HW6

xw-zeng

2022-11-03

HW6.

1、(1) 求先验概率 P_k 的 MLE. $P_k = P(Y=c_k)$.

$$L(P_k) = \prod_{i=1}^{N} P_k^{1(y_i=c_k)} (1-P_k)^{1(y_i\neq c_k)}.$$

$$= P_k^{\sum 1(y_i=c_k)} (1-P_k)^{N-\sum 1(y_i=c_k)}.$$

$$l(P_k) = \log L(P_k) = \sum 1(y_i=c_k) \log P_k - (N-\sum 1(y_i=c_k)) \log(1-P_k).$$

$$l'(P_k) = \frac{1}{P_k} \sum 1(y_i=c_k) + \frac{1}{1-P_k}(N-\sum 1(y_i=c_k))$$

令 $l'(P_k)=0$.

$$\sum 1(y_i=c_k)(1-P_k) + (N-\sum 1(y_i=c_k))P_k = 0.$$

$\therefore \hat{P_k} = \frac{\sum 1(y_i=c_k)}{N}$, 得证.

(2) 求条件概率 P_c 的 MLE. $P_c = P(x^{(j)}=a_{jl} \mid Y=c_k)$.

$$L(P_c) = \prod_{i=1}^{N} P_c^{1(x_i^{(j)}=a_{jl},\, y_i=c_k)} (1-P_c)^{1(x_i^{(j)}\neq a_{jl},\, y_i=c_k)1(y_i=c_k)}.$$

$$= P_c^{\sum_{i=1}^{N} 1(x_i^{(j)}=a_{jl},\, y_i=c_k)} (1-P_c)^{\sum_{i=1}^{N}1(y_i=c_k) - \sum_{i=1}^{k}1(x_i^{(j)}=a_{jl},\, y_i=c_k)}.$$

$$l(P_c) = \log L(P_c) = \sum_{i=1}^{N} 1(x_i^{(j)}=a_{jl},\, y_i=c_k) \log P_c + (\sum_{i=1}^{N} 1(y_i=c_k) - \sum_{i=1}^{N} 1(x_i^{(j)}=a_{jl},\, y_i=c_k)) \log(1-P_c)$$

令 $l'(P_c)=0$.

$$\frac{1}{P_c} \sum_{i=1}^{N} 1(x_i^{(j)}=a_{jl},\, y_i=c_k) - \frac{1}{1-P_c}(\sum_{i=1}^{N} 1(y_i=c_k) - \sum_{i=1}^{N} 1(x_i^{(j)}=a_{jl},\, y_i=c_k)) = 0.$$

$\therefore \hat{P_c} = \frac{\sum_{i=1}^{N} 1(x_i^{(j)}=a_{jl}, y_i=c_k)}{\sum_{i=1}^{N} 1(y_i=c_k)}$, 得证.

2、即求 $\arg\max_{c_k} P(Y=c_k \mid X=(2,M)^T)$.

(1) MLE 估计.

① 先验概率 $P(Y=1) = \frac{10}{15} = \frac{2}{3}$, $P(Y=-1) = \frac{1}{3}$.

② 条件概率 $P(X^{(1)}=2 \mid Y=1) = \frac{2}{10}$. $P(X^{(2)}=M \mid Y=1) = \frac{4}{10}$.

$P(X^{(1)}=2 \mid Y=-1) = \frac{3}{5}$. $P(X^{(2)}=M \mid Y=-1) = \frac{2}{5}$.

③ 后验概率 $P(Y=1 \mid X=(2,M)^T) = \frac{2}{3} \times \frac{2}{10} \times \frac{4}{10} = \frac{4}{75}$.

$P(Y=-1 \mid X=(2,M)^T) = \frac{1}{3} \times \frac{3}{5} \times \frac{2}{5} = \frac{6}{75}$.

$\Rightarrow$ 类标记为 -1.

(2) 贝叶斯估计. $\lambda=1$.

① 先验概率 $P(Y=1)=\frac{10+1}{15+2}=\frac{11}{17}$, $P(Y=-1)=\frac{5+1}{15+2}=\frac{6}{17}$.

② 条件概率 $P(X^{(1)}=2|Y=1)=\frac{2+1}{10+3}=\frac{3}{13}$. $P(X^{(2)}=M|Y=1)=\frac{4+1}{10+3}=\frac{5}{13}$.

$P(X^{(1)}=2|Y=-1)=\frac{3+1}{5+3}=\frac{4}{8}$. $P(X^{(2)}=M|Y=-1)=\frac{2+1}{5+3}=\frac{3}{8}$.

③ 后验概率 $P(Y=1|X=(2,M)^T)=\frac{11}{17}\times\frac{3}{13}\times\frac{5}{13}\approx 0.05$

$P(Y=-1|X=(2,M)^T)=\frac{6}{17}\times\frac{4}{8}\times\frac{3}{8}\approx 0.066$

$\Rightarrow$ 类标记为 -1.

∴ 两种估计方法类标记都为 -1.

市长电话分析

分析任务 1

读入市长电话训练集和测试集。

```
data_train <- read.csv('train_set.csv', encoding = 'utf-8')
data_test <- read.csv('test_set.csv', encoding = 'utf-8')
```

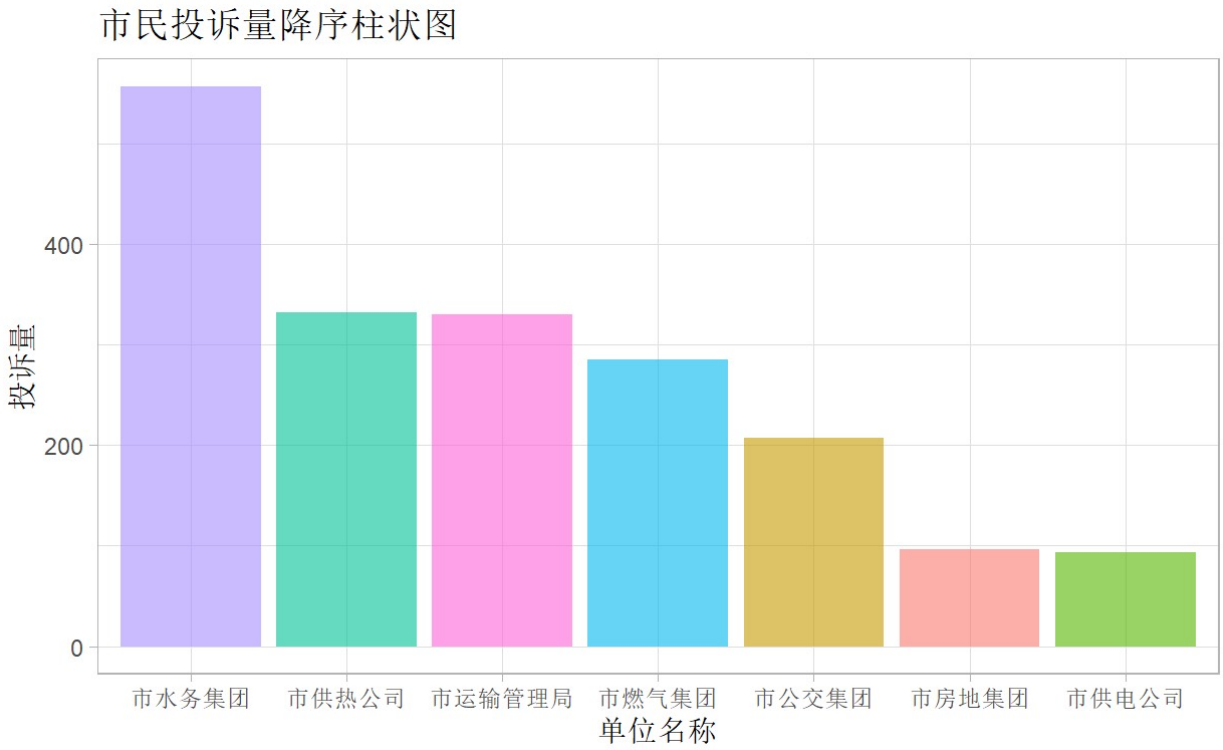
统计训练集中各个政府单位接到的市民投诉量。

```
complaints <- data_train %>% group_by(单位名称) %>% summarise(投诉量 = n())
complaints
```

```
## # A tibble: 7 x 2
##   单位名称      投诉量
##   <chr>          <int>
## 1 市房地集团        96
## 2 市公交集团       207
## 3 市供电公司        93
## 4 市供热公司       332
## 5 市燃气集团       285
## 6 市水务集团       557
## 7 市运输管理局     330
```

按照投诉量降序，绘制柱状图。

```
complaints <- complaints[order(complaints$投诉量, decreasing = TRUE), ]
ggplot(complaints, mapping = aes(x = 单位名称, y = 投诉量, fill = 单位名称)) +
  geom_bar(stat = 'identity', alpha = 0.6) +
  scale_x_discrete(limits = complaints$单位名称) +
  labs(title = '市民投诉量降序柱状图', x = '单位名称', y = '投诉量') +
  guides(fill = 'none') +
  theme_light()
```



由上图可知，市水务集团收到的投诉量最多，有 557 条；投诉量较多的第二梯队 (300-400 条) 为市供热公司、市运输管理局，投诉量分别为 330 条、332 条；第三梯队 (200-300 条) 为市燃气集团、市公交集团，投诉量分别为 285 条、207 条；投诉量最少的政府单位为市房地集团、市供电公司，投诉量分别为 96 条和 93 条，都小于 100 条。

分析任务 2

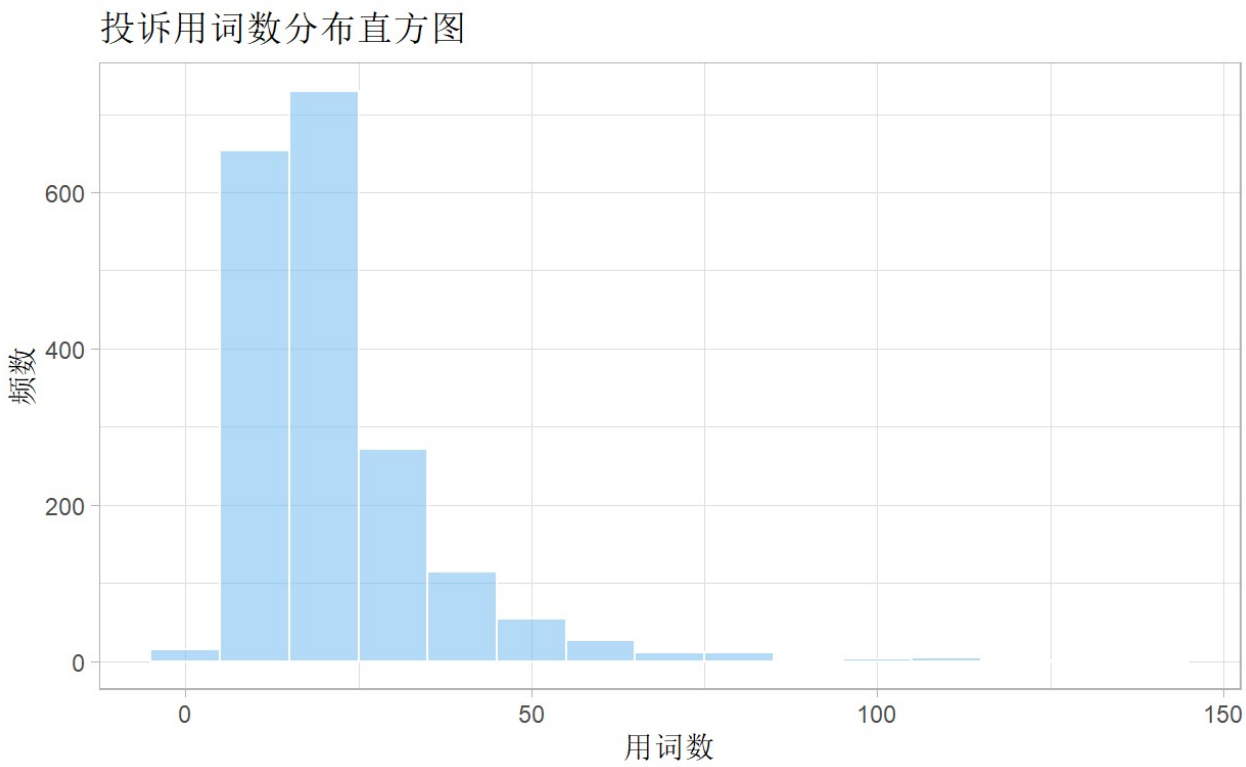
统计每条投诉用词数。

```
vols <- data_train[, c(1, 2)]
vols['用词数'] <- rowSums(data_train[, 2:6236])
vols <- vols[, -2]
head(vols)
```

```
##      单位名称 用词数
## 1 市供热公司     8
## 2 市水务集团     7
## 3 市水务集团    33
## 4 市燃气集团    19
## 5 市公交集团    11
## 6 市供热公司    11
```

绘制投诉用词数的分布直方图。

```
ggplot(data = vols, mapping = aes(x = 用词数)) +
  geom_histogram(color = 'white', fill = 'skyblue2', binwidth = 10,
                 mapping = aes(y = ..count..), alpha = 0.6) +
  labs(title = '投诉用词数分布直方图', y = '频数', x = '用词数') +
  theme_light()
```

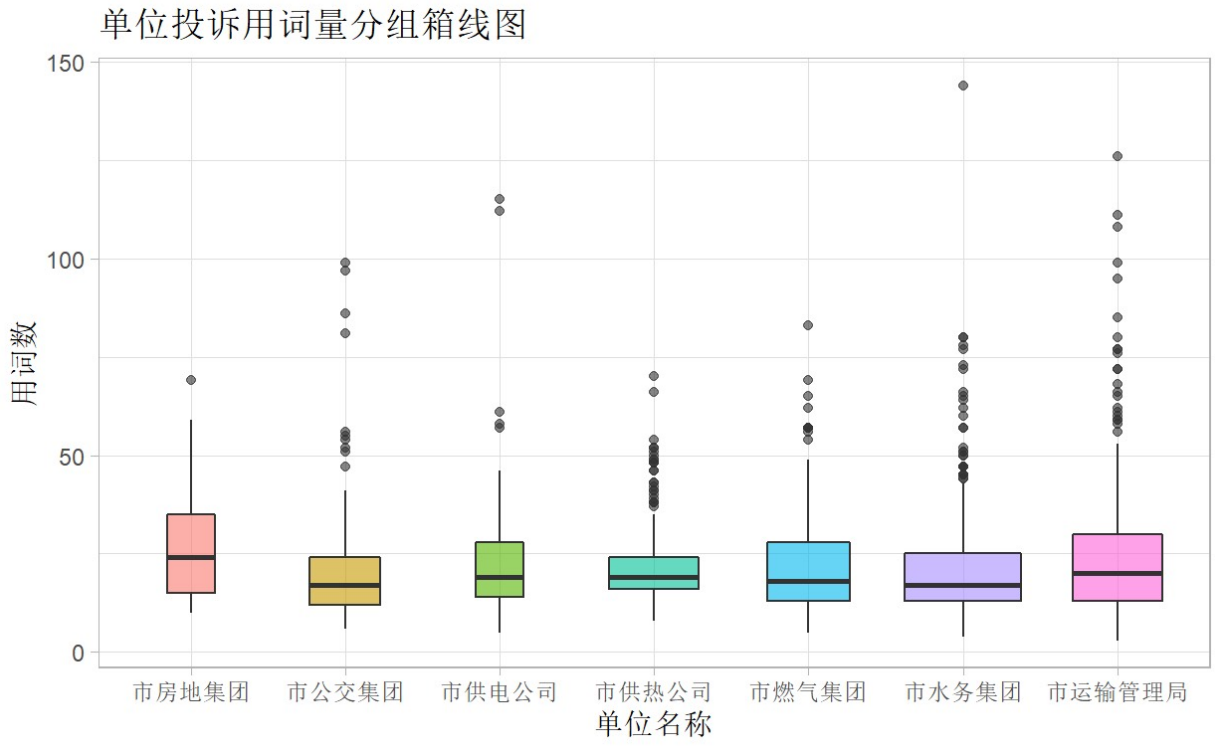


用词数的分布呈右偏分布，大部分的投诉用词数都在 10-50 个之间，只有极少数一部分的投诉超过了 50 个词，说明大部分市民投诉都比较简短。

分析任务 3

用箱线图表示各单位收集到投诉信息分词后总词数的差异。

```
ggplot(vols, mapping = aes(x = 单位名称, y = 用词数, fill = 单位名称)) +
  geom_boxplot(varwidth = TRUE, alpha = 0.6) +
  labs(title = '单位投诉用词量分组箱线图') +
  guides(fill = 'none') +
  theme_light()
```



由上图可以得到以下结论：

- 市房地集团投诉用词数的中位数最高，约为 25 个词，说明该单位收到的投诉用词量较多，但是该部门的投诉数量却在所有单位中排名倒数第二，这与我们日常生活是相符的，因为房地产相关的问题往往比较复杂，需要用较多词汇才能描述清楚。
- 市水务集团、市公交集团、市燃气集团的用词数中位数最低，所以即使这些单位投诉数量比较多，但是用词数都是比较少的。
- 市房地集团、市供电公司用词数的离群值较少，说明大家的问题可能都差不多复杂；其他单位的离群值较多，说明市民问题的异质性可能较大。

分析任务 4

将训练集与测试集转化为 0-1 矩阵。

```
dtm_train <- apply(data_train[, -1], 2, as.factor)
dtm_test <- apply(data_test[, -1], 2, as.factor)
```

修剪词汇表，取前 200 个高频词作为分类器特征。

```
wordfreq <- colSums(data_train[, 2:6236])
index <- order(wordfreq, decreasing = TRUE)[1:200]
```

以训练集中的政府单位为因变量类别，电话文本为自变量，建立朴素贝叶斯分类器。

```
nb <- naiveBayes(x = dtm_train[, index], y = as.factor(data_train$单位名称))
```

使用该模型对测试集电话文本进行预测。

```
pred <- predict(nb, dtm_test[, index], type = 'class')
```

计算准确率。

```
accuracy <- sum(as.character(data_test$单位名称) == as.character(pred)) /
  nrow(data_test)
print(paste0('准确率为: ', accuracy))
```

```
## [1] "准确率为: 0.98"
```

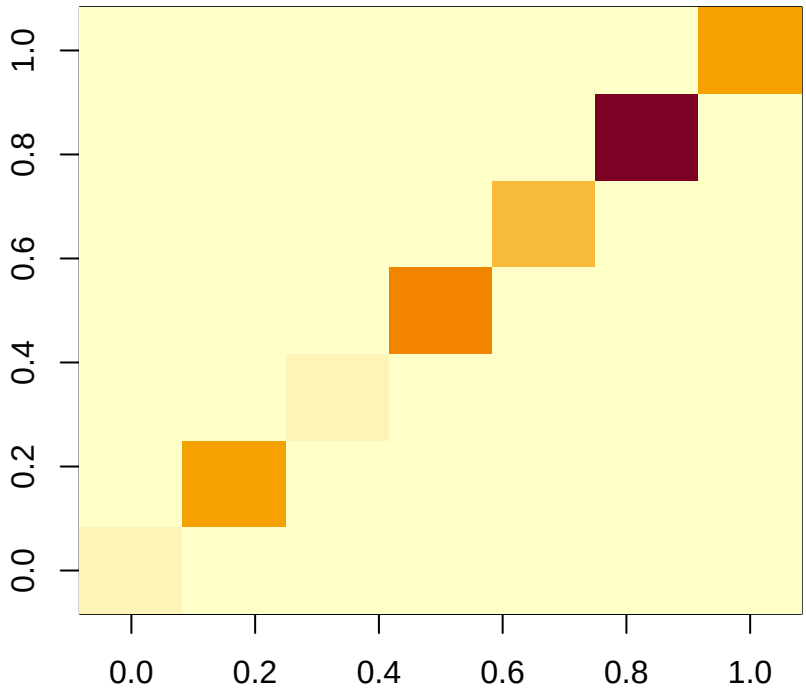
计算混淆矩阵。

```
table(as.character(data_test$单位名称), as.character(pred))
```

```
##
##                 市房地集团 市公交集团 市供电公司 市供热公司 市燃气集团
##   市房地集团             5          0          0          1          0
##   市公交集团             0         15          0          0          0
##   市供电公司             0          0          3          0          0
##   市供热公司             0          0          0         17          0
##   市燃气集团             0          0          0          0         12
##   市水务集团             0          0          0          0          0
##   市运输管理局           0          0          0          0          0
##
##                 市水务集团 市运输管理局
##   市房地集团             0            0
##   市公交集团             0            0
##   市供电公司             1            0
##   市供热公司             0            0
##   市燃气集团             0            0
##   市水务集团            32            0
##   市运输管理局           0           14
```

绘制混淆矩阵图像。

```
image(table(as.character(data_test$单位名称), as.character(pred)))
```



模型准确率高达 98%，即在 100 个测试样本中只发生了 2 次分类错误，混淆矩阵的深色块都集中在对角线处，说明模型效果很好。

THE END. THANKS! ^_^