Machine Learning in Data Science Project Week 13

Eva Aßmann, Paul Vogler, Katrin Böhler

Januar 2020

1 Data

The Wisconsin breast cancer data set was loaded from the scikit-learn datasets module. The data set stores information about cell shapes from cancer diagnostics images, that are labeled with being either malign or benign. The cells are described by 30 features, that show 10 different properties of the cell shape. Namely these are radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry and fractal dimension. All of these appear three times with a mean, standard error and worst value, all of which are real valued entries. In total there are 569 entries in this dataset.
The MNIST data set was loaded using the scikit-learn datasets fetch_openml function to get the mnist_784 data set from the OpenML database. MNIST is a subset of a larger data set from the National Institute of Standards and Technology (NIST). MNIST stores 70.000 images of handwritten digits in range 0 to 9 from approximately 250 writers. The digits in the data set came size normalized to fit in a 20x20 pixel box and centered in a 28x28 image. Each digit image is described by 784 pixel values ranging from 0 meaning background (white) to 255 meaning foreground (black). Each data set entry is assigned a string-formatted label storing the digit presented in the image ('0','1','2','3','4','5','6','7','8','9').
Both data sets contained no missing values and all feature columns were already in the desired format, therefore no further preprocessing was needed. Within the scope of the active learning procedure the dataset was normalised and denormalised again multiple times, which is further described in section 3.

2 Methods

Fully supervised learning is a machine learning strategy that aims at learning to generalize from feature-described, fully labelled data in order to correctly predict labels for new subjects with unknown label information. During training procedure, the complete training set is used for learning.

Active learning arises from the difficulty of retrieving labelled data in many allday scenarios, which can be time-consuming, expensive and in some cases dangerous which creates the need for alternative learning approaches. In active learning, a small subset of a large unlabelled data set is selected and labelled by an external entity, e.g. an expert or an oracle. The resulting seed is used for training an initial classifier. In the next step, the seed model iteratively gets retrained by selecting one or more unlabelled instances based on a specific sampling strategy and using it as learning input. The training is finished by one or more stopping criteria. e.g. a maximum number of training iterations or training instances.
There are three strategies to select samples for learning. In membership query synthesis, items are selected based on some underlying distribution and sent to the oracle for labelling. Stream-based selective sampling considers each unlabelled instance separately and decides, whether it should be labelled or rejected based on its informativeness. The most commonly applied strategy is pool-based sampling which selects instances from a large pool of smaller subsets of unlabelled data based on some informativeness measure. Within stream-based selective and pool-based sampling there are at least four ways to select instances: The Least Confidence (LC) strategy compares the label probabilities for each of the potential query instances and chooses the one with the least confidence for the most likely label. Margin sampling does not only consider the most probable label assignment, but selects the instance with the smallest difference between the first and second most probable labels. Entropy sampling actually includes all label probabilities into its informativeness measure and selects the instance with the largest entropy value over all label likelihoods.

In contrast to fully supervised learning, the active approach works for larger and unlabelled data sets. It does not use the complete available data set for training, but aims at yielding at least the same performance as fully supervised learning by training on a small seed and iteratively improving the model using a labelled subsample as validation set. The informativeness measures aim at improving the model by not only learning with items that are more or less similar, but specifically introducing items to the model which would lead to better generalization properties. Thus, active learning takes more control in the learning by setting a focus on generating a diverse training set.

In the following experiments, three learning algorithms were employed: Support vector machines (SVM) are used for supervised classification or regression tasks. The main principle of an SVM is to find a hyperplane that best separates a set of labelled data points into their respective label categories. The optimal hyperplane shows the highest possible margin to all training points in order to enable a correct classification of new data points. Random Forest (RF) is an ensemble supervised learning method for classification and regression tasks. Multiple decision trees are trained on a set of labelled instances. A new instance is then classified by assigning it the label which is the mode of all trees' outputs. Naïve Bayes (NB) is a probabilistic supervised learning method that aims at maximizing an instance's probability to belong to a specific class based on its attribute values. Within the NB method, the Gaussian NB algorithm assumes the feature values associated with each class to be distributed according to a a normal distribution.

In the following experiments, hold-out sets were used for evaluation. The MNIST and breast cancer data sets were split into a training and test set, respectively (85% train, 15% test). For each experiment, the training was performed on the training set, while the test set was used for evaluation. To assess a classifier's performance and compare it to the trained models from other experiments, the predictive accuracy and the F1 score were computed. A model's accuracy describes the number of correct classifications divided by the total number of test cases. The F1 score is defined by the harmonic mean of model's precision and recall. A model's precision is defined by the number of samples classified correctly as positive divided by all test cases that were classified as positive. The Recall measures the amount of samples classified correctly as positive out of all actually positive test cases. The weighted F1 score is used in case of multi-label classification and is defined by the average F1 score over all labels, weighted by the support for each label to account for label imbalance. The specific evaluation strategies and calculation of measures for each experiment was described in the section 3.

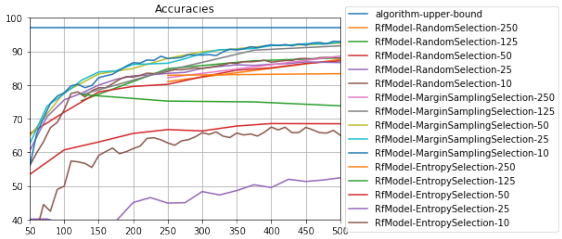
3 Experiments & Results

3.1 Experiments

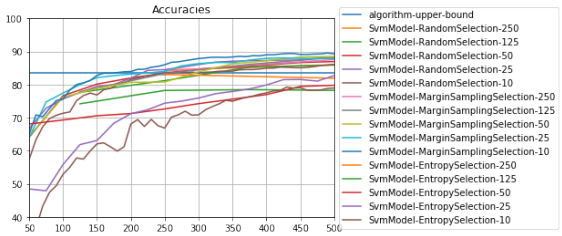
For both of the datasets, three fully supervised learning experiments were conducted with three different classifiers and a 85%/15% training/testing set split. At first, a SVM with a linear kernel was used, that was capped at 10000 iterations for the sake of saving runtime, after SVM took by far the longest to converge in test runs. The second classifier was a random forest with 500 estimator trees, a gaussian naive bayes classifier with no additional settings was used as the third one.
For both data sets, also three active learning experiments were run using the same classification methods as described above. Each of the three classifiers was executed with three different sample selection strategies, i.e. random, entropy and margin, whereby each sample selection strategy was run for 5 different training seed set sizes k (k=[10,25,50,125,250]), accumulating a total of 6 fully supervised and 90 active learning experiments. The list of k values was slightly different than it was required in the exercise description, but that could not be fixed in time, because of the exceeding runtime of up to 7 hours. The resulting trajectories are now just missing one of the k values. For each of the 90 classification runs in the active learning algorithm, the subsample was normalized with the sklearn 'minmaxscaler' before each step and then the transformation was reversed after the classification.
For both the fully supervised and active learning experiment, the trained models were evaluated on the test sets by calculating accuracy and the weighted F1 score.

3.2 Results

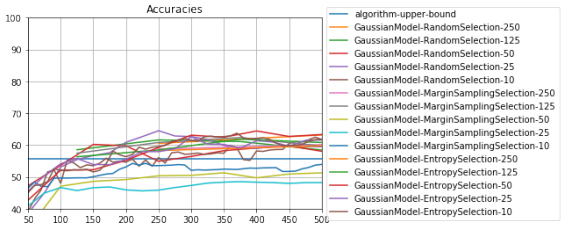
The active learning experiments yielded the following results for the MNIST data set:
While the Gaussian NB and SVM models that were trained on the MNIST data set yielded a maximum accuracy of approximately 65% and 90%, respectively, the trained RF models yielded a maximum accuracy of over 90% over all seed sizes k (see Fig. 1). Also, with almost 100%, the RF models showed the highest upper bound accuracy in comparison to the SVM models with an upper bound of around 85% and the SVM models with an upper bound of around 55%. While the performances within SVM and Gaussian NB trials converged jointly, the performance of trained RF models appeared to be more distributed over the complete accuracy range. The Gaussian NB models performed best on random sample selection and seed size. The SVM and RF models both performed best for margin sample selection and seed sizes 25 and 50. Evaluating the weighted F1 scores for each active learning experiments yielded the same performance hierarchy of RF, SVM and Gaussian NB as comparing the accuracies did (see Fig. 2).
The fully supervised learning experiments on the MNIST data set yielded the following results:
While the trained Gaussian NB and SVM models yielded accuracy values of 55.58% and 83.63%, respectively, the RF model achieved the highest accuracy of 97.14% (see Fig. 3). Also regarding the F1 score, the Gaussian NB model scored worst with 51.7%, the SVM model scored 83.64% and the RF model yielded the highest F1 score of 97.14%. The result of Gaussian NB implies a performance that is hardly better than random guessing. These results are highly similar to the maximum performance results achieved in the active experiments on the MNIST data set and represent the same performance rankling of RF, SVM Gaussian NB from best to worst performing.



a) MNIST with random forest



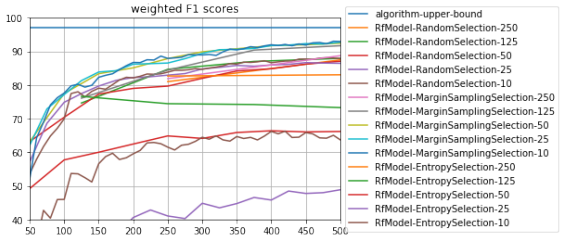
b) MNIST with SVM



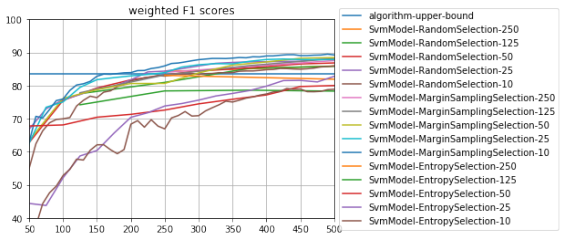
c) MNIST with Gaussian model

Figure 1: Accuracy values for the active learning experiments on the MNIST data set: converging performance is plotted for every seed size and sampling strategy.

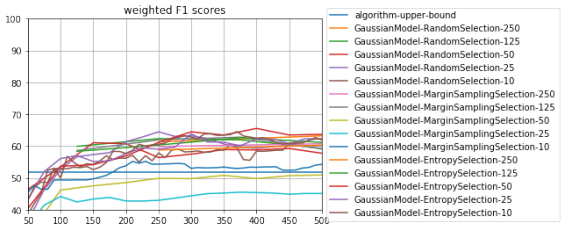
The active learning experiments yielded the following results for the breast cancer data set:
All three classification methods yielded very high maximum accuracy values between 97.5% and around 98.5% (see Fig. 4). The trained RF models yielded a maximum accuracy of around 98.5% for margin sampling strategy and seed size 10. The trained SVM models yielded a maximum accuracy of around 98.5% for entropy sampling strategy and seed size 50. The trained Gaussian NB models yielded a maximum accuracy of 97.5% using margin sampling strategy and seed size 25. The performance convergence showed to proceed quite uniformly over the range of the test set for all three classification models. Evaluating the weighted F1 scores for each active learning experiments yielded the same performance hierarchy of RF, SVM and Gaussian NB as comparing the accuracies did (see Fig. 5).
The fully supervised learning experiments yielded the following results for the breast cancer experiments:



a) MNIST with random forest



b) MNIST with SVM



c) MNIST with Gaussian model

Figure 2: Weighted F1 scores for the active learning experiments on the MNIST data set: converging performance is plotted for every seed size and sampling strategy.

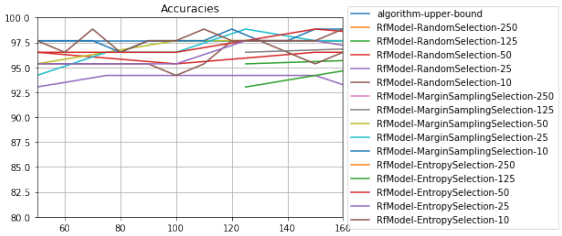
All three classification methods yielded very high accuracy values: the trained Gaussian NB model yielded an accuracy value of 95.35%, the SVM achieved 96.51% and the RF yielded the highest accuracy of 97.67% (see Fig. 6). Also regarding the F1 score, the Gaussian NB model scored 'worst' with 95.35%, the SVM model scored 96.48% and the RF model yielded the highest F1 score of 97.71%. These results are highly similar to the maximum performance results achieved in the active experiments on the breast cancer data set, yet, the active learning approach yielded slightly better performance results.

4 Discussion

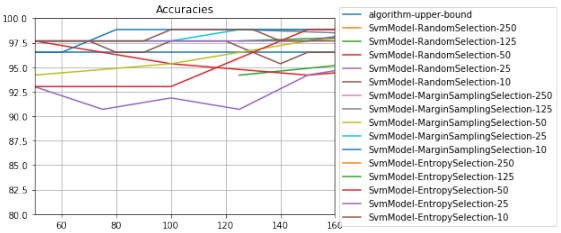
By using an active-learning approach, with a limited amount of labelled data an accuracy very close to a fully supervised approach was achieved. Overall random forest seemed to be the most stable at predicting the labels well between both datasets. SVM was highly reliant on finding an optimal solution in the defined number of iterations, so for the MNIST the resulting score was better with the smaller sample size and the full dataset may have slowed down finding the optimal solution significantly. The gaussian model was very inaccurate for the large MNIST dataset in both the active as well as the fully supervised learning run, maybe because of the dataset dimensionality of over 700 hindered the prediction. But for the breast cancer dataset the prediction was very close to the other two algorithms.
So the best algorithm in this use case should be the one that can scale accurate predictions to high dimensionality and to a high number of data entries, where the dimensionality scaling is more important here, because the data set is subsampled in case of the active learning algorithm.

```
train: (60000, 784) (60000,)
test : (10000, 784) (10000,)
unique classes 10
train  linear SVM  Classifier:...
[LibLinear]/usr/local/lib/python3.6/dist-packages/sklearn/svm/_base.py:947: Conve
  "the number of iterations.", ConvergenceWarning)
predict labels...
linear SVM  accuracy score:  83.63 %
linear SVM  weighted f1 score:  83.64 %
train  Random Forest  Classifier:...
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 500 out of 500 | elapsed:  4.2min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
predict labels...
[Parallel(n_jobs=1)]: Done 500 out of 500 | elapsed:    2.3s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 500 out of 500 | elapsed:    2.1s finished
Random Forest  accuracy score:  97.14 %
Random Forest  weighted f1 score:  97.14 %
train  Gaussian Naive Bayes  Classifier:...
predict labels...
Gaussian Naive Bayes  accuracy score:  55.58 %
Gaussian Naive Bayes  weighted f1 score:  51.7 %
```

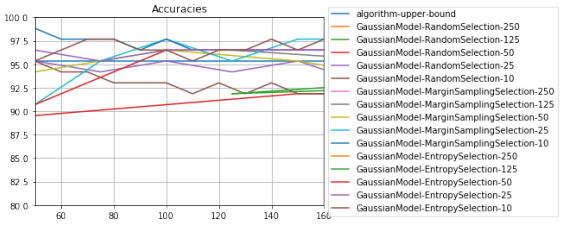
Figure 3: Accuracy values and weighted F1 scores for the fully supervised experiments on the MNIST data set



a) BC with random forest

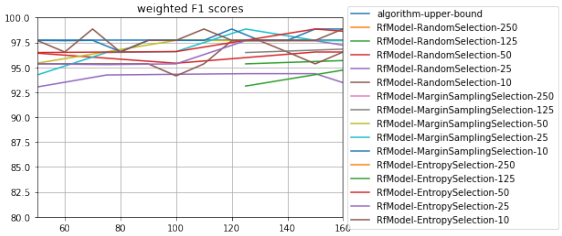


b) BC with SVM

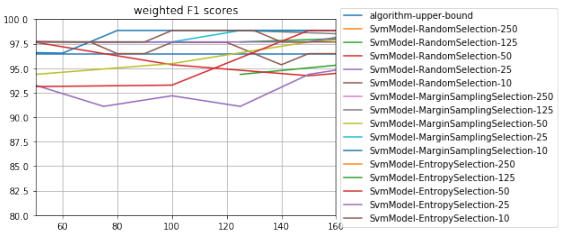


c) BC with Gaussian model

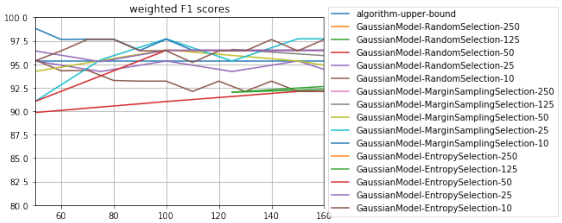
Figure 4: Accuracy values for the active learning experiments on the breast cancer data set: converging performance is plotted for every seed size and sampling strategy.



a) BC with random forest



b) BC with SVM



c) BC with Gaussian model

Figure 5: Weighted F1 scores for the active learning experiments on the breast cancer data set: converging performance is plotted for every seed size and sampling strategy.

```
train: (483, 30) (483,)
test : (86, 30) (86,)
unique classes 2
train  linear SVM  Classifier:...
[LibLinear]/usr/local/lib/python3.6/dist-packages/sk
  "the number of iterations.", ConvergenceWarning)
predict labels...
linear SVM  accuracy score:  96.51 %
linear SVM  weighted f1 score:  96.48 %
train  Random Forest  Classifier:...
predict labels...
Random Forest  accuracy score:  97.67 %
Random Forest  weighted f1 score:  97.71 %
train  Gaussian Naive Bayes  Classifier:...
predict labels...
Gaussian Naive Bayes  accuracy score:  95.35 %
Gaussian Naive Bayes  weighted f1 score:  95.35 %
```

Figure 6: Accuracy values for the fully supervised experiments on the breast cancer data set