机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

ML 机器学习

MACHINE LEARNING-MIN ZHANG

Unit.02

机器学习实验方法与原则（I）

*图片均来自网络或已发表刊物

机器学习实验方法与原则

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- K折交叉验证
- 统计有效性检验

机器学习实验方法与原则 1

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- K折交叉验证
- 统计有效性检验

评价指标

在不同任务下衡量模型的性能，有不同的评价指标，例如：

- 回归任务
  - 平均绝对误差（MAE）、均方误差（MSE）、均方根误差（RMSE）等
- 分类任务
  - 准确率（Accuracy）、精度（Precision）、召回率（Recall）等
- 特定任务
  - 个性化推荐：前K项精度（Precision@K）、前K项召回率（Recall@K）、前K项命中率（Hit@K）等
  - 对话系统：BLEU、ROUGE、METEOR等
  - ……

常用评价指标 – 1.回归任务(MAE, MSE,RMSE)

预测值p_i常为连续值，需要衡量与真实值y_i之间的误差

- 平均绝对误差（MAE）

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i|$$

- 均方误差（MSE）：预测误差较大的样本影响更大

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2$$

- 均方根误差（RMSE）：与预测值、标签单位相同

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}$$

常用评价指标 – 2.分类任务 (Accuracy, ER)

预测值一般为离散的类别，需要判断是否等于真实类别

- 准确率（Accuracy）

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i = p_i)$$

- 错误率（Error Rate）

$$Error\ Rate = 1 - Accuracy = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i = p_i)$$

常用评价指标 – 2.二分类任务 (P,R,F)

针对二分类任务的评价指标

- 精准率 / 精度（Precision）
  - 预测为正例的样本中有多少确为正例

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

- 召回率（Recall）
  - 找到的真实正例占所有正例中的比例

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

- F_β精准率和召回率的加权调和平均

真实标签	预测标签	
	$p = 1$	$p = 0$
$y = 1$	TP (True Positive，真正例，真阳性)	FN (False Negative，假反例，假阴性)
$y = 0$	FP (False Positive，假正例，假阳性)	TN (True Negative，真反例，真阴性)

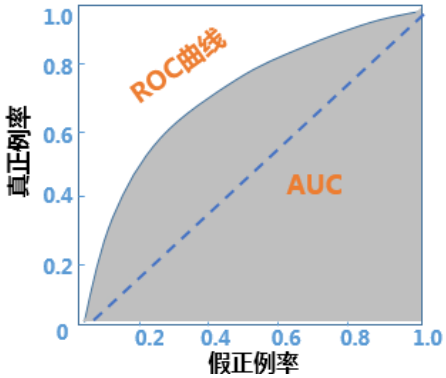
$$F_\beta = 1/[\frac{1}{1+\beta^2}\left(\frac{1}{P} + \frac{\beta^2}{R}\right)] = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2TP}{2TP + FP + FN}$$

常用评价指标 – 2. 二分类任务(AUC)

考虑二分类时划分正负的阈值

- ROC曲线
  - 根据预测值对样本排序
  - 以该样本的预测值为阈值
  - 大于或等于阈值记正例，否则记负例
  - 可得到一组结果及评价指标，共有样本数n组结果
  - 假正例率（False Positive Rate, FPR）为横轴
  - 真正例率（True Positive Rate, TPR, 也即召回率）为纵轴



真正例率

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

假正例率

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

- 随机猜测模型的ROC曲线为(0,0) 到 (1,1)的对角线
- 理想模型的ROC曲线为(0,0)-(0,1)-(1,1)，所有正例预测值大于所有负例预测值
- AUC：（Area Under ROC Curve）ROC曲线下的面积，越大越好

常用评价指标 – 2. 二分类任务（AUC)

- T：阈值

对于一个模型 f, 每个阈值的取值 T 都对应于ROC空间上的一个点

Actual class	score
1	0.98
0	0.80
1	0.67
1	0.65
0	0.54
0	0.32

T=0.7

Actual class	score	class
1	0.98	1
0	0.80	1
1	0.67	0
1	0.65	0
0	0.54	0
0	0.32	0

predicted \ actual	0	1
0	2	2
1	1	1

FPR = 1/(2+1)=0.33
TPR = 1/(2+1)=0.33

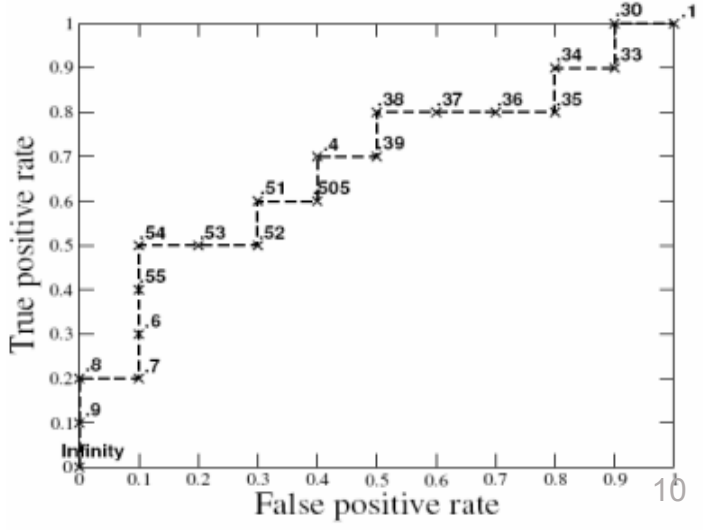
T=0.5

Actual class	score	class
1	0.98	1
0	0.80	1
1	0.67	1
1	0.65	1
0	0.54	1
0	0.32	0

predicted \ actual	0	1
0	1	0
1	2	3

FPR = 2/(1+2)=0.67
TPR = 3/(0+3) = 1

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



常用评价指标 – 2. 二分类任务（AUC)

AUC的简便计算方法：

- 把测试样例以预测值从大到小排序，其中有*n1*个真实正例，其中n0个真实负例
- 设 r_i 为第 i个真实负例的秩（排序位置），$S_0 = \sum r_i$
- AUC可以计算为:

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1}$$

(Hand & Till, 2001, MLJ)

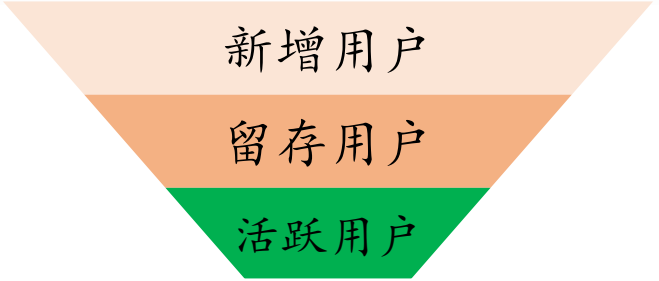
	+ + – + – + – +
i	1 2 3
r_i	3 5 7

Ranklist 1	+	+	+	+	–	+	–	–	–	–
Ranklist 2	–	+	+	+	+	–	–	–	–	+

AUC1: $\frac{(5+7+8+9+10)-5\times6/2}{5\times5} = \frac{24}{25}$

AUC2: $\frac{(1+6+7+8+9)-5\times6/2}{5\times5} = \frac{16}{25}$

常用评价指标 – 3.特定任务



一些特定任务有其特有评价指标

- 个性化推荐
  - 前K项精度（Precision@K）：模型排序给出的前K个推荐中，用户喜欢的项目（正例）的比例
  - 前K项召回率（Recall@K）：模型排序给出的前K个推荐中，正例数占候选集中所有正例的比例
  - 前K项命中率（Hit@K）：模型排序给出的前K个推荐中，是否有正例
  - nDCG@K、点击率、用户留存、利润转化等
- 对话系统
  - BLEU、ROUGE、 METEOR：基于词、n-gram匹配衡量预测句子与目标句子之间的相似度
  - 基于词向量计算预测句子与目标句子之间的相似度
  - 用户与系统对话的时长、次数
  - 人工评价

常用评价指标 – 3.特定任务（DCG）

- DCG: Discounted Cumulative Gain
- 检测一个文档，用*分级的相关性*来衡量有用性，或者增益(Gain)
  - $rel_1 + rel_2 + rel_3 + \ldots$
- 增益从排序列表的开头开始累积，随着位次增加，增益可能会减弱(Discounted)
  - $rel_1 + \text{discounted}(rel_2) + \text{discounted}(rel_3) + \ldots$
  - 典型的折损函数有$1/log\ (rank)$
  - 底数为2时，位次4的折损为1/2，位次8为1/3
  - $rel_1 + rel_2 / \log_2 2 + rel_3 / \log_2 3 + \ldots$

13

常用评价指标 – 3.特定任务（DCG）

- DCG: Discounted Cumulative Gain
- *DCG* 是对一个特定位次*p*的累积增益(Cumulative):

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- 或:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

14

常用评价指标 – 3.特定任务（DCG）举例

- 10 个文档的展示列表，相关性分级0-3:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0

- 折扣增益: $(1/\log_2 i)$
  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0

- 累积折扣增益 (DCG@n):
  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

15

常用评价指标 – 3.特定任务（NDCG）

- NDCG：归一化DCG （Normalized Discounted Cumulative Gain）
- 文档： 3, 2, 3, 0, 0, 1, 2, 2, 3, 0

DCG@n: 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

- 归一化，通过对比理想排序的DCG

理想排序: 3, 3, 3, 2, 2, 2, 1, 0, 0, 0

理想DCG@n: 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88

- NDCG@n (除以理想值):

$$\left(\frac{3}{3}, \frac{5}{6}, \frac{6.89}{7.89}, \frac{6.89}{8.89}, \frac{6.89}{9.75}, \frac{7.28}{10.52}, \frac{7.99}{10.88}, \frac{8.66}{10.88}, \frac{9.61}{10.88}, \frac{9.61}{10.88}\right)$$

1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

16

常用评价指标 – 3.特定任务（NDCG）

- 通过与理想排序的对应位置的DCG进行对比来归一化
- 对有不同数量相关文档的搜索结果求均值时更科学简洁
- 在任何位置都有NDCG ≤ 1
- 考虑了分级相关性和位置信息
- 搜索引擎等与排序相关的应用中相当常用的评价指标之一

17

常用评价指标 – 3.特定任务（BLEU）

- BLEU: bilingual evaluation understudy 双语替代评价
- 最早多用于机器翻译，后来也被其他任务借鉴（如对话生成等）
- 检测译文中的每个n-gram是否在参考译文中出现
- Precision没有考虑词出现的次数限制，结果偏高
- 某个词在译文中的有效频次不应超过参考译文中的频次

你好	1-gram	2-gram
参考译文：how are you	how, are, you	how are, are you
模型译文：you you	you, you	you you
Precision	(1+1)/2 = 1.0	0/1 = 0
Precision – 修正	(1+0)/2 = 0.5	0/1 = 0

18

BL常用评价指标 – 3.特定任务（BLEU）

- 译文太短时精度高但翻译不一定准确：译文较参考译文更长时，$BP = 1$
  - 译文较参考译文更短时，$BP = \exp\left(1 - \frac{\text{参考译文长度}}{\text{模型译文长度}}\right)$
- n-gram中n越大时的精度高表示句子越流畅，对n几何加权平均

你好	1-gram	2-gram	BLEU(n=2)
参考译文：how are you	how, are, you	how are, are you	$BLEU = BP * \exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln(P_i)\right)$ $= 0.61*\exp(0.5(ln0.5 + ln0)))$ $= 0$
模型译文：you you	you, you	you you	
Precision – 修正(P_i)	(1+0)/2 = 0.5	0/1 = 0	
BP	$e^{1-\frac{3}{2}} \approx 0.61$		

19

常用评价指标 – 3.特定任务（BLEU）

- 精度log可能出现为0的情况 → 置BLEU=0
- 也可对精度做平滑
- Google的参考实现（扩展：多个句子的翻译、多个参考译文）
  - https://github.com/tensorflow/nmt/blob/master/nmt/scripts/bleu.py

你好	1-gram	2-gram	BLEU(n=2)
参考译文：how are you	how, are, you	how are, are you	$BLEU = BP * \exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln(P_i)\right)$
模型译文：you you	you, you	you you	$= 0.61\exp(0.5\left(\ln\left(\frac{2}{3}\right)+\ln\left(\frac{1}{2}\right)\right))$
Precision – 修正平滑(P_i)	(1+0+1)/(2+1) = 2/3	(0+1)/(1+1) = 0.5	≈ 0.35
BP	$e^{1-\frac{3}{2}} \approx 0.61$		

机器学习实验方法与原则 2

- 评价指标
- 训练集、验证集与测试集
- 随机重复实验
- K折交叉验证
- 统计有效性检验