机器学习导论习题课

第五次作业

第一题：朴素贝叶斯分类器

数据集表示如下：

Table 1: 数据集

编号	x_1	x_2	x_3	x_4	y
样本1	1	1	1	0	1
样本2	1	1	0	0	0
样本3	0	0	1	1	0
样本4	1	0	1	1	1
样本5	0	0	1	1	1

第一题：朴素贝叶斯分类器

Learning And Mining from DatA
http://lamda.nju.edu.cn

(1) 计算：

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} \text{ 与} \Pr\{y=0|\mathbf{x}=(1,1,0,1)\}$$

解：

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} = \frac{P(y)}{P(\mathbf{x})}\prod_{i=1}^{4} P(x_i|y)$$

$$= \frac{\Pr\{y=1\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}} \Pr\{x_1=1|y=1\}\Pr\{x_2=1|y=1\}\Pr\{x_3=0|y=1\}\Pr\{x_4=1|y=1\}$$

$$= \frac{3/5}{P(\mathbf{x})} \times \frac{2}{3} \times \frac{1}{3} \times \frac{0}{3} \times \frac{2}{3}$$

$$= 0/P(\mathbf{x})$$

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\} = \frac{P(y)}{P(\mathbf{x})}\prod_{i=1}^{4} P(x_i|y)$$

$$= \frac{\Pr\{y=0\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}} \Pr\{x_1=1|y=0\}\Pr\{x_2=1|y=0\}\Pr\{x_3=0|y=0\}\Pr\{x_4=1|y=0\}$$

$$= \frac{2/5}{P(\mathbf{x})} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$$

$$= \frac{1}{40}/P(\mathbf{x})$$

第一题：朴素贝叶斯分类器

(2) “拉普拉斯修正”之后，再重新计算上一问

很多样本的取值可能并不在训练集上，这不并代表这种情况发生的概率为0，因为未被观测到，并不代表出现的概率为0。拉普拉斯修正避免出现先验概率为零。

先验概率：

$$P(c)=\frac{Dc}{D} \Rightarrow P(c)=\frac{Dc+1}{D+N}$$

类的条件概率

$$P(x_i \mid c)=\frac{D_{c,xi}}{Dc} \Rightarrow P(x_i \mid c)=\frac{D_{c,xi}+1}{Dc+Ni}$$

第一题：朴素贝叶斯分类器

(2) [10pts]“拉普拉斯修正”之后，再重新计算上一问。

解：

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\}=\frac{\hat{P}(y)}{\hat{P}(\mathbf{x})}\prod_{i=1}^{4}\hat{P}(x_i|y)$$

$$=\frac{\hat{\Pr}\{y=1\}}{\hat{\Pr}\{\mathbf{x}=(1,1,0,1)\}}\hat{\Pr}\{x_1=1|y=1\}\hat{\Pr}\{x_2=1|y=1\}\hat{\Pr}\{x_3=0|y=1\}\hat{\Pr}\{x_4=1|y=1\}$$

$$=\frac{4/7}{P(\mathbf{x})}\times\frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{3}{5}$$

$$=\frac{72}{4375}/P(\mathbf{x})$$

$$=\frac{0.0164}{P(\mathbf{x})}$$

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}=\frac{\hat{P}(y)}{\hat{P}(\mathbf{x})}\prod_{i=1}^{4}\hat{P}(x_i|y)$$

$$=\frac{\hat{\Pr}\{y=0\}}{\hat{\Pr}\{\mathbf{x}=(1,1,0,1)\}}\hat{\Pr}\{x_1=1|y=0\}\hat{\Pr}\{x_2=1|y=0\}\hat{\Pr}\{x_3=0|y=0\}\hat{\Pr}\{x_4=1|y=0\}$$

$$=\frac{3/7}{P(\mathbf{x})}\times\frac{2}{4}\times\frac{2}{4}\times\frac{2}{4}\times\frac{2}{4}$$

$$=\frac{3}{112}/P(\mathbf{x})$$

$$=\frac{0.268}{P(\mathbf{x})}$$

第二题：贝叶斯最优分类器

假设同先验，证明当二分类任务中两类数据满足高斯分布且方差相同时，线性判别分析产生贝叶斯最优分类器。

答：贝叶斯最优分类器：

$$h^*(x) = \arg\max_{c \in y} P(x|c)P(c)$$

$$= \arg\max_{c \in y} log(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu_c)^T \Sigma^{-1}(x-\mu_c))) + log(P(c))$$

$$= \arg\max_{c \in y} -\frac{1}{2}(x-\mu_c)^T \Sigma^{-1}(x-\mu_c) + log(P(c))$$

$$= \arg\max_{c \in y} x^T \Sigma^{-1}\mu_c - \frac{1}{2}\mu_c^T \Sigma^{-1}\mu_c + log(P(c))$$

第二题：贝叶斯最优分类器

所以，贝叶斯决策边界为：

$$g(x) = x^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}\mu_0 - (\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0) + log(\frac{P(1)}{P(0)})$$

$$= x^T\Sigma^{-1}(\mu_1 - \mu_0) - \frac{1}{2}(\mu_1 + \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0) \text{（同先验）}$$

对于LDA，其决策边界为：

$$\boldsymbol{w}^{\mathrm{T}}\left(x - \frac{\mu_0 + \mu_1}{2}\right) = 0 \quad \text{其中，} \boldsymbol{w} = (2\Sigma)^{-1}(\mu_0 - \mu_1) = \frac{1}{2}\Sigma^{-1}(\mu_0 - \mu_1)$$

代入可知，两者决策边界相同。

编程题：集成算法

本次实验选用UCI数据集Adult，此数据集为一个二分类数据集。因为Adult是一个类别不平衡的数据集，所以采用AUC作为测试分类器性能的评价指标。

1) 实现AdaBoost算法。可以参考教材8.2节中图8.3所示的算法伪代码来实现。基分类器选用决策树，可以直接调用sklearn中决策树的实现。

2) 实现Random Forest算法。可以参考教材8.3.2节所述来实现。基分类器仍可以直接调用sklearn。

编程题：集成算法

3)根据交叉验证来分析基学习器数量对分类器训练效果的影响。这一步是希望大家根据交叉验证算法评估性能，来选取最优的超参数-基学习器的数量。可以自己先确定一个基分类器数目的取值范围，然后根据交叉验证来选取最优的超参。

4) 根据参数调查结果，对AdaBoost和随机 森林选取最好的基分类器数目，在训练数据集上进行训练，在实验报告中报告在测试集上的AUC指标。

在实验报告中，除了报告上述要求报告的内容外还需要展现实验过程，实验报告需要有层次和条理性，能让读者仅通过实验报告便能了解实验的目的，过程和结果。