Final Project Proposal

Instructor: Prof. Cho-Jui Hsieh

Department of Statistics, University of California, Davis

2018 年 5 月 7 日

1 Members

Names	Email Address
Wangqian Miao	wqmiao@ucdavis.edu
Mingyi Xue	myxue@ucdavis.edu
Rui Wang	ruiang@ucdavis.edu

2 Project

Project Infomation

Topic	Kobe Bryant Shot Selection
Input Format	tabular data
Output	binary classification
Tools	pandas, sklearn
Algorithms	logistic regression, SVM, neural networks

1. The **dataset** is from `https://www.kaggle.com/c/kobe-bryant-shot-selection/data` .
2. Our **goal** is to perform varied classification algotithms mentioned above to predict which shots Kobe sank, comparing the efficiency and accuracy of these methods.
3. One **difficulty** we will confront is *feature engnieering*. Because this dataset involves 25 explanatory variables, types of which contain numeric, categoric and datetime, we are supposed to deal with different types of feature. Firstly, after we transform categorical variables to dummy variables, what to do if there are too many variables after this transformation. Secondly, how to deal with datetime variables, whether to treat them as categorical variables or not.
4. Another **difficulty** is memory capacity. Since the training dataset has more than 30 thousand samples with 25 features each, will personal computer consumes excessive time running classification algorithms on the dataset?

3 Reference

1. `https://dnc1994.com/2016/04/rank-10-percent-in-first-kaggle-competition/`
2. `https://www.zhihu.com/question/23987009`
3. `http://www.cnblogs.com/jasonfreak/p/5448385.html`
4. 机器学习 周志华