Executive Summary of Commercial Analysis

Bingxi Li

May 2016

Abstract

The k-nearest neighbor, logistic regression and Support Vector Machine (SVM) algorithm with kernel method are tuned to predict customer behavior. The performance of these models are compared in terms of accuracy and receiver operating characteristic (roc). The SVM using radial basis function kernel gives most reliable prediction.

1 Introduction

A robust model should be trained based on the training data from a company to predict if a customer will buy the product in the future or not. The training data set has **4** features and **3089** pieces of records. The predictor will be either **1** or **0** given a new **4** dimension feature.

2 Methodology and Result

a preprocessing

The features are standardized before training. 30% of the records are hold out as **test set** while the rest records are use as **train set** to tune knn and logistic model and SVM model. The latter two of the three models are tuned with different kernel through 5-fold cross validation. The best model are selected among these tuned ones based on their predictabilities to the same test set. The details about implementation are included in the attached “training.py”.

b K-Nearest Neighbor

Ranging from 1 to 24 stepped by 1, the number of nearest neighbors, K, at **1** yields highest accuracy. The KNN model with K = 1.0 is then trained with whole **train set** to predict **test set**. The accuracy is **0.95361**.

c logistic regression

Ranging from 0.5 to 50 with step size being 0.5, the inverse of regularization strength, C, at **7.0** yields highest mean score through 5-fold cross-validation. The Logistic regression model with C = 1.0 is then trained with whole **train set** and to predict **test set**. The following is summary of precision, recall, F1 score for each class.

```
              precision    recall  f1-score   support
        0.0     0.93354   0.92767   0.93060       318
        1.0     0.96236   0.96552   0.96393       609
avg / total     0.95247   0.95254   0.95250       927
```

The accuracy given by the best logistic model is **0.95254**. The quality of this model is also determined with receiver operating characteristic (ROC) curve and precision-recall (PR) curve in Figure 1 (left). The area under curve (AUC) is 0.98941, which indicates very high probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.

d support vector machine

d.1 with linear kernel

Ranging from 0.1 to 5.0 with step size being 0.1, penalty parameter C of the error term at **1.1** and yield highest mean score through 5-fold cross-validation. The linear SVM model with C = 1.1 is then trained with whole **train set** and to predict **test set**. The following is summary of precision, recall, F1 score for each class.

```
              precision    recall  f1-score   support
        0.0     0.94654   0.94654   0.94654       318
        1.0     0.97209   0.97209   0.97209       609
avg / total     0.96332   0.96332   0.96332       927
```

The accuracy given by the best linear SVM model is **0.96332**. The quality of this model is also determined with ROC and PR curve in Figure 2. The AUC is **0.99034**, which indicates the reliable prediction of this model.

d.2 with radial basis function(rbf) kernel

Separately ranging from 0.1 to 5.0 stepped by 0.1 and from 0.1 to 1.9 stepped by 0.2, penalty parameter C of the error term at **4.7** and kenel coefficient γ at **1.1** yield highest mean score through 5-fold cross-validation. The rbf SVM model with C = 4.7 and $\gamma = 1.1$ is then trained with whole **train set** and to predict **test set**. The following is summary of precision, recall, F1 score for each class.

```
              precision    recall  f1-score   support
        0.0     0.97134   0.95912   0.96519       318
        1.0     0.97879   0.98522   0.98200       609
avg / total     0.97624   0.97627   0.97623       927
```

The accuracy given by the best rbf SVM model is **0.97627**. The quality of this model is also determined with ROC and PR curve in Figure 3. The AUC is **0.99592** under ROC curve, which indicates the perfect predictability of the model.

d.3 with polynomial kernel

Separately ranging from 0.1 to 5.0 stepped by 0.1 and from 1 to 4 stepped by 1, penalty parameter C of the error term at **3.7** and kenel coefficient γ at **3** yield highest mean score through 5-fold cross-validation. The polynomial SVM model with C = 3.7 and $\gamma = 3$ is then trained with whole **train set** and to predict **test set**. The following is summary of precision, recall, F1 score for each class.

```
              precision    recall  f1-score   support
        0.0     0.98039   0.94340   0.96154       318
        1.0     0.97101   0.99015   0.98049       609
avg / total     0.97423   0.97411   0.97399       927
```

The accuracy given by the best polynomial SVM model is **0.97411**. The quality of this model is also determined with ROC and PR curve in Figure 4. The AUC is **0.99563** under ROC curve, which indicates that SVM with polynomial kernel is also a perfect predict model.

d.4 with sigmoid kernel

Separately ranging from 0.1 to 5.0 stepped by 0.1 and from 1 to 4 stepped by 1, penalty parameter C of the error term at **0.2** and kenel coefficient γ at **0.1** yield highest mean score through 5-fold cross-validation. The sigmoid SVM model with C = 0.1 and $\gamma = 0.2$ is then trained with whole **train set** and to predict **test set**. The following is summary of precision, recall, F1 score for each class.

```
              precision    recall  f1-score   support
        0.0     0.88554   0.92453   0.90462       318
        1.0     0.95966   0.93760   0.94850       609
avg / total     0.93424   0.93312   0.93345       927
```

The accuracy given by the best polynomial SVM model is **0.93312**. The quality of this model is also determined with ROC and PR curve in Figure 5. The AUC is **0.98352** under ROC curve, which indicates that SVM with polynomial kernel is also a perfect predict model.

3 Conclusion

Through this study, the SVM with polynomial kernel and radial basis function(rbf) kernel have similarly perfect prediction. SVM with linear kernel is also good. The KNN and logistic seems less favorable here while sigmoid SVM is probably not a good choice for this prediction.

The SVM using rbf kernel is implemented in **testing.py** with the optimal parameters given above, C = 4.7 and $\gamma = 1.1$. For future prediction of any new 4-feature data, this model is trained with the data from exam.dat.txt. The attached “testing.py” implemented the above idea using a vector like “F1 F2 F3 F4” or a test file like “test.dat.txt” as input. When used a test file as the input, the file should be in the same format of “exam.dat.txt”. Output will be an either 1 or 0 predictor or an array of either 1 or 0 predictors.

```
# use a four feature vector as input
python testing.py F1 F2 F3 F4
#[1.]

# use a test data file as input
python testing.py ”test.dat.txt”
#[1. 1. 0. ....]
```

4 Reference

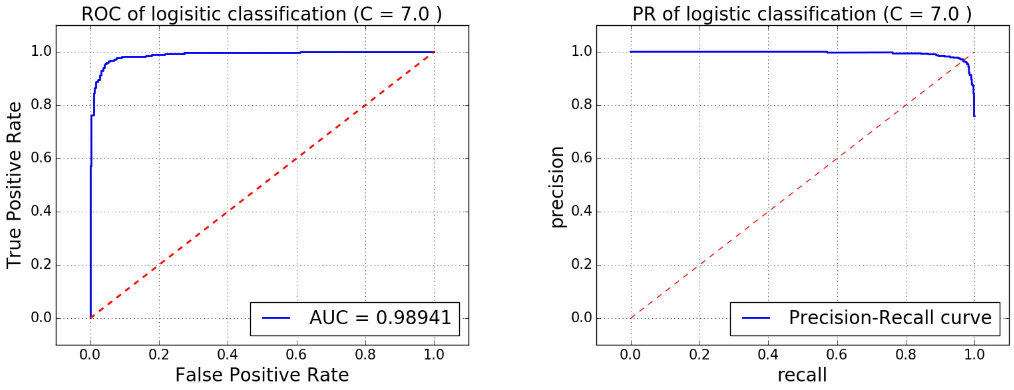

Figure 1: ROC curve(left) and PR curve(right) by logistic prediction

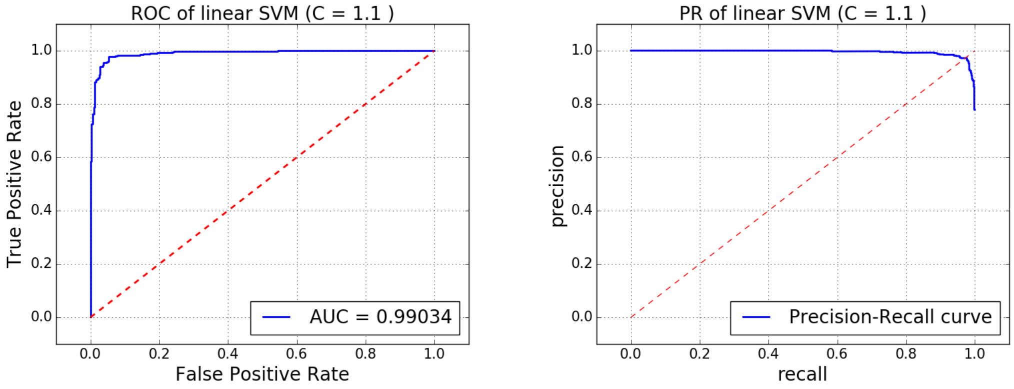

Figure 2: ROC (left) and PR curve(right) by linear SVM prediction

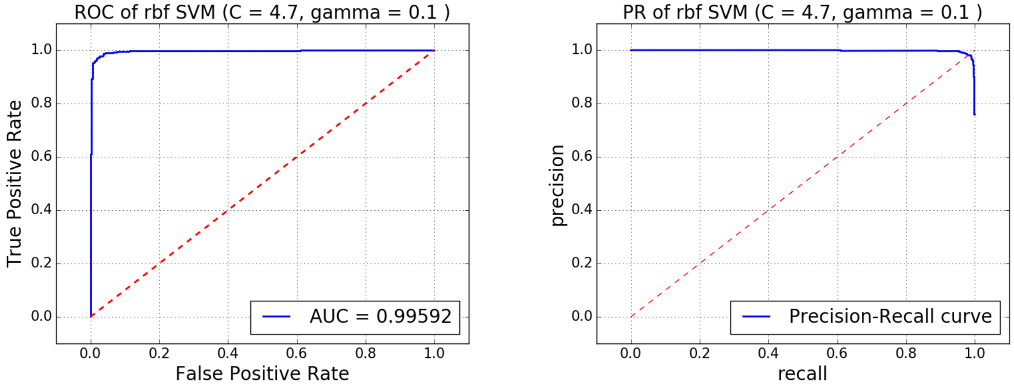

Figure 3: ROC (left) and PR curve(right) by rbf SVM prediction

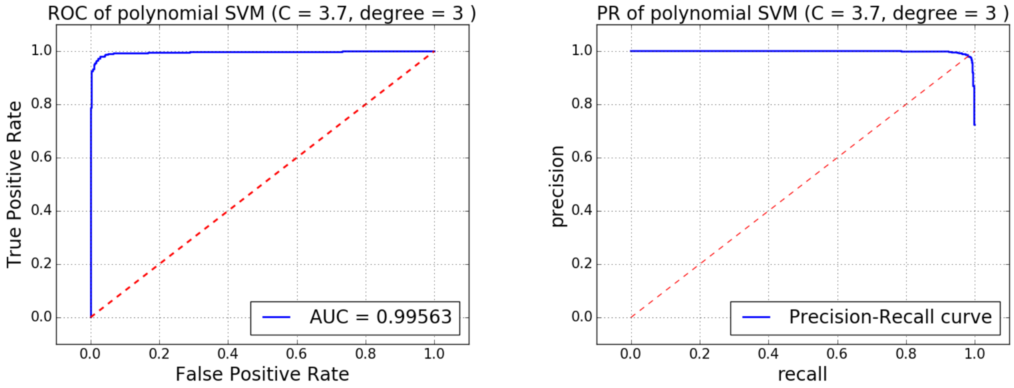

Figure 4: ROC (left) and PR curve(right) by polynomial SVM prediction

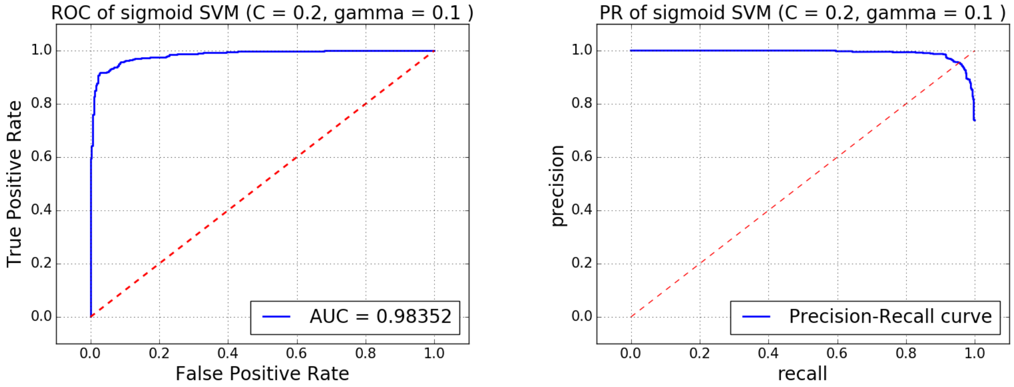

Figure 5: ROC (left) and PR curve(right) by sigmoid SVM prediction