Performance Measures

Performance Measures

- The accuracy of a classification method is the ability of the method to correctly determine the class of a randomly selected data instance.
- The most obvious criterion to use for estimating the performance of a classifier is *predictive accuracy*.
- Error rate = (T-C)/T
  – where T is total objects in test data, C objects are correctly classified our of T objects.

Performance Measures

- A more difficult trade-off occurs when the classes are *severely unbalanced*. Suppose we are considering investing in one of the leading companies quoted on a certain stock market.
- Can we predict which companies will become bankrupt by the next two years (so we can avoid investing in them)?

Performance Measures

- The proportion of such companies is obviously small, lets say 0.02, so on average out of every 100 companies 2 will become bankrupt.
- Call these “bad” and “good” companies.
- If we have a very *trusting* classifier that always predicts “good” under all circumstances its predictive accuracy will be 98 %, a very high value.
- Looked at only in terms of predictive accuracy this is a very successful classifier.

Performance Measures

- BUT, it will give us no help at all in avoiding investing in "bad" companies.
- Alternatively, if we want to be very safe we could use a very "cautious" classifier that always predicted "bad".
- Though, we would never loose our money in a bankrupt company BUT would never invest in a good one either.
- It is clear from this example that predictive accuracy on its own is not a reliable indicator if classes are severely unbalanced.

Performance Measures

- A “confusion matrix” is sometimes used to represent the result of testing in more detail.
- The advantage of using this matrix is that it not only tells us how many got misclassified but also what misclassifications occurred.
- When there are two classes, positive (+) and negative (-), the confusion matrix consists of four cells, i.e., TP, FP, FN and TN.

Performance Measures

		Predicted Class	
		+	-
Actual Class	+	**TP**	**FN**
	-	**FP**	**TN**

TP: True Positive. The number of positive instances that are classified as positive.
FP: False Positive. The number of negative instances that are classified as positive.
FN: False Negative. The number of positive instances that are classified as negative.
TN: True Negative. The number of negative instances that are classified as negative.

Performance Measures

- In our “bad” company problem we would like the number of false positives to be as small as possible, ideally zero.
- We would probably be willing to accept a high proportion of false negatives since there are large number of possible companies to invest in.

‘False Positives’ are Bad

- Here we would like the number of false positives to be fairly small.

- We would probably be willing to accept a high proportion of false negatives.

Performance Measures

- Medical Screening Application. Its not feasible to screen the entire population for a condition that occurs only rarely e.g. brain tumor.
- Instead doctor uses his/her experience to judge which patients are most likely to be suffering from a brain tumor and sends them to a hospital for screening.

‘False Negatives’ are Bad

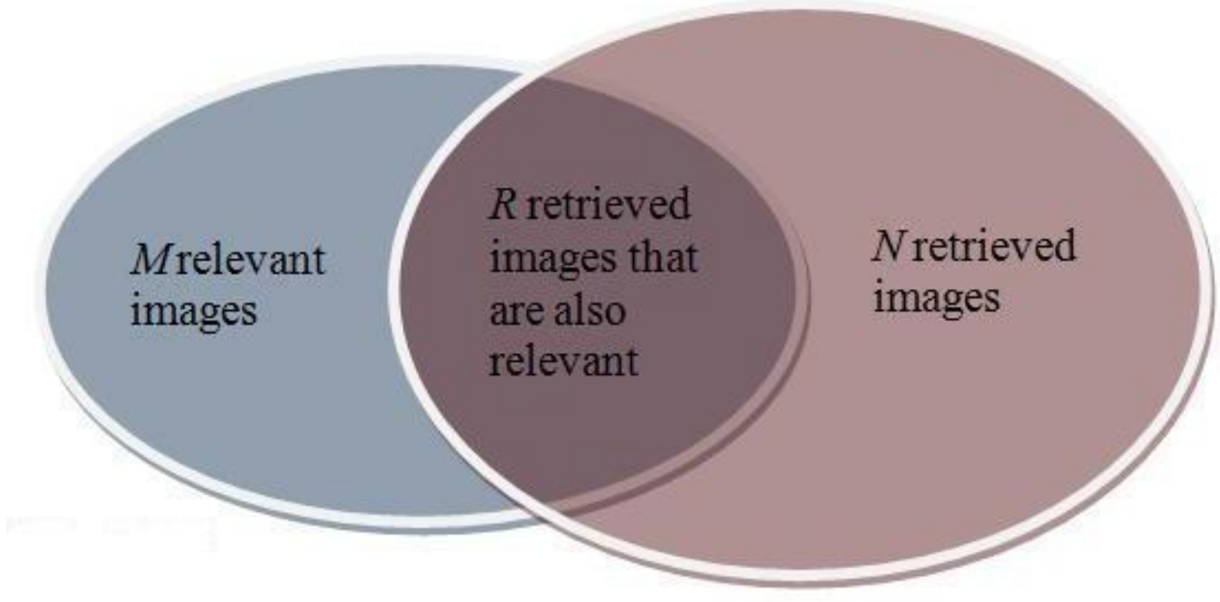
- For this application we might be willing to accept quite a high proportion of false positives e.g. 90% i.e. 1/10 patients screened has a brain tumor or even higher.
- However we would like the proportion of false negatives to be as small as possible.

So It Depends

- A web search engine can be looked at as a kind of classifier.
- Given a specification, it effectively classifies all pages on the web that are known to it as either “relevant” or “not relevant”.
- Here we may be willing to accept a high proportion of false negatives e.g. 30% or more, but probably do not want too many false positives e.g. 10% or less.
- Recall and Precision (IR students !!!)

Performance Measures

- Recall: It is the fraction of relevant instances that are retrieved. R/M
- Precision: It is fraction of retrieved instances that are relevant. R/N



Performance Measures

- These examples illustrate that, leaving aside the ideal of perfect classification accuracy, there is no single combination of FP and FN that is ideal for every application.

Performance Measures

		Predicted Class		
		A	B	C
Actual Class	A	8	2	0
	B	1	9	0
	C	1	2	7

- Consider Class A. There are 10 objects that belong to this class and 20 that don't. Out of 10, only 8 are classified correctly.
- In total 24 objects are classified correctly.
- Class A: TP=8, TN=18, FN=2, FP=2.
- Class B: TP=9, TN=16, FN=1, FP=4.
- Class C: TP=7, TN=20, FN=3, FP=0.

Performance Measures

- Sensitivity = TP/(TP+FN) = 24/30 = 80%
  - It specifies the proportion of positive instances that are correctly classified as positive.

- Specificity = TN/(TN+FP) = 54/60=90%
  - It specifies the proportion of negative instances that are correctly classified as negative.

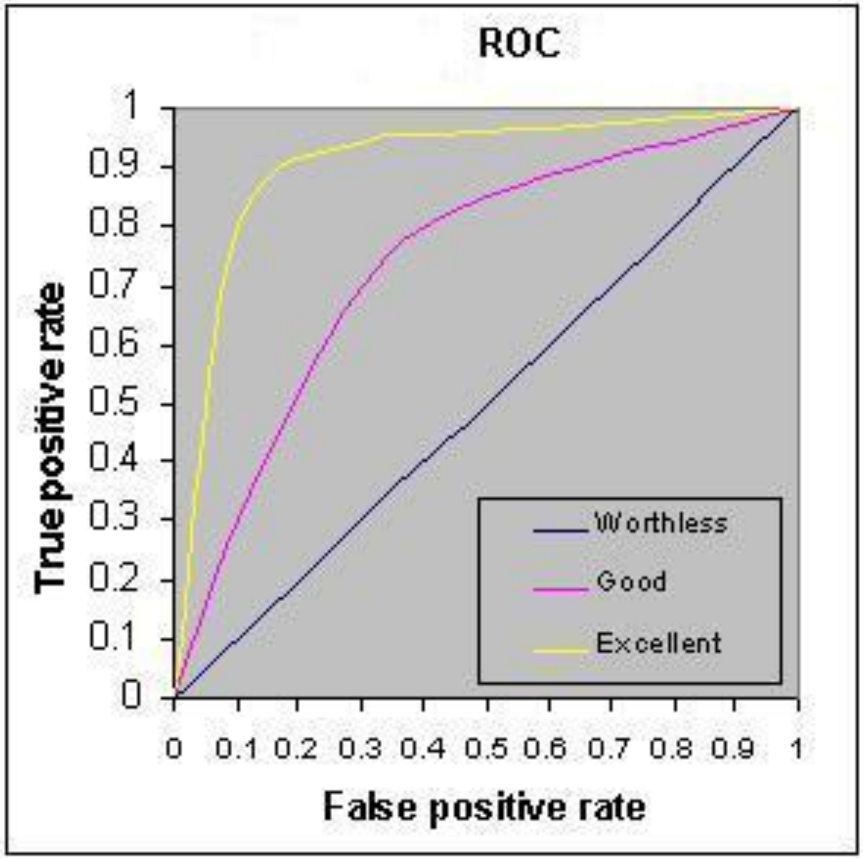
Receiver Operating Characteristics Graph

- The TP Rate and FP Rate values of different classifiers on the same test set are often represented diagrammatically by ROC Graph.
- The value of FP Rate is plotted on the horizontal axis, with TP Rate plotted on the vertical axis.
- If all the classifiers are good ones, all the points on the ROC Graph are likely to be around the top left hand corner.

ROC Graph

- One classifier is better than another if its corresponding point on the ROC Graph is to the ‘north-west’

  of the other’s.



Estimating accuracy of a model

- Holdout Method: Requires a test set and training set, both are mutually exclusive.
- Random sub-sampling Method: It is much like holdout method except it doesn't rely on a single test set. Essentially, the holdout method is repeated several times and the accuracy estimate is obtained by computing the mean of the several trails.

Estimating accuracy of a model

- K-fold Cross Validation Method: In this method the available data is randomly divided into k disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining k-1 sets are used for building the classifier. The test set is then used to estimate the accuracy. This is done repeatedly k times so that each subset is used as a test subset once. Then mean is calculated of all the k estimates.

Estimating accuracy of a model

- N-Fold Cross Validation: It is an extreme case of k-fold cross-validation, often known as ‘leave-one-out’.
- Where the dataset is divided into as many parts as there are instances, each instance effectively forming a test set of one.

Other Evaluation Criteria

- Speed: It is not just the time or computation cost of constructing a model it also includes the time required to learn to use the model.
- Robustness: Data errors are common, in particular when data is being collected from a number of sources and errors may remain even after data cleaning. It is therefore desirable that a method be able to produce good results in spite of some errors and missing values in datasets.

Other Evaluation Criteria

- Scalability: Many data mining methods were originally designed for small datasets. Given that large datasets are becoming common, it is desirable that a method continues to work efficiently for large disk-resident databases as well.
- Goodness of the Model: For a model to be effective, it needs to fit the problem that is being solved. For example in a Decision Tree classification, it is desirable to find a decision tree of the "right" size and compactness with high accuracy.

Other Evaluation Criteria

- Interpretability: An important task of a data mining professional is to ensure that the results of data mining are explained to the decision makers. It is therefore desirable that the end-user be able to understand and gain insight from the results produced by the classification method.

Estimating accuracy of a model

- Holdout Method: Requires a test set and training set, both are mutually exclusive.
- Random sub-sampling Method: It is much like holdout method except it doesn't rely on a single test set. Essentially, the holdout method is repeated several times and the accuracy estimate is obtained by computing the mean of the several trails.

Estimating accuracy of a model

- K-fold Cross Validation Method: In this method the available data is randomly divided into k disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining k-1 sets are used for building the classifier. The test set is then used to estimate the accuracy. This is done repeatedly k times so that each subset is used as a test subset once. Then mean is calculated of all the k estimates.

Estimating accuracy of a model

- <u>N-fold Cross Validation Method</u>: It is an extreme case of k-fold method, also called “leave-one-out”. Where the dataset is divided into as many parts as there are instances, each instance effectively forming a test set of one.

The Perfect Classifier

- A: The Perfect Classifier
  - Here every instance is correctly classified. TP=P, TN=N and following is its Confusion Matrix

		Predicted Class	
		+	-
Actual Class	+	P	0
	-	0	N

The Worst Possible Classifier

- B: The Worst Possible Classifier
  - Here every instance is wrongly classified. TP=0, TN=0 and following is its Confusion Matrix

		Predicted Class	
		+	**-**
Actual Class	**+**	**0**	**P**
	-	**N**	**0**

The Ultra-Liberal Classifier

- C: The Ultra-Liberal Classifier
  - This Classifier always predicts the positive class. The TP rate = 1, but so is the FP rate.

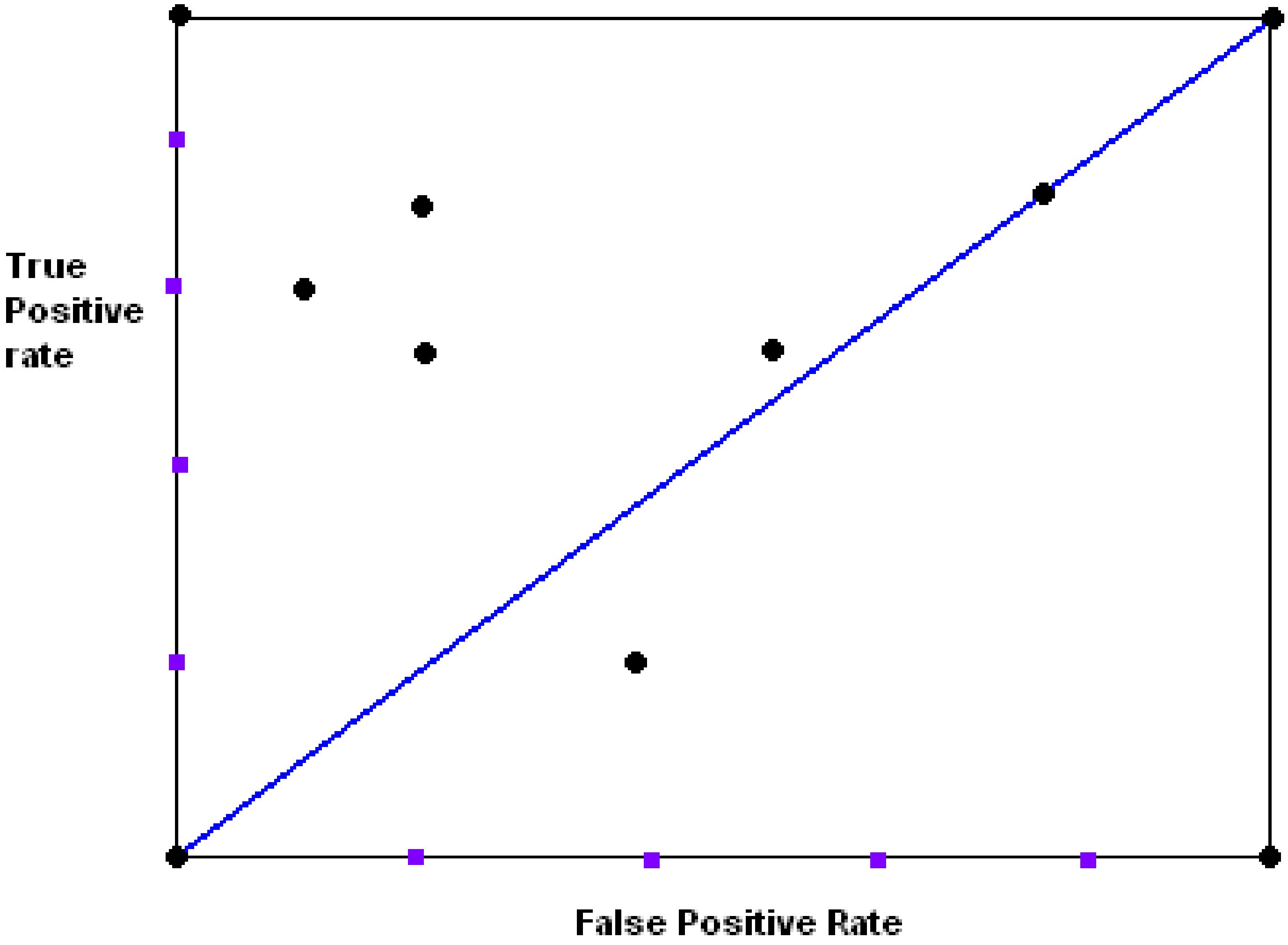
		Predicted Class	
		+	-
Actual Class	+	**P**	**0**
	-	**N**	**0**

The Ultra-Conservative Classifier

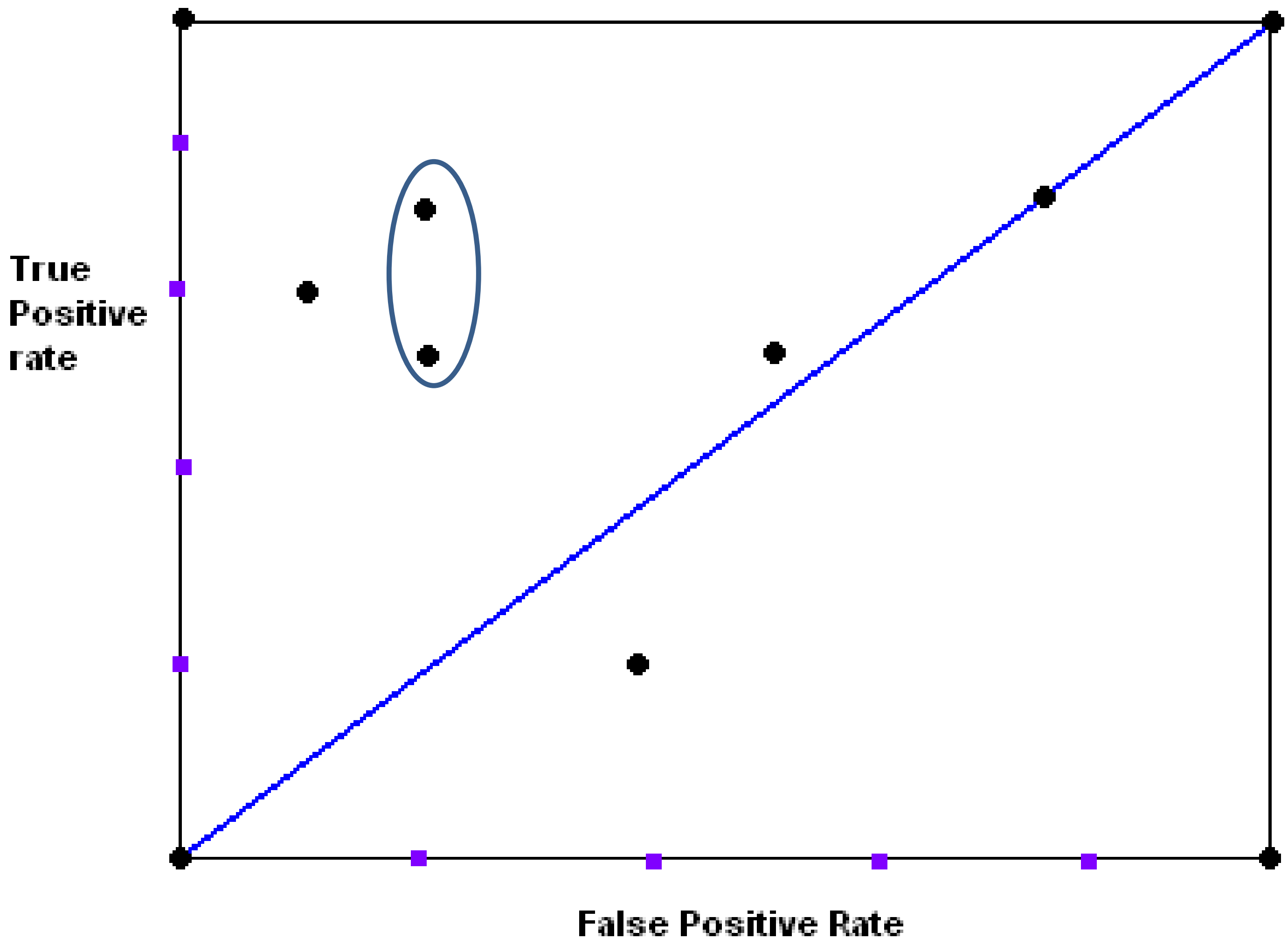
- D: The Ultra-Conservative Classifier
  - This Classifier always predicts the negative class. The FP rate = 0, but so is the TP rate.

		Predicted Class	
		+	-
Actual Class	+	**0**	**P**
	-	**0**	**N**

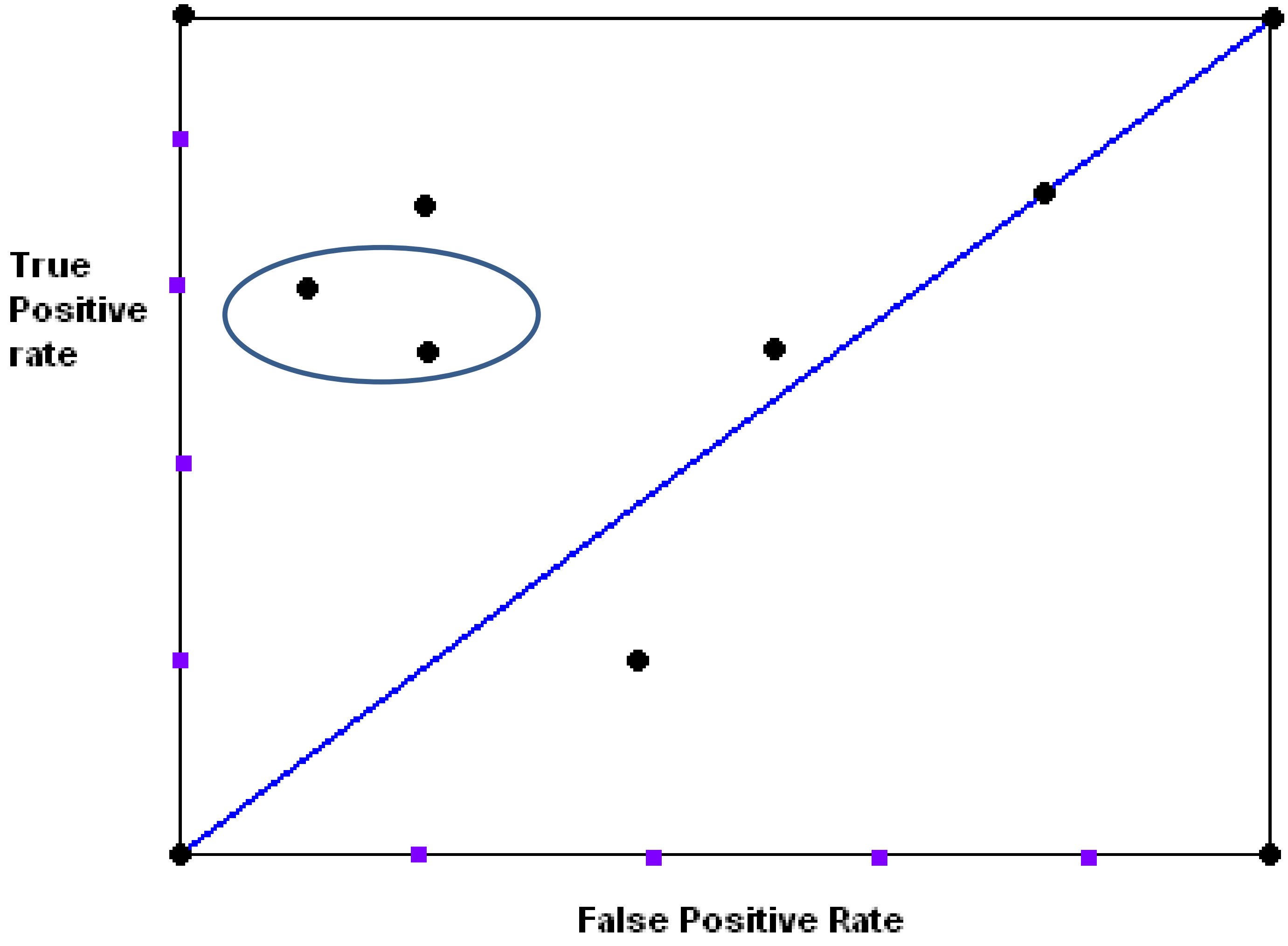
ROC Graph



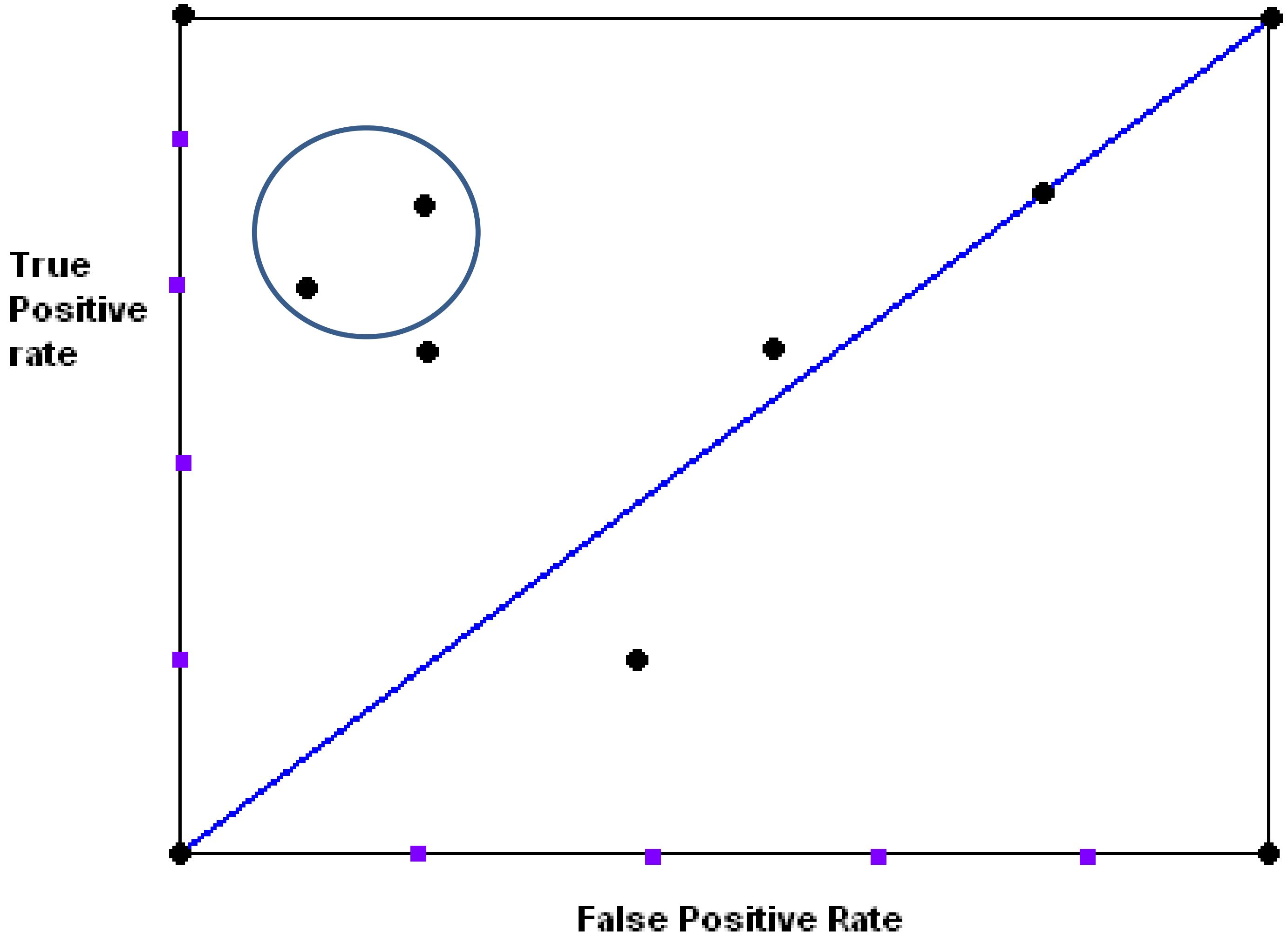
ROC Graph



ROC Graph



ROC Graph



Other Evaluation Criteria

- **Speed**

  It is not just the time or computation cost of constructing a model it also includes the time required to learn to use the model.

- **Robustness**

  Data errors are common, in particular when data is being collected from a number of sources and errors may remain even after data cleaning. It is therefore desirable that a method be able to produce good results in spite of some errors and missing values in datasets.

Other Evaluation Criteria

- **Scalability**

  Many data mining methods were originally designed for small datasets. Given that large datasets are becoming common, it is desirable that a method continues to work efficiently for large disk-resident databases as well.

- **Goodness of the Model**

  For a model to be effective, it needs to fit the problem that is being solved. For example in a Decision Tree classification, it is desirable to find a decision tree of the "right" size and compactness with high accuracy.

Other Evaluation Criteria

- **Interpretability**

  An important task of a data mining professional is to ensure that the results of data mining are explained to the decision makers. It is therefore desirable that the end-user be able to understand and gain insight from the results produced by the classification method.

Accuracy

- We saw how pruning can be applied to decision tree induction to help improve the accuracy of the resulting decision trees. Are there general strategies for improving classifier and predictor accuracy?
  - Yes... (Ensemble Methods)