Homework 3 Basic Machine Learning

For the deadline see Canvas

Version: Mon 17$^{\text{th}}$ Sept, 2018 at 10:50.

Introduction

1. This is a group assignment, so sign up in groups of two students.
2. Each group has to submit a **pdf** with their answers and explanation. **Please put your names and group number at the top of the hand in**.
3. For questions about this homework assignment use the Discussion Board on Canvas.
4. Of course you may use a calculator or a programming environment such as Matlab or Python. **But your report should not contain any code. Explain your computations and results in English!**
5. It is allowed to incorporate handwritten notes or derivations or drawings in your submission as long as these are readable!
6. Explain your answers!

Exercise 1: Questions about K-CV, over- and under-fitting (10 points)

Part a

Describe in your own words under-fitting and over-fitting. Give typical plots of under-fitting and over-fitting and clearly explain these plots. What is on the different axis, what is plotted, what should the reader observe?

Part b

Describe on your own words why we need validation set(s). Why can we not simply split the available data in a training and test set and train the different models on the training set and test each model on the test set in order to select the best model?

Part c

Why do we use K-fold cross validation (K-CV) instead of just one validation set. What are the advantages and disadvantages of using K-CV?

Exercise 2: ROC Curve (10 points)

Suppose we have a two class problem (positive versus negative) and use a random classifier as a classification model. To classify a data sample x one draws a random number from a uniform distribution on the interval $[0, 1]$. If the value of this random number is larger than θ then the to be classified sample x is classified as negative. Denote by P (N) the total number of positive (negative) examples in the dataset. You may assume that P and N are large.

Part a

Compute for a given θ the TP and FP rate in terms of P, N and θ.

Part b

Draw the ROC curve for this classifier.

Part c

What is the area under the ROC curve (AUC) for this classifier?

Part d

Consider the following ROC curve. Explain clearly which point on the ROC curve corresponds to an equal error rate, i.e. the error rate on positive examples is equal to the error rate on the negative examples.

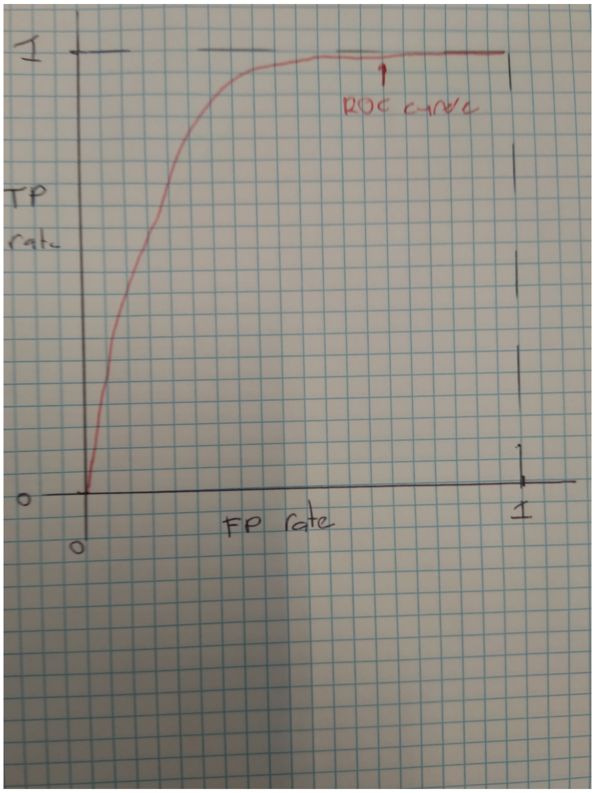

Figure 1: A typical ROC curve.

Exercise 3: Logistic classification and regularization (10 points)

Consider a 3-dimensional classification problem and assume that the weights for the logistic classifier are given by $w = (w_0, w_1, w_2, w_3) = (0, 1, -2, 2)$.

Part a

Consider the data point $x = (1, -1, -2)$ and assume that this data point is misclassified. What is the formula for updating the weights if one applies L_1 regularization and what will be the new weights if one applies stochastic gradient descent (one iteration) with learning parameter 0.7 and regularization parameter 0.2.

Part b

Same setting as in part a but now with L_2 regularization.

Exercise 4: Multiple choice questions (3 bonus points)

Design two multiple choice (MC) questions. Clearly indicate what knowledge or skill you want to test with the MC questions.