Homework Five Proposal

CMPS242 - Fall 2015

Andrei Ignat Sabina Tomkins Guangyu Wang

Preprocessing

Before we begin with any sort of prediction for the **Titanic** dataset, we must first preprocess the data. For this we want to turn the `sex`, `age`, `sibsp`, `parch`, and `embarked` variables into numerical variables by using some sort of numerical encoding scheme. The `name` and `ticket` variables do not bear any relevant information for our model since each data point is already labeled by a unique id. The price variable should be exciting to work with, it will be interesting to see if we can confirm the assertion that social status(as approximated by ticket price) was an indicator of survival.

Data Exploration

What kinds of passengers have similar characteristics? How does the probability of survival change with gender? We would like to uncover basic relationships between attributes before prediction, and to do so will use unsupervised methods covered in class, such as **clustering**. Understanding the features, which should enable us to do feature selection.

Model

For the purpose of prediction, we would like to build an **Ensemble Learning** model. As we saw in class ensembles can reduce variance and provide stability to predictions. As part of this model, we will include the *Random Forest Classifier* since this is the benchmark used by the authors of the sample script. The goal of the ensemble is to try to find techniques which complement each other. So that one may detect outliers which are lost on the others, while some may be very good and predicting the bulk of the data well. While it may be difficult to ensure that our ensemble is diverse[3] we will attempt to achieve a good ensemble by including as many techniques as possible, and by testing all combinations of classifiers to measure which subsets contribute the most. Furthermore, we would like to integrate the following techniques in the ensemble :

- **Decision Trees -** Though decision trees may not be the most powerful classifiers, they provide some of the most interpretable results, and in addition to this reason, can be useful when combined (such as in the random forest classifier) in an ensemble.
- **Logistic Regression -** This is an all time favorite in ML competitions and for good reason. We want a model which can learn how to predict binary features, representing whether or not the passenger survived. Logistic regression can handle outliers(which are to be expected in a dataset this small) and is interpretable.
- **Support Vector Machines -** As we saw in class Support Vector Machines[2] are powerful classifiers, especially when used with the data-appropriate kernel function. We are excited to gain more exposure to kernel functions through this project, by testing as many as we can and finding the ones which capture the data the best.
- **Neural Networks -** Training a neural network on the dataset also seems to be a promising idea. The features and their potential internal interactions make the dataset suitable for a neural network approach.
- **Deep Neural Networks -** the last method in the ensemble is also a bit of a reach method. Nonetheless, given the recent successes of deep neural networks, it would almost be a shame not to try and see how

they fare against the other methods in the ensemble. While such a method may be unnecessary for the scope of the data, we are eager to try this exciting tool.

In order to combine all the results, we will test a variety of voting schemes, beginning with majority vote, progressing to weighted majority vote and cascading, and also researching more sophisticated schemes[1]. We would also like to try AdaBoost both on it's own and as a method within the ensemble.

References

[1] Roberto Battiti and Anna Maria Colla. Democracy in neural nets: Voting schemes for classification. *Neural Netw.*, 7(4):691–707, April 1994.

[2] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998.

[3] LudmilaI. Kuncheva and ChristopherJ. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.