计算机应用数学第一次作业

SA25011049 李宇哲

计算题

T1

1、已知某工厂某批次水泥重量服从正态分布，总体方差为 2.65 公斤， 从该工厂随机抽取 18 袋水泥，其平均重量为 24.9 公斤， 试求该工厂水泥平均重量的 95%和 99%的置信区间。（10 分）

$$\sigma = 2.65, n = 18, \bar{x} = 24.9$$

95%置信区间：

$$\text{误差范围} E_{95\%} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{2.65}{\sqrt{18}} = 1.96 \cdot 0.625 \approx 1.225$$

$$[24.9 - 1.225, 24.9 + 1.225] = [23.675, 26.125]$$

$$99\%\text{置信区间：} E_{99\%} = z_{0.005} \cdot \frac{\sigma}{\sqrt{n}} = 2.576 \cdot 0.625 \approx 1.610$$

$$[24.9 - 1.610, 24.9 + 1.610] = [23.2990, 26.510]$$

95%置信区间$[23.68, 26.12]$公斤

99%置信区间$[23.29, 26.51]$公斤

T2

2、从某学校抽取 20 名高一学生，经测量，这 20 名学生的平均身高为 155cm，标准差为 10cm，假设平均身高服从正态分布，试求该学校高一学生总平均身高的 95%和 99%的置信区间。（10 分）

$$n = 20, \bar{x} = 155, s = 10$$

当总体方差未知，使用标准差估计：

$$\bar{x} - t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}}$$

$$SE = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{20}} \approx 2.236$$

95%置信区间

$$E_{95\%} = t_{0.025}(19) \cdot SE = 2.093 \cdot 2.236 \approx 4.680$$

$$[155 - 4.680, 155 + 4.680] = [150.32, 159.68]$$

$$99\%\text{置信区间} E_{99\%} = t_{0.005}(19) \cdot SE = 2.861 \cdot 2.236 \approx 6.397$$

$$[155 - 6.397, 155 + 6.397] = [148.60, 161.40]$$

95%置信区间为$[150.32, 159, 68]cm$

99%置信区间为$148.60,\ 161.40cm$

T3

例 设$x_1, x_2, \ldots, x_3$是均值为x、方差为σ^2的随机变量X的n个观测值的随机样本，证明：样本方差s^2是总体方差σ^2的一个无偏估计，其中：
a) 被抽样总体为正太分布
b) 被抽样总体的分布未知

3、1.2 节 P19，证明被抽样总体分布未知情况下的无偏估计。（10 分）

即需要证明：$E(S^2) = E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2] = \sigma^2$

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}[(x_i-\mu)-(\bar{x}-\mu)]^2 \\ &= \sum_{i=1}^{n}[(x_i-\mu)^2 - 2(x_i-\mu)(\bar{x}-\mu) + (\bar{x}-\mu)^2] \\ &= \sum_{i=1}^{n}(x_i-\mu)^2 - 2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu) + n(\bar{x}-\mu)^2\end{aligned}$$

注意到：$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}x_i - n\mu = n\bar{x} - n\mu = n(\bar{x}-\mu)$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i-\mu) = n(\bar{x}-\mu)^2$$

对两侧取期望：

$$E[\sum_{i=1}^{n}(x_i-\bar{x})^2] = E[\sum_{i=1}^{n}(x_i-\mu)^2] - E[n(\bar{x}-\mu)^2]$$

$$E[\sum_{i=1}^{n}(x_i-\mu)^2] = n\sigma^2$$

由于$x_1, x_2, \ldots, x_n$独立同分布：

$$Var(\bar{x}) = Var(\frac{1}{n}\sum_{i=1}^{n}x_i) = \frac{1}{n^2}$$

$$E[n(\bar{x}-\mu)^2] = n\cdot\frac{\sigma^2}{n} = \sigma^2$$

因此：$E[\sum_{i=1}^{n}(x_i-\bar{x})^2] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n-1}\cdot(n-1)\sigma^2 = \sigma^2$$

T4

> 4、证明 Lemma 1.3.1。（20 分）
>
> **Lemma 1.3.1**. Let P be the transition probability matrix for a connected Markov Chain. The $n \times (n+1)$ matrix $A = [P - I, 1]$ obtained by augmenting the matrix $P - I$ with an additional column of ones has rank n.

步骤1：证明 $\mathrm{rank}(P-I) = n-1$

$(P-I)\mathbf{1} = \mathbf{0}$ 说明 $\mathbf{1}$ 在 $P-I$ 的零空间中，因此：
$rank(P-I) \leq n-1\mathrm{rank}(P-I) \leq n-1rank(P-I) \leq n-1$

根据秩-零化度定理：
$rank(P-I) + dim(Null(P-I)) = n\mathrm{rank}(P-I) + \dim(\mathrm{Null}(P-I)) = nrank(P-I) + dim(Null(P-I)) = n$

由于 $\dim(\mathrm{Null}(P-I)) = 1$（零空间由 $\mathbf{1}$ 张成），因此：
$rank(P-I) = n-1\mathrm{rank}(P-I) = n-1rank(P-I) = n-1$

矩阵 $A = [P-I, \mathbf{1}]$ 是由 $P-I$ 和列向量 $\mathbf{1}$ 组成的 $n \times (n+1)$ 矩阵。

假设列向量 $\mathbf{1}$ 在 $P-I$ 的列空间中，即存在向量 $\mathbf{v}$ 使得：
$(P-I)v = 1(P-I)\mathbf{v} = \mathbf{1}(P-I)v = 1$

这等价于：
$Pv = 1 + vP\mathbf{v} = \mathbf{1} + \mathbf{v}Pv = 1 + v$

对两边求和（左乘行向量 $\mathbf{1}^T$）：
$1TPv = 1T(1+v) = n + 1Tv1^TP\mathbf{v} = \mathbf{1}^T(\mathbf{1}+\mathbf{v}) = n + \mathbf{1}^T\mathbf{v}1TPv = 1T(1+v) = n + 1Tv$

因此
$1TPv = 1Tv1^TP\mathbf{v} = \mathbf{1}^T\mathbf{v}1TPv = 1Tv$

但这与 $\mathbf{1}^TP\mathbf{v} = n + \mathbf{1}^T\mathbf{v}$ 矛盾（除非 $n=0$）。

所以有向量 $\mathbf{1}$ 不在 $P-I$ 的列空间中

由于：

- $\mathrm{rank}(P-I) = n-1$
- 向量 $\mathbf{1}$ 线性独立于 $P-I$ 的列向量（不在其列空间中）

因此，增加列向量 $\mathbf{1}$ 后，秩增加1：
$rank(A) = rank([P-I, 1]) = rank(P-I) + 1 = (n-1) + 1 = n\mathrm{rank}(A) = \mathrm{rank}([P-I, \mathbf{1}]) = \mathrm{rank}(P-I) + 1 = (n-1) + 1 = nrank(A) = rank$

所以 矩阵A是满秩的

实验题

T1 PageRank

实现见代码

结果统计

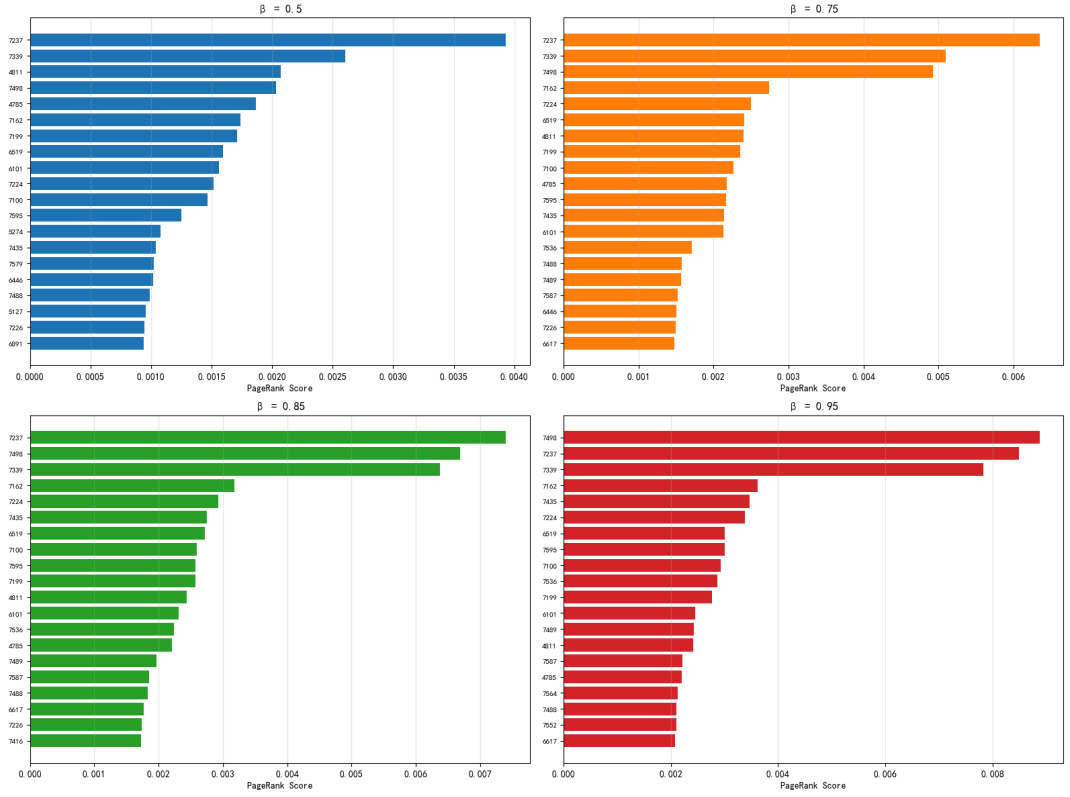
```
PageRank 分数最高的 20 个节点
 1. Node 7237: 0.00740342
 2. Node 7498: 0.00669226
 3. Node 7339: 0.00637877
 4. Node 7162: 0.00317726
 5. Node 7224: 0.00293076
 6. Node 7435: 0.00275113
 7. Node 6519: 0.00271756
 8. Node 7100: 0.00259440
 9. Node 7595: 0.00257613
10. Node 7199: 0.00257569
11. Node 4811: 0.00243469
12. Node 6101: 0.00230890
13. Node 7536: 0.00223911
14. Node 4785: 0.00221205
15. Node 7489: 0.00197029
16. Node 7587: 0.00184996
17. Node 7488: 0.00183503
18. Node 6617: 0.00176548
19. Node 7226: 0.00173878
20. Node 7416: 0.00172525
```

在 β 为 0.85（by default）的情况相爱，PageRank 在第20次迭代收敛，节点总数 7624，其他信息如下：

```
节点总数: 7624
平均 PageRank: 0.000131
最大 PageRank: 0.007403
最小 PageRank: 0.000055
标准差: 0.000225
PageRank 总和: 1.000000
```

Optional

我尝试了不同参数（β）对结果的影响（0.5,0.75,0.85,0.95）



β越小，迭代收敛速度越快，不同参数下 top参数有所区别，但pagerank最大的几个结点几乎相同。

β越大，top结点 PageRank 更明显。

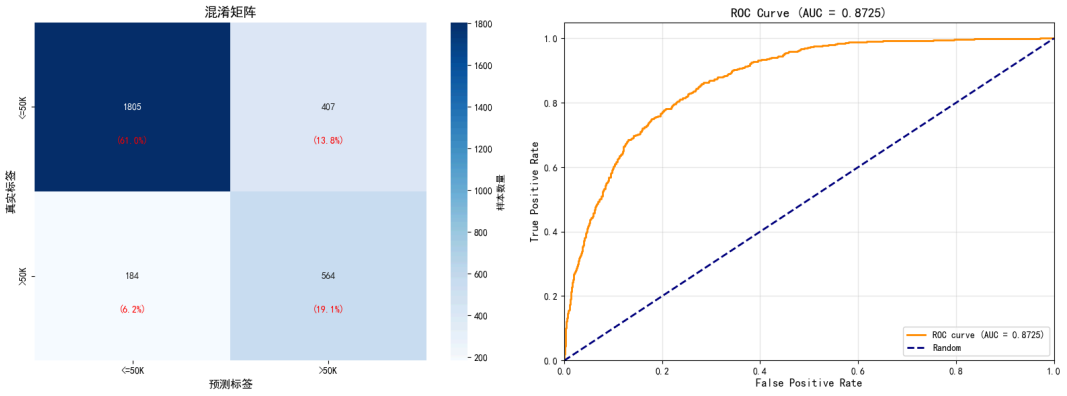
T2 朴素贝叶斯分类器

实现见代码部分

这里我设置了 10 个 features，最后一个 feature 是国籍，虽然这个 feature 并没有什么作用。

结果统计

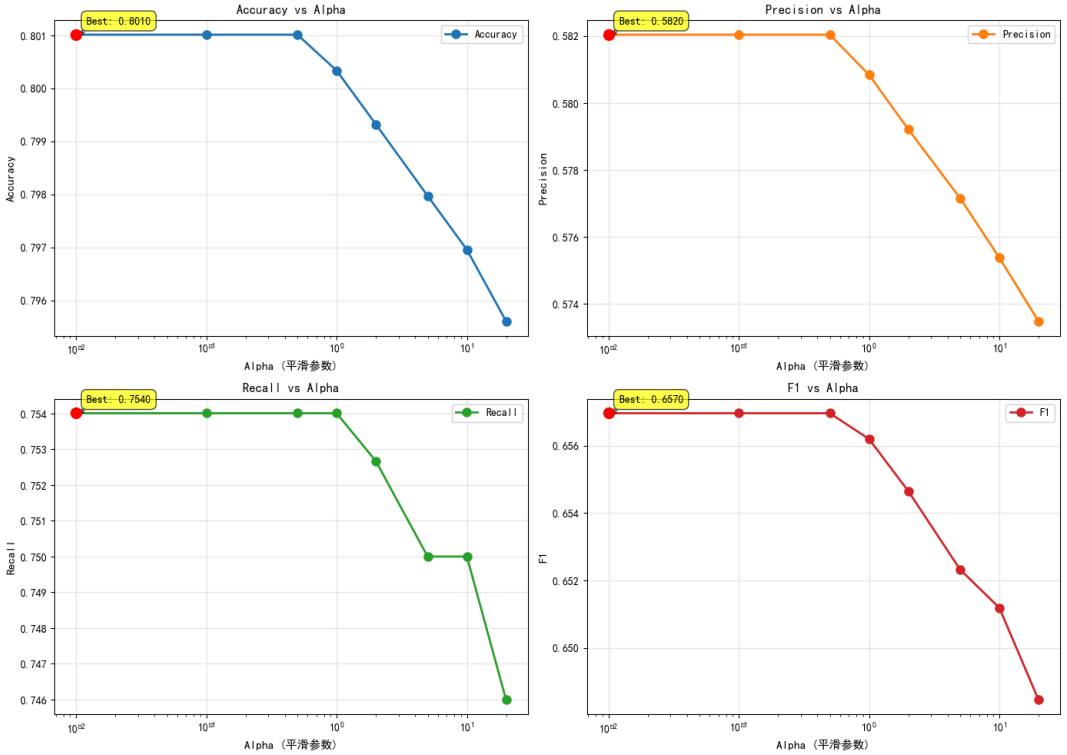
```
Accuracy:  0.8003
Precision: 0.5808
Recall:    0.7540
F1 Score:  0.6562

混淆矩阵:
[[1805  407]
 [ 184  564]]
TN=1805, FP=407, FN=184, TP=564

详细分类报告:
              precision    recall  f1-score   support

       <=50K       0.91      0.82      0.86      2212
        >50K       0.58      0.75      0.66       748

    accuracy                           0.80      2960
   macro avg       0.74      0.79      0.76      2960
weighted avg       0.82      0.80      0.81      2960


AUC Score: 0.8725
```



csv文件

处于debug需求，我并没有直接修改 test 文件，而是将预测结果和 test 的标签一起写入一个新文件 `bayesian_detailed_results.csv`，存放每一个数据对应预测结果和test实际结果，correct表示结果是否正确。

Optional



上图是随着 alpha 变化，4个不同指标的变化情况，可以看出，alpha越大，模型预测准确率，精度，recall和F1都有所下降。