Report - Practical 2

Important Information

Daniela Flüeli left our group at the end of the semester and she was assigned to this practical but did not do anything (Prof. Olteanu was informed by email). So, this practical was done by Christian Birchler (60%) and Wenjie Tu (40%) in a rush during our examination phase. We would appreciate it if you take this into account for grading.

Distributions

The implementation of the distributions for the Naïve Bayes Classifier was almost straight forward. For the continuous and binary features, we could use the functions provided by 'scipy.stats'. Those functions allow to fit the parameters and also provide the possibility to calculate the density or probability mass. For the categorical features a multinoulli distribution is applied. The parameters for this distribution had to be calculated manually since there is no function for multinoulli in 'scipy.stats'.

Naïve Bayes

For the Naïve Bayes Classifier, we implemented a class 'NBC' that provides a method 'fit' for learning and a method 'predict' for doing prediction on new data. For the 'fit' function we first calculated the prior probabilities for each class. This was quite easy since we only look at the labels. For the probabilities of the conditional distributions we iterated over each class and for each class, we iterated also over all features. Corresponding to the feature type we stored a new instance of the Distribution so that we can refer to them in the 'predict' method.

In the 'predict' method we calculate the probability for the new data belonging to each class, which is the idea of generative models. We assume independence between the features (therefore the **Naïve** Bayes), which makes the computation simpler. We calculate kind of a conditional probability for each feature. We don't need the exact probability since we are only interested in the right classification and not in the probability itself. Further, we do our calculation in log space to avoid underflow.

Our implementation of the Naïve Bayes Classifier reaches an accuracy of over 90% in the provided test code.

Comparison NBC and LR

Fort the comparison between the NBC and the LR we trained six data sets with these classifiers. We used the same preprocessed data sets for both classifiers. The preprocessing was done so that the categorical features are one-hot encoded. This was done since the base of a logistic regression classifier is a linear model. Furthermore, with one-hot encoding, it is also possible to encode the missing values in the data. In general, we can observe from the plots that the logistic regression performs steadily better if we increase the training size. The improvement of the NBC is not that large as compared to the logistic regression. So, we conclude that the logistic regression generally outperforms the NBC by using more training data.