HW8

xw-zeng

2022-11-28

理论推导

1. 对偶问题

软间隔最大化问题的原问题为:

$$\begin{array}{ll}\min_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i^2 \\ \text{s.t.} & y_i\left(w\cdot x_i + b\right) \geqslant 1-\xi_i, \quad i=1,2,\cdots,N \\ & \xi_i \geqslant 0, \quad i=1,2,\cdots,N\end{array}$$

写出拉格朗日函数:

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left\{y_i\left(w\cdot x_i+b\right)-1+\xi_i\right\} - \sum_{i=1}^{N}\mu_i\xi_i$$

其中 $\alpha_i \geq 0, \mu_i \geq 0$，转化为极大极小问题，首先求解极小问题，关于 w, b, ξ 求偏导:

$$\nabla_w L(w,b,\xi,\alpha,\mu) = w - \sum_{i=1}^{N}\alpha_i y_i x_i = 0$$

$$\nabla_b L(w,b,\xi,\alpha,\mu) = -\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\nabla_{\xi_i} L(w,b,\xi,\alpha,\mu) = C - \alpha_i - \mu_i = 0$$

化简后可得:

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$C = \alpha_i + \mu_i$$

代入拉格朗日函数:

$$\begin{aligned}
&\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\left(1-\xi_i-y_i\left(w^T x_i+b\right)\right) - \sum_{i=1}^{N}\mu_i\xi_i \\
=&\frac{1}{2}\|w\|^2 + \sum_{i=1}^{N}\alpha_i\left(1-y_i\left(w^T x_i+b\right)\right) + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\mu_i\xi_i \\
=&-\frac{1}{2}\sum_{i=1}^{N}\alpha_i y_i x_i^T\sum_{i=1}^{N}\alpha_i y_i x_i + \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}C\xi_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\mu_i\xi_i \\
=&-\frac{1}{2}\sum_{i=1}^{N}\alpha_i y_i x_i^T\sum_{i=1}^{N}\alpha_i y_i x_i + \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}\left(C-\alpha_i-\mu_i\right)\xi_i \\
=&\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
=&\min_{w,b,\xi} L(w,b,\alpha,\xi,\mu)
\end{aligned}$$

再求解极大问题，即为对偶问题：

$$\begin{aligned}\max_{\alpha,\mu}\min_{w,b,\xi} L(w,b,\alpha,\xi,\mu) &= \max_{\alpha,\mu}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &= \max_{\alpha}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j\end{aligned}$$

约束条件为:

$$\begin{aligned}&\alpha_i \geq 0\\ &\mu_i \geq 0\\ &C=\alpha_i+\mu_i\\ \Rightarrow &0\leq \alpha_i \leq C, 0\leq \mu_i\leq C\end{aligned}$$

2. 核方法

首先看两个例子：

1. 对 $p=2, n=2$，有 $\phi_2(x)=\left(x_1^2,\sqrt{2}x_1x_2,x_2^2\right)$，使得：

$$\begin{aligned}K(x,z) &= \left(x^T z\right)^2=\left(\left(x_1,x_2\right)^T\left(z_1,z_2\right)\right)^2\\ &=\left(x_1z_1+x_2z_2\right)^2\\ &=x_1^2z_1^2+2x_1z_1x_2z_2+x_2^2z_2^2\\ &=\left(x_1^2,\sqrt{2}x_1x_2,x_2^2\right)^T\left(z_1^2,\sqrt{2}z_1z_2,z_2^2\right)^T\\ &:=\phi_2(x)\cdot\phi_2(z)\end{aligned}$$

2. 对 $p=3, n=2$，有 $\phi_3(x)=(x_1^3,\sqrt{3}x_1^2x_2,\sqrt{3}x_1x_2^2,x_2^3)$，使得：

$$\begin{aligned}K(x,z) &= \left(x^T z\right)^3=\left(\left(x_1,x_2\right)^T\left(z_1,z_2\right)\right)^3\\ &=\left(x_1z_1+x_2z_2\right)^3\\ &=x_1^3z_1^3+3x_1^2z_1^2x_2z_2+3x_1z_1x_2^2z_2^2+x_2^3z_2^3\\ &=\left(x_1^3,\sqrt{3}x_1^2x_2,\sqrt{3}x_1x_2^2,x_2^3\right)^T\left(z_1^3,\sqrt{3}z_1^2z_2,\sqrt{3}z_1z_2^2,z_2^3\right)\\ &:=\phi_3(x)\cdot\phi_3(z)\end{aligned}$$

因此，我们猜测对于任意 p,n，都能为 $K(x,z)$ 找到从 $\mathcal{X}$ 到希尔伯特空间 $\mathcal{H}$ 的映射 ϕ，使得：

$$K(x,z)=\phi(x)\cdot\phi(z)$$

令 $x,z\in\mathcal{R}^n$，使用数学归纳法进行证明：

当 $p=1$ 时，有 $\phi_1(x)=x$ 使得：

$$K(x,z)=\phi_1(x)\cdot\phi_1(z)=x\cdot z$$

假设 $p=k$ 时，存在 ϕ_k 使得：

$$K(x,z)=\phi_k(x)\cdot\phi_k(z)$$

则 $p=k+1$ 时：

$$K(x,z)=(x\cdot z)^{k+1}=(x\cdot z)^k(x\cdot z)=\left(\phi_k(x)\cdot\phi_k(z)\right)(x\cdot z)$$

不妨假设 ϕ_k 由以下形式变换得到：

$$\phi_k(x)=\left(f_1(x),f_2(x),\cdots,f_m(x)\right)^T$$

则有：

$$\begin{aligned}
K(x,z)&=\left(f_1(x)f_1(z)+f_2(x)f_2(z)+...+f_m(x)f_m(z)\right)\left(x_1z_1+x_2z_2+...+x_nz_n\right)\\
&=f_1(x)f_1(z)\left(x_1z_1+x_2z_2+...+x_nz_n\right)+f_2(x)f_2(z)\left(x_1z_1+x_2z_2+...+x_nz_n\right)+...\\
&\quad+f_m(x)f_m(z)\left(x_1z_1+x_2z_2+...+x_nz_n\right)\\
&=\left(f_1(x)x_1\right)\left(f_1(z)z_1\right)+\left(f_1(x)x_2\right)\left(f_1(z)z_2\right)+...+\left(f_1(x)x_n\right)\left(f_1(z)z_n\right)+\left(f_2(x)x_1\right)\left(f_2(z)z_1\right)+...\\
&\quad+\left(f_2(x)x_n\right)\left(f_2(z)z_n\right)+\left(f_m(x)x_1\right)\left(f_m(z)z_1\right)+...+\left(f_m(x)x_n\right)\left(f_m(z)z_n\right)\\
&:=\phi_{k+1}(x)\cdot\phi_{k+1}(z)
\end{aligned}$$

故有 $\phi_{k+1}(x)=\left(f_1(x)x_1,...,f_1(x)x_n,f_2(x)x_1,...,f_2(x)x_n,......,f_m(x)x_1,...,f_m(x)x_n\right)^T$ 使得 $K(x,z)=(x\cdot z)^{k+1}=\phi_{k+1}(x)\cdot\phi_{k+1}(z)$，结论在 $p=k+1$ 时也成立，得证。

对任意 $x_i\in\mathcal{X}(i=1,...,n)$，构造 $K(x,z)$ 关于 $x_1,x_2,...,x_n$ 的 Gram 矩阵：

$$K=[K_{ij}]_{n\times n}=[K(x_i,x_j)]_{n\times n}$$

则对任意 $a_1,a_2,...,a_n\in\mathcal{R}$，有：

$$\begin{aligned}
\sum_{i,j=1}^{n}a_ia_jK\left(x_i,x_j\right)&=\sum_{i,j=1}^{n}a_ia_j\left(\phi\left(x_i\right)\cdot\phi\left(x_j\right)\right)\\
&=\left(\sum_i a_i\phi\left(x_i\right)\right)\cdot\left(\sum_j a_j\phi\left(x_j\right)\right)\\
&=\left\|\sum_i a_i\phi\left(x_i\right)\right\|^2\geq 0
\end{aligned}$$

因此，$K(x,z)$ 关于 $x_1,x_2,...,x_n$ 的 Gram 矩阵为半正定矩阵。

内积的正整数幂函数满足对称性（因为内积本身就具有对称性）：

$$K(x,z)=(x\cdot z)^p=(z\cdot x)^p$$

综上所述，内积的正整数幂函数 $K(x,z)=(x,z)^p$ 满足正定核的充要条件，是正定核函数，问题得证。

征信系列-用户行为数据分析

载入 R 包。

```
library(ggplot2) # plot beautiful graphs
library(rpart) # decision tree
library(e1071) # svm
library(pROC) # draw ROC curve
library(patchwork) # merge two ggplots
```

定义画 ROC 曲线的函数。

```
show_roc <- function(true, pred, train_or_test){
  roc_curve <- roc(true, pred)
  plot(roc_curve, print.auc = TRUE, auc.polygon = TRUE, legacy.axes = TRUE,
       grid = c(0.1, 0.2), grid.col = c('green', 'red'), max.auc.polygon = TRUE,
       auc.polygon.col = 'skyblue', print.thres = TRUE, main = paste0('ROC Curve on ',
          ifelse(train_or_test == 'train', 'Training', 'Test'), ' Data'))
}
```

定义画混淆矩阵 (confusion matrix) 的函数。

```
show_cm <- function(df, modelname){
  df$pred <- df[, 2]
  out <- ggplot(df, aes(pred, true_class, fill = Freq)) +
    geom_tile() + geom_text(aes(label = Freq)) +
    scale_fill_gradient(low = 'white', high = '#3575b5') +
    labs(x = 'Pred', y = 'True', title = paste('Confusion Matrix of', modelname),
         fill = '是否违约') + guides(fill = 'none') + theme_light()
  return (out)
}
```

分析任务 1

读入数据。

```
data <- read.csv('simudata.csv')
```

将变量 black (是否违约) 转化为因子型变量。

```
data$black <- factor(data$black, levels = c(0, 1), labels = c('未违约', '违约'))
```

分析任务 2

按照 7:3 的比例划分训练集和测试集。

```
set.seed(1234)
idx_train <- sample(1:nrow(data), 0.7 * nrow(data))
data_train <- data[idx_train, ]
data_test <- data[-idx_train, ]
```

用决策树模型在训练集上进行建模。

```
fit_rpart <- rpart(black ~ ., data_train)
```

用 SVM 模型在训练集上进行建模。

```
fit_svm <- svm(black ~ ., data_train, probability = TRUE)
```

使用建立好的模型对测试集进行预测，输出预测概率，便于绘制 ROC 曲线。

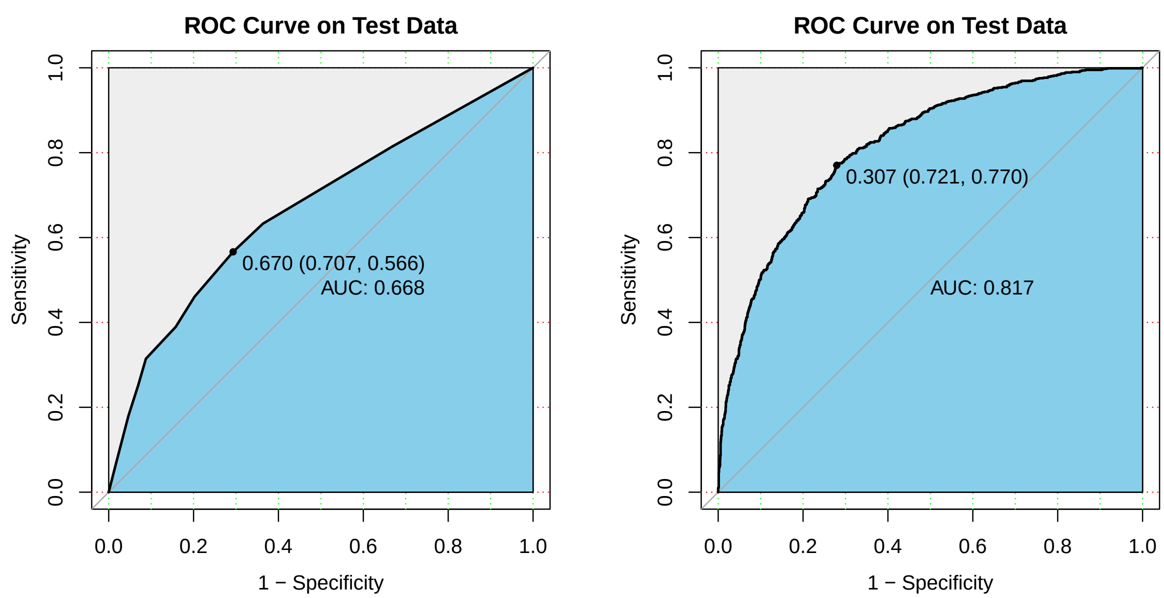
```
pred_rpart_prob <- predict(fit_rpart, data_test, type = 'prob')
pred_svm_prob <- attr(predict(fit_svm, data_test, probability = TRUE), 'probabilities')
```

再输出预测类型，便于绘制混淆矩阵。

```
pred_rpart_class <- predict(fit_rpart, data_test, type = 'class')
pred_svm_class <- predict(fit_svm, data_test)
```

分别画出决策树模型、SVM 模型在测试集上的 ROC 曲线。

```
par(mfrow=c(1,2))
show_roc(data_test$black, pred_rpart_prob[, 1], 'test')
show_roc(data_test$black, pred_svm_prob[, 1], 'test')
```



由上图，决策树的 AUC 大小为 0.668，SVM 的 AUC 大小为 0.817，这说明 SVM 模型在该问题上的分类能力远高于决策树。

分别计算两个模型的分类准确率 accuracy。

```
true_class <- data_test$black
pred_rpart <- table(true_class, pred_rpart_class)
pred_svm <- table(true_class, pred_svm_class)
acc_rpart <- sum(diag(pred_rpart)) / nrow(data_test)
acc_svm <- sum(diag(pred_svm)) / nrow(data_test)
print(paste0('决策树的准确率为：', round(acc_rpart * 100, 2), '%'))
```

```
## [1] "决策树的准确率为：71.04%"
```

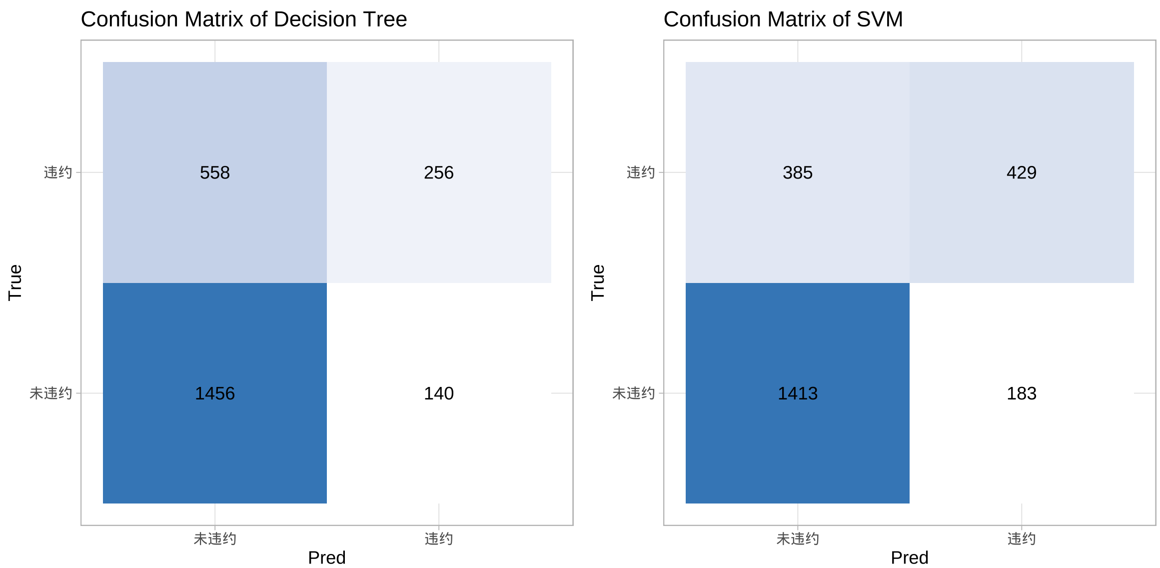
```
print(paste0('SVM 的准确率为：', round(acc_svm * 100, 2), '%'))
```

```
## [1] "SVM的准确率为：76.43%"
```

故 SVM 分类准确率比决策树的准确率更高，效果更好。

最后分别画出决策树模型、SVM 模型在测试集上的混淆矩阵。

```
p1 <- show_cm(data.frame(pred_rpart), 'Decision Tree')
p2 <- show_cm(data.frame(pred_svm), 'SVM')
p1 + p2
```



由上图可知，SVM 在预测的真正例比决策树更多，即更有能力分辨出违约用户；而两个模型对负例的预测效果相差不大。

综上所述，SVM 模型的效果比决策树模型好。

THE END. THANKS! ^_^