Advanced Machine Learning (Topics) CW1

Objective. As stated in the course syllabus, this module assumes and builds upon the expertise developed through undergraduate-level machine learning courses. The goal of this coursework is a quick recap of basic machine learning techniques via a simple classification task: We will conduct an experimental study of linear and nonlinear (kernel-based) support vector machines, decision trees, random forests, and k-nearest neighbor classifiers.

Datasets. We will use the CIFAR-10 and CIFAR-100 classification datasets. These datasets represent image classification problems with 10 and 100 categories, respectively. Each dataset provides 50,000 training entries and 10,000 testing entries. The datasets were originally used in [6] and they can be downloaded from the Website of [6].[1] See the dataset website for details.

Task. Perform experiments with 1) decision trees, 2) random forests, 3) linear and nonlinear support vector machines, and 4) k-nearest neighbor classifiers,[2] and provide discussions on

- classification accuracies and run-times (for both training and testing) of individual algorithms;
- details of hyperparameter selection including the considered hyperparameters, their search ranges, and selection strategies (e.g. cross-validation vs. validation on a separate validation set);

You are expected to support your discussion with plots and/or tables: Examples include but are not limited to visualization of different hyperparameter combinations, error rates with respect to a varying number of training data points (in addition to the case of using all 50,000 training instances).

These constitute the basic task. You can earn additional marks via advanced tasks including but not limited to

- experiments using other classification algorithms, e.g. convolutional neural networks and/or advanced feature extraction methods, e.g. features extracted by a pretrained ResNet101.
- experiments on other datasets e.g. Fashion MNIST[3] and Caltech-256 datasets.[4]
- ablation studies, e.g. analyzing the effect of removing and/or adding specific components and design choices in each algorithm.

What to hand on? Please submit a single zip file containing a pdf document of your report and a zip file including your code. You need to implement decision trees, random forests, and k-nearest neighbor classifiers. You can use basic math libraries e.g. NumPy. However, you should implement the training and testing steps of these algorithms by yourself. For (linear and nonlinear) support vector machines (SVMs) you can use existing software packages. You are not allowed to use code provided by other UNIST students.

Please format the zip file submission as in **'StudentID_Name.zip'**, e.g. 20221234_KwangInKim.zip. Format your report similarly: **'StudentID_Name.pdf'**. If your submission does not meet this file name requirement, the final mark will become 80% of the initial mark.

Report format and contents. Please use ICML2022 style files (available at `https://icml.cc/Conferences/2022/CallForPapers`). Your report should be maximum of 8 pages long including figures and tables but excluding the references. The report should discuss the main findings (e.g. comparisons of the classification accuracies and run-times of different algorithms) and support these discussions with the corresponding evidence (e.g. figures and tables summarizing the experimental results). It is expected that your discussions

[1] `https://www.cs.toronto.edu/~kriz/cifar.html/`

[2] Details of these algorithms can be found in standard machine learning textbooks including [1, 2].

[3] `https://github.com/zalandoresearch/fashion-mnist`

[4] `https://authors.library.caltech.edu/7694/`

refer to figures and/or tables. Simply placing figures and tables in the report is not enough: Any table or figure that is not accompanied by a discussion will be ignored during marking.

Example discussions can be found at 'Experiments' sections of academic publications, e.g. Section 4 of [3], Section 4 of [5], and Section 4 of [4] among others: Basically, your report is expected to be similar to Section 4's of these papers.

Grading. Sixty marks are assigned for the basic task: Among them, 50% will be reserved for code submission. You do not need to optimize your code submission: We will earn complete 50% marks if your code package correctly implements decision trees, random forests, and k-nearest neighbor classifiers. The remaining 40 marks (out of total 100 marks) are reserved for the advanced task. For the report, we will assess how convincingly the main findings are communicated. Please format your document professionally: E.g., low-quality figures or illegible (or too small) fonts will lower your mark.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[4] Kwang In Kim, Juhyun Park, and James Tompkin. High-order tensor regularization with application to attribute ranking. In *CVPR*, pages 4334–4357, 2018.

[5] Kwang In Kim, James Tompkin, and Christian Richardt. Predictor combination at test time. In *ICCV*, pages 3553–3561, 2017.

[6] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.