Title

—Kaggle 201X

Yuan, Jumao

04/28/2015

Contents

1 Abstract

2 Introduction

3 Data Description

4 Evaluation

Entries will be evaluated using the area under the **receiver operator curve** (AUC). AUC was first used by the American army after the attack on Pearl Harbour, to detect Japanese aircraft from radar signals. Today, it is a commonly used evaluation method for binary choose problems, which involve classifying an instance as either positive or negative. Its main advantages over other evaluation methods, such as the simpler misclassification error, are:
1. It's insensitive to unbalanced datasets (datasets that have more installeds than not-installeds or vice versa).
2. For other evaluation methods, a user has to choose a cut-off point above which the target variable is part of the positive class (e.g. a logistic regression model returns any real number between 0 and 1 - the modeler might decide that predictions greater than 0.5 mean a positive class prediction while a prediction of less than 0.5 mean a negative class prediction). AUC evaluates entries at all cut-off points, giving better insight into how well the classifier is able to separate the two classes.

Understanding AUC

To understand the calculation of AUC, a few basic concepts must be introduced. For a binary choice prediction, there are four possible outcomes:

- true positive - a positive instance that is correctly classified as positive;
- false positive - a negative instance that is incorrectly classified as positive;
- true negative - a negative instance that is correctly classified as negative;
- false negative - a positive instance that is incorrectly classified as negative);

These possibilities can be neatly displayed in a confusion matrix:

	P	N
P	true positive	false positive
N	false positive	true positive

The true positive rate, or recall, is calculated as the number of true positives divided by the total number of positives. When identifying aircraft from radar signals, it is proportion that are correctly identified. The false positive rate is calculated as the number of false positives divided by the total number of negatives. When identifying aircraft from radar signals, it is the rate of false alarms.

5 Data Preprocessing

5.1 Missing values and typos

Usually, we use data imputation to make up missing values; while, in this problem, not many missing values adn typos exist, so we can just remove them.

5.2 Remove redundant variables

In the original dataset, we can see variables “P8”, “V7” and “V9” are redundant. Therefore, we can delete these 3 columns in the data preprocessing.

5.3 Split into training and testing datasets

We randomly split data into 70% test datasets and 30% training datasets and repeate all models with a few iterations.

6 Build Moldes

R code can be accessible on my github repository `https://github.com/jyuan4/Final_Project`.

6.1 GLM

6.2 Logistic Regression

6.3 3

6.4 4

6.5 5

6.6 6

7 Result

8 Discussion

9 References

[1] “An Introduction to Statistical Learning” by James, G., Witten, D., Hastie, T., Tibshirani, R.

[2]