实验数据与结果分析

杨沐昀

教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

序

问题：分词实验中，只有1份数据，既要统计参数，又要检验模型性能？

- 从技术上看，NLP前期主要学习了统计建模；
  - 相对于经典的MLE，我们了解了无参估计、参数优化
- 不失一般地，我们将统计建模成为统计机器学习模型。
- 本节在机器学习这个大概念下，讨论数据集、评价和结果分析。

目录

- 机器学习基础和数据分配
- 性能评估
- 结果分析

机器学习一般流程

- **机器学习**：从大量实例中学习经验，然后使用经验去解决新问题
  - 例：买来10000根香蕉尝一尝，发现表面没有黑斑的香蕉大多比较好吃，对于一根没有品尝过的新香蕉，观察到它的表面没有黑斑，则认为它是好吃的
- **数据集**：大量实例（样本）组成的集合
- **训练（学习）**：从数据集中学习模型的过程。所使用的数据集称为**训练集**
- **测试（预测）**：使用训练好的模型对未知实例进行预测的过程。数据集称为**测试集**

机器学习中的样本

- **样本**：数据集中的一条记录。一般由**特征**和**标签**两部分组成
- **特征**：用于反映样本的某方面性质。数据集中的样本往往具有多个特征，从各方面对样本加以描述
- **标签**：用于表示样本在当前任务下的结果信息
- 比如在前面的例子中，一根香蕉就是一个**样本**，香蕉有没有黑斑是一个**特征**，而香蕉是否好吃是样本的**标记**

分类任务

- **分类任务**：输出离散值的机器学习任务，一般而言是样本所属的类别
  - **例**：确定香蕉是否好吃
- 只涉及两个类别的分类任务称为**二分类任务**，其中一个类称为“**正类**”，另一个类别称为“**反类**”
- 涉及多个类别的分类任务称为**多分类任务**

回归任务

- **回归任务**：输出连续值的机器学习任务
  - **例**：预测明天的气温
- 分类任务与回归任务可以在一定程度上互相转换
  - 判断香蕉是否好吃是个分类任务，但如果先计算香蕉好吃的概率，然后根据概率判断香蕉是否好吃，就可以使用回归模型处理分类任务

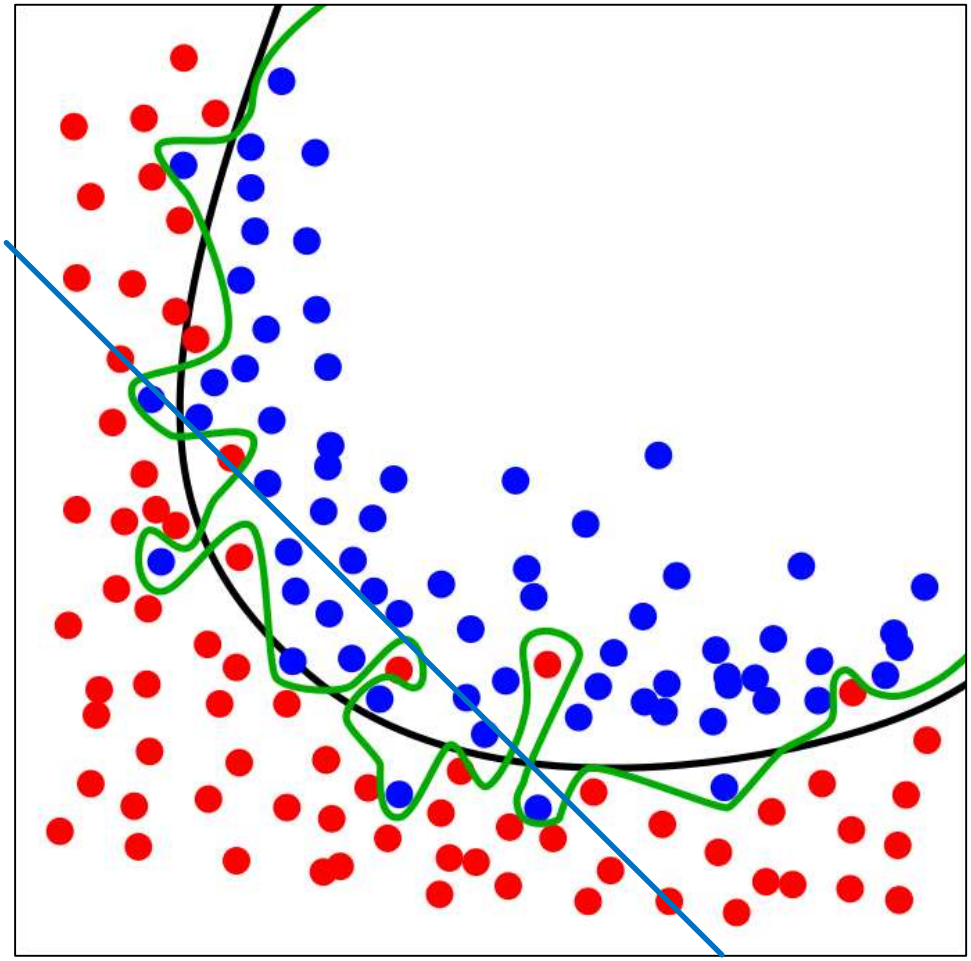
机器学习目标：泛化

- 学习器的**泛化**能力：除了能够在训练样本上工作的很好，在处理训练集中没有出现过的样本时也应该有不错的表现
- 如何保证泛化能力？
  - 训练样本数量足够多，更多地反映样本空间的分布~见多识广
  - 选用合适的模型和算法~游刃有余
  - 防止**过拟合**与**欠拟合**~恰如其分
- 泛化能力如何度量？
  - **泛化误差**：学习器的预测输出与新样本的真实情况之间的差异

过拟合和欠拟合

- **一种恰当的方案**：学习器从训练样本中尽可能学出适用于完整样本空间的普遍规律
  - 随着训练的进行，学习器的拟合能力不断增强，泛化误差逐渐减小
- **过拟合**：学习器把训练样本自身的一些特点当做了普遍规律，导致泛化性能不佳
  - 训练充分之后继续训练，学习器逐渐学到了只在训练数据中出现的噪声信息，泛化误差逐渐增大，表现为**过拟合**
- **欠拟合**：学习器没能找出足够深入的规律，同样导致泛化性能不佳
  - 训练不足时，学习器的拟合能力不够强，泛化误差较大，此时表现为**欠拟合**

过拟合和欠拟合-示例

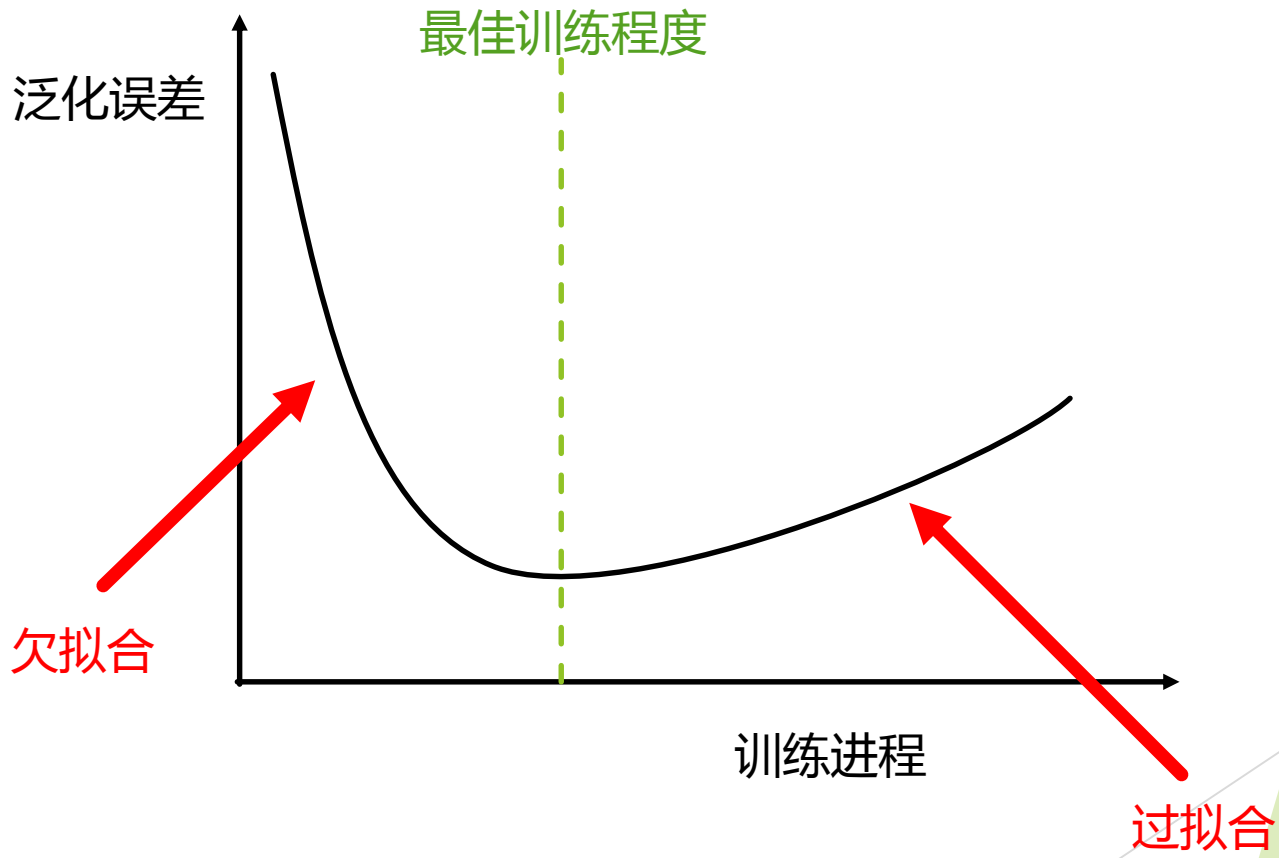


欠拟合

过拟合

适当拟合

训练过程示意图



验证集

- 在训练时，如何对泛化误差进行估计？
- 从可用数据中取样一部分样本，组成**验证集（开发集）**
- 验证集不参与训练，也不能与训练集有所重复
- 验证集的作用：用验证集上的误差近似泛化误差
  1. 对不同模型的泛化误差进行对比，可以帮助进行模型的选择
  2. 监控学习器的训练进程，在一定程度上防止欠拟合与过拟合

验证集的构造方法

- 留出法
- 交叉验证法
- 自助法
- 随机采样
  - 验证集从所有可用样本中采样得到，应使用合理的随机抽样方法进行抽样，保证数据分布的一致性

留出法

- 直接将数据集D 划分为两个互斥的集合，其中一个集合作为训练集S，另一个作为验证集T
- $S \cup T = D,\ S \cap T = \emptyset$
- 使用S训练模型，然后在T上评估误差，做为泛化误差的估计
- S和T的划分应当尽量保证数据分布的一致性，同时验证集中样本数量不应太少
- 只能利用数据集中的一部分进行训练，可能导致性能下降，因此训练集中样本数量也不应太少

交叉验证

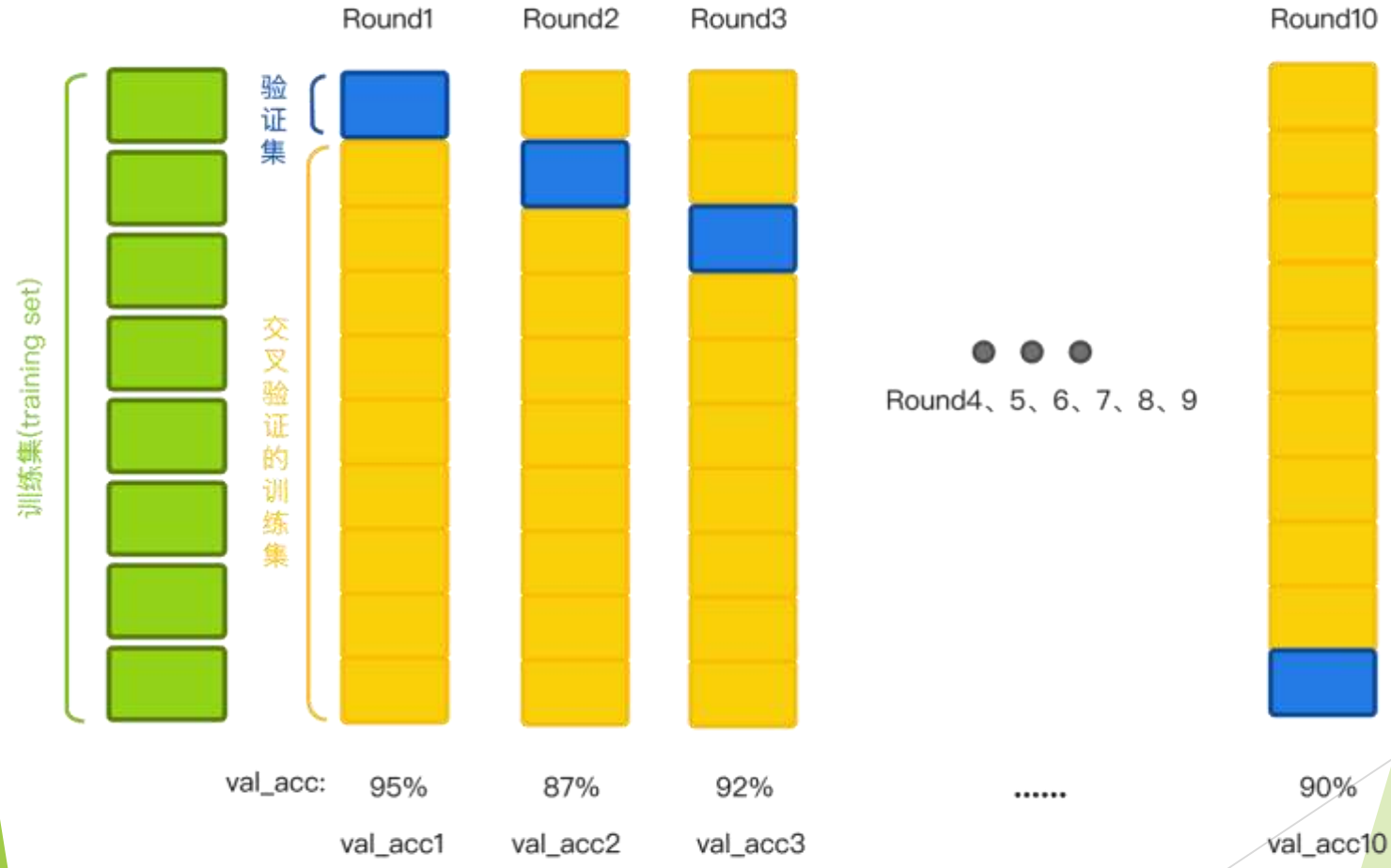
- **k折交叉验证**：先将D分为k个互斥的子集$D_1, D_2, \ldots, D_k$，每次取一个子集作为验证集，其他子集做训练集进行训练。所有子集都做过一次验证集之后，平均得到D上的测试误差
- 运行一次k折交叉验证，需要进行k次模型训练
- 常用的k值是10
- 使用了所有数据作为验证集，但每次训练时仍然只能使用一部分数据进行训练

交叉验证-留一法

- **留一法**：将n个样本组成的数据集D分为n组，每组只有一个样本，进行n折交叉验证

- **优点**：最大程度降低了对训练集数目的影响，同时不会受到随机抽取方式的制约
- **缺点**：当样本数量比较多时，留一法需要训练很多次模型，计算开销是难以接受的

交叉验证-示例


Round1
Round2
Round3
Round10
验证集
交叉验证的训练集
训练集(training set)
Round4、5、6、7、8、9
val_acc:
95%
87%
92%
.....
90%
val_acc1
val_acc2
val_acc3
val_acc10
Final Validation Accuracy = mean(val_acc1 + val_acc2 + ... + val_acc10)

自助法

- 自助法抽样步骤：
  1. 对n个样本组成的数据集D进行有放回抽样，抽出的样本放入D'，至D'中包括了n个样本为止
  2. 此时有些样本在D'中出现了多次，而有些样本在D'没有出现。样本在n次采样中始终不被取到的概率：

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n}=\frac{1}{e}\approx 0.368$$

  3. 则D中约有36.8%的数据没有被采样到D'中，使用D'做为训练集，D中没有被采样的样本做为验证集
- 能保证训练集不会因采样而变小，但是改变了初始数据集的分布

随机抽样方法（1）

- 如何保证抽样出的验证集与原始数据集同分布？
- **简单随机抽样**：对于数据集中的样本，按照一定概率进行随机抽样
- **分层抽样**：将样本按照几种显著特征分为多层，然后从每层中随机抽取
  - **例**：整个数据集由10000根香蕉组成，其中60%是好吃的香蕉（**正样本**），40%是不好吃的香蕉（**负样本**）。假设我们要抽取10%的样本作为验证集，则应当从正负样本中分别抽取10%，即600根好吃的香蕉和400根不好吃的香蕉

随机抽样方法-系统抽样

- **系统抽样**：将总体编号，根据待抽取的数目和样本总数计算抽样距离，然后按照抽样距离**等距**抽样
  - **例**：把10000根香蕉按1~10000编号，假设我们要抽取10%（1000根香蕉），则应当每隔10个样本抽取一次。从1~9十个数字中选取一个随机数，例如4，则抽取编号为4、14、24、……、9994的香蕉作为验证集
  - **注意**：系统抽样要求样本不存在周期性

随机抽样方法-整群抽样

- **整群抽样**：按照某种标准对总体分群，然后以群为单位进行抽样
- 与分层抽样层内差异尽量小的特点不同，整群抽样分群时应当保证群内差异尽量大，子群对整体数据分布有足够的代表性
- **例**：对于前面所说的香蕉数据集，将其随机分成100个群，每个群由60个正样本和40个负样本组成，然后抽取10个群作为验证集

目录

- 机器学习基础和数据分配
- **性能评估**
- 结果分析

为什么需要准确与召回

- 错误率与精度不能完全满足任务需求
  - **例**：某黑心商家卖的香蕉99%都是不好吃的，学习器只要简单地把所有的香蕉都分类为不好吃即可达到99%的精度，但这样的模型没有能力找出好吃的香蕉，不能说是一个好的学习器
- 准确率（precision）与召回率（recall）可以解决上面的问题

分类器的分类结果

- 对于二分类问题，根据真实情况与预测结果的不同组合，分类结果可以分为下面4类：
  - 真正例 (true positive)：真实情况为正，预测结果也为正
  - 假正例 (false positive)：真实情况为反，预测结果为正
  - 真反例 (true negative)：真实情况为反，预测结果也为反（前面的例子有大量的真反例）
  - 假反例 (false negative)：真实情况为正，预测结果为反
- $TP + FP + TN + FN = m$

准确率（precision）

	预测结果为正	预测结果为负
真实情况为正	TP	FN
真实情况为负	FP	TN

$$precision = \frac{TP}{TP + FP}$$

- 准确率表示分类器预测出的正例中准确预测的比例
  - 例：分类器觉得好吃的香蕉中真的好吃的有多少

召回率（recall）

	预测结果为正	预测结果为负
真实情况为正	TP	FN
真实情况为负	FP	TN

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

TN在什么地方用呢？

- 召回率表示分类器成功找出了所有正例中的多少
  - 例：分类器检索出的好香蕉占全部好香蕉的比例

准确与召回的矛盾

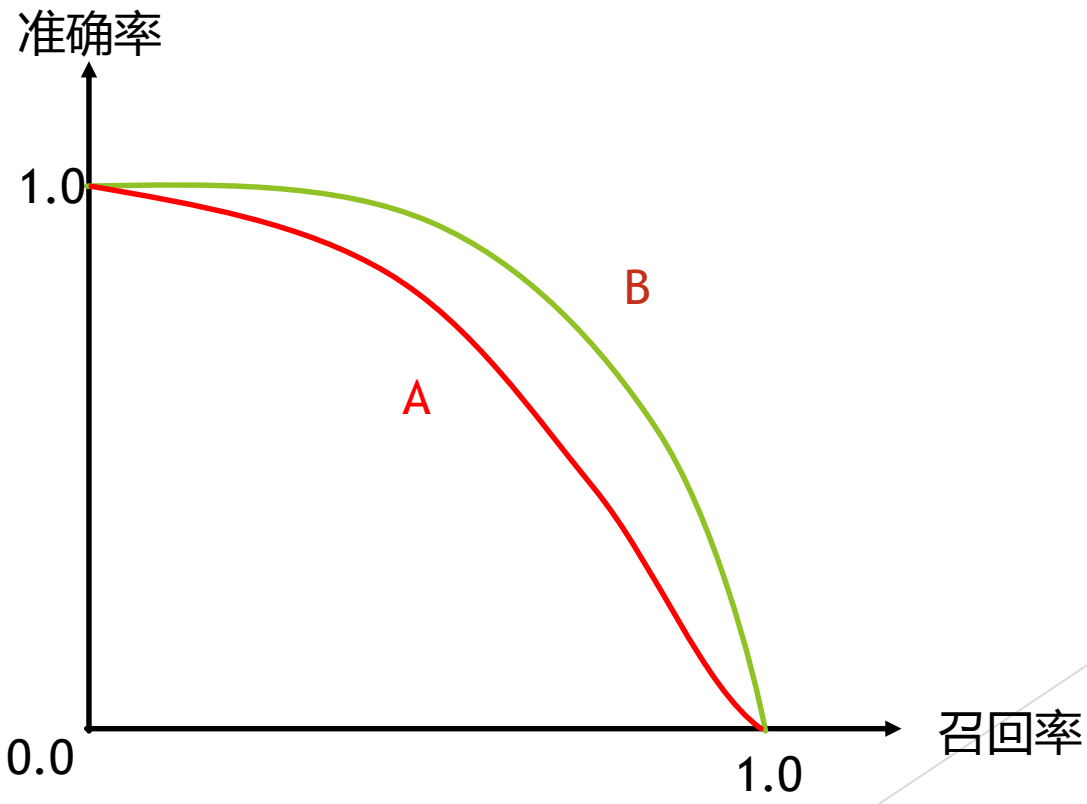
- 准确率与召回率存在矛盾：

  准确率偏高时，召回率往往偏低；召回率偏高时，准确率往往偏低

- 原因：
  - 为了提高召回率，可以选择宽松的筛选方案，多选择一些正例出来例子，如果把所有的香蕉都判断为好吃，则召回率可以达到100%，但准确率就会相应降低
  - 如果要提高准确率，可以选择严格的筛选方案，只把把握较大的香蕉判断为好吃。此时虽然可以提高准确率，但是会造成召回率的下降

P-R 曲线

- 准确率与召回率通常形成下面的曲线（示意图）



不同任务中的准确与召回

- 准确率与召回率从不同的方面反映模型性能，有的任务偏重准确，有的任务偏重召回
- 两个例子：
  - **偏重准确**：广告推荐系统向用户推送广告，为了防止推送量太多使用户反感，只选择用户最可能感兴趣的几条进行推送，也就是选择准确率较高的模型
  - **偏重召回**：对于地震预测任务，系统应该将所有可能的地震全部报出，宁可错报也不能漏报，此时应当选择召回率较高的模型

综合准确与召回

- 更多的时候任务对准确和召回没有明显的偏向，应该综合准确与召回进行考虑
- 一种方法：使用P-R曲线下的面积（**AUC**，area under curve）衡量模型性能，比如前面图中B模型的性能优于A
- 这个方法有没有问题？

  AUC计算非常困难！
- 另一种方法：平均准确率与召回率，得到**F值**

F值

- **$F1$值**是准确率与召回率的调和平均数：

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

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{m + TP - TN}$$

- **F_β值**：加权调和平均

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

- 与算术平均、几何平均不同，调和平均更注重较小值

还有一些任务...

- 之前提到的任务都是有“标准答案”的
  - 对香蕉进行分类时，香蕉要么好吃，要么不好吃
  - 预测气温时也有确定的数值
- 但是现实生活中也有很多任务没有“标准答案”
  - 机器翻译任务中一句英文句子可以有多个正确的中文翻译，只要语义相同，可以有多种不同的表述
  - 对话任务中一个句子可以从各种不同的角度加以回答，它们的语义都不一定相同
- 如何评价这些任务模型性能的好坏？

人工评测

- 对于人工智能领域的很多任务，**人工评测**都是最精确的评价标准

- 以BLEU为例的评价指标，其核心设计前提都是**使得评价指标的判断与人工判断尽量相似**
- 在尽量近似人工评价的前提下，设计这些评价指标，可以**大大减少模型评估的耗时，同时避免人类主观想法带来的标准不统一**

目录

- 机器学习基础和数据分配
- 性能评估
- **结果分析**

特征贡献度分析

- **特征**：反映事物某一方面的属性，机器学习系统根据样本的特征来做出预测
  - 例如：根据香蕉的表面有没有黑斑判断香蕉是否好吃
- 在进行预测时，不同的特征有不同的贡献度
  - 例如：在判断香蕉是否好吃时，香蕉表面的黑斑比香蕉的弯曲程度贡献更大
- 对于复杂的机器学习模型，不同的特征之间可能存在各种复杂的依赖、覆盖关系，一种简单的特征贡献度分析方法是**ablation study**

ablation study

- **主要思想**：在进行模型训练时，去除一个或一些特征，检查去除特征之后模型性能受到了多大的影响
- 去除某个特征之后，如果模型性能下降比较大，说明这个特征贡献度较高
- 不同的特征之间可能不是简单的叠加关系

统计显著性

- 当一个模型的评测结果超过了另一模型，我们能否说这个模型一定更优呢？
- 不能！
  - 例：在某种指标下，模型A的性能为99.0，模型B的性能为99.1，但是模型B只在0.01%的样本上远远超过模型A，在其它样本上的表现相比模型A并没有优势
  - 明显这时不能说模型B一定优于模型A
- 使用统计学方法进行测试，对不同模型的性能进行比较

回忆——统计假设检验

- **基本思想**：小概率事件在一次试验中基本不可能发生
- **检验方法**：如果原假设成立，其对应的检验统计量在某个区域内取值的概率α应该足够小。如果样本的观测数值落在这个小概率区间内，则原假设不正确，拒绝原假设；否则，接受原假设
- **核心技术**：人为构造一个小概率事件

符号检验

- 假设使用两个模型A和B分别对k个样本组成的测试集D进行预测，对D中的单个样本，它们的性能分别为$x_1, x_2, \dots, x_k$和$y_1, y_2, \dots, y_k$
- 基本思想：
  1. 对测试集中的每个样本i，求两个模型的性能之差$x_i - y_i$
  2. 统计$x_i - y_i > 0$的样本数n_+，$x_i - y_i < 0$的样本数n_-
  3. 如果两个模型性能相近，则n_+与n_-也应该相近（假设）；当n_+与n_-有明显区别时（二项检验），就认为两个模型有显著差异
- 符号检验只考虑数据变化的性质，即是变大了还是变小了，但没有考虑变化幅度，即大了多少，小了多少，因而对数据利用是不充分的

Wilcoxon符号秩检验

- 与符号检验类似，但是可以利用变化幅度的信息
- 基本流程：
  1. 同样求性能之差$x_i - y_i$和这个差值的符号$sgn(x_i - y_i)$
  2. 将n个不为0的差值$x_i - y_i$按绝对值由小到大排序，并找出它们的秩R_i
  3. 计算$W = \sum_{i=1}^{n}[sgn(x_i - y_i) \cdot R_i]$
  4. 如果两个模型性能相近，则W应该接近于0；否则，认为两个模型性能有较大区别

课下思考题

- NLP每人任务的显著性检验都有各自的惯例。请学习下列材料：
  - Statistical Significance Tests for Machine Translation Evaluation
  - https://aclanthology.org/W04-3250.pdf
  - 代码：www.cs.cmu.edu/~ark/MT/*paired_bootstrap_v13a*.tar.gz
- 尝试回答：机器翻译的统计显著性检验是怎么实现的（给出伪代码流程），尝试分析这么这么实现的原因。