Artificial intelligence and knowledge engineeringList 4

Katarzyna Fojcik, Joanna

SzołomickaMay 9, 2023

1 Purpose of the list

The purpose of the exercise is to familiarize with simple machine learning algorithms and the basic steps of implementing a machine learning project: data mining and problem definition, data preparation, selection of algorithms and hyperparameters, evaluation of algorithms,improvement of results, presentation of results.

2 Theoretical introduction

Below you will find information and basic definitions to help you complete your task list.

2.1 Machine Learning

machine learning*(Machine Learning)*, one of the fields of artificial intelligence, a broad category of algorithms whose goal is to perform specific tasks automatically. Algorithms that take in data learn from that data, and then apply what they learn to infer and complete the task.

Machine learning types:

- **Supervised learning***(supervised learning)*— a method of learning in which the set of training data on which the algorithm learns contains an attached solution to the problem, the so-called labels or classes. The two main areas that use supervised learning are solving a regression problem (value prediction) and a classification problem (class prediction).
- **Unsupervised learning***(unsupervised learning)*– a model learning method in which the training data are unlabelled, i.e. they do not have labels. In colloquial terms, we have raw data that we throw into the model and leave all the work of finding connections between the data to the algorithm (so-called learning without a teacher). These include, among others clustering algorithms – in other words, clustering or grouping, and various types of algorithms visualizing data and reducing dimensionality.
- **Reinforcement learning***(RL - reinforcement learning)*– its task is to interact with the environment on the basis of the collected information. Unlike the types mentioned earlier, in reinforcement learning, you do not prepare a set of training data, but an environment from which the model will collect data automatically. Its goal is to maximize the reward it returns. The environment may depend on the learning objective. In the case of teaching a program that plays games, it will be a game with its rules, or the real world, in the case of a program that learns to control a rover.

2.2 Data division into training and validation sets

Training set– is a set of data that we use to learn the algorithm. Based on this data, the model learns to properly classify, builds all dependencies. You can say it anticipatespossible outcomes and makes decisions based on the data provided to it.

Validation set- is such a dataset that we use to run unbiasedtest of the model that we trained on the training (training) data. We run this testwhen selecting a model or by selecting a set of hyperparameters. It is important that the data contained in the validation set have not been used for model training before, as they will then be unsuitable for objective, unbiased testing.

Test set– once we have chosen the model, it is time to test it on the data from the test set. It is very important that this data has not been used for model training or validation before, because we want to know how the selected algorithm performs on data that it has never dealt with before.

Cross validation(cross-validation) is an alternative way to prepare training and test sets for the model. It is characterized by greater efficiency in assessing the actual accuracy of the model, but it is less efficient than the naive method. This method consists in dividing the data set into subsets and creating validation and training data sets from them.

There are many types of cross-validation. Basic K-fold cross-validation assumes that the data are independent, and identically distributed (IID - Independent and Identically Distributed). Divides all samples into k groups, if possible of equal size. The ML algorithm is trained using k 1 groups and the skipped group is used for validation. Therefore, k trainings are performed, and the obtainedthe results from the k validations performed are averaged.

The ratio of splitting data into training, validation and test sets should depend on the sample size. Different division practices are distinguished, but it is generally assumed that for smaller sets we prioritize the need to train the algorithm on as many samples as possible. Then the validation (and test) set is relatively smaller (in the case of cross-validation, we choose a larger number of k, e.g. 10). However, having several thousand samples at our disposal, we can decide to increase some of the validation data (alternatively, reducing the number of groups to$k$= 5).

2.3 Naive Bayes Classifier

The naive Bayes classifier is a simple probabilistic classifier that naively assumesdependence and the same significance of features (predictors) for a fixed class label, is based on TheoremBayes.

Definition(Bayes' theorem) Let A and B be events, P (B) > 0, P (AB) denotes conditional probability, then:

$$P(AND \wedge b) = P(AND|b)P(b) = P(b|AND)P(AND)$$
$$P(B|A) = \frac{P(AND|b)P(b)}{P(AND)} \quad (1)$$

Definition(Naive Bayes Classifier) Let X1, . . . , Xn are features, then:

$$P(x_1|x_2, Y)$$
$$= P(x_1|Y)P(x_1, X_2, Y) = P(x_1|x_2, Y)P(x_2, Y) = P(x_1|x_2, Y)P(x_2|Y)P(Y)$$
$$P(x_1, X_2|Y) = P(x_1|Y)P(x_2|Y) \quad (2)$$
$$P(x_1, x_2, \ldots, Xn|Y) = \prod_{and=1}^{n} P(Xi|Y)$$

2.4 Decision tree

A decision tree in machine learning is a representation of a classifier in the form of a tree supporting the decision-making process. Building such a tree is based on various algorithms and their parameters, which are calculated for each new decision tree node. Example parameters are entropyand knowledge growth.

Definition(Entropy) Entropy is a measure of the variability of the data given by the formula3.

$$h(S) = -\sum_{x \in x} p(x) \log_2 p(x) \tag{3}$$

where S is the current data set for which the entropy is calculated (for each node of the tree it will be a different - respectively smaller data set), X is the set of classes in the set S, $p(x)$ is the ratio of the number of elements from the class x *to the number of elements in the set S.*

Definition(Increase of knowledge - Information Gain) Gain of knowledge G(A) is a measure of the difference in entropy before and after breaking the data set with a given attribute (feature).

$$g(AND) = h(S) - \sum_{v \in Vales(AND)} \frac{|S_v|}{|S|} h(S)_\iota \tag{4}$$

where $h(S)$ is the entropy for the set S, $Vales(AND)$ is the set of all attribute values AND, *Sv is a subset* S *such that* $St = \{p \in S : AND(p) = v\}$, $h(Saint)$ is the entropy of the subset $Saint$.

Decision trees are very susceptible to the overshooting phenomenonoverfitting. As the depth increases, the trees are able to adapt very well to the training data, but at the same time their ability to generalize deteriorates (the error on the test set increases).

2.5 PCA (Principal Component Analysis)

PCA (principal component analysis) is a statistical method of factor analysis: a data set consisting of N observations, each containing K variables, can be interpreted as a cloud of N points in K-dimensional space. The goal of PCA is to rotate the coordinate system so as to maximize the variance of the first coordinate, then the second, and so on. The transformed coordinate values are the principal components (the initial ones explain the most variation). In machine learningnew is used to reduce the dimension of the data.

2.6 Metrics

To evaluate the classification carried out, the results of various metrics are used (usuallyfor validation data only):

2.6.1 Error matrix

Depending on the operation of the classifier, four cases are distinguished (Table1):

- **TP (True Positive)**– the number of truly diagnosed positive cases
- **FP (False Positive)**– number of falsely diagnosed positive cases
- **FN (False Negative)**– the number of falsely diagnosed negative cases
- **TN (True Negative)**– the number of truly diagnosed negative cases

		Reality	
		Positive class	Negative class
prediction	Positive class	TP	FP
	Negative class	FN	TN

Table 1: Confusion Matrix.

2.6.2 Accuracy (aka accuracy)

ACC accuracy is the most frequently quoted indicator that allows us to assess the quality of data classification. We find out what part of all classified items has been classified correctly. That is, the sum of correct classifications from a given category (TP) and the correct classification from other categories (TN) is divided by the number of all classified cases.

$$ACC = \frac{TP + TN}{P+n} \tag{5}$$

2.6.3 Sensitivity (or recall/sensitivity/true positive rate)

TPR sensitivity tells us what is the share of correctly predicted positive cases (TP) among all positive cases (including those that were incorrectly classifiednegative – FN).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{6}$$

2.6.4 Precision (aka precision/positive predictive rate)

We divide the number of correctly predicted positive values (TP) by the sum of all positively predicted values (including those incorrectly predicted in this way), i.e. TP+FP. As a result, we find out how many positively predicted examples actually arepositive.

$$\text{pay-per-view} = \frac{TP}{TP + FP} \tag{7}$$

2.6.5 F1-score

The F1-score is the harmonic mean between precision and recall. The closer it is to one, the better it is for the classification algorithm. In the best case, it takes the value of 1 when we are dealing with perfect sensitivity and precision.

$$f_1 = \frac{2TP}{2TP+FP+FN} \tag{8}$$

3 Task

Face the problem of identifying the type of glass. Use one of the datasets for this purposeUCI-GLASS. Tasks should be done with Python and/or WEKI. Punctation:

1. data mining - provide basic statistical data and notes on features and labelsdataset. (10 points)
2. data preparation - divide the data into training and validation sets (alternatively, use cross-validation), examine the impact of various types of data processing on the classification results (pro-repeated: normalization, standardization, discretization, feature selection, PCA) - i.e. compare the results without data processing with the results after data processing, using at least 2 methods of different types (separately). (30 points)

   Bonus - remove 5% of the feature values and prepare the data using methods to deal with missing data. (5 points)
3. classification – test the classifiers and examine the impact on the results: naive Bayes classifier and decision tree using at least 3 different sets of hyperparameters. (40 points)

   Bonus - Test (with understanding!) more advanced algorithms such as Random Forestor Support Vector Machines (SVM). (5 points)
4. classification evaluation – to compare the results of various types of data preparation andusedclassifier, use the known classification evaluation metrics and interpret the results. (20 points)

For the task, prepare a report containing a short description of all performed steps andthe results of the task (preferably collected tables) along with the interpretation. In the report, indicate the source materials used and briefly describe the libraries used in the implementation. Send the report to the teacher at least 24 hours before handing in the list.

4 Appendix - filters and classifiers in Weka Explorer

4.1 Filters

- **attribute discretization:**filters → supervised → attribute → Discretize or filters → unsupervised → attribute → Discretize
- **attribute normalization:**filters → unsupervised → attribute → Normalize
- **PCA:**filters → unsupervised → attribute → PrincipalComponents
- **deleting an attribute:**filters → unsupervised → attribute → Remove
- **completing the attribute value:**filtersunsupervisedattribute ReplaceMissingVa-lues
- **standardizationattribute:**filters → unsupervised → attribute → Standardize
- **deleting instances with given attribute values:**filters → unsupervised → instance → RemoveWithValues

4.2 Attribute selection

Select attributes tab - selection with Attribute Evaluator button:

- GainRatioAttributeEval
- InfoGainAttributeEval
- PrincipalComponents

4.3 Classifiers

Classify tab:

- **naive Bayes classifier:**classifiersbayes → NaiveBayes or NaiveBayesMultinomial(depending on the needs)
- **decision tree C4.5:**classifiers → trees → J48
- **group of classifiers:**
  - **boosting:**classifiers → meta → AdaBoostM1
  - **bagging:**classifiers → meta → Bagging
  - **Various kinds ofclassifiers:**classifiers → meta → Stacking

5 Appendix - useful libraries for python

- from sklearn.preprocessing import Normalizer, StandardScaler
- from sklearn.decomposition import PCA
- from sklearn.model_selection import train_test_split
- from sklearn.metrics import accuracy_score, precision_score, recall_score, confusion_matrix
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.naive_bayes import GaussianNB