PART1 总论

软件学院

2020-6-1

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？
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，我们学了哪些模型，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？你会使用这些模型了吗？

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

参考：

人工智能

答案1：

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。

答案2：

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

参考答案1：

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。

参考答案2：

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$\{(\boldsymbol{x}_i, y_i), \mathrm{i}=1,\cdots,N\}$.

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

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$\boldsymbol{x}$进行类别决策 。

(2)聚类

给定样本集$\{\boldsymbol{x}_i, \mathrm{i}=1,\cdots,N\}$.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.

以上部分分别 3 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、非监督式学习？举例说明。

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？

3. 给定已知类别标记的样本集$\boldsymbol{D}=\{(\boldsymbol{x}_i, y_i), \mathrm{i}=1,\cdots,N\}$. 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) 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$\boldsymbol{D}$随机打乱，均分成K个子集：$\boldsymbol{D}=\boldsymbol{D_1}\cup\boldsymbol{D_2}\cup\cdots\cup\boldsymbol{D_K}$；

STEP2.对于$i=1,2,\ldots,\boldsymbol{K}$，完成如下工作：

从给定样本集$\boldsymbol{D}$内留出$\boldsymbol{D_i}$作为测试集，其余$\mathrm{K}-1$个子集构成训练集$\boldsymbol{D}\backslash\boldsymbol{D_i}$，以样本集$\boldsymbol{D}\backslash\boldsymbol{D_i}$学习一个分类模型，利用该分类模型对测试集$\boldsymbol{D_i}$进行预测，得错误率$\mathrm{Err_i}$.

STEP3.输出K-倍交叉验证的总体预测错误率：$\mathrm{Err}=\frac{1}{\mathrm{K}}\sum_{i=1}^{K}\mathrm{Err_i}$

(2)当样本数N过小时。

STEP1. 将正确分类的样本数目num初始化为0；

STEP2. 对于$i = 1,2,\ldots,N$，重复完成如下工作：

从给定样本集$\boldsymbol{D}$内留出第i个样本作为测试样本，其余$\mathrm{N}-1$个样本构成训练样本集$\boldsymbol{D_i}$，以样本集$\boldsymbol{D_i}$学习一个分类模型，利用该分类模型对留出的测试样本进行预测，若错误预测，则$\mathrm{num} = num + 1$.

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

5. 对于实值函数$\mathbf{y} = \mathbf{f}(\mathbf{x})$的回归问题，设已知正确答案的样本集为$\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i), \mathrm{i} = 1,\cdots,N\}$，其中$\boldsymbol{x}_i \in R^d$，$y_i \in R$请分别基于如下两种交叉验证方式，以平均绝对预测误差评价回归函数$\mathbf{y} = \mathbf{f}(\mathbf{x})$的预测性能，并明确两种交叉验证方式的适用场合。

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

(2)留一法交叉验证.

6. 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	$\mathbf{n_{11}}$	$\mathbf{n_{12}}$	$\mathbf{n_{13}}$
	2	$\mathbf{n_{21}}$	$\mathbf{n_{22}}$	$\mathbf{n_{23}}$
	3	$\mathbf{n_{31}}$	$\mathbf{n_{32}}$	$\mathbf{n_{33}}$

该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 3 类的查准率；
(3)第 3 类的查全率；
(4)第 3 类的 F1 值。
(5)各类预测正确率的算术均值；
(6)召回率的宏平均值。

7. 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)**	$\mathbf{n_{11}}$	$\mathbf{n_{12}}$
	Negative(阴性)	$\mathbf{n_{21}}$	$\mathbf{n_{22}}$

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)F1 值.
(7)真阴性率 灵敏度 特异度 马修相关系数

解：

(1)查准率： $Precision = \dfrac{\text{正确决策为阳性类的样本总数}}{\text{决策为阳性类的样本总数}} \times 100\% = \dfrac{n_{11}}{n_{11}+n_{21}} \times 100\%$

(2)查全率： $Recall = \dfrac{\text{正确决策为阳性类的样本总数}}{\text{参与决策的阳性类样本总数}} \times 100\% = \dfrac{n_{11}}{n_{11}+n_{12}} \times 100\%$

(3)真阳性率： $Recall = \dfrac{\text{正确决策为阳性类的样本总数}}{\text{参与决策的阳性类样本总数}} \times 100\% = \dfrac{n_{11}}{n_{11}+n_{12}} \times 100\%$

(4) 假阳性率: $\dfrac{\text{错误决策为阳性类的样本总数}}{\text{参与决策的阴性类样本总数}} \times 100\% = \dfrac{n_{21}}{n_{21}+n_{22}} \times 100\%$

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

7. 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)第 1 类查准率；
(2)第 1 类查全率；
(3)第 1 类 F1 值.
(4)第 1 类 F_β 值.

8. 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}
	⋮	⋮	⋮	⋱	⋮
	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；
(2)第 i 类的查全率；
(3)第 i 类的 F1 值、F_β 分值；
(4)分类模型的总体预测正确率；
(5)分类模型关于各类别预测错误率的算数平均值。

Macro-平均 Micro-平均