人工智能基础

作业 5

注意：

1） 请在网络学堂提交**电子版**；

2） 请在 **11 月 10 日晚 23:59:59** 前提交作业，**不接受补交**；

3） 4 道题目中任选 3 道解答(多做不加分；4 题全做则按题目的解答顺序，只计前 3 题的分数，如提交作业中题目解答顺序是 1、3、2、4,则第 4 题不计分)。

4） 如有疑问，请联系助教：

杨鹏帅：yps18@mails.tsinghua.edu.cn

鄞启进：yqj17@mails.tsinghua.edu.cn

崔雪建：cuixj19@mails.tsinghua.edu.cn

高子靖：gzj21@mails.tsinghua.edu.cn

鲁永浩：yonghao.lu@foxmail.com

牛家赫：njh20@mails.tsinghua.edu.cn

江　澜：jiangl20@mails.tsinghua.edu.cn

尹小旭：yxx21@mails.tsinghua.edu.cn

1. 证明课件中一元线性回归的平方和分解公式。对于n组观测点$(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$，通过最小二乘法求得线性回归方程为 $\hat{y}=\hat{\omega}x+b$，试证明下式成立：

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

2. 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：

重金属含量(mg/L)	0.47	0.64	1.00	1.47	1.60	2.86	3.21	4.71
患病率（%）	22.37	23.31	25.32	22.29	28.57	35.00	46.07	46.08

(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（写出计算过程，计算结果保留小数点后 4 位）；

*(c).（此问不做要求，不计入作业分数）对回归方程的回归系数进行假设检验。（假设回归方程的残差服从的正态分布方差未知，可以采用 t 分布进行假设检验，α取 0.05，即$p-value$为 0.05）

注：自由度$\nu=6$时，$T_{0.01}\approx 3.143, T_{0.025}\approx 2.447, T_{0.05}\approx 1.943, T_{0.1}\approx 1.440$

3. 在使用最小二乘法对观测数据做线性回归时，常常假设预测变量的观测值没有误差，但有些情况下预测变量的观测值也会存在误差，即对于一组观测数据$(x_i,y_i), i=1,\dots n$，有

$$y_i = \omega\xi_i + b + \varepsilon_i; \varepsilon \sim N(0, \sigma_\varepsilon^2)$$
$$x_i = \xi_i + \delta_i; \delta_i \sim N(0, \sigma_\delta^2)$$

(a).该情况下最小二乘法是否适用，为什么？

(b).可以考虑用正交最小二乘法来对该类问题做线性回归。该方法通常使用回归直线的单位法向量$\boldsymbol{n}$和一个直线上的点$\boldsymbol{x_0}$来表示，为了统一表示，我们用$\boldsymbol{x_i} = (x_i, y_i)^T$来表示观测点，用每个观测点到回归直线的距离平方和来表示此时的平方误差损失，请写出该损失的表达式；（形式正确即可，系数不做过多要求）

(c).什么情况下(b)求得的平方损失表达式值最小？

（提示：考虑在直线参数固定的情况下，损失值取最小时直线上的点$\boldsymbol{x_0}$和数据均值点$\overline{\boldsymbol{x}}$的关系）

4. 对于多元回归问题$\boldsymbol{Y} = \boldsymbol{X\beta}$，如果采用最小二乘法优化平方误差损失，即

$$J(\boldsymbol{\beta}) = ||\boldsymbol{X\beta} - \boldsymbol{Y}||_2^2$$

可以得到

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

而当$\boldsymbol{X}$不是列满秩或某些列之间线性相关性较高时，$\boldsymbol{X}^T\boldsymbol{X}$接近奇异，用上述函数计算得到参数$\boldsymbol{\beta}$的值误差较大，缺乏稳定性和可靠性。为了防止参数$\beta$值出现异常现象，我们在损失函数中加入一个正则化项$\lambda\left|\left|\boldsymbol{\beta}\right|\right|_2^2$，用损失一定的无偏性来换取参数$\beta$数值的高稳定性，即损失函数为

$$J(\boldsymbol{\beta}) = \left|\left|\boldsymbol{X\beta} - \boldsymbol{Y}\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{\beta}\right|\right|_2^2, \lambda > 0$$

此时，线性回归问题转化为了岭回归问题(Ridge Regression)。

(a).试证明当上述损失函数最小时，参数$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$。其中，$\boldsymbol{I}$是单位矩阵，$\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$可逆。

（提示：对于长度为n的列向量w，$\frac{\partial \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\partial \boldsymbol{w}} = 2\boldsymbol{A}\boldsymbol{w}, \frac{\partial \boldsymbol{w}^T\boldsymbol{A}}{\partial \boldsymbol{w}} = \frac{\partial \boldsymbol{A}^T\boldsymbol{w}}{\partial \boldsymbol{w}} = \boldsymbol{A}$）

(b).令$\boldsymbol{X} = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix}, \boldsymbol{Y} = (1.3, -0.5, 2.6, 0.9)^T$。分别计算当损失函数中惩罚项$\lambda = 1, 5, 10$的情况下，参数$\boldsymbol{\beta}$的取值。