

PART1 总论

2023-06-05

1. 什么是人工智能？什么是机器学习？人工智能与机器学习的关系？

2. 机器学习中存在一些典型的任务，如：分类、聚类、回归等。

应在概念上区分：分类、聚类、回归。

3. 应能够结合一些典型的评价方式，对分类、回归模型的性能进行评价、或者选择模型。

4. 在样本的使用之前，往往要进行样本的规范化处理，为什么要进行处理？

通常有哪些方式？

注意：预处理不只是适用输入特征，有时还用于目标答案的预处理。

注意：虽然我们还没有学习神经网络

BP 神经网络既可用于分类，也可用于实值函数回归

BP 模型学习时，需要同时对训练集的输入与输出进行预处理

BP 模型使用之前，需要对样本的输入预测；模型产生的输出，还要进行反向后处理，才能作为最终的预测结果。

5. 应能区分两种典型的模型学习方式(监督式学习、非监督式学习)；

并能理解它们的适用场景。

6. 本学期，我们陆续接触了面向不同任务的典型模型，这个时候，你能把这些模型一一说出来吗？这些模型背后的原理？针对每一种模型，从原始数据集的获取开始，应掌握模型学习、模型评价、模型使用的流程。

7. 结合本学期学到的模型，如何结合 K-Fold Cross Validation 进行模型的超参数优选？

例如：K 近邻回归/分类模型的 K 值选择

决策树(分类树, 回归树)的深度

随机森林中的单棵树的深度

8. 如何基于 K-Fold Cross Validation, Leave-One-Out Cross Validation 进行分类模型的性能评价？总体错误率、正确率？

9. 如何基于 K-Fold Cross Validation, Leave-One-Out Cross Validation 进行回归模型的性能评价？ MAE, RMSE

10. 如何面向两类别(两种情况)、多类别分类问题，基于测试集得到的混淆矩阵进行模型有

关评价指标的估计？

11. 当采用训练集完成了基于 bagging 或 RF 的分类模型的学习之后，如何充分利用这个训练集，采用包外错误率评价该模型的性能？

练习

1. 结合课程学习，给出关于人工智能、机器学习的定义？

例：

- 中国《人工智能标准化白皮书(2018)》关于人工智能的定义：

人工智能是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

- 谭铁牛院士在2019《求是》定义“人工智能”：

人工智能是研究开发能够模拟、延伸和扩展人类智能的理论、方法、技术及应用系统的一门新的技术科学。研究目的是促使智能机器：会听(语音识别、机器翻译等)、会看(图像识别、文字识别等)、会说(语音合成、人机对话等)、会思考(人机对弈、定理证明等)、会学习(机器学习、知识表示等)、会行动(机器人、自动驾驶汽车等)。

机器学习

- 机器学习是人工智能的一个分支，是一门科学学科，涉及算法的开发与设计，该算法以经验数据为输入，并产生(被认为是生成数据的潜在机制特征的)模式或预测。
- 机器学习（Machine Learning）是一门涉及统计学、系统辨识、逼近理论、神经网络、优化理论、计算机科学、脑科学等诸多领域的交叉学科，研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，是人工智能技术的核心。

2. 什么是分类？什么是聚类？请给出二者的区别与联系。什么是回归？

解：

(1)分类

给定带有类别标记的训练样本集 $\{(x_i, y_i), i=1, \dots, N\}$ 。

其中： \mathbf{x}_i 为第*i*个观测样本的特征向量， $\mathbf{x}_i \in \mathbf{X} \subseteq \mathbb{R}^d$

y_i 为第*i*个观测样本的类别标号

基于上述样本集，**监督式学习**，设计分类模型；

对特征空间的任意观测 \mathbf{x} 进行类别决策。

(2) 聚类

给定样本集 $\{\mathbf{x}_i, i=1, \dots, N\}$. 寻找一种最优划分结果，以便对该数据集的内在结构进行合理描述，其直接结果是得到关于该数据集的划分。

二者区别：前者是基于已知答案的数据集,监督式学习一种划分模型，以便对特征空间的划分；后者是得到关于数据集的直接划分结果，是无监督式的学习的结果。

二者联系：可以借助聚类实现数据集的划分，以便实现自动式标注；进而，基于这种标注结果，学习分类模型，以实现关于整个特征空间划分。

搞清楚二者区别即可得全分。

3. 什么是监督式学习？什么是非监督式学习？请举例说明。

答案：

监督式学习

模型学习中，不仅用到训练样本的输入部分，还用到除了样本输入部分以外的、关于样本标签信息(即：模型学习的目标答案)。以目标答案为指导，进行模型学习，称之为监督式模型学习。例如：分类模型的学习，实值函数回归模型的学习，均为监督式学习。

非监督式学习

模型学习过程中，只用到了训练样本的输入部分，而没有使用除了样本输入部分以外的其它信息。这样的学习方式就是非监督学习。

例如：聚类模型的学习就是一种非监督式的学习。

4. 以分类或回归任务为例，结合模型的学习，能够区分：训练集、测试集、验证集、估计集的作用？

5. 给定**已知类别标记**的样本集 $D = \{(\mathbf{x}_i, y_i), i=1, \dots, N\}$ 。请分别基于如下两种交叉验证方式，估计某分类模型的总体预测错误率(或总体预测正确率)：

(1) 单轮K-倍交叉验证(K-Fold Cross Validation，也称K-折交叉验证)；

(2) 留一法交叉验证.

若是回归模型呢？参数优选？

解：

(1)当样本数目 N 不够多时，为确保模型性能预测更为客观，采用交叉验证方式评价。

STEP1.将样本集 D 分层随机打乱，均分成 K 个子集： $D = D_1 \cup D_2 \cup \dots \cup D_K$ ；

STEP2.对于 $i = 1, 2, \dots, K$ ，完成如下工作：

从给定样本集 D 内留出 D_i 作为测试集(验证集)，其余 $K - 1$ 个子集构成训练集(估计集) $D \setminus D_i$ ，以样本集 $D \setminus D_i$ 学习一个分类模型，利用该分类模型对测试集 D_i 进行预测，得错误率 Err_i 。

STEP3.输出 K -倍交叉验证的评价结果：

总体预测错误率均值： $\mu_{Err} = \frac{1}{K} \sum_{i=1}^K Err_i$

标准差： $\sigma_{Err} = \left[\frac{1}{K} \sum_{i=1}^K (Err_i - \mu_{Err})^2 \right]^{1/2}$

最终， K -倍交叉验证的总体预测错误率的均值 \pm 标准差：

$$\mu_{Err} \pm \sigma_{Err}$$

(2)当样本数 N 过小时，采用留一法交叉验证方式评价。

STEP1. 将错误分类的样本数目 num 初始化为0；

STEP2. 对于 $i = 1, 2, \dots, N$ ，重复完成如下工作：

(1)从给定样本集 D 内留出第 i 个样本作为测试样本，其余 $N - 1$ 个样本构成训练样本集 D_i ；

(2)以样本集 D_i 学习一个分类模型；

(3)利用该分类模型对留出的测试样本进行预测；

(4)若错误预测，则 $num = num + 1$ 。

STEP3. 输出留一法交叉验证的总体预测错误率： $Err = \frac{num}{N} \times 100\%$

6.给定已知类别标记的样本集 $D = \{(x_i, y_i), i = 1, \dots, N\}$ ，其中： $x_i \in R^d$ ， $y_i \in \{1, 2, \dots, C\}$ 。请结

合留一法交叉验证，以预测正确率为评价指标，对分类模型的性能进行评价。

参考答案：当样本数 N 过小时，采用留一法交叉验证方式评价。

STEP1. 将正确分类的样本数目 num 初始化为0；

STEP2. 对于 $i = 1, 2, \dots, N$ ，重复完成如下工作：

- (1) 从给定样本集 D 内留出第 i 个样本作为测试样本，其余 $N - 1$ 个样本构成训练样本集 D_i ;
- (2) 以样本集 D_i 学习一个分类模型;
- (3) 利用该分类模型对留出的测试样本进行预测;
- (4) 若正确预测，则 $num = num + 1$.

STEP3. 输出留一法交叉验证的总体预测正确率: $Acc = \frac{num}{N} \times 100\%$

6. 对于实值函数 $y = f(x)$ 的回归问题，设已知正确答案的样本集为

$$D = \{(x_i, y_i), i=1, \dots, N\}, \text{ 其中 } x_i \in R^d, y_i \in R$$

请分别基于如下两种交叉验证方式，以平均绝对误差评价回归函数 $y = f(x)$ 的预测性能，并明确两种交叉验证方式的适用场合。

(1) 单轮K-倍交叉验证(K-Fold Cross Validation，也称K-折交叉验证)方式;

当样本数目 N 不够多时，为确保模型性能预测更为客观，采用交叉验证方式评价。

STEP1. 将样本集 D 随机打乱，均分成 K 个子集: $D = D_1 \cup D_2 \cup \dots \cup D_K$;

STEP2. 对于 $i = 1, 2, \dots, K$ ，完成如下工作:

- 从给定样本集 D 内留出 D_i 作为测试集，其余 $K - 1$ 个子集构成训练集 $D \setminus D_i$
- 以样本集 $D \setminus D_i$ 学习一个预测模型
- 利用该模型对测试集 D_i 进行预测，得平均绝对误差 $MAE_i = \frac{1}{|D_i|} \sum_{(x,y) \in D_i} |y - \hat{y}|$

STEP3. 计算K-倍交叉验证的总体预测MAE均值 μ_{MAE} 及标准差 σ_{MAE} ，得评价结果:

$$\mu_{MAE} \pm \sigma_{MAE}$$

其中:

$$MAE_i = \frac{1}{|D_i|} \sum_{(x,y) \in D_i} |y - \hat{y}| \quad | \quad i = 1, \dots, K$$

$$\mu_{MAE} = \frac{1}{K} \sum_{i=1}^K MAE_i$$

$$\sigma_{MAE} = \left[\frac{1}{K} \sum_{i=1}^K (MAE_i - \mu_{MAE})^2 \right]^{1/2}$$

$$\mu_{MAE} \pm \sigma_{MAE}$$

(2) 留一法交叉验证.

当样本数 N 过小时。

STEP1. 对于 $i = 1, 2, \dots, N$, 重复完成如下工作:

- (1) 从给定样本集 D 内留出样本 x_i 作为测试样本, 其余 $N - 1$ 个样本构成训练样本集 D_i ;
- (2) 以样本集 D_i 学习一个实值函数预测模型;
- (3) 利用该模型对留出的测试样本 x_i 的输出进行预测, 得 \hat{y}_i .

STEP2. 输出留一法交叉验证的平均绝对误差: $MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$

7. 设某分类模型对已知类别标记的测试样本集进行分类(其中: 总类别数=3), 得到如下表所示的混淆矩阵:

		预测类别		
		1	2	3
真实类别	1	n_{11}	n_{12}	n_{13}
	2	n_{21}	n_{22}	n_{23}
	3	n_{31}	n_{32}	n_{33}

上述混淆矩阵的元素值为样本数。请给出如下指标的计算结果:

- (1) 测试样本集的总体预测错误率、正确率;
- (2) 第 i 类的查准率;
- (3) 第 i 类的查全率;
- (4) 第 i 类的 F1 值。
- (5) 各类预测正确率的算术均值;
- (6) 宏召回率
- (7) 宏查准率
- (8) 宏 F1 值
- (9) 第 i 类的样本数目? 其中被错分的样本数?
- (10) 总样本数、正确预测的样本总数、错误预测的样本总数

8. 给定用于 C 类别划分的分类模型, 利用测试集对该模型进行性能评价, 得到如下表所示的混淆矩阵:

		预测类别			
		1	2	...	C
真	1	n_{11}	n_{12}	...	n_{1c}

实 类 别	2	n_{21}	n_{22}	...	n_{2c}
	\vdots	\vdots	\vdots	\ddots	\vdots
	C	n_{C1}	n_{C2}	...	n_{CC}

请基于该混淆矩阵，进行如下内容估计：

- (1) 真实类别为第 i 类的测试样本数；
- (2) 预测类别为第 i 类的测试样本数；
- (3) 第 i 类的查准率？精度？
- (4) 第 i 类的查全率？召回率？预测正确率
- (5) 第 i 类的 F_β 值；
- (6) 各类平均预测正确率？总体预测正确率？

解：

- (1) 真实类别为第 i 类的测试样本数 $\sum_{j=1}^C n_{ij}$

- (2) 预测类别为第 i 类的测试样本数 $\sum_{j=1}^C n_{ji}$ ；

- (3) 第 i 类查准率：

$$Precision_i = \frac{n_{ii}}{\sum_{j=1}^C n_{ji}} \times 100\%$$

- (4) 第 i 类查全率： $Recall_i = \frac{n_{ii}}{\sum_{j=1}^C n_{ij}} \times 100\%$

- (5) 第 i 类的 F_β 值： $F_\beta = \frac{(\beta^2 + 1) Precision_i \cdot Recall_i}{\beta^2 Precision_i + Recall_i}$

- (6) $\overline{Acc} = \frac{1}{C} \sum_{i=1}^C Recall_i$

9. 给定用于两个类别划分的分类模型，其中阳性类为感兴趣的类别。利用测试集对该模型进

行性能评价，得到如下表所示的混淆矩阵：

		预测类别	
		Positive(阳性)	Negative(阴性)
真实类别	Positive(阳性)	n_{11}	n_{12}
	Negative(阴性)	n_{21}	n_{22}

若该混淆矩阵的元素值为样本数，请基于该混淆矩阵，进行如下指标的估计：

- (1)查准率；
- (2)查全率；
- (3)真阳性率；
- (4)假阳性率；
- (5) F1 值.
- (6) F_β 值.
- (7) 真阴性率 灵敏度 特异度 马修相关系数
- (8) 精度；
- (9) 召回率；
- (10)总体预测正确率；
- (11)平均正确率

解：

$$(1) \text{查准率: } Precision = \frac{\text{正确决策为阳性类的样本总数}}{\text{决策为阳性类的样本总数}} = \frac{n_{11}}{n_{11} + n_{21}}$$

$$(2) \text{查全率: } Recall = \frac{\text{正确决策为阳性类的样本总数}}{\text{参与决策的阳性类样本总数}} = \frac{n_{11}}{n_{11} + n_{12}}$$

$$(3) \text{真阳性率: } Recall = \frac{\text{正确决策为阳性类的样本总数}}{\text{参与决策的阳性类样本总数}} = \frac{n_{11}}{n_{11} + n_{12}}$$

$$(4) \text{假阳性率: } \frac{\text{错误决策为阳性类的样本总数}}{\text{参与决策的阴性类样本总数}} = \frac{n_{21}}{n_{21} + n_{22}}$$

$$(5) \text{F1 值: } F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

10. 给定用于两个类别划分的分类模型，利用测试集对该模型进行性能评价，得到如下表所示的混淆矩阵：

		预测类别	
		第1类	第2类
真实类别	第1类	n_{11}	n_{12}
	第2类	n_{21}	n_{22}

若该混淆矩阵的元素值为样本数，请基于该混淆矩阵，进行如下指标的估计：

- (1) 各类查准率；
- (2) 各类查全率；
- (3) 各类 F1 值.
- (4) 各类 F_{β} 值.

11. 给定用于多类别划分的分类模型，利用测试集对该模型进行性能评价，得到如下表所示的混淆矩阵：

		预测类别			
		1	2	...	C
真实类别	1	n_{11}	n_{12}	...	n_{1c}
	2	n_{21}	n_{22}	...	n_{2c}
	\vdots	\vdots	\vdots	\ddots	\vdots
	C	n_{c1}	n_{c2}	...	n_{cc}

该混淆矩阵的元素值为样本数。请基于该混淆矩阵，进行如下指标的估计：

- (1) 第 i 类的查准率(精度, precision);
- (2) 第 i 类的查全率(召回率, 预测正确率, Recall);

- (3)第 i 类的 F1 值、 F_{β} 值;
- (4)分类模型的总体预测正确率、错误率;
- (5)分类模型关于各类别预测错误率的算数平均值;
- (6)由该混淆矩阵提取“第 2 类 VS.非第 2 类”的混淆矩阵。