Wenhan Li

[wenhanli1010@gmail.com](mailto:wenhanli1010@gmail.com)

Alejandro Martinez

School of Mathematical and Statistical Sciences

**Machine Learning Through Cross Validation on NCAA March Madness Results**

**Abstract**

This project attempts to predict the probabilities of match results (noted as a win or lose) in National Collegiate Athletic Association (NCAA) March Madness through simulated machine learning by a ten-fold cross-validation. The program, written in Python, will process and “learn” data from match results of March Madness from past years to predict correlations between data and the probabilities of results for upcoming seasons. The program compares different data analysis models to determine the integrity across models and evaluate their performance.

**Introduction**

With more advanced mechanics, data analysis has become more efficient and competent as driven by the increasing demand for giant data. While some data, specifically numerical data, can be represented by an algorithm, such as the fluctuation of economic exchanges, others cannot. Data like human behavior cannot be represented through an algorithm, but they can be recorded and analyzed, and through machine learning, which uses computation to recognize patterns and regularities in data, can be represented in “useful approximations” (Alpaydin, 2014). Often, machine learning is achieved using cross-validation.

Cross-validation is a statistical method of comparing and validating learning algorithms through the division of data into sections for learning and training. While the field is relatively new, it has grown quickly along with the growth of technology. The simplest form of cross-validation is the *k*-fold cross-validation, which divides the data into *k* sets, as equally as possible, and repeats *k* iterations of training and validating, using *k-1* sets to train and one set for validation each iteration (Refaeilzadeh, 2016). Other forms of cross-validation are also available, such as the “leave one out” method, which repeats *n* iterations of training and validating, where *n* is the cardinality of the set of data, with each iteration trains *n-1* data sets and is validated by only one data. In addition, cross-validation must also address accuracy and validity of models derived, which must not over-fit the data, where the model represents the data but does not function in predictions, and must represent the data.

This project aims to gain a deeper understanding of machine learning through cross-validation by attempting to predict the probability of win/loss of NCAA March Madness games. March Madness is an NCAA basketball league that includes 68 teams in a bracketed single-elimination tournament that is hosted annually around March (Wonderopolis). On a predictive analytics competition website named Kaggle, there is a competition by Google Cloud and NCAA that challenges competitors to 1) build and test models based on past data on win/loss records and 2) predict future outcomes of matchups (Kaggle). The competition has a prize of $50,000 for the past year, but it is closed at the moment. This project is aimed to challenge this competition in order to gain a better understanding of the process behind machine learning.

**Objectives**

1. To gain a better understanding of machine learning through cross-validation.
2. To predict the probability of win/loss of future NCAA matchups through past data.
3. To compare learning models for integrity and validity.

**Hypothesis**

The mean log loss for logistic classification will be minimal.

**Materials and Methods**

A computer, data provided by Kaggle, Python, Pandas, Scikit Learn, and Git will be used for the project. The dataset includes results from past March Madness matchups

The project will test the following models:

1. Logistic Classification
2. Support Vector Machines
3. Linear Discriminant Analysis
4. Decision Trees
5. Naïve Bayes
6. K Nearest Neighbors

**Experimental/Simulations Design and Data Analysis**

The project focuses on cross-validation to score different workflows. Workflows will take the form of

*Coder -> imputer -> filter -> feature engineering -> feature selector -> model training -> cross-validation score*

This is to ensure the separation of different jobs in creating an effective model, and the sections in the workflow will run through various combinations.

The models will be trained and the data and scores will be presented in different forms of visualization through codes written in Python. Pandas, a Python library for data analysis tools, will be used to visualize and manipulate the data. Scikit Learn, a Python machine learning library, will be used to train the models and compare them using cross-validation.

**Expected Results**

We expect to find the best workflows and evaluation them with the 2018 data, as well as the data that will be generated next year. We expect the logistic classification model to perform the best.

**Works Cited**

Alpaydin, Ethem. Introduction to Machine Learning. MIT Press, 2014.

Git, git-scm.com/.

“Google Cloud & NCAA® ML Competition 2018-Men's | Kaggle.” Kaggle, [www.kaggle.com/c/mens-machine-learning-competition-2018](http://www.kaggle.com/c/mens-machine-learning-competition-2018).

“Python Data Analysis Library¶.” *Pandas: Powerful Python Data Analysis Toolkit - Pandas 0.22.0 Documentation*, pandas.pydata.org/.

Refaeilzadeh P., Tang L., Liu H. (2016) Cross-Validation. In: Liu L., Özsu M. (eds) Encyclopedia of Database Systems. Springer, New York, NY

“Scikit-Learn.” *1.4. Support Vector Machines - Scikit-Learn 0.19.1 Documentation*, scikit-learn.org/stable/index.html.

“What Is March Madness?” Wonderopolis, wonderopolis.org/wonder/what-is-march-madness.