VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Fundamentals of Machine Learning

Credibility and Algorithm evaluation

Jan Platos

November 25, 2021

Credibility and Algorithm evaluation

Credibility and Algorithm evaluation

- The presented algorithms need to be compared and the results evaluated to get valuable information.
- The evaluation differs for classification and/or clustering algorithms.
- The evaluation need to be systematic.
- The evaluation also need to be defined according the evaluated metric.

Credibility and Algorithm evaluation - Classification

- Dataset composition:
  - Training data
  - Testing data
  - Validation data
- Not all sets are required.
- Evaluation may be done on the Training data only.

Credibility and Algorithm evaluation - Classification

- The using of the same data for training and testing is not possible due to over-fitting and overestimation.
- The validation part is used for parameter tuning or model solution.
- When the parameter tuning is done, the model is reconstructed on the whole dataset.
- The knowledge from the testing dataset should not be used in parameter tuning.

Credibility and Algorithm evaluation - Classification

- *Accuracy* - the fraction of test instances in which the predicted value matched the ground-truth value.

$$Accuracy = \frac{1}{N}\sum_{i=0}^{N} 1 \; if \; (y^i_{pred} == y^i_{truth}) \; else \; 0$$

Credibility and Algorithm evaluation - Classification

- *Cost-sensitive accuracy*
  - Not all cases are equally important in all scenarios while comparing the accuracy, e.g. Imbalanced data, ill vs. healthy patients, etc.
  - This is frequently quantified by imposing different costs $c_1, \ldots, c_k$ on the misclassification on the different classes.
  - Let $n_1, \ldots, n_k$ be the number of test instances belonging to each class.
  - Let $a_1, \ldots, a_k$ be the accuracy (expressed as a fraction) on the subset of test instances belonging to each class.
  - The overall accuracy A can be computed as a weighted combination of the accuracy over the individual labels:

$$A = \frac{\sum_{i=1}^{k} c_i n_i a_i}{\sum_{i=1}^{k} c_i n_i}$$

Credibility and Algorithm evaluation - Classification

- *Confusion matrix*

	Ground truth	
Predicted	True	False
True	TP	FP
False	FN	TN

Credibility and Algorithm evaluation - Classification

Holdout

- The labeled data is randomly divided into two disjoint sets (training and testing).
- Typically 60% to 75% is used for training set.
- This partition may be repeated several times to get he final estimation.
- The over-presented samples in the training set are under-presented in the testing sets.
- Due to not using of the whole data set for training the estimation are pessimistic.

Credibility and Algorithm evaluation - Classification

Holdout

- By repeating the process over *b* different holdout samples the mean and the variance of the error estimates may be determined.
- These information may be used for building the confidence intervals on the error.
- In case of imbalanced data, an independent sampling (for each class separately) have to be used to ensure the similarity between whole dataset and the testing dataset.

Credibility and Algorithm evaluation - Classification

Cross-Validation

- The data is divided into m disjoint subsets of equal size n/m.
- A typical choice for m is around 10.
- One segment is used as a testing set the the remaining $m - 1$ as a training set.
- This process is repeated by selection each of the m subsets as a testing sets.
- The average accuracy over the m different test sets is reported.
- The size of the training set is $(m - 1) * n/m$.

Credibility and Algorithm evaluation - Classification

Cross-Validation

- When m is chosen large, the training set size is close to the whole dataset and the reported prediction is very close to the whole data set.
- The estimate of the accuracy tends to be highly representative but pessimistic.
- A special case is when $m = n$, this is called a *leave-one-out* cross-validation.
- *Stratified cross-validation* uses proportional representation of each class in the different folds and usually provides less pessimistic results.

Credibility and Algorithm evaluation - Classification

Bootstrap

- The labeled data are sampled uniformly with replacement to create a training set that may contain a duplicates.
- The labeled data of size n is sampled n times with replacement.
- The probability that a particular data point is not included in a sample is given by $(1 - 1/n)$
- The probability that the point is not included in n samples is then $(1 - 1/n)^n$.

Credibility and Algorithm evaluation - Classification

Bootstrap

- For large values of n the expression is approximately $1/e$.
- The fraction of labeled points included included at least once in the dataset is $1 - 1/e = 0.632$.
- The training model is constructed on the bootstrapped sample with duplicates.
- The overall accuracy is computed using the whole dataset.
- The estimate is highly optimistic due to large overlap between training and testing set.

Credibility and Algorithm evaluation - Clustering

- Internal Validation Criteria
  - Sum of Square Distances to Centroids
  - Intra-cluster to Inter-cluster distance ratio.
  - Silhouette coefficient
  - Probabilistic measure
- External Validation Criteria
  - Purity
  - Gini index
  - Entropy

Credibility and Algorithm evaluation - Clustering

Internal Validation Criteria

- Useful when no external criteria is available.
- The major problem if internal criteria is that they are biased toward one or another algorithms.
- The criteria is usually borrowed from the objective function used by certain algorithms.
- The main usage of these criteria is for comparison of the algorithm from the same class or different run of the same algorithm.

Credibility and Algorithm evaluation - Clustering

Sum of Square Distances to Centroids

- Useful when centroids are determined – mainly distance-based algorithms.
- The sum of squared distances of each point to corresponding centroid is used as a quality measure.
- The smaller value indicate better clustering quality.

$$SSQ = \sum_{X \in D} dist(X, C)^2$$

- Where C is the closest centroid to X.

Credibility and Algorithm evaluation - Clustering

Intra-cluster to Inter-cluster distance ratio

- Based on sets of random pairs of objects.
- The P is a set of pairs that belong to the same cluster.
- The Q is a set of pairs that does not belong to the same cluster.
- The average distances are defined as follows:

$$Intra = \frac{1}{|P|} \sum_{(X_i, X_j) \in P} dist\,(X_i, X_j) \qquad Inter = \frac{1}{|Q|} \sum_{(X_i, X_j) \in Q} dist\,(X_i, X_j)$$

- The ratio $Intra/Inter$ is a quality measure. Smaller values means higher quality.

Credibility and Algorithm evaluation - Clustering

Silhouette Coefficient

- Compares similar distances as the previous one.
- $Davg_i^{in}$ is the average distance of X_i to data points within the cluster.
- $Dmin_i^{out}$ is the minimum of the average distances to all other clusters.

$$S_i = \frac{Dmin_i^{out} - Davg_i^{in}}{\max\left\{Dmin_i^{out}, Davg_i^{in}\right\}}$$

- The overall silhouette coefficient is the average of the data point-specific coefficients.
- The value is in the range $\{-1, 1\}$. Large positive values indicate highly separated clustering, large negative value indicate a ”mixing” between clusters.

Credibility and Algorithm evaluation - Clustering

External Validation Criteria

- These criteria are available when the ground truth is know.
- In the real datasets, the ground truth is usually not know.
- An approximation may be achieved using available class labels.
- These labels should not correspond to the natural clusters.
- Despite these problems, external evaluation criteria are preferable.
- The number of natural clusters may not reflect the number of classes.

Cluster Validation - External Validation Criteria

- When the number of determined clusters and the number of classes is equal, a confusion matrix is useful.

Cluster Indices	1	2	3	4
1	97	0	2	1
2	5	191	1	3
3	4	3	87	6
4	0	0	5	195

Cluster Indices	1	2	3	4
1	33	30	17	20
2	51	101	24	24
3	24	23	31	22
4	46	40	44	70

Questions