DataScience06_QuestionsOneThruFive

Jim Stearns for UW PCE Data Science 1, Fall 2014

22 November 2014

1. Why are performance metrics better on training data than on test data?

Measuring results on training data is analogous to playing poker with all cards-up. More data –> better predictions, for that data set.

Measuring performance on a test set that has not been part of training is like adding a "blind" to an experiment in order to negate observer bias.

2. How do you determine which data are training data and which are test data?

By placing the "determination" into hands other than the modeler, typically to a random number generator.

3. Confusion Matrix Metrics

(Beware, this problem contains irrelevant data while some important numbers are not explicitly presented.)

A model was trained on 300 individuals where 149 had the cold and 151 were healthy. The model was tested on 100 individuals where 10 were ill. The model correctly predicted that 85 of the healthy individuals were indeed healthy and correctly predicted that 7 of the ill individuals were indeed ill. The other predictions were incorrect.

Consult Wikipedia: http://en.wikipedia.org/wiki/Precision_and_recall

Construct a confusion matrix

Interpretation: Construct a confusion matrix of the **test** dataset results.

I choose to use R to construct the confusion matrix by specifying actual and predicted results as two factor vectors. Convention (arbitrary): 'ill' is positive result, 'healthy' is negative result.

Actual results: 10 'ill', 90 'healthy':

```
actual=factor(c(rep('ill',10),
                rep('healthy',90)),
              levels=c('ill','healthy'))
```

Predicted results: of the 10 actually ill, 7 were predicted to be ill (TP) and 3 were predicted to be healthy (FN). Of the 90 actually healthy, 85 were predicted to be healthy (TN) and 5 were predicted to be ill (FP):

```
predicted=factor(c(rep('ill',7), rep('healthy',3),
                   rep("healthy",85), rep('ill',5)),
                 levels=c('ill','healthy'))
```

Conventions for display of confusion matrix:

- actual 'ill'/'healthy' are columns
- predicted 'ill'/'healthy' are rows

- TP is upper left cell, TN is lower right.

```
confusionMatrix <- table(predicted, actual)
confusionMatrix
```

```
##          actual
## predicted ill healthy
##   ill       7       5
##   healthy   3      85
```

```
# Set up variables for calculating sensitivity etc.
TP=confusionMatrix[1,1]
FP=confusionMatrix[1,2]
FN=confusionMatrix[2,1]
TN=confusionMatrix[2,2]
N=100
```

Then calculate the following:

a. Sensitivity (aka True Positive Rate, Recall)

Sensitivity = TP/(TP+FN) = 0.70

b. Specificity (aka True Negative Rate)

Specificity = TN/(TN+FP) = 0.94

c. Accuracy

Overall, how often is the classifier correct?

(TP+TN)/N = 0.92

d. Precision (aka Positive Predictive Value)

When yes is predicted, how often is it correct?

"[H]igh precision means that an algorithm returned substantially more relevant results than irrelevant" Wiki-Precision and recall

TP/(TP+FP) = 0.58

e. Recall (aka Sensitivity, True Positive Rate)

"[H]igh recall means that an algorithm returned most of the relevant results." Wiki-Precision and recall

Sensitivity = TP/(TP+FN) = 0.70

4. The probability threshold for a classification varies in an ROC chart from 0 to 1.

a. What point of the graph corresponds to a threshold of zero?

(1, 1) A threshold of zero means that the model classifies (predicts) all observations as true. The results are either TP or FP - no TN or FN.

FPR is 1 (FPR == FP/(FP+TN(0)) == FP/FP == 1). TPR is 1 (TPR == TP/(TP+FN(0)) == TP/TP == 1).

b. What point of the graph corresponds to a threshold of one?

(0,0) A threshold of one means that the model classifies (predicts) all observations as false. The results are either TN or FN - no TP or FP.

FPR is 0: The numerator - FP - is zero. TPR is also 0: The numerator - TP is zero.

c. What point of the graph corresponds to a threshold of 0.5? (trick question)

It depends on the model and the data. All that can be said about the location is that it's a point on the curve.

5. Confusion Matrix/ROC

A Classification is tested on 1000 cases. In the middle of its ROC chart, where the false positive rate is 0.4, the true positive rate is 0.8. The accuracy is 0.7.

a. What does the confusion matrix look like?

Given:

- T + F = 1000
- A point on the curve: (FPR=0.4, TPR=0.8).
- At the same point accuracy is 0.7.

```
ACC = (TP + TN) / (T + F) ==> 0.7 = (TP + TN) / 1000 ==>    TP + TN = 700
                                                           FP + FN = 300

TPR = TP / P = TP / (TP + FN) ==> TP = 0.8 (TP + FN) ==>   TP = 4 FN

FPR = FP / N = FP / (TN + FP) ==> FP = 0.4 (TN + FP) ==>   TN = 1.5 FP
```

Four equations with four unknowns yields this confusion matrix:

```
                Actual
                T       F
Pred    T       TP=400  FP=200
        F       FN=100  TN=300
```

b. What can you say about the probability threshold at that point? (trick question)

Since we have negatives predicted, the threshold isn't 0.

Since we have positives predicted, the threshold isn't 1.

There are twice as many false positives as false negatives. This ratio hints at threshold is shaded more towards zero than to one. Such a shading may make sense for some applications (e.g. a medical diagnostic).