Machine Learning in Data Science Project Week 6

Eva Aßmann, Paul Vogler

November 2019

1 Description of the data and preprocessing steps

The Contraceptive Method Choice Data Set (CMC) is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey and was obtained from the UCI Machine Learning Repository. The CMC data set comprised information on 1473 married women who were either not pregnant or did not know if they were at the time of data collection. For every woman, 9 attributes describing her demographic and socio-economic characteristics as well as her contraceptive method use were recorded (Table 1).
There were no missing values to handle. Categorical variables were binary encoded in order to prevent later machine learning methods from inducing relations or orders between a variable's classes.

Attribute	**Values**	**Type**
Wife's age	range 16-49	numerical
Wife's education	1=low, 2, 3, 4=high	categorical
Husband's education	1=low, 2, 3, 4=high	categorical
Number of children ever born	range 0-16	numerical
Wife's religion	0=Non-Islam, 1=Islam	binary
Wife's now working?	0=Yes, 1=No	binary
Husband's occupation	1, 2, 3, 4	categorical
Standard-of-living index	1=low, 2, 3, 4=high	categorical
Media exposure	0=Good, 1=Not good	binary
Contraceptive method used	1=No-use, 2=Long-term, 3=Short-term	class attribute

Table 1: List of attributes that were collected for every woman in the CMC data set.

2 Results for the Naïve Bayes classifier

Naïve Bayes classification was used to predict a woman's contraceptive method use based on her demographic and socio-economic information. Employing three different validation strategies, for every approach Naïve Bayes (NB) model was trained and evaluated by determining the model's efficiency in the training and prediction process, its robustness under different noise levels as well as the following classification measures:

- Accuracy
  The accuracy of a model describes the proportion of correctly predicted values among the total number of instances it was applied to.
- Confusion matrix
  After applying a model to a set of instances, the confusion matrix stores the number correctly and falsely classified instances for every target class.
- Precision and recall
  The group-specific precision is the number of correctly classified objects divided by the total number of

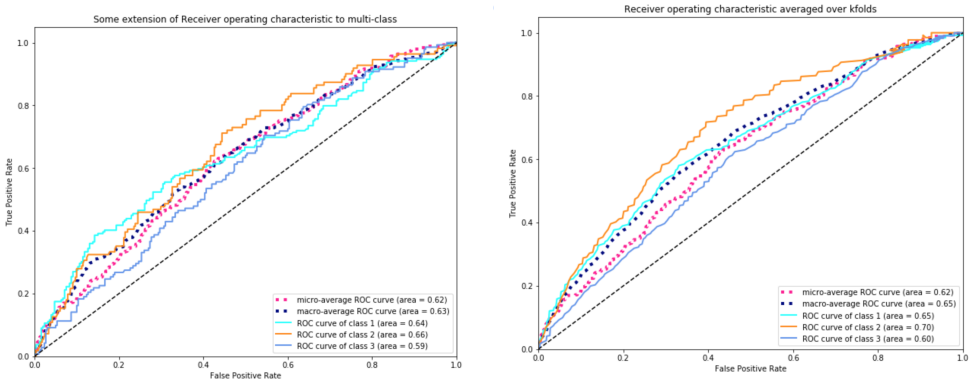
objects that are classified into the respective target group. Recall is defined by the number of objects that are correctly classified into a specific target class divided by the total number of objects that actually belong to that class (the same as sensitivity).

- F1 score
  The F1 score is the harmonic mean of precision and recall. It is calculated from two times the product of precision and recall devided by the sum of precision and recall. The F1 score allows pairwise comparison of two classifiers.

- Sensitivity and specificity
  Sensitivity and specificity are defined as the true positive rate and true negative rate of a classififer, respectively. Sensitivity and recall describe the same measure.

- ROC curve and AUC value
  The ROC curve is a plot of the true positive rate against the false positive rate using various thresholds to assess the power of a decision rule in binary classification. The area under the ROC curve is called AUC and it is a measure of seperability. The larger the AUC, the better the model is at predicting the binary target.

In a first approach, the processed CMC data was split into 70% training and 30% test data. A NB model was trained on the training set and evaluated on the test set. All of the above metrics were applied to the classification result of the test set (see Figure 1 and Table 2).
The second approach employed a 5-fold cross-validation. The data was partitioned into 5 equal-sized disjoint subsets. In 5 iterations, each subset was once used as the test set, the respective remaining groups were combined as one training set. For each iteration the evaluation measure mentioned above were produced and averaged (see Figure 1 and Table 2).
The last evaluation strategy was Leave-one-out. It functions similar to n-fold cross-validation with the slight difference that in each iteration the test set comprises only a single data instance. Thus, Leave-one-out results in n train and test iterations and n values for each evaluation measure which were then averaged, respectively (see Table 2). When looking at the pairwise correlation between all CMC data set feature



(a) ROC curve for NB classifier trained on 0.7 hold-out training set.

(b) Average ROC curve for NB classifier trained on 5-fold cross validation.

Figure 1

variables, a weak correlation was observed between a woman's standard-of-living index and her education, as well as the husband's education. The number of children and a woman's age showed moderate correlation and the strongest correlation was yielded for a woman's education and her husband's education (and vice versa) (see Figure 2). A weak negative correlation could be observed for a woman's education and her husband's occupation. Because there was a strong positive relationship between a woman's and her husband's education, feature selection could be applied in order to make the predicted variable more accurate or eliminate irrelevant information to prevent it from decreasing the model's accuracy and quality.

	Accuracy	Sensitivity			Specificity			Precision	Recall
		1	2	3	1	2	3		
70/30 train/test	43%	67%	34%	39%	15%	61%	44%	43%	43%
5fold CV	43%	61%	35%	42%	18%	58%	43%	43%	43%
Leave one out	43%	63%	35%	42%	17%	59%	43%	43%	43%
	Roc Auc		Robustness				Efficiency		F1 score
	Micro	Macro	0	0.1	1	2	fit	predict	
70/30 train/test	62%	63%	43%	44%	47%	48%	0.0037	0.0027	43%
5fold CV	62%	65%	43%	44%	49%	45%	0.0037	0.0353	43%
Leave one out	66%		43%	44%	46%	44%	0.0035	6.6209	43%

Table 2: Metric results for Gaussian NB.

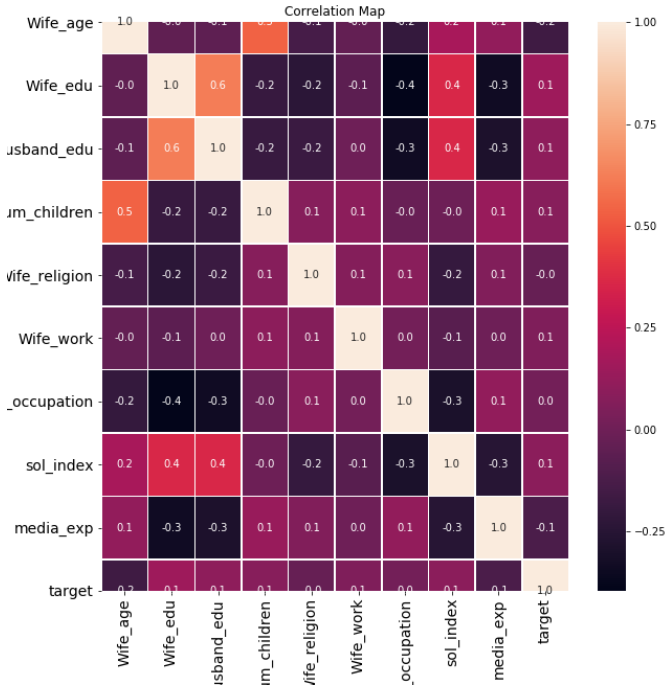
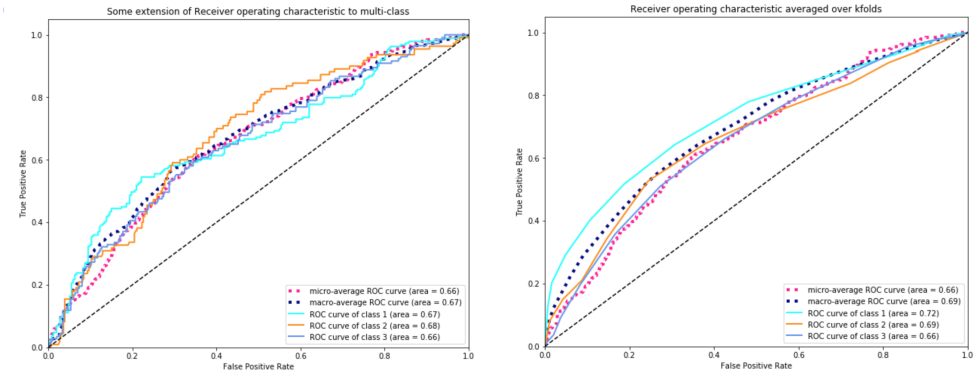


Figure 2: Pairwise correlation of the CMC data set features.

3 Results for the k-nearest Neighbor classifier

The same holdout sets and metrics used for the Naïve Bayes where also applied to a k-nearest Neighbor classifier (KNN). At first a k, as the number of neighbors per classification cluster had to be set. The k with the best accuracy score after a onefold cross validation was set as the best k. One k chosen was 30, but it may vary, depending on the random split of the testing and training set.
The hold out sets used were a 70/30 training/testing one-fold cross-validation, a 5-fold cross-validation and leave-one-out validation (see Figure 3 and Table 3). The metrics to analyse the performance between the holdout sets and classifiers were the accuracy, sensitivity, specificity, precision, recall, F1 score, Roc-curve AUC, robustness (accuracy after noise addition) and the time efficiency. The results for one run of the hold out methods and for all metrics can be seen in Table 3.



(a) ROC curve for kNN classifier trained on hold-out training set.

(b) Average ROC curve for kNN classifier trained on 5-fold cross validation.

Figure 3

	Accuracy	Sensitivity			Specificity			Precision	Recall
		1	2	3	1	2	3		
70/30 train/test	55%	60%	44%	53%	58%	13%	27%	51%	51%
5fold CV	53%	62%	39%	50%	40%	20%	37%	53%	53%
Leave one out	52%	61%	41%	49%	42%	19%	39%	52%	52%
	Roc Auc		Robustness				Efficiency		F1 score
	Micro	Macro	0	0.1	1	2	fit	predict	51%
70/30 train/test	63%	65%	56%	54%	45%	44%	0.0056s	0.0256s	53%
5fold CV	66%	69%	53%	53%	47%	42%	0.0036s	0.095s	52%
Leave one out	68%		52%	52%	47%	43%	0.003s	7.36s	

Table 3: Metric results for KNN.

4 Brief discussion of the results

For the Gaussian Naïve Bayes classifier the accuracy, precision, recall and f1 score all resulted in a rather low score of 43%. Notably in all three holdout sets the specificity for target 1 was below 20%. The major difference between the holdout sets can be seen in the prediction time, that increased about tenfold between 70/30 and 5fold CV, and then again by 200 between 5fold CV and Leave-one-out. The accuracy increased for all three holdout sets with increasing the standart deviation of added noise, up to a SD of 1.
Most of the measures were close to equal between hold out methods for the KNN classifier. Notable trends are that the accuracy for all sets is only slightly above 50%, and that the specificity for the targets 2 and 3 are generally really low with percentages under 40%. Also the Robustness accuracy drops more, the more noise is introduced into the dataset. One significant difference between the three hold out sets is again that the prediction time increases with the number of cross-validations.
Overall The KNN resulted in the higher accuracy, precision, recall and f1 scores than Naïve Bayes. For both classifiers the prediction time went up significantly with an increased number of cross-validations. While Naïve Bayes had a really low specificity for target 1, KNN had a lower specificity for targets 2 and 3. The last trend is that the addition of noise increased the accuracy of the Naïve Bayes classifier, but decreased the KNN accuracy.