生物信息学概论 第三次作业

1. 多元线性回归模型的系数与单变量线性回归模型的系数的区别是什么？
   (可提交纸板作业)
   a. 考虑 $y = \boldsymbol{x}^t\boldsymbol{\beta} + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ 线性模型，x 为 p 维已知列向量量，y 为观测值标量，$\boldsymbol{\beta}$为 p 维系数列向量，ε 为随机误差项。
   请说明极大似然估计与最小二乘法，在求解该模型$\boldsymbol{\beta}$时是等价的
   b. 单变量回归：对于 x 向量某一元$\boldsymbol{x}_i$，我们可以考虑这样的模型$y = \boldsymbol{x}_i\boldsymbol{\beta}_i^* + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ 线性模型。$\boldsymbol{\beta}_i^*$为此模型通过极大似然或者最小二乘法求解时，得到的系数估计。
   请说明 $\boldsymbol{\beta}_i^*$ 与 $\boldsymbol{\beta}_i$ 的区别是什么？可以从理论推导或者实验模拟角度。
   c. *(此问题为附加题目，较难，慎选)
   生物信息中，很多数据的表现为样本数 n 远小于样本的维度 p. 我们会在传统的线性回归模型中，引入 L1 (1 范数正则化，即 LASSO),L2(2 范数正则化，即 Ridge) 正则化项。考虑 L1，或者 L2 或者无约束下，你能否仅利用$\boldsymbol{\beta}_i^*$的估计情况，来提前判断哪些维度肯定无效(无效的含义为，相应的系数估计为 0 或者系数估计的置信区间含有 0)，不需要带入模型中求解。这样的好处是，在处理大规模数据中，我们可以减小内存消耗，加快算法运行，提高模型精度。
2. 实现一种求解 LASSO 的算法，并在一组数据集上进行计算。
   求解 LASSO 的方法有很多，你可以选择课上讲解的，也可以利用 LARS 等其他算法，但尽量自己实现算法的细节。
   数据集: statweb.stanford.edu/~tibs/ElemStatLearn/ 选择 Data 中的 Prostate。
   为了验证你的算法实现正确，请利用 R package glmnet 进行比较；也可以下载上述网站提供的书籍，参考第三章 Figure 3.10 (correction 10th version)，与之进行比较。
   注：该书是机器学习领域中的经典书籍。