Introduction to Machine Learning
Homework 5

2020 年 12 月 26 日

1 [30pts] Naive Bayes Classifier

We learned about the naive Bayes classifier using the "property conditional independence hypothesis". Now we have a data set as shown in the following table:

表 1: Dataset

	x_1	x_2	x_3	x_4	y
Instance1	1	1	1	0	1
Instance2	1	1	0	0	0
Instance3	0	0	1	1	0
Instance4	1	0	1	1	1
Instance5	0	0	1	1	1

(1) [15pts] Calculate: $\Pr\{y = 1|\mathbf{x} = (1, 1, 0, 1)\}$ and $\Pr\{y = 0|\mathbf{x} = (1, 1, 0, 1)\}$.
(2) [15pts] After using Laplacian Correction, recalculate the value in the previous question.

解：(1) 首先估计类先验概率 $P(c)$，有

$$P(y = 0) = \frac{2}{5}, P(y = 1) = \frac{3}{5}$$

然后，为每个属性估计条件概率 $P(x_i|c)$：

$$
\begin{aligned}
&P(x_1=1|y=0)=\frac{1}{2}, P(x_1=1|y=1)=\frac{2}{3}\\
&P(x_2=1|y=0)=\frac{1}{2}, P(x_2=1|y=1)=\frac{1}{3}\\
&P(x_3=0|y=0)=\frac{1}{2}, P(x_3=0|y=1)=\frac{0}{3}=0\\
&P(x_4=1|y=0)=\frac{1}{2}, P(x_4=1|y=1)=\frac{2}{3}
\end{aligned}
$$

利用全概率公式计算 $P(\mathbf{x}=(1,1,0,1))$

$$
\begin{aligned}
&P(\mathbf{x}=(1,1,0,1))\\
&=P(y=0)\prod_{i=1}^{4}P(x_i|y=0)+P(y=1)\prod_{i=1}^{4}P(x_i|y=1)\\
&=\frac{2}{5}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}+\frac{3}{5}\times\frac{2}{3}\times\frac{1}{3}\times 0\times\frac{2}{3}\\
&=\frac{1}{40}
\end{aligned}
$$

于是，有

$$
\begin{aligned}
&\Pr\{y=1|\mathbf{x}=(1,1,0,1)\}=\frac{P(y=1)}{P(\mathbf{x}=(1,1,0,1))}\prod_{i=1}^{4}P(x_i|y=1)\\
&=\frac{1}{P(\mathbf{x})}\times P(y=1)\times P(x_1=1|y=1)\times P(x_2=1|y=1)\times P(x_3=0|y=1)\times P(x_4=1|y=1)\\
&=40\times\frac{3}{5}\times\frac{2}{3}\times\frac{1}{3}\times 0\times\frac{2}{3}\\
&=0
\end{aligned}
$$

$$
\begin{aligned}
&\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}=\frac{P(y=0)}{P(\mathbf{x})}\prod_{i=1}^{4}P(x_i|y=0)\\
&=\frac{1}{P(\mathbf{x})}\times P(y=0)\times P(x_1=1|y=0)\times P(x_2=1|y=0)\times P(x_3=0|y=0)\times P(x_4=1|y=0)\\
&=40\times\frac{2}{5}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\\
&=1
\end{aligned}
$$

(2) 拉普拉斯修正之后，类先验概率 $P(c)$ 可估计为

$$
\hat{P}(y=0)=\frac{2+1}{5+2}=\frac{3}{7}, \hat{P}(y=1)=\frac{3+1}{5+2}=\frac{4}{7}
$$

每个属性的条件概率 $P(x_i|c)$ 可估计为

$$\hat{P}(x_1=1|y=0)=\frac{1+1}{2+2}=\frac{1}{2},\hat{P}(x_1=1|y=1)=\frac{2+1}{3+2}=\frac{3}{5}$$
$$\hat{P}(x_2=1|y=0)=\frac{1+1}{2+2}=\frac{1}{2},\hat{P}(x_2=1|y=1)=\frac{1+1}{3+2}=\frac{2}{5}$$
$$\hat{P}(x_3=0|y=0)=\frac{1+1}{2+2}=\frac{1}{2},\hat{P}(x_3=0|y=1)=\frac{0+1}{3+2}=\frac{1}{5}$$
$$\hat{P}(x_4=1|y=0)=\frac{1+1}{2+2}=\frac{1}{2},\hat{P}(x_4=1|y=1)=\frac{2+1}{3+2}=\frac{3}{5}$$

利用全概率公式计算 $\hat{P}(\mathbf{x}=(1,1,0,1))$

$$\begin{aligned}&\hat{P}(\mathbf{x}=(1,1,0,1))\\&=\hat{P}(y=0)\prod_{i=1}^{4}\hat{P}(x_i|y=0)+\hat{P}(y=1)\prod_{i=1}^{4}\hat{P}(x_i|y=1)\\&=\frac{3}{7}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}+\frac{4}{7}\times\frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{3}{5}\\&=\frac{3027}{70000}\end{aligned}$$

于是，有

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\}=\frac{\hat{P}(y=1)}{\hat{P}(\mathbf{x})}\prod_{i=1}^{4}\hat{P}(x_i|y=1)$$
$$=\frac{1}{\hat{P}(\mathbf{x})}\times\hat{P}(y=1)\times\hat{P}(x_1=1|y=1)\times\hat{P}(x_2=1|y=1)\times\hat{P}(x_3=0|y=1)\times\hat{P}(x_4=1|y=1)$$
$$=\frac{70000}{3027}\times\frac{4}{7}\times\frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{3}{5}$$
$$\approx 0.381$$

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}=\frac{\hat{P}(y=0)}{\hat{P}(\mathbf{x})}\prod_{i=1}^{4}\hat{P}(x_i|y=0)$$
$$=\frac{1}{P(\mathbf{x})}\times P(y=0)\times\hat{P}(x_1=1|y=0)\times\hat{P}(x_2=1|y=0)\times\hat{P}(x_3=0|y=0)\times\hat{P}(x_4=1|y=0)$$
$$=\frac{70000}{3027}\times\frac{3}{7}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}$$
$$\approx 0.619$$

2 [70pts] Ensemble Methods in Practice

Due to their outstanding performance and robustness, ensemble methods are very popular in machine community. In this experiment we will practice ensemble learning methods based on two classic ideas: Boosting and Bagging.

In this experiment, we use an UCI dataset Adult. You can refer to the link[1] to see the data description and download the dataset.

Adult is an class imbalanced dataset, so we select AUC as the performance measure. You can adopt sklearn to calculate AUC.

(1) [20pts] You need finish the code in Python, and only have two files: AdaBoost.py, RandomForestMain.py. (The training and testing process are implemented in one file for each algorithm.)

(2) [40pts] The is experiment requires to finish the following methods:

- Implement AdaBoost algorithm according to the Fig(8.3), and adopt decision tree as the base learner (For the base learner, you can import sklearn.)
- Implement Random Forest algorithm. Please give a pseudo-code in the experiment report.
- According to the AdaBoost and random forest, analysis the effect of the number of base learners on the performance. Specifically, given the number of base learners, use 5-fold cross validation to obtain the AUC. The range of the number of base learners is decided by yourself.
- Select the best number of base classifiers for AdaBoost and random forests, and obtain the AUC in the test set.

(3) [10pts] In the experimental report, you need to present the detail experimental process. The experimental report needs to be hierarchical and organized, so that the reader can understand the purpose, process and result of the experiment.

[1] http://archive.ics.uci.edu/ml/datasets/Adult

参考文献

[1] 周志华. 机器学习 [M]. 清华大学出版社, 2016.

[2] sklearn Adaboost 类库使用小结
https://www.cnblogs.com/pinard/p/6136914.html

[3] sklearn 缺失值处理
https://blog.csdn.net/qq_40773512/article/details/82662191

[4] sklearn 决策树算法类库使用小结
https://www.cnblogs.com/pinard/p/6056319.html

[5] sklearn 交叉验证
https://www.cnblogs.com/jiaxin359/p/8552800.html

[6] sklearn LabelEncoder 的使用
https://www.cnblogs.com/sench/p/10134094.html