Pilab PhD Qualifying Written Exam Questions
Fall 2012

Advisor(s): A. Taylan Cemgil

Jan, 2013

Table 1: Important information for this exam

Given on:	Jan 02, 2013
Submit no later than:	Jan 14, 2013, 9:30
Advisors:	Cemgil
Jury:	Ethem Alpaydın (alpaydin@boun.edu.tr) Arzucan Ozgur (arzucan.ozgur@boun.edu.tr) Ilker Birbil (sibirbil@sabanciuniv.edu) A. Taylan Cemgil (taylan.cemgil@boun.edu.tr)
Send answers to :	A. Taylan Cemgil (taylan.cemgil@boun.edu.tr)

Contents

1 Rules

1.1 First steps

This PhD Qualifying Exam Questions document is sent from the advisor of the student, to the student with cc to the jury and cc to the member of the PhD Committee responsible from Written Exams. Please see the emails at the table. Please confirm, by replay all, that you have received the questions.

In principle, there should not be any communication during the exam period. In case there is a need for any further communication, it should be in written and include all the parties mention in Table 1.

1.2 Exam

This exam has two parts; an in class close notes exam and a take-home exam.

For the in class exam please prepare the answers for each question on separate sheets. Submit you results at the end of the exam.

Please show your own work. You may use any material as long as you do not get any help from anybody. No questions can be asked to the jury during the exam. If a point is not clear, just state your assumption and proceed.

Each answer should start at a new page. Please have the question first, then its answer on the answer page. There is no page limitation for the answer pages but keep in mind that the answers should be clear and concise.

You may also submit your improved answers to the questions asked during the in class exam with your take home exam, if you were not fully satisfied with your original answers.

1.3 How to prepare your answers to the take home exam

PhDQualifying-Year-Semester-YourLastName is the general naming convention of the files where Semester is "Fall" or "Spring". Note that this document is named accordingly as "*PhDQualifying*-Year-Semester-YourLastName-Questions.pdf". For example if Ali Velioglu is taking the qualifying exams in the fall semester of 2009, the filenames should be *PhDQualifying-2009-Fall-Velioglu*.

Then the returned document should be a single electronic file in zip format whose name should be "PiLab-PhDQualifying-2012.zip". When the zip file is extracted, it should create a directory with name "PiLab-PhDQualifying-2012". Under this directory there should be two directories:

- "LaTeX" directory
  - Answers as a pdf file ("PiLab-PhDQualifying-2012.pdf")
  - All the necessary files to regenerate the answer pdf including "PiLab-PhDQualifying-2012.tex"
- "SupplementaryMaterials" directory. If there is any other material that would be necessary in grading. Such as:
  - A "ReadMe.pdf" file explaining the content of the supplementary material directory.
  - "PhDQualifying-2009-Fall-Velioglu-Questions.pdf" sent to you by email.
  - Source codes
  - Data files
  - Papers and other references in pdf format.

  All the files should have self-explanatory file namings.

1.4 How to submit your answers

Once "PhDQualifying-Year-Semester-YourLastName.zip" is ready, please email it to your advisor, cc to the jury and cc to the member of the PhD Committee responsible from Written Exams as given in Table 1.

Good luck.

2 Take-Home, EA

Derive an algorithm for online principal components analysis. That is, we want to decrease the dimensionality from d to k (given) and we need to update the model as we see instances one by one.

Implement the algorithm using Matlab and show its convergence on example data.

3 Take-home, AO

In this problem you will use supper vector machines (SVMs) to develop a spam email classifier. You can use any of the available SVM packages such as LIBSVM[1]. You will use a subset of the Ling-Spam corpus[2] to train and test your system. The provided training and the test sets (included in the *dataset.zip file*) each contain 240 spam and 240 legitimate email messages. Each email message is provided as a separate file. All files start with a "subject:" heading. Stopword removal and lemmatization have already been performed.

Preprocess the files by extracting the individual words from them. Assume that a word consists of letters from the English alphabet. Discard all tokens that contain different characters such as digits, punctuation marks, or other special symbols (e.g. $). Represent your documents as TF-IDF weighted, length normalized vectors.

(A) Train SVM classifiers with linear and with Gaussian kernels. Apply 5-fold cross-validation on the training set to learn the parameters of your models. Report the average micro-averaged F-measure results as well as the standard deviations for the different kernel and parameter settings. Train your classifier using the training set with the parameters that you learned. Evaluate it by reporting the micro-averaged F-measure over the test set.

(B) Discuss why one-class SVMs might be appropriate for spam email filtering. You can refer to the paper by Manevitz and Yousef, 2002[3] where one-class SVMs have been applied for document classification. Train one-class SVM classifiers with linear and Gaussian kernels. Apply 5-fold cross-validation on the training set to learn the parameters of your models. Report the average micro-averaged F-measure results as well as the standard deviations for the different kernel and parameter settings. Train your classifier using the training set with the parameters that you learned. Evaluate it by reporting the micro-averaged F-measure over the test set. Discuss your results.

4 Take-home, IB

Suppose that we try to solve the following problem:

$$\begin{aligned} \text{minimize} \quad & x^\intercal B x, && (1)\\ \text{subject to} \quad & \mu^\intercal x = c, && (2)\\ & e^\intercal x = 1, && (3)\\ & x \geq 0, && (4) \end{aligned}$$

where $x \in \mathbb{R}^n$, B is a negative semi-definite matrix and e is a vector of ones. This is a special concave minimization problem over a polytope. Propose a **polynomial** algorithm to solve this problem.

Hint: Two facts are well-known for the *generic* problem with an *arbitrary* polyhedron: (i) It is $\mathcal{NP}$-hard; (ii) its *global* optimum solution resides at an extreme point. The major difficulty for working with a general

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[2] Androutsopoulos, Ion, et al. "An Evaluation of Naive Bayesian Anti-Spam Filtering". In Potamias, G., Moustakis, V. and van Someren, M. (Eds.), Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain, pp. 9-17, 2000.

[3] Manevitz, Larry M., and Malik Yousef. "One-class SVMs for document classification." the Journal of machine Learning research 2 (2002): 139-154.

polyhedron is that many extreme points are local optimum solutions, and hence, they cause premature convergence for general nonlinear programming algorithms.

5 Take Home, ATC

Derive and implement a valid Monte Carlo method for sampling from

$$p(H|X)$$

and estimate $p(X)$. Here X is an observed binary $(0-1)$ $M \times N$ matrix and W is a known $M \times K$ matrix

$$\begin{aligned} X(i,j) &\sim \text{Bernoulli}(\sigma(\sum_{k=1}^{K} W(i,k)H(k,j))) \\ H(k,j) &\sim \mathcal{N}(0,\Sigma) \end{aligned}$$

for $i = 1 \ldots M$ and $j = 1 \ldots N$ where $\sigma(x) = 1/(1+\exp(-x))$. Hence H is $K \times N$.