模型评估与选择

2.1&2.2

- 函数空间：$\{y1,y2,,,yn\}^{|x|}$
- 精度=1-错误率，训练集上的误差称为训练误差，新样本上的误差称为泛化误差
- 过拟合：训练集上分类精度太高，欠拟合：训练样本的一般性质尚未学好
- 分层采样：保留类别比例的采样
- 数据集划分方法:
  - 留出法
  - 交叉验证法：把数据集划分为K个子集，每次选一个子集用于测试
  - 留出法，交叉验证法的缺点：评估用D训练出的模型，保留一部分样本用于测试，评估效果受训练样本规模变化影响较大，而留一法不适合于超大规模样本
  - 自助法适用于数据集较小和数据集特别大的情况

    自助法的步骤:

    1. 设数据集包含m个样本，每次随机从D中挑选一个样本，一共进行m次，形成包含m个数据的训练集D'
    2. 令测试集为D-D',亦称包外估计
- 超参数与模型参数
  - 超参数：算法的参数，用户自己设置
  - 模型参数：学习得到的参数
- 训练集，验证集，测试集的区别：训练集用于训练出模型，验证集用于对模型的参数进行微调，测试集用于评估模型的泛化能力。训练集与验证集是同根同源的
- 回归问题又称聚类问题，性能度量常用均方误差
- 分类问题常用错误率与精度，查准率查全率和F1度量
- **ROC曲线和PR曲线的应用场景**
  ROC曲线的应用场景：
  ROC曲线主要应用于测试集中的样本分布的较为均匀的情况，且当测试集中的正负样本的分布变化的时候，ROC曲线能够保持不变。这也是ROC曲线一个很好的特性。
  但ROC曲线在出现类不平衡现象的数据集中时，即负样本比正样本多很多（或者相反），而且测试数据中的正负样本的分布也可能随着时间变化。ROC曲线是不敏感的，其曲线能够基本保持不变。**ROC的面对不平衡数据的一致性表明其能够衡量一个模型本身的预测能力**，而这个预测能力是与样本正负比例无关的。但是这个不敏感的特性使得其较难以看出一个模型在面临样本比例变化时模型的预测情况。此时ROC曲线最大的优点在面对不平衡数据集时便成为了它最大的一个缺点。

  PR曲线的应用场景：
  PRC因为对样本比例敏感，因此能够看出分类器随着样本比例变化的效果，而实际中的数据又是不平衡的，这样有助于了解分类器实际的效果和作用，也能够以此进行模型的改进。
  在面对出现类不平衡现象数据集时，可以根据PRC表现出来的结果衡量一个分类器面对不平衡数据进行分类时的能力，从而进行模型的改进和优化。

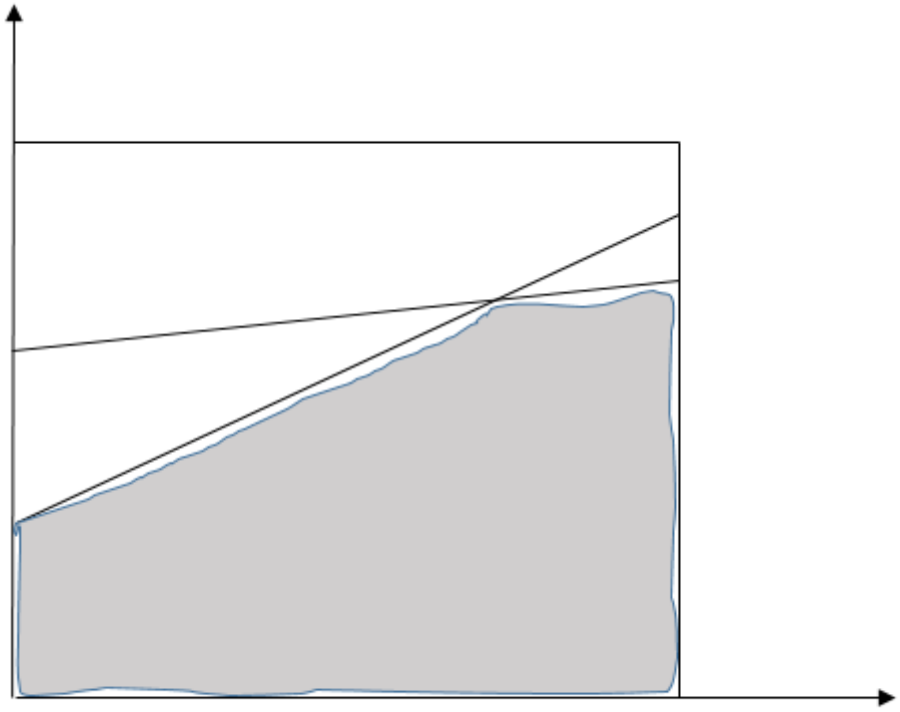
2.3

- 代价曲线:

期望损失代价$cost_{norm}$=FNR×P(+)+FPR×(1-P(+))，因此ROC曲线上一点便确定一对(FNR,FPR),从而确定一条代价直线的一部分。

绘制方法：

对于给定的正例概率p，取所有代价曲线在该频率下的最小值，然后求该最小值与正例概率所围成的面积，得到期望总体代价阴影曲线。



假设检验

这部分先放一放，暂时有需要在加进来，着重注意二项分布的假设检验