Selected Topics

Julie Chou

5/8/2020

1. INTRODUCTION

This paper presents the analysis and evaluation on data sets using machine learning techniques. In this paper I set out to apply the Machine learning techniques like KNN, Logistic Regression, Decision Tree and SVM (support vector machine) in the analysis of sample data sets from Scikit-learn. I used the “Wisconsin Breast Cancer” which is default, preprocessed and cleaned datasets comes with scikit-learn. The target is to classify tumor as ‘malignant’ or ‘benign’ and code is written in python using jupyter notebook Our method in this assignment consist of working with any health or medicine related datasets and implement basic machine learning algorithms and understand more about the field of machine learning and its applications in the real world.

1. PROBLEM AND DATA SETS

***Breast Cancer, logistic regression, SVM, decision tree***

The objective is to classify the tumor as malignant or benign. Breast cancer dataset is a classification dataset. The output variable of the cancer data was either malignant or benign.

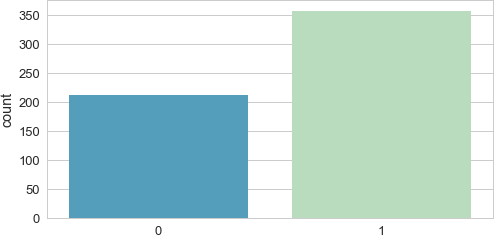
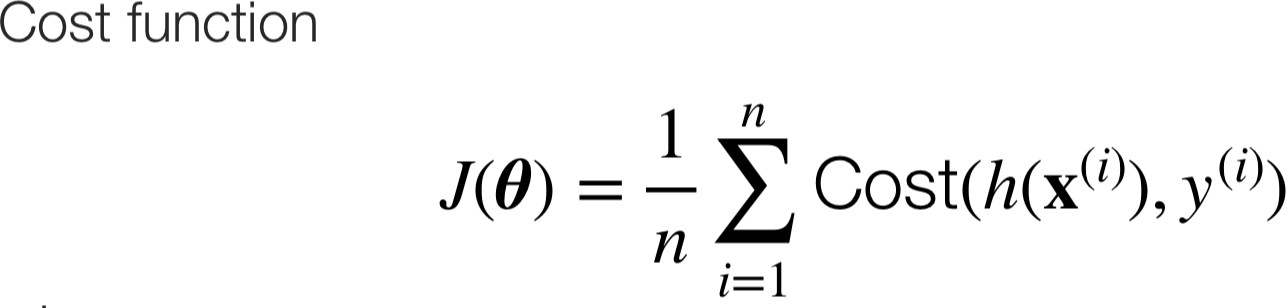


Fig. 1. 0: Malignant 1: Benign

I chose logistic regression model in this case. The idea of logistic regression is that there is an optimal decision boundary that separates the two classes of cancer.

For the given training data, we want to find parameters Θ that are most likely be maximizing L(Θ). This is the maximum likelihood estimator.



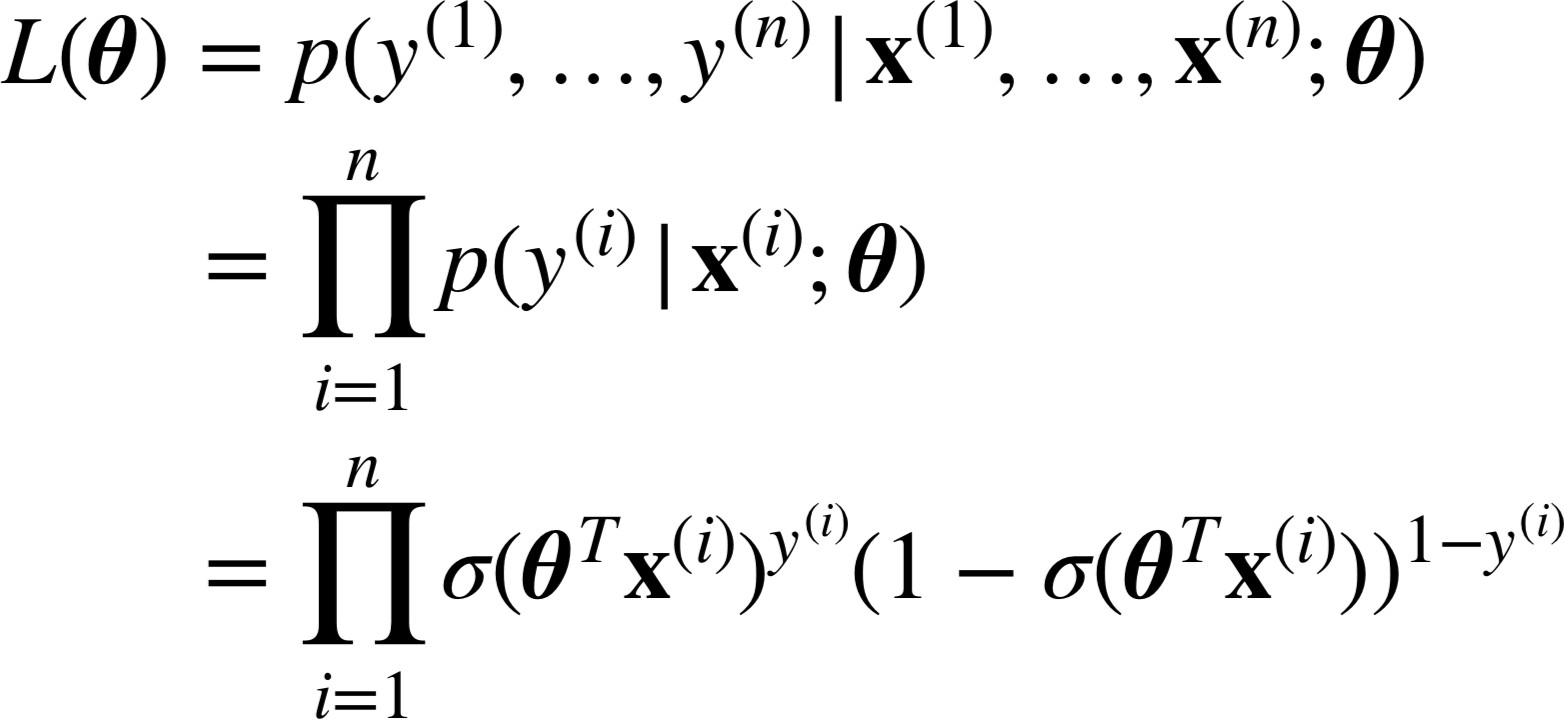


Fig. 2. maximum likelihood estimator function

Knowing the likelihood function for this given problem, we look for such Θ that maximizes the probability of obtaining the data we have. To find the Θ, there is an optimization algorithm to find that Θ.

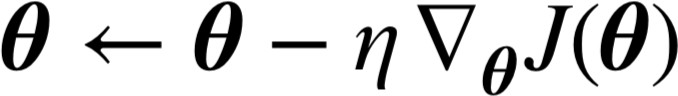


Fig. 3. gradient descend where n is the learning rate

Firstly, we will import the dataset from scikit-learn

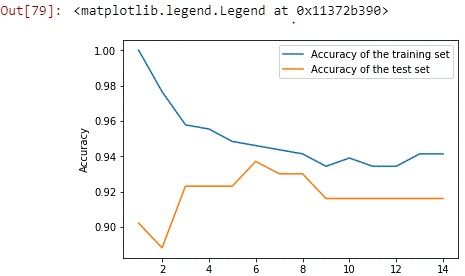
Import the dataset included inside the sklearn.datasets.

In next step we divide our actual training dataset into subsets viz. train and test subset. This is called cross validation. We make 75% of our dataset into train subset and remaining i.e., 25% into test subset which makes train\_size=0.75, we can use any proportions like 0.80 for 80%, 0.90 for 90%. Similarly, we can assign value to test\_size for defining size of test subset. Random\_state used here indicates the random selection data for splitting. If random\_state is not included, then we get different set of train and test dataset every time we split the datasets.

Next is to import the logistic regression classifier in the program from sklearn.linear\_model. To use the classifier, we need to create its instance.

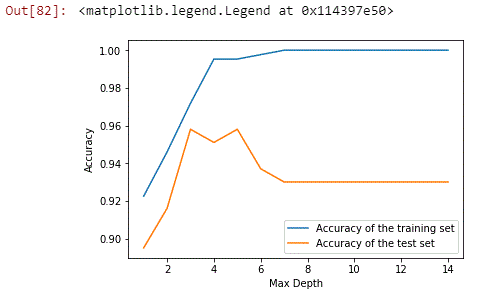
Next is where we test the accuracy of model. The function score(), returns the accuracy of the classifier. It is also predefined function.

By using KNN classifier from sklearn we try to find the different k nearest neighbor from the range of 1 to 15. By looking at plot, best result occurs when n\_neighbors is 6.



Number of neighbors

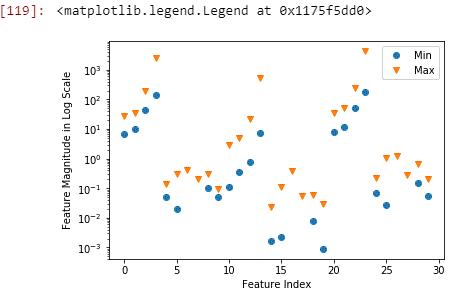
By having larger max\_depth (>5), we overfit the model into training data, so the accuracy for training set increased but the accuracy of test set decreases.



Another technique that was used in SVM (support vector machine)

Support vector machine or SVM is another method for classification of problems such as breast cancer. This works efficiently on datasets that are linearly separable, and SVM attempts to separate the data by a ‘separating hyperplane’. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predict to belong to a category based on which side of the gap they fall.

The methodology for SVM was to setup the data frame and use a train test 40:60 split. Then using the support vector classifier, or SVC model, I went to predict the accuracy of the model.



As we can see that there is huge difference between min and max, the accuracies are as follows

Acc on train dataset: 0.948

Acc on test dataset: 0.951

Now by underfitting to fix it we need to change hyper parameters. And after that accuracies are as follows

Acc to train dataset: 0.988

Acc to test dataset is: 0.972

SVM can work well on high dimensional data with smaller sample size but do not perform well on high dim with lots of samples(>100k).

1. CONCLUSION

This was a great introductory project to the world of machine learning, and I would definitely like to continue to build data to get more practice and experience in the model that I have learned. On the side, this project could extend to doing research and problem solving in other areas aside from the data sets we worked on. I would have loved to do a research project on how students past grades affect future classes, but that is for the future and a different time.