成合成数据（人工样本）
- 不平衡数据的解决方法2：重新考虑训练目标
- 天下没有免费午餐定理

Reference:

- 刘志荣4
- Bishop 1.3, 1.5；
- Elements 7.1-7.3, 7.10；
- 实战 7.7
- 吴恩达10,11,15

扩展阅读：

- https://developer.aliyun.com/article/228637

*** 如何保存和恢复scikit-learn训练的模型.mht

- https://www.jiqizhixin.com/articles/2018-11-30-6

开源一年多的模型交换格式ONNX，已经一统框架江湖了？.mht

- https://www.jianshu.com/p/be343414dd24

*** 机器学习：如何解决机器学习中数据不平衡问题？.mht

- https://www.cnblogs.com/princecoding/p/6714216.html

机器学习常用性能指标总结.mht

- https://www.cnblogs.com/xuexuefirst/p/8858274.html

衡量机器学习模型的三大指标：准确率、精度和召回率.mht

- http://www.valleytalk.org/wp-content/uploads/2012/11/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0%E9%82%A3%E4%BA%9B%E4%BA%8B.pdf

机器学习那些事.pdf

- https://www.jiqizhixin.com/articles/amazon-aclu-facial-recognition-controversy

美国参众两院议员中有28名罪犯？亚马逊AI人脸识别系统遭质疑.mht

- https://news.sina.com.cn/c/2020-09-01/doc-iivhvpwy4339323.shtml

新冠病毒核酸检测：瑞典政府太阴了.mht

- https://xw.qq.com/cmsid/20210601A0E8P200

《噪声》，犯错背后的另一个原因.mht

有两个人打枪，一个人总是打歪，而且歪的角度，歪的程度都差不多。另一个人也打歪，但是歪上歪下都有，偶而不歪了还能命中一枪。

如果真要让你选择一个，你会选哪一个？

《思考快与慢》的作者丹尼尔卡尼曼，在今年以87岁的高龄又出了一本神书《噪声》，在人的决策和判断领域又填上了一块拼图。

如果说《思考快与慢》说的是人在做决策时容易陷入的各种认知偏误，那么《噪声》说的就是人在做判断时候的各种噪声。

错误=偏差+噪声。这是卡尼曼在《噪声》里补全的决策和判断领域地图后的公式。

- https://www.jiqizhixin.com/articles/2023-08-21-11

预测热门歌曲成功率 97%？这份清单前来「打假」.mhtml

本文的关键错误就在于，这种训练 - 测试分离是在数据已经过采样之后进行的。

模型存在机器学习中最常见的缺陷之一：数据泄漏。

谢谢大家！