Machine Learning

JACOBS UNIVERSITY

Spring Semester 2022
Prof. Dr. Peter Zaspel
Nitesh Khatiwada, Tin Nguyen, Diwas Pandey

Assignment Sheet 9.

Submit on **Tuesday, April 19, 2022, 10:00**.

Excercise 1. (Classification by linear regression on the indicator matrix)

We would like to solve a simple classification problem using paper and pencil. To this end, you are given the training data

$$\mathcal{T} = \{(0.3, 1), (1.8, 1), (1.5, 1), (4.8, 2), (2.6, 2)\} \ .$$

The objective is to predict the class labels for the two evaluation points $x_1 = 2.4$ and $x_2 = 6.2$. Use linear regression on the indicator matrix to build the necessary predictor and evaluate it at x_1 and x_2. In addition, evaluate the *training error* using the 0-1 loss.

(4 Points)

Excercise 2. (Classification by linear discriminant analysis)

We again start from the training data set

$$\mathcal{T}_{train} = \{(-2.1, 1), (-0.9, 1), (0.6, 2), (1.5, 2), (2.7, 2)\}$$

for a paper and pencil classification task. In addition, you are given the validation set

$$\mathcal{T}_{val} = \{(-1.2, 1), (0.5, 1), (1.4, 2)\} \ .$$

a) Use linear discriminant analysis to build a classifier based on the training data.

b) Evaluate the generalization error for the just constructed predictor using the 0-1 loss and the validation set approach.

(4 Points)

Excercise 3. (Training of logistic regression)

Prove Lemma 8.2 from the lecture.

(4 Points)

Programming Exercise 1. We would like to implement and use linear discriminant analysis for classification. To this end, complete the following tasks:

a) Start from Example 8.1 from the lecture notes for which you have the Python source code available as Jupyter notebook. Ignore kNN classfication and linear regression on indicator matrix and (only) re-implement the linear discriminant analysis by yourself. Verify the correctness of your implementation by cross-checking it with Example 8.1.

b) Now, we would like to compare the performance of the just implemented linear discriminant analysis to the performance of kNN classification (based e.g. on Scikit-learn) on SPAM data. Use as data set the Spambase Data Set from the UCI Machine Learning Repository. To carry out the comparison, implement the validation set approach with the 0-1 loss. Randomly split the data set into $N_{train} = 1000$ training samples and $N_{val} = 100$ validation samples and use the same split to evaluate the generalization error for LDA and kNN (with $k = 3$) classification.

Reference solutions will only be provided in Python+Matplotlib. The submission format for Python is a Jupyter notebook. The submission format for C/C++ is standard source files. Choose an appropriate format for the Gnuplot-related submission.

(4 Points)