人工智能原理

作业 5

注意：

1）请在网络学堂提交电子版；

2）请在 5 月 16 日晚 23:59:59 前提交作业，不接受补交；

1. 调查某地引用水源中某重金属元素含量和该地人群患病率之间的关系，收集到的调查结果如下：（写出计算过程或推导过程，如有计算结果请保留小数点后 4 位）。

重金属含量 (mg/L)	0.47	0.64	1.00	1.47	1.60	2.86	3.21	4.05	4.71	5.12
患病率 (%)	20.23	27.90	23.77	28.85	27.62	36.08	34.52	37.45	40.71	46.58

a) 画出患病率关于重金属含量的散点图；

b) 利用表中数据求患病率y关于重金属含量x的一元线性回归方程和确定系数r^2，并说明确定系数的含义；

c) 计算平均绝对误差(MAE)与均方误差(MSE)，并用这两个指标评估该线性模型的好坏；

d) 推导一般一元线性回归模型$\hat{y} = \hat{w}x + b$中总离差平方和$\sum_{i=1}^{n}(y_i - \bar{y})^2$，总误差平方和$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$与不同响应的离差平方和$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$的关系，并用题中数据验证该结论。

2. 使用二元 Logistic 回归公式的对数模型进行扩展，推导 Softmax 回归模型，即下式：

$$P\left(y^{(i)} = k\right) = \frac{e^{\boldsymbol{\beta}_k x_i}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k x_i}}$$

其中$P\left(y^{(i)} = k\right)$表示第$i$个样例被预测为第$k$类的概率，$x_i$为输入的第$i$个样例数据，$\boldsymbol{\beta}_k$为权重，二者都是向量。

提示：使用对数线性模型对预测结果为k的概率进行建模，这里的$-logZ$为归一化因子，用于保证模型预测的所有类别的概率集合构成一个概率分布，即模型预测的所有类别的概率之和为 1。

$$\log P\left(y^{(i)}=k\right)=\boldsymbol{\beta}_k x_i-\log Z$$

3. 输入数据$\{(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)\}$有$l$个类别，即$y\in\{1,2,\dots,l\}$。设 Softmax 回归模型对应的$l$个类别权重分别为$w_1,w_2,\dots,w_l$，偏置为$b_1,b_2,\dots,b_l$，损失函数为交叉熵损失。简洁起见，假设$x_i$为标量。

   a) 请以单数据输入(x_1,y_1)为例，给出使用梯度下降求解该模型时，由输入计算输出和由输出计算梯度的过程；（学习率用α表示）

   b) 说明在梯度下降的过程中，参数是如何被更新的。