IML: Exercises Weeks 5

October 11, 2021

1 Introduction

This document contains the exercises for week 5 of the IML course. The exercises are split into three parts: revision, homework and practicum. The only questions you have to hand in are the homework questions. You do this every 3 weeks. Note that we will select a subset of the questions to be graded. If anything is unclear or you have any questions, please feel free to use the forum on Brightspace or send an email to the TAs.

2 Supervised Learning 3 - wk5

2.1 Revision questions

1. Why do we want to have a large margin separating observations in the training set?
2. Is it correct that maximum-margin hyperplane maximises the distance from this hyperplane to the furthest data point of the training set?
3. What are support vectors?
4. What property of SVMs allows us to work in the high-dimensional space with the enriched hypothesis class?
5. Can SVMs be used for regression? How?
6. What are kernel methods?
7. Why is the outlier detection problem considered difficult?

2.2 Homework questions

Exercise 1 For this exercise you will use the breast cancer dataset[1] from `sklearn.datasets`. Split the dataset in train set and test set. Train a linear

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html`

SVM classifier on the training data (predicting if the diagnosis is Malignant or Benign, using all other features except ID). Measure the performance of your classifier on the test data. Now train two more SVM classifiers with different kernels. Measure the performance of these classifiers as well. Report the results of your classifiers and explain their differences.

Your submission should contain both the used code, as well as a table containing the results. Make sure to clearly explain what the used kernels are, and why they lead to different / similar results.

Exercise 2 In what ways does an SVM generalize the concept of separating hyperplanes? How does this relate to the concepts (e.g. PAC-learning) seen in Statistical Learning 1? *Hint: what happens to the XOR-example?*

Exercise 3 Why is it sufficient to only specify the kernel function instead of the inner product? Can any function serve as a kernel?

2.3 Practicum questions

Exercise 1 Generate a dataset for binary classification which is clearly linearly separable. Train the and compare the performance (train and test) of the following methods from sklearn on it:

- LinearSVM
- SVC
- SGDClassifier

Exercise 2 For this exercise, you will use the MNIST data set you used last week (*sklearn.datasets.load_digits* in sklearn). Since SVM classifiers are binary classifiers, you will need to use one-vs-all or one-vs-one to classify all 10 digits. Train the three methods from the previous exercise on this dataset. How do their performances compare on this task? Why is this different than / the same as the previous exercise?

Exercise 3 For this exercise, you will use the California housing dataset (*sklearn.datasets.fetch_california_housing*). Split the dataset in train set and test set. Train a SVM regressor on the training data and predict the price-variable on the test data. Measure the performance of your model in Mean Squared Error.