Machine Learning

Yuh-Jye Lee

Lab of Data Science and Machine Intelligence
Dept. of Applied Math. at NCTU

March 14, 2017

How to Evaluate What's been Learned
Cost is Not Sensitive and *Bainary Classification*

- Measure the performance of a classifier in terms of error rate or accuracy

$$\textit{Error rate} = \frac{\textit{Number of misclassified point}}{\textit{Total number of data point}}$$

Main Goal: Predict the unseen class label for new data

- We have to asses a classifier's error rate on a set that play no rule in the learning class
- Split the data instances in hand into two parts:
  1. Training set: for learning the classifier.
  2. Testing set: for evaluating the classifier.

How to Evaluate What's been Learned
Cost is Not Sensitive and *Multi-class Classification* with One-vs.-Rest

- Generalize the performance evaluation in term of *error* to *multi-class classification*
  1. Micro average: Calculate the performance from each *individual*

$$\color{red}{Error_{micro}} = \frac{Number\ of\ misclassified\ point}{Total\ number\ of\ data\ point}$$

  2. Macro average:

$$\color{red}{Error_{macro}} = \frac{Sum\ of\ error\ from\ each\ class}{k}$$

Note: We decompose a k categories classification problem into a series of k *binary* classification problems.

How to Evaluate What's been Learned
Regression Problems

- Measure the performance of a classifier in terms of error
  1. MSE: Mean Squares Error

$$MSE = \frac{\sum_{1=1}^{\ell}(f(\mathbf{x^i}) - \mathbf{y_i})^2}{\ell}$$

  2. RMSE: Root of Mean Squares Error

$$RMSE = \sqrt{\frac{\sum_{1=1}^{\ell}(f(\mathbf{x^i}) - \mathbf{y_i})^2}{\ell}}$$

  3. MAE: Mean Absolute Error

$$MSE = \frac{\sum_{1=1}^{\ell}|f(\mathbf{x^i}) - \mathbf{y_i}|}{\ell}$$

k-fold Stratified Cross Validation
Maximize the Usage of the Data in Hands

- Split the dataset into k approximately equal partitions.
- Each in turn is used for testing while the remainder is used for training.
- The labels $(+/-)$ in the training and testing sets should be in about right proportion.
  - Doing the random splitting in the positive class and negative class respectively will guarantee it.
  - This procedure is called stratification.
- Leave-one-out cross-validation if $k = \#$ of data point.
  - No random sampling is involved but nonstratified.

How to Compare Two Classifiers? Testing Hypothesis: Paired t-test

- We compare two learning algorithms by comparing the average error rate over several cross-validations.
- Assume that the same cross-validation splits can be used for both methods:

$$H_0 : \bar{d} = 0 \ vs. \ H_1 : \bar{d} \neq 0$$

where $\bar{d} = \frac{1}{k} \sum_{i=1}^{k} d_i$ and $d_i = x_i - y_i$

- The t-statistic:

$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2 / k}}$$

How to Evaluate What's been Learned? When Cost is Sensitive

- Two types error will occur: False Positive(FP) & False Negative(FN)
- For binary classification problem, the results can be summarized in a 2×2 confusion matrix.

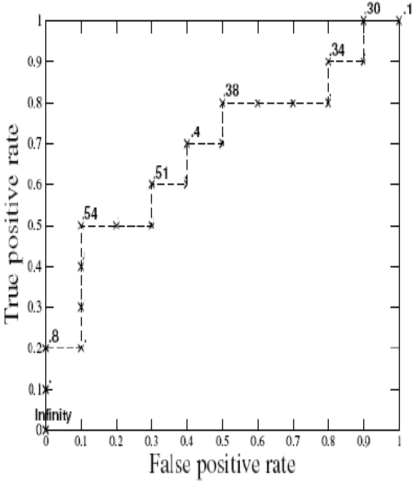
	Predicted Class	
Actual Class	True Pos. (TP)	False Neg. (FN)
	False Pos. (FP)	True Neg. (FN)

Note:The *confusion matrix* can be extended to multi-class classification problem

ROC Curve
Receiver Operating Characteristic Curve

- An evaluation method for learning models.
- What it concerns about is the Ranking of instances made by the learning model.
- A Ranking means that we sort the instances w.r.t the probability of being a positive instance from high to low.
- ROC curve plots the true positive rate (TPr) as a function of the false positive rate (FPr).

An Example of ROC Curve



Inst ID	Class	Score
1	P	0.51
2	P	0.8
3	P	0.3
4	P	0.55
5	P	0.4
6	P	0.34
7	P	0.9
8	P	0.54
9	P	0.38
10	P	0.6
11	N	0.35
12	N	0.52
13	N	0.36
14	N	0.37
15	N	0.7
16	N	0.1
17	N	0.39
18	N	0.53
19	N	0.5
20	N	0.33

Sort →

Inst ID	Class	Score
7	P	0.9
2	P	0.8
15	N	0.7
10	P	0.6
4	P	0.55
8	P	0.54
18	N	0.53
12	N	0.52
1	P	0.51
19	N	0.5
5	P	0.4
17	N	0.39
9	P	0.38
14	N	0.37
13	N	0.36
11	N	0.35
6	P	0.34
20	N	0.33
3	P	0.3
16	N	0.1

Using ROC to Compare Two Methods

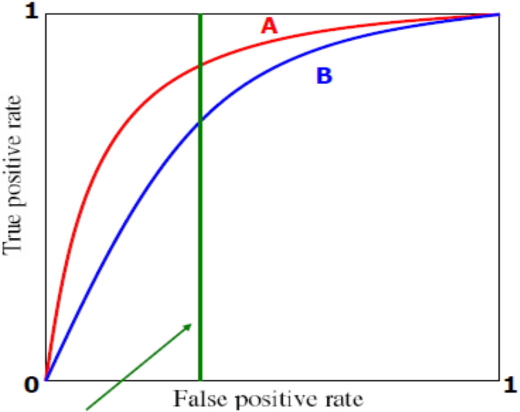
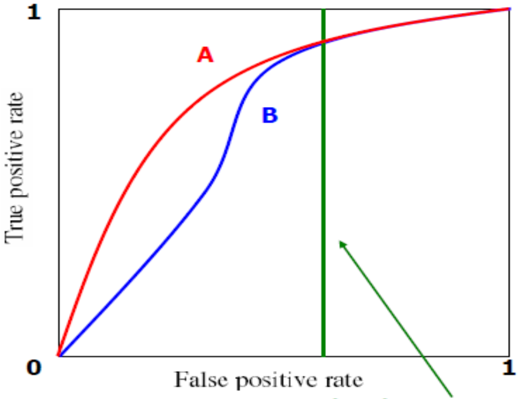


Figure: Under the same FP rate, method A is better than B.

What if there is a Tie?



Which one is better?

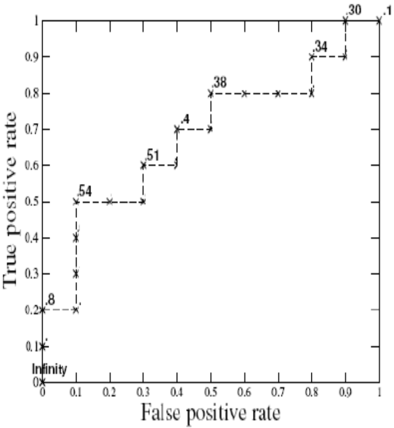
Area under the Curve (AUC)

- An index of ROC curve with range from 0 to 1.
- An AUC value of 1 corresponds to a perfect Ranking (all positive instances are ranked high than all negative instance).
- A simple formula for calculating AUC:

$$AUC = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{f(x_i) > f(x_j)\}}}{m \times n}$$

where m: number of positive instances.
n: number of negative instances.

An Example of ROC Curve



Inst ID	Class	Score
1	P	0.51
2	P	0.8
3	P	0.3
4	P	0.55
5	P	0.4
6	P	0.34
7	P	0.9
8	P	0.54
9	P	0.38
10	P	0.6
11	N	0.35
12	N	0.52
13	N	0.36
14	N	0.37
15	N	0.7
16	N	0.1
17	N	0.39
18	N	0.53
19	N	0.5
20	N	0.33

Sort →

Inst ID	Class	Score
7	P	0.9
2	P	0.8
15	N	0.7
10	P	0.6
4	P	0.55
8	P	0.54
18	N	0.53
12	N	0.52
1	P	0.51
19	N	0.5
5	P	0.4
17	N	0.39
9	P	0.38
14	N	0.37
13	N	0.36
11	N	0.35
6	P	0.34
20	N	0.33
3	P	0.3
16	N	0.1

Performance Measures in Information Retrieval (IR)

- An IR system, such as Google, for given a query (keywords search) will try to retrieve all relevant documents in a corpus.
  - Documents returned that are NOT relevant: FP.
  - The relevant documents that are NOT return: FN.
- Performance measures in IR, Recall & Precision.

$$Recall = \frac{TP}{TP + FN}$$

and

$$Precision = \frac{TP}{TP + FP}$$

Balance the Trade-off between Recall and Precision

- Two extreme cases:
  1. Return only document with 100% confidence then precision=1 but recall will be very small.
  2. Return all documents in the corpus then recall=1 but precision will be very small.
- F-measure balances this trade-off:

$$F - measure = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$