SEMINAR

Topic on: Metrics for Evaluating Classifier Performance,crossvalidation

submitted by,
Alan Cyril
roll no:801

1.Metrics for Evaluating Classifier Performance

- Some terminology:
- Positive tuples (tuples of the main class of interest) and negative tuples (all other tuples).6 Given two classes, for example, the positive tuples may be buys computer = yes while the negative tuples are buys computer = no. Suppose we use our classifier on a test set of labeled tuples. P is the number of positive tuples and N is the number of negative tuples. For each tuple, we compare the classifier's class label prediction with the tuple's known class label.

- True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- True negatives(TN): These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- False positives (FP): These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class buys computer = no for which the classifier predicted buys computer = yes). Let FP be the number of false positives.
- False negatives (FN): These are the positive tuples that were mislabeled as negative (e.g., tuples of class buys computer = yes for which the classifier predicted buys computer = no). Let FN be the number of false negatives.

- The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$accuracy = \frac{TP + TN}{P + N}.$$

- In the pattern recognition literature, this is also referred to as the overall recognition rate of the classifier, that is, it reflects how well the classifier recognizes tuples of the various classes.

- The error rate or misclassification rate of a classifier, M, which is simply 1 − accuracy(M), where accuracy(M) is the accuracy of M. This also can be computed as,

$$error\ rate = \frac{FP + FN}{P + N}.$$

- The training set (instead of a test set) to estimate the error rate of a model, this quantity is known as the resubstitution error.
- The class imbalance problem, where the main class of interest is rare. That is, the data set distribution reflects a significant majority of the negative class and a minority positive class.
- The sensitivity and specificity measures can be used.
- Sensitivity is also referred to as the true positive (recognition) rate (i.e., the proportion of positive tuples that are correctly identified), while specificity is the true negative rate (i.e., the proportion of negative tuples that are correctly identified). These measures are defined as

$$sensitivity = \frac{TP}{P}$$

$$specificity = \frac{TN}{N}.$$

It can be shown that accuracy is a function of sensitivity and specificity:

$$accuracy = sensitivity\frac{P}{(P+N)} + specificity\frac{N}{(P+N)}.$$

Cross-Validation

- In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” D1, D2,..., Dk , each of approximately equal size.
- Training and testing is performed k times. In iteration i, partition Di is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets D2,..., Dk collectively serve as the training set to obtain a first model, which is tested on D1; the second iteration is trained on subsets D1, D3,..., Dk and tested on D2; and so on.
- Unlike the holdout and random subsampling methods, here each sample is used the same number of times for training and once for testing. For classification, the accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data.

- Leave-one-out is a special case of k-fold cross-validation where k is set to the number of initial tuples. That is, only one sample is “left out” at a time for the test set.
- In stratified cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data.

THANK YOU