Take-home exam: Research Topics in Data Science

Isak Samsten
isak-kar@dsv.su.se

HT, 2019

Introduction

This exam is likely to amount to a lot of work. Mainly because the work required is not limited (other than by the word-count) you can always revisit the questions and go over your answers. Knowing when to stop and what is *"good enough"* is an important quality. As a result, it is easy but misleading to think that the quality of an answer is related to the time you spent preparing it.

When answering the questions of this exam, you should use the literature presented during the course. This includes papers, excerpts from books and the lectures. You are also strongly encouraged to seek additional references to supplement your answers. You may, however, **not** copy any answer and every thought that is not your own needs to be *properly referenced*! Moreover, you may **not** discuss these question with anyone else, or allow anyone to aide you in answering these questions. *Your answers must be a product of your own mind and understanding!*

Note. The page limit (A4 with reasonable font-size) imposed on the questions must **not** be exceeded.

Grading

Each question will be awarded a maximum of 3 points. For 1 point, the answer must be correct, well-presented and properly referenced. To be awarded the full points, the answer should be of well discussed and show a deep understanding of the material.

Handing in

Deadline is January 22, 2020. Submissions are handed in through iLearn (*each exam will be scanned for plagiarism by Urkund*).

Questions

Question I **(1 page, 3 points)**

In class, we discussed the bias-variance trade-off and its implication for ensemble based machine learning methods.

(a) Explain (in your words) the bias/variance trade-off.

(b) Compare and contrast bagging, boosting and stacking in terms of the bias/variance trade-off.

(c) Discuss how the bias/variance trade-off can be used to diagnose the performance of machine learning models.

Question II **(1 page, 3 points)**

In class, we discussed various activation functions for hidden layers in neural networks.

(a) Carefully and in detail explain (in your words) and define the *sigmoid*, *rectified linear unit* and *leaky rectified linear unit* activation functions.

(b) Elaborate on the pros and cons of these activation functions in relation to neural network architectures, weight optimization and backpropagation.

(c) Discuss the reason for using activation functions in neural networks.

Question III **(1 page, 3 points)**

In class, we discussed different distance measures for time series mining.

(a) Explain why the dynamic time warping distance measure is usefull and how it differs from the euclidean distance.

(b) To determine an optimal warping path one could test every possible warping path between a time series T and query Q. Discuss the drawback of this, and describe, in your words, a dynamic programming algorithm for finding an optimal warping path between T and Q.

(c) Discuss the usefulness of lower bounding and describe and contrast **two** measures for measuring the effectiveness of a lower bound.

Question IV **(2–3 pages, 3 × 3 points)**

In class, we discussed several statistical tests for comparing the performance of predictive models. In this question, you will use the results produced by comparing 39 classifiers over 85 datasets. These results can be downloaded from `http://timeseriesclassification.com/Resamples.csv`

(a) Using the provided data, define a null-hypothesis for testing whether there is any significant difference between `BOSS` and `HIVE-COTE` using the paired **t-test**[1] at the $\alpha = 0.01$ significance level[2]. You should carefully state the null-hypothesis and the alternative hypothesis and show the steps required for performing the significance test.

(b) Using the provided data, define a null-hypothesis for testing whether there are any significant differences between `BOSS`, `HIVE-COTE`, `Flat-COTE`, `LS`, `DDTW_R1_1NN` and `TSBF` using a **Friedman test**, followed by a post-hoc Nemenyi test at the $\alpha = 0.05$ significance level[3]. You should carefully state the null-hypothesis and alternative hypothesis and show the steps required for performing the significance test.

[1] You can assume that all assumptions hold.

[2] Use the table of critical values: `https://www.danielsoper.com/statcalc/calculator.aspx?id=10`

[3] Find the critical values for the F-distribution here: `https://www.danielsoper.com/statcalc/calculator.aspx?id=4`