模型评估与选择

泛化误差：在"未来"样本上的误差，이건 적으면 적을수록 좋겠지

经验误差：在训练集上的误差，亦称"训练误差" .하지만 经验误差는 적다고 좋은게 아니다

왜냐하면 过拟合（overfitting）이 발생할 수 있기 때문에!

过拟合가 무엇이냐: 经验误差小，泛化误差大。훈련데이터에서는 퍼포먼스가 좋은데, 새롭고 접해보지 못한 데이터에서는 퍼포먼스가 좋지 못한 현상, 보통 모델이 너무 복잡해서 훈련데이터의 디테일이나 噪声을 기억할 수 있어서, 无法泛化到新数据上

방지방법은 그럼 뭐냐: 交叉验证，增加数据集，数据多样性，正则化，集成学习，验证集에서 성능이 더 이상 좋아지지 않을 때 미리 중지 시키기

模型选择的三个关键问题

评估方法：评估模型的泛化性能的方法

关键：测试集应该与训练集"**互斥**"

常用方法：

1)留出法（데이터셋을 두개의 서로 배척하는 집합으로 나누는것, 그 중 하나를 훈련데이터셋, 하나를 테스트 데이터셋으로 하는것）：保持数据分布一致性원래 데이터에서 类别비율을 나눈 데이터 셋에서도 맞추는것（如：分层抽样），多次重复划分，测试集不能太大，不能太小（1/5~1/3）

→ 빠르지만 데이터를 충분히 이용하지 못하고, 결과도 안정적이지 못함

2)k-折交叉验证法(시험에 나왔음)：데이터셋을 k개의 互斥子集로 나눈다. 子集들은 分布一致性을 유지해야한다(**分层抽样得到**). 그리고 k-1개의 합을 훈련 데이터셋으로 만들고, 나머지 하나만 테스트 데이터 셋으로 사용한다. 이렇게 k번의 훈련과 테스트를 할 수 있고, k개의 테스트 결과가 나오는데, 이걸 평균값으로 낸다. 좋은점이 무엇이냐, 결과의 안정성은 k의 값에 따라 달라짐. 보통 k는 10을 취한다.

→ 评估稳定，数据利用率高，计算成本高，设用于数据量小或提高评估精度

3)自助法（bootstrapping）：**有放回采样** 원래 데이터셋에서 데이터 하나씩 새로운 데이터 셋으로 옮기는데, 이미 사용된 데이터가 다시 뽑혀 갈 수 있음. 훈련 데이터 셋과 원래 데이터셋의 크기는 같아지고, 데이터 분포는 변화가 생긴다. 데이터 양이 부족할 때 사용할 수 있는 방법.

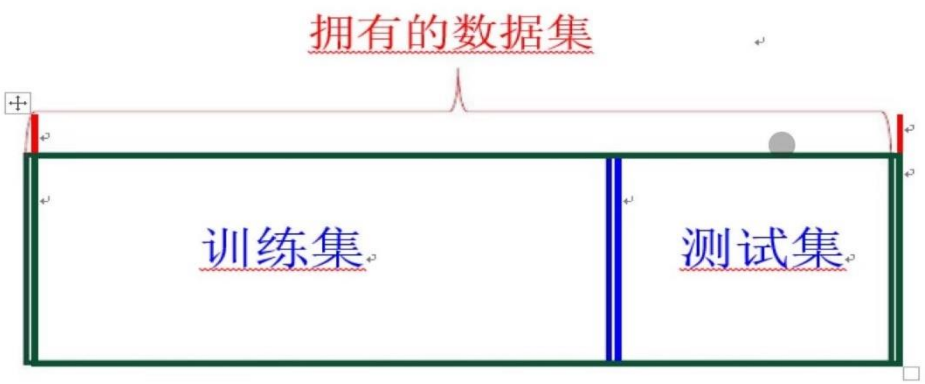
→ 작은 데이터에 대처할 수 있고, 泛化误差를 예측 할 수 있음, 근데 모델 偏差를 증가 시킬수있고 대표성이 부족하다.

데이터 시간 순서일때는?:

固定窗口划分法(앞에 부분을 훈련, 뒤에 부분을 테스트),

滚动窗口划分法(점차적으로 훈련데이터 셋을 크기를 키우고, 서로 다른 시간대에서의 퍼포먼스를 확인),

滑动窗口划分法(고정이랑 방법 비슷한데, 다른점은 훈련 데이터 크기를 고정시켜 놓고 새로 데이터가 들어 올때마다 가장 오래된 데이터 빼고 새로운 데이터 셋을 만듬)



调参은 모델 효율을 높이는데 필수적인 부분이다

超参数: 모델 훈련전에 설정해놓은 参数, 훈련을 통해 자동으로 업데이트 하는건 불가능 (如: 决策树中的最大深度, 学习率, 正则化系数)

模型参数: 모델 훈련중에 데이터에서 학습해 얻은 参数

调参(超参数调优): 모델을 조정하는 超参数을 사용해서 优化模型性能하는 과정

调参方法:

1) 网络搜索(Grid Search): 가능한 모든 参数组合를 遍历해서 가장 좋은 参数를 찾는다. 전면적이지만 비용 높음
2) 随机搜索(Random Search): 랜덤으로 参数선택해서 테스트해봄, 계산 효율이 더 높음

2번째 관건 요소인 性能度量은 뭘까: 衡量模型泛化能力的评价标准

여기서 그 유명한 错误率&准确率, 查准率&查全率가 나옴

错误率: 잘못 분류된 데이터의 비율, 准确率는 정확히 분류된 데이터 비율

查准率: 正类로 분류된 데이터 중에서 실제로 正类인 데이터 비율

查全率: 실제로 正类인 데이터 중에서 正类로 분류된 데이터 비율

表 2.1 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)

- 查准率: $P = \frac{TP}{TP + FP}$
- 查全率: $R = \frac{TP}{TP + FN}$

F1度量:

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

만약에 混合矩阵을 여러 개 얻을 수 있으면(여러 번 훈련/테스트 해서):

宏(macro-)查准率、查全率、F1

$$\text{macro-}P = \frac{1}{n}\sum_{i=1}^{n} P_i ,$$

$$\text{macro-}R = \frac{1}{n}\sum_{i=1}^{n} R_i ,$$

$$\text{macro-}F1 = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R} .$$

微(micro-)查准率、查全率、F1

$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} ,$$

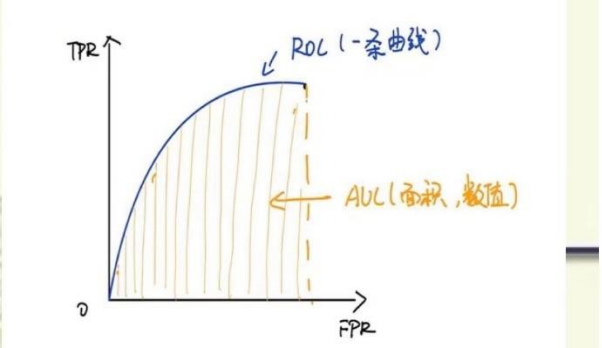
$$\text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} ,$$

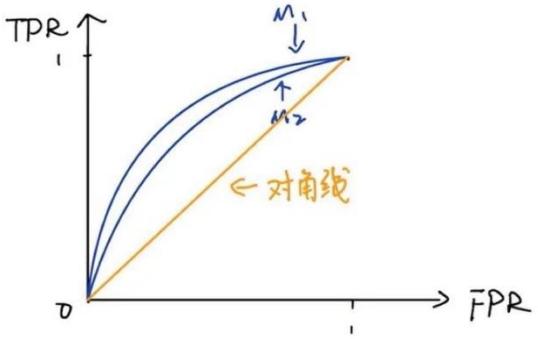
$$\text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R} .$$

ROC(受试者工作特征)와 AUC = 用来衡量二分类问题中的模型性能

TPR = 真正率（정확한 예측이 正인 수량 / 正 총 개수）

FPR = 假正率（잘못된 예측이 正인 수량 / 负 총 개수）





AUC = 1; 완벽한 分类器

AUC (0.5 ~1); 랜덤보다 우월한 ,비교적 우수한

AUC = 0.5 거의 랜덤 추측

AUC <0.5 ; 랜덤보다 못한