University of Victoria Computer Science

CSC475 Music Information Retrieval

Data Mining

George Tzanetakis

University of Victoria

2014

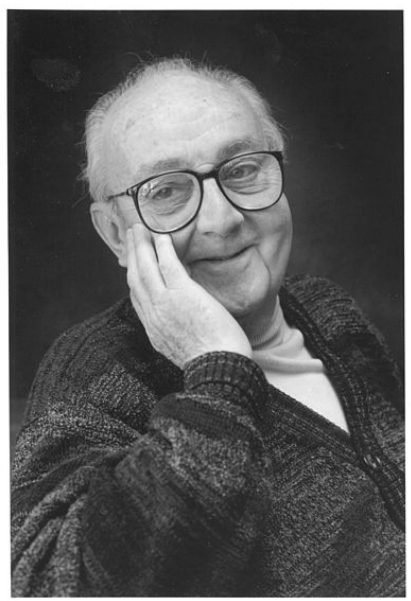
Table of Contents I

In a nutshell

Quote

Essentially, all models are wrong but some are useful.

George Box (1919, 2013). He got interested in statistics after performing experiments on the effect of poison gas to small animals in WW II.



Terminology

- Machine Learning (when cs theory people do it)
- Data Mining (when database people get involved)
- Statistics (when applied mathematicians are involved)
- Artificial Intelligence (when applied cs people use it)

The underlying ideas behind all these categories are the same and there is significant overlap although each term has its own flavor. A radically different way to "program" a computer to perform a particular task.

Problem Setting

University of Victoria Computer Science

Definition

A computer program "learns" from experience E with respect to tasks T measured by performance measure P means that the performance at tasks in T, as measured by P, improves with experience E (Tom M. Mitchell)

Example

A chess program "learns" from playing chess with human players (T, E) as measured by the number of games it wins (P) if the number of games it wins (P) increases as it plays more games.

Problem Setting

University of Victoria Computer Science

Definition

A computer program "learns" from experience E with respect to tasks T measured by performance measure P means that the performance at tasks in T, as measured by P, improves with experience E (Tom M. Mitchell)

Example

A chess program "learns" from playing chess with human players (T, E) as measured by the number of games it wins (P) if the number of games it wins (P) increases as it plays more games.

Definition

Classification (or Supervised Learning) is the machine learning task of inferring the label (class) of an object (typically represented as a vector of numbers) by analyzing training data consisting of feature vectors (representing objects) and associated labels.

Example

You are given the height, weight of 1000 people as well as a binary attribute that indicates whether they are professional basketball (or if you prefer hockey) players or not (training set). You are then given the height and weight of 100 new people and are asked to predict whether they are professional basketball players or not (testing set). The number of correct predictions is the performance metric.

Basketball players

Basketball players

Train Set and Test Set

$$\left(\begin{array}{cc|c} H(cm) & W(kgr) & NBA \\ 190 & 100 & 0 \\ 195 & 110 & 0 \\ 180 & 85 & 0 \\ \dots & & \\ 178 & 75 & 1 \end{array}\right) \qquad \left(\begin{array}{cc|c} H(cm) & W(kgr) & NBA \\ 193 & 102 & ? \\ 180 & 87 & ? \\ 200 & 150 & ? \\ \dots & & \\ 168 & 65 & ? \end{array}\right)$$

Definition

A feature matrix is a 2D matrix where each row (instance or sample) is a feature vector consististing of numbers (attribute or feature) characterizing a particular object and an associated class label (the ground truth). A column corresponds to attribute or feature.

Train Set and Test Set

$$\left(\begin{array}{cc|c} H(cm) & W(kgr) & NBA \\ 190 & 100 & 0 \\ 195 & 110 & 0 \\ 180 & 85 & 0 \\ \dots & & \\ 178 & 75 & 1 \end{array}\right) \qquad \left(\begin{array}{cc|c} H(cm) & W(kgr) & NBA \\ 193 & 102 & ? \\ 180 & 87 & ? \\ 200 & 150 & ? \\ \dots & & \\ 168 & 65 & ? \end{array}\right)$$

Definition

A feature matrix is a 2D matrix where each row (instance or sample) is a feature vector consististing of numbers (attribute or feature) characterizing a particular object and an associated class label (the ground truth). A column corresponds to attribute or feature.

Train Set and Test Set

University of Victoria Computer Science

$$\left(\begin{array}{cc|c} H(cm) & W(kgr) & NBA \\ 190 & 100 & 0 \\ 195 & 110 & 0 \\ 180 & 85 & 0 \\ \dots & & \\ 178 & 75 & 1 \end{array}\right) \qquad \left(\begin{array}{cc|c} H(cm) & W(kgr) & NBA \\ 193 & 102 & ? \\ 180 & 87 & ? \\ 200 & 150 & ? \\ \dots & & \\ 168 & 65 & ? \end{array}\right)$$

Definition

A feature matrix is a 2D matrix where each row (instance or sample) is a feature vector consististing of numbers (attribute or feature) characterizing a particular object and an associated class label (the ground truth). A column corresponds to attribute or feature.

Train Set and Test Set

$H(cm)$	$W(kgr)$	NBA
190	100	0
195	110	0
180	85	0
...		
178	75	1

$H(cm)$	$W(kgr)$	NBA
193	102	?
180	87	?
200	150	?
...		
168	65	?

Definition

A feature matrix is a 2D matrix where each row (instance or sample) is a feature vector consististing of numbers (attribute or feature) characterizing a particular object and an associated class label (the ground truth). A column corresponds to attribute or feature.

Classification Formulation

University of Victoria Computer Science

Input

Training vectors $\mathbf{x_i} \in R^d, i = 1, \ldots, n$ where i is the instance index, n is the number of instances, and d is the dimensionality of the feature vector. Associated ground truth classification labels can be written as integers y_i.

Classification Operations

University of Victoria Computer Science

Classifier

A classification algorithm typically supports 2 operations:

- **Train** takes as input a labeled training set (a 2D matrix of features and 1D vector of associated labels) and outputs a model (a representation of the classifier for that particular problem).
- **Predict** takes as input a trained model and an unlabeled testing set (2D feature matrix) and produces a predicted labeled test set (1D vector of labels).

Classification Operations

University of Victoria Computer Science

Classifier

A classification algorithm typically supports 2 operations:

- **Train** takes as input a labeled training set (a 2D matrix of features and 1D vector of associated labels) and outputs a model (a representation of the classifier for that particular problem).
- **Predict** takes as input a trained model and an unlabeled testing set (2D feature matrix) and produces a predicted labeled test set (1D vector of labels).

Table of Contents I

Evaluation

University of Victoria Computer Science

Evaluator

- **Evaluate** takes as input a vector of ground truth labels and a corresponding vector of predicted labels and calculates various classification performance metrics. Classifier is treated as a black box.

Example

Classification accurancy can be computed as the percentage of labels that were correctly predicted i.e the ground truth label and the prediction label match.

Evaluation

University of Victoria Computer Science

Evaluator

- **Evaluate** takes as input a vector of ground truth labels and a corresponding vector of predicted labels and calculates various classification performance metrics. Classifier is treated as a black box.

Example

Classification accurancy can be computed as the percentage of labels that were correctly predicted i.e the ground truth label and the prediction label match.

Evaluating Classifiers

University of Victoria Computer Science

In order to evaluate classification performance, ground truth is needed not only for the training set but also for the testing set. The amount of labeled data is provided by the specific problem and somehow needs to be partitioned into a training set and a testing set for evaluation. Suppose you are a company interested in a particular classification problem. You contract two other companies to provide you with trained classifiers based on labeled data you provide to them for training. The trained classifiers are proprietary and you can only use them as black boxes to predict labels for feature vectors but you have no knowledge of how they perform this task. How do you choose which company to use ?

Classification Accuracy and Generalization

University of Victoria Computer Science

Training data

Labeled data provided to the contract companies. Let's call EA the classification accuracy somehow estimated by the contract companies.

Testing data

Labeled data withheld by company contracting to keep the other two honest. Let's call GA the classification accuracy estimated on the witheld testing data.

Challenge

How to estimate EA so that it is as close to GA as possible ?

Generalization and overfitting

University of Victoria Computer Science

Definition

Generalization in machine learning is the ability of an algorithm to perform well on new examples of data that are not part of the data that it was trained on. Estimating generalization performance is not trivial.

Definition

Overfitting in machine learning occurs when a model ends up describing errors and noise in the training data instead of the actual underlying data. Complex models (having many parameters relative to the number of observations) are particularly prone to overfitting. When overfitting occurs the estimated performance of a machine learning algorithm can be misleading and far from the "true" generalization performance.

Schemes for splitting train/test

University of Victoria Computer Science

Use training set for testing

Results can be misleading as a result of overfitting. Learning the parameters of a prediction function and testing it on the same data is a methodological mistakes. A perfect memorizer would work perfectly but fail miserably on new samples (i.e it would overfit the training data).

Schemes for splittin train/test

University of Victoria Computer Science

Percentage Split

Split labeled data into training and testing using a particular percentage (for example 75%). Tradeoff: more training data results in better models but more testing data increases confidence in generalization capability.

Main idea

Do multiple splits to training and test estimating multiple metrics in the process that then are combined to produce the final performance metrics.

A sidenote about permutations/shuffling

Frequently it is necessary to shuffle the instances of a feature matrix typically during multiple iterations of splitting into training and testing. Random permutations are provided in some programming environments but sometimes you need to implement them. It is relatively simple but not trivial (google Knuth Shuffles for details). As typically feature matrices are really large, permutations are typically performed using external permutation vectors of indices i.e the feature matrix is left unchanged in memory and accessed through the indices in permuation vectors.

Permutation Example

Feature Matrix

$$\left(\begin{array}{cc|c} H(cm) & W(gr) & NBA \\ 190 & 100 & 0 \\ 195 & 110 & 0 \\ 180 & 85 & 0 \\ \dots & & \\ 178 & 75 & 1 \\ 165 & 80 & 1 \\ 185 & 90 & 1 \end{array}\right)$$

Permutation vectors

$$\left(\begin{array}{c|c} original & permutation \\ 1 & 70 \\ 2 & 8 \\ 3 & 89 \\ \dots & \\ 97 & 22 \\ 98 & 20 \\ 99 & 45 \end{array}\right)$$

K-Fold Cross-Validation

University of Victoria Computer Science

- Do a shuffle
- Partition the data into K folds. Each fold will have approximately the same number of instances
- Use one fold for testing and the remaining k-1 folds used for training. This will yield predictions for 1 fold.
- Repeat K times resulting in predictions for K folds.
- Calculate the classification metrics you need based on the ground truth and the predictions
- Bias-variance tradeoff: K large: lower bias (large training sets), higher variance (training sets are similar)
- Note that there is only one initial shuffle, a fixed number of iterations, and no estimate of estimation variance
- When the number of folds equals the number of instances k-fold cross-validation is called leave-one-out evaluation.

Bootstrap estimate

University of Victoria Computer Science

- Randomly select (with replacement) N samples from the N instances. Notice that the same sample might be selected multiple times and some samples might not be selected. With some math you can show that approximately $0.632 * n$ distinct instances will be selected and $0.368 * n$ will not be selected. The boostrap sample is used for training and the remaining instances are used for testing.
- Can be repeated as many times as desired giving a handle to the variance of the estimates.

Variants

University of Victoria Computer Science

Repeated Cross-Validation

Shuffle, perform K-Fold cross-validation, estimate performance metrics and repeat for as many times as desired for variance estimation.

Variants

University of Victoria Computer Science

Stratification

When the number of instances in each class is unequal the partitioning process can result in partitions in which some classes are under-represented or even completely absent. In stratified cross-validation (or boot-strapping) the training set is partitioned based on class, folding (or boostrap sampling) is performed in each partition and then the results are merged. This ensures that the number of instances for each class in each fold is approxixately the same as the distribution of classes in the training set.

More variants

University of Victoria Computer Science

Domain specific folds

It is possible to provide domain specific folds rather than arbitrary folds (for example if multiple collection experiments are conducted or time information is included). In this case the analyst provides an extra vector of integers specifying the fold index for each instance.

Subsampling

If there are many instances (common with big data sets) then the training time can become prohibitive. In such cases it is common to use a random sample of smaller size for training and testing. For iterative algorithms which are common in classification this random subsample can change during each iterations providing fast training times while at the same time utilizing the majority of the labeled data.

More variants

University of Victoria Computer Science

Domain specific folds

It is possible to provide domain specific folds rather than arbitrary folds (for example if multiple collection experiments are conducted or time information is included). In this case the analyst provides an extra vector of integers specifying the fold index for each instance.

Subsampling

If there are many instances (common with big data sets) then the training time can become prohibitive. In such cases it is common to use a random sample of smaller size for training and testing. For iterative algorithms which are common in classification this random subsample can change during each iterations providing fast training times while at the same time utilizing the majority of the labeled data.

Measuring classification performance

University of Victoria Computer Science

Input

For each instance we have one label for the ground truth and one predicted label by the classifier. All classification performance metrics are based on this information. For some the only thing that matters is whether the prediction is correct or not.

Output

A single number, or a matrix, or a plot either characterizing the entire experiment or a specific class.

Common Summary Metrics

University of Victoria Computer Science

Classification accuracy

$$accuracy = \frac{\#correct\ \ predictions}{\#instances}$$

Typically it is expressed as a percentage. Can also be calculated for each class separately.

Confusion Matrix

This is an C by C matrix M in which each element $M_{i,j}$ shows the percentage of instances with ground truth class label i that were predicted as class label j. The diagonal element correspond to the correctly classified instances for each class. The confusion matrix reveals information about how classification and misclassification are distributed among different combinations of classes.

Example confusion matrix

Linear SVM classifier for automatic music genre classificaiton (80%) classification accuracy.

a	b	c	d	e	f	g	h	i	j	$\leftarrow$ *classified as*
79	0	4	4	0	2	6	0	3	2	$a = bl$
0	94	1	0	0	3	0	0	0	2	$b = cl$
5	0	74	1	0	1	3	0	3	13	$c = co$
1	0	3	74	2	0	1	4	6	9	$d = di$
2	0	0	5	78	0	4	3	8	0	$e = hi$
2	4	3	0	0	89	1	0	0	1	$f = ja$
3	0	0	5	1	0	83	1	0	7	$g = me$
0	0	10	5	4	0	0	72	2	7	$h = po$
3	0	5	8	11	1	0	2	68	2	$i = re$
3	0	17	9	0	0	8	2	6	55	$j = ro$

Common Class (Binary) Metrics

University of Victoria Computer Science

Terminology

If we consider only a single class c there are four possible outcome combination of ground truth and prediction. Using terminology from binary testing we can term these true positives (TP): ground truth and prediction are both c, false positives (FP): ground truth is not c but prediction is c, false negatives (FN): ground truth is c but prediction is not c, and true negatives (TN): both ground truth and prediction are not c.

Information Retrieval Metrics

Based on this terminology one can define the following metrics.

Precision and Recall

precision (p), recall (r) and accuracy (a) as follows:

$$p = \frac{tp}{tp + fp} \qquad r = \frac{tp}{tp + fn}$$

There is a trade-off between precision and recall and it is easy to come up with simple schemes that maximize one at the expense of the other. For example consider the following scenarios:

- Classifier that predicts only two instances correctly as positive (precision of 1, low recall)
- Classifier that predicts all instances as positive (low precision, recall of 1)

Information Retrieval Metrics

University of Victoria
Computer Science

The F-measure attempts to balance precision and recall and is defined as the harmonic mean of them:

F-Measure

$$F = 2 \cdot \frac{p \cdot r}{p + r}$$

Accuracy can also be defined similarly and is informative.

Accuracy

$$a = \frac{tp + tn}{tp + tn + fp + fn}$$