Homework 3

CS420 Machine learning 2019 Spring∗
Department of Computer Science and Engineering
Shanghai Jiao Tong University

Submission deadline: 23:59, June 20, 2020, Saturday

Submission to:

Please submit your homework in pdf/doc format to Canvas platform.

1 (30 **points +** 20 **points) SVM vs. Neural Networks**

(30 points) Select at least two data sets from the link (Data A) below, and then investigate classification performances of Support Vector Machine (SVM) and neural networks (e.g., MLP) on the selected data sets. You may try different experimental settings, e.g., varying the sample size of the training set, trying data sets with different dimensions, and other configurations that may affect the performance in your mind. You may also try different kernels for SVM.

(20 points) Select at least one data set from the link (Data B) below, and applied SVM on it. Compare the SVM performance with the deep learning algorithm benchmarks (if any, by searching the literature on the selected data set). Discuss the strengths and weaknesses of SVM on big data sets.

Links to the data sets:

- Data A: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
- Data B: `http://deeplearning.net/datasets/`

Links to the codes for your references:

- SVM: `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`
- SVM: `http://scikit-learn.org/stable/modules/svm.html`
- MLP: `http://scikit-learn.org/stable/modules/neural_networks_supervised.html`

2 (30 **points) Causal discovery algorithms**

Apply one causal discovery algorithm on a real world problem. You need to specify the details of the problem, collect the data by yourself or from a public website, briefly summarize what algorithm you use, and explain the results.

You may use any causal discovery algorithm described in the following paper [Spirtes et al., 2016], and use the software packages in Page 26 of the paper.

- Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. Applied Informatics, 3:3, 2016
  `https://applied-informatics-j.springeropen.com/track/pdf/10.1186/s40535-016-0018-x`

∗tushikui@sjtu.edu.cn

Page 26:
The following software packages are available online:

- The Tetrad project webpage (Tetrad implements a large number of causal discovery methods, including PC and its variants, FCI, and LiNGAM): `http://www.phil.cmu.edu/tetrad/`.

- Kernel-based conditional independence test Zhang et al. (2011): `http://people.tuebingen.mpg.de/kzhang/KCI-test.zip`.

- LiNGAM and its extensions, Shimizu et al. (2006, 2011): `https://sites.google.com/site/sshimizu06/lingam`.

- Fitting the nonlinear additive noise model Hoyer et al. (2009): `http://webdav.tuebingen.mpg.de/causality/additive-noise.tar.gz`

- Distinguishing cause from effect based on the PNL causal model, Zhang and Hyvärinen (2009, 2010): `http://webdav.tuebingen.mpg.de/causality/CauseOrEffect_NICA.rar`

- Probabilistic latent variable models for distinguishing between cause and effect, Mooij et al. (2010): `http://webdav.tuebingen.mpg.de/causality/nips2010-gpi-code.tar.gz`

- Information-geometric causal inference, Daniusis et al. (2010); Janzing et al. (2012): `http://webdav.tuebingen.mpg.de/causality/igci.tar.gz`