ISYE 6420 Fall 2020 Project

Xiao Nan
GT Account: nxiao30
GT ID: 903472104

1 Introduction

In this project, we study the classical regression problem – the Iris dataset, with the methods introduced in ISYE6420 Bayesian Statistics. This dataset consists of 150 observations with 4 features each, and 3 target species – setosa, versicolor, virginica.

We are going to compare the performance between linear model and naive bayes methods like Gaussian Naive Bayes, and Multinomial Naive Bayes.

By using Naïve Bayes method, we have the naïve conditional independence assumption that

$$P(x_i \mid y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i \mid y)$$

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \dots, x_n)}$$

$$P(y \mid x_1, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

2 Experiment

Logistic Regression: We use L2 regularization

Gaussian Naive Bayes: the likelihood of the features is assumed to be Gaussian[1]

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Support Vector Classifier: We use linear kernel for our support vector classifier.

Using the scikit-learn library, we have the following result:

Model	Logistic Regression	Gaussian Naive Bayes	Multinomial Naive Bayes	Support Vector Classifier
Accuracy	96.00%	98.67%	97.33%	98.67%

3 Discussion

From the result we can see that Naïve Bayes methods provide a better distribution comparing with logistic regression. But from the study[2], we can see that logistic regression (discriminative model) performs better than Naïve Bayes (generative model) when the training size reaches infinity. However, naïve bayes is much faster (O(log n)) than logistic regression (O(n)). But we need to keep in mind that Naïve Bayes assumes the attributes are

conditionally independent. Also, we can see that support vector machine gives a solid performance on this task. Though most of the times, it is more suitable for high dimensional data. Gaussian Naive Bayes classifier gives a better performance than Multinomial Naive Bayes classifier, since the latter one is more suitable for classification with discrete features.

4 Reference

[1] Scikit-learn Naïve Bayes (https://scikit-learn.org/stable/modules/naive_bayes.html)
[2] Ng, A., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 14, 841-848.