机器学习

朴素贝叶斯

假设我们正在构建一个分类器，该分类器说明文本是否与运动(Sports)有关。我们的训练数据有5句话：

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

我们想要计算句子“A very close game”是 Sports 的概率以及它不是 Sports 的概率。
即$P(\text{Sports} \mid \text{a very close game})$这个句子的类别是Sports的概率

28、根据下列数据判定(3，S)的类别，用朴素贝叶斯法（极大似然估计）。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

注：计算过程用分数表示即可，无需算成小数。

分类的评价指标

有100张照片，其中，猫的照片有60张，狗的照片是40张。

输入这100张照片进行二分类识别，找出这100张照片中的所有的猫。

正例（Positives）：识别对的

负例（Negatives）：识别错的

识别结果的混淆矩阵

		预测值	
		Positive	Negtive
实际值	Positive	TP=40	FN=20
	Negtive	FP=10	TN=30

正确率（Accuracy）=(TP+ TN)/S

TP+ TN =70，S= 100，则正确率为：

Accuracy =70/100=0.7

精度（Precision）=TP/(TP+ FP)

TP=40，TP+ FP=50。

Precision =40/50=0.8

召回率（Recall）=TP/(TP+ FN)

TP=40，TP+FN =60。则召回率为：

Recall =40/60=0.67

项目	符号	猫狗的例子
识别出的正例	TP+FP	40+10=50
识别出的负例	TN+FN	30+20=50
总识别样本数	TP+FP+TN+FN	50+50=100
识别对了的正例与负例	真正例+真负例=TP+TN	40+30=70
识别错了的正例与负例	伪正例+伪负例=FP+FN	10+20=30
实际总正例数量	真正例+伪负例=TP+FN	40+20=60
实际总负例数量	真负例+伪正例=TN+FP	30+10=40

1. 基于混淆矩阵的评价指标计算

（1）已知混淆矩阵

真实情况＼预测情况	正例	反例
正例	TP = 60	FN = 80
反例	FP = 20	TN = 40

（1）查全率（Recall, R）与真正例率（TPR）的关系，并计算查全率

（2）查准率（Precision, P）与假正例率（FPR）的关系，并计算查准率

SVD计算案例

设矩阵 A定义为：$A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$

PCA的算法

PCA的算法案例

$$X = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

以这个为例，我们用PCA的方法将这组二维数据降到一维

K-均值

假设有 8 个点：$A_1(3,1)$、$A_2(3,2)$、$A_3(4,1)$、$A_4(4,2)$、$A_5(1,3)$、$A_6(1,4)$、$A_7(2,3)$、$A_8(2,4)$，请采用 k - 均值聚类算法将它们聚成两类（距离度量方法采用欧式距离），设初始聚类中心分别为 $D_1(4,4)$ 和 $D_2(3,3)$，写出详细的计算过程。

DBSCAN算法

- 对每个点计算其邻域Eps=3内的点的集合。
- 集合内点的个数超过MinPts=3的点为核心点。

举例：有如下13个样本点，使用DBSCAN进行聚类

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
X	1	2	2	4	5	6	6	7	9	1	3	5	3
Y	2	1	4	3	8	7	9	9	5	12	12	12	3

4. 已知有如下表的某数据集D，采用 DBSCAN 算法对其进行密度聚类分析，取$\varepsilon = 1$、$MinPts = 4$、$n = 14$。

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
属性 A	1	4	0	1	2	3	4	5	0	1	4	1	5	5
属性 B	0	0	1	1	1	1	1	1	2	2	2	3	0	2

KNN算法

29、一维数据（0.6.0.7.0.8,0.9,1.1,1.3,1.32,1.33,1.34,1.35,1.36,1.37）为第一类数据，（1.6, 1.8, 1.9, 2. 2.1, 2.2, 2.3, 2.4）为第二类。用 KNN 算法，当 K=3,6,9 时，分别判断点“1.7”为第一类还是第二类？

Apriori算法

10. 以下某超市得到的交易数据库，其项集I={a,b,c,d,e},设最小支持度 MinS=04,使用Apriori算法找出所有频繁项集。(10分)

信息ID	顾客ID	购买的商品
T_1	I_1	a,b,c,d
T_2	I_2	b,c,e
T_3	I_2	a,b,c,e
T_4	I_3	b,d,e
T_5	I_4	a,b,d,e

ID.3算法

21.请根据ID.3算法的原理计算下列数据集中属性色泽的信息增益。(15分)

编号	色泽	根蒂	做声	织理	触感	好瓜
1	青绿	蜷缩	浊响	清晰	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	硬滑	是
3	乌黑	蜷缩	浊呀	消除	硬滑	是
4	青经	蜷缩	沉闷	清晰	硬滑	总
5	浅白	蜷缩	性别	清晰	硬路	蔡
6	青绿	稍蜷	注明	清晰	软粘	是
7	乌黑	稍娇	独响	徐颖	软黏	是
8	乌墨	稍蟾	浊响	清晰	硬滑	是
9	乌黑	稍蜷	沉闷	稍稠	硬滑	否
10	青绿	硬挺	清脆	清晰	软贴	否
11	浅白	硬挺	清脆	模糊	硬纸	否
12	浅白	蜷缩	浊咱	模糊	软粘	蛋
13	青绿	稍蜂	浊响	稍微	硬滑	否
14	浅白	稍蜡	沉闷	稻制	硬塑	否
15	马黑	稍蜷	浊响	消费	软粘	蛋
16	浅白	蜷缩	浊响	library	硬屑	否
17	青绿	蜷缩	沉闷	稍糊	硬置	否

10

一、单选题(本大题共15小题，每小题1分，共15分)

1. 一般来说，下列哪种方法不属于监督学习?(B)
   A.线性回归 ✓
   B. K均值聚类 ✗
   C.决策树 ✓
   D.支持向量机 ✓

2. 在监督学习中，以下哪个任务属于回归问题?(B)
   A.根据肿瘤的体积和患者的年龄来判断良性或恶性
   B.如何预测上海浦东的房价
   C.将教室里的学生按爱好和身高划分为几类
   D.用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化的问题

3. 在ID3 决策树算法中，选择划分属性的依据是(B)
   A.属性的取值数量越多越好。
   B.信息增益越大越好。
   C.划分后的子集样本数量越多越好。
   D.属性的取值与目标变量越相关越好。

4.已知在所有男子中有 5%，在所有女子中有 0.25%患有色盲症。随机抽一人发现患色盲症，问其为男子的概率是多少?(设男子和女子的人数相等)(A)
   A.95%
   B.5%
   C.75%
   D.25%

$$\frac{0.05}{0.0025+0.05} = \frac{5}{5.25}$$

5. 支持向量机(SVM)使用什么方法来找到最佳分类超平面?()
   A.最小化误差
   B.最大化间隔
   C.最小化损失
   D.最大化准确率

在神经网络(Perceptron) 的训练中，任务顺序是什么?(A)
1.初始化随机权重 2.输入一个样本，计算输出值 3.如果预测值和输出不致，改变权重
4.进行数据集的下一批(batch)
A.1,2,3,4 B.4,3,2,1 C.3,1,2,4 D.1,4,3,2

7. 在训练一个线性回归模型时，我们通常最小化哪个损失函数?(C)
   A.绝对误差
   B.交叉熵
   C.均方误差
   D.对数损失

8. 在神经网络中，下列哪个选项最准确地描述激活函数的作用?(C)
   A.激活函数用于在前向传播过程中对输入数据进行缩放，以适应网络的权重。
   B.激活函数通过调整神经元的偏置项，控制神经元的激活水平。
   C.激活函数引入非线性变换，使神经网络能够学习复杂的非线性关系。
   D.激活函数用于在反向传播中计算损失函数关于权重的梯度。

9. 在神经网络训练过程中，何时会使用 Dropout技术?(B)
   A.数据预处理阶段
   B.模型训练阶段
   C.模型评估阶段
   D.模型预测阶段

10.在梯度下降算法中，如果学习率 α 过大，可能会导致什么结果?(C)
   A.参数更新过慢
   B.参数永远不收敛
   C.参数更新过快而不稳定
   D.参数停滞不前

11. 设X={1，2，3，4}是频繁项集，则可由X产生 ()个关联规则。
   A.6 B.7 C.8 D.9

12.在 Apriori算法中，关联规则的生成过程是基于什么原则的()
   A.频繁项集的支持度不低于设定的阈值。
   B.关联规则的置信度不低于设定的阈值。
   C.频繁项集的长度不超过设定的阈值。
   D.关联规则的支持度不低于设定的阈值。

13.在 K-means聚类算法中，以下哪个步骤描述了 K-means算法的收敛条件?
   A.选择K个点作为初始质心
   B.将每个点指派到最近的质心，形成K个簇
   C.质心停止移动，即它们不再改变自己的位置
   D.计算质心与 数据点之间的距离

13

B

A. 簇的形成依赖于数据点之间的距离和密度。

B簇的形成仅基于数据点的距离。

C、簇的形成依赖于数据点的标签信息。

D、簇的形成与数据点的分布无关，完全随机。

A 15、以下关于深度网络训练的说法正确的是()

A、训练过程中的梯度衡量了损失函数相对于模型参数的变化率

B.损失函数只能采用模型预测结果与真实值之间的差异 X

C.模型的使用过程往往用到反向传播 X

D.其他选项都正确

二、简答题(本大题共3小题，共20分)

16.简述K近邻算法(KNN)的基本原理，并讨论其优缺点。(6分)

17、请简述机器学习中的过拟合和欠拟合现象，并解释如何通过交叉验证来检合(6分)

18.什么是梯度下降算法?可以结合线性回归简述其基本原理。(8分)

四、综合题(本大题共2小题，共30分)

22.监督学习和非监督学习是一个常见的概念、请简述：1)、非监督学习的优势和劣势；2)、各举三个一般情况下的监督学习和非监督学习的算法例子；3)尝试举例讨论一下是否能将一个监督学习方法改为一个非监督学习方法，如果不能，为什么，如果能，请说明原因。(15分)。

23.请选取一个你熟悉的机器学习应用场景，详细描述从 1)数据准备。2)数据处理。3)模型选择、4)模型评估到5)模型部署的完整过程。(15分)