XXXXXXXXXX 学院

2020 至 2021 学年第 一 学期

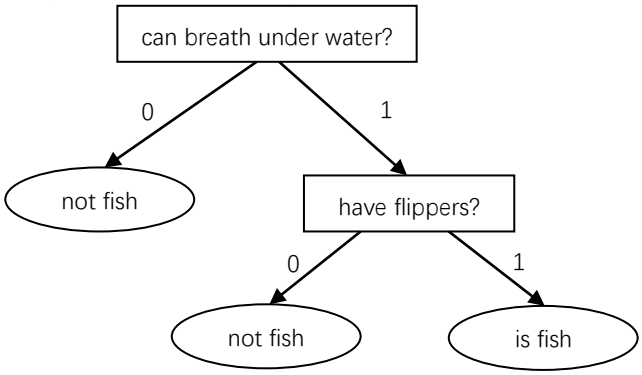
《机器学习》 期末考试试题评分标准（B 卷）

一、单选题（本题共 25 小题，满分 50 分）

题号	1	2	3	4	5	6	7	8	9	10
答案	D	A	A	B	B	A	B	C	D	C
题号	11	12	13	14	15	16	17	18	19	20
答案	B	C	C	B	C	B	B	D	D	B
题号	21	22	23	24	25					
答案	D	D	B	B	A					

二、计算题（本题共 5 小题，满分 50 分）

1.（本小题 4 分，每错一处扣一分，扣完为止）



2.（本小题 8 分）

（1）（每空 1 分，共 2 分）

clusterAssement

数据	对应样本所在的簇的序号 P1(第 0 簇质心)P2(第 1 簇质心)	样本距离所在簇质心的距离平方 $dist(x,C_i)^2$
P4	1/第 1 簇	2

（2）P1 P3 P6 属于第 0 簇（2 分）； P2 P4 P5 属于第 1 簇。（2 分）

（3）第一次迭代结束后，更新簇的质心，新的质心是什么？（2 分）

（2，5/3） （6,5）

3. （本小题 13 分）岭回归算法实现。

（1）岭回归算法采用 L2 正则化来简化模型。（1 分）

岭回归算法的损失函数是 $J(w)=\frac{1}{2m}\sum_{i=1}^{m}(h_w(x_i)-y_i)^2+\lambda\sum_{j=1}^{n}w_j^2$ （1 分）

（2）$\hat{\omega} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T y$ （1 分）

（3）（10 分）

```
def ridgeRegres(xArr, yArr, lam):
    xMat = np.mat(xArr) （1 分）
    yMat = np.mat(yArr).T（1 分）
    xTx = xMat.T * xMat （1 分）
    denom = xTx + np.eye(np.shape(xMat)[1]) * lam（2 分）
    if np.linalg.det(denom) == 0.0: （1 分）
        print(“矩阵为奇异矩阵,不能求逆") （1 分）
        return （1 分）
    ws = denom.I * (xMat.T * yMat) （1 分）
    return ws （1 分）
```

4. （本小题 15 分）

（1）写出 KNN 算法思想的基本步骤。（5 分）

1 计算已知类别中数据集的点与当前点的距离。

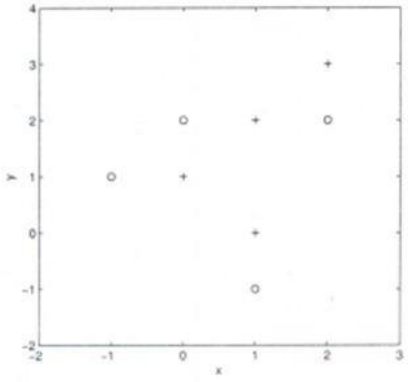
2 按照距离递增次序排序。

3 选取与当前点距离最小的 k 个点。

4 确定前 k 个点所在类别的出现频率。

5 返回前 k 个点出现频率最高的类别作为当前点的预测分类。

（2）使用 2D 空间显示上述数据，画出训练数据的散点图，用‘○’表示负样本，‘+’表示正样本。（2 分）



（3）假设你要使用 KNN（k=3）中的欧氏距离来预测新数据点 x = 1 和 y = 1 的类别。该数据点属于哪个类别？（需按照问题（1）写出计算过程）（5 分）

Step1:

x	y	x=1 y=1 距离点的距离
-1	1	2
0	1	1
0	2	1.414
1	-1	2
1	0	1
1	2	1

2	2	1.414
2	3	2.236

Step2-3: k=3

x	y	Class
0	1	+
1	0	+
1	2	+

Step4-5: 数据点 x = 1 和 y = 1 属于 + 类

（4）在问题（3）中，使用 KNN 算法时令 k=7，那么 x = 1 和 y = 1 属于哪个类别？为什么？
数据点 x = 1 和 y = 1 属于 - 类别（1 分）
k 值取值过大，结果受到样本不均衡影响。（2 分）

5. （本小题 10 分）
计算目标分类的信息熵 H（D）= -(1/3 $\log_2$1/3+2/3 $\log_2$2/3) = 0.918（1 分）
计算“天气”属性的条件熵 H（D|天气）= -1/3(2/5 $\log_2$2/5+3/5 $\log_2$3/5)-1/3(1*$\log_2$1+0*$\log_2$0)-1/3(2/5 $\log_2$2/5+3/5 $\log_2$3/5)= (0.971+0+0.971)/3 = 0.647（1 分）
计算“天气”属性的信息增益 g（天气，D）= 0.918-0.647 = **0.271**（1 分）
计算“气温”属性的条件熵 H（D|气温）= -4/15(1/2 $\log_2$1/2+1/2 $\log_2$1/2)-2/5(1/3 $\log_2$1/3+2/3 $\log_2$2/3)-1/3(1/5 $\log_2$1/5+4/5 $\log_2$4/5)= (1*4/15+0.918*2/5+0.722/3) = 0.875（1 分）
计算“气温”属性的信息增益 g（气温，D）= 0.918-0.875 = **0.043**（1 分）
计算“湿度”属性的条件熵 H（D|湿度）= -8/15(1/2 $\log_2$1/2+1/2 $\log_2$1/2)-7/15(1/7 $\log_2$1/7+6/7 $\log_2$6/7)= (1*8/15+0.591*7/15) = 0.809（1 分）
计算“湿度”属性的信息增益 g（湿度，D）= 0.918-0.809 = **0.109**（1 分）
计算“风力”属性的条件熵 H（D|风力）= -8/15(1/4 $\log_2$1/4+3/4 $\log_2$3/4)-7/15(3/7 $\log_2$3/7+4/7 $\log_2$4/7)= (0.811*8/15+0.985*7/15) = 0.892（1 分）
计算“风力”属性的信息增益 g（风力，D）= 0.918-0.892 = **0.026**（1 分）
比较信息增益值，“天气”属性的信息增益 g（天气，D）最大，“天气”属性作为根节点。（1 分）