Foundations of Machine Learning

STAT 535 Autumn Quarter 2022

Home

Course Description

Syllabus

Books and other resources

Class mailing list

Assignments

Handouts/Course notes

UW Statistics

Project

[Generalities] [Data sets] [Methods] [Software] [Time line] [Report] [Results]

Each project will have the same training set as starting point. You will perform binary class classification on this data. You will train an assigned predictor as well as a predictor of your choice for this task, and do all you can to obtain a low expected classification error.

The classification loss L is valued at 1 for misclassifying a true negative example and 100 for misclassifying a true positive example.

You will have to submit a report (approximately 10 pages) about what you did, submit your code (excluding the packages you may have used), and make a 1-2 slide summary. Besides the written project, you will have a short presentation 1-2 minutes followed by 1-2 questions from the instructors and the audience. A few days before the last day of classes, we will provide a test set with hidden labels. You will run your predictor on the test set and submit the results, which I and Steve will evaluate. In the same class, we'll unveil and compare the results.

Data sets The data is a subset of HIGGS, containing 100,000 examples subsampled to reflect a different positivity rate than the UCI repository. The data sets for training are available **on canvas** under "Files Project" folder. It is a subset of size 100,000 of the HIGGS dataset with some preprocessing. Each column represents a different physical kinematic properties measured by the particle detectors in the accelerator, where the first row is whether a Higgs Boson was produced. Note that the dataset is presented in a .txt format to reduce file sizes. More instructions will come later. Use these data as you wish to obtain your predictor. Later, we will post an unlabeled test set, with the same format, on which you will test the predictor you obtained.

Methods for classification You will use the data made available to construct your predictor. You will use the data made available to construct your predictor. You need to register (more instructions later) by Nov 19. Below is a list of possible predictors. find the list, with short clarifications for each model. No matter what method you choose, **you are responsible for knowing how this method works, and for explaining how you chose parameters for training**. Demonstrating that you understand how to use a predictor is the most important goal of this project.

List of Predictors

- Decision tree: single decision tree with branches predicting labels [0,1]
- Bagged decision tree: an ensemble of decision trees obtained by either randomizing the construction of the trees, or by resampling the training set
- Neural net >=2 layers: a multilayer neural network
- Boosting: a boosted weak classifier of your choice [not included this Fall]
- K-NN: K-nearest neighbors [you can choose the distance]
- Naive Bayes [you can choose the features]
- SVM - RBF kernel: multiclass SVM with the Gaussian kernel
- SVM - other: multiclass SVM with a different kernel
- Linear regression -- you can include higher order terms
- Logistic regression -- you can include higher order terms
- Generative model: train a P(XIdigit) model separately for each class, predict the class label by Bayes' rule

For any method, you should explore the data first, and do some preprocessing. In particular, you can derive new features from the existing ones, or you can define a particular type of "distance". In addition, whenever it makes sense, it is highly recommended that you also use the raw features in the same classifier, for comparison. Tell us in the project how the raw features fared compared to the features "engineered" by you.

Software resources You are allowed to download software for this project. In this case, you must know intimately what the software is doing in the context of your project. You must also demonstrate by your project that you mastered various issues of the process of data analysis/prediction. You will be graded mostly (this will become more precise eventually) on your intellectual contribution to the project and only secondarily on the performance/sophistication of the methods borrowed from others.

Generic machine learning packages

- scikit-learn (Python)
- Weka (Java)
- pmtk3 (matlab)
- SVM packagages: SVM-torch, SVM-light, LibSVM
- TensorFlow (especially for neural nets, but other methods included) (Python)
- more TB posted

Time line

Data available	Nov 17
Choose method	Nov 19
Test set available	Dec 5 noon

Test results due	Dec 7, noon
Award ceremony	Dec 8 lecture
Submit report	Dec 12 midnight

Report outline

- Preprocessing, what feature set you used
- Predictor(s): complete model description, parametrization
- Basic training algorithm(s): what algorithm, what parameters, anything unusual you did. Do not reproduce the algorithms from books or lecture unless you make modifications.
- Training strategy. Reproducible description of what you did for training (e.g training set sizes, number epochs, how initialized, did you do CV or model selection)
- Experimental results, e.g learning curve(s), training (validation) losses, estimated parameter values if they are interpretable and plottable. Be selectivein what you show! Credit will given for careful analysis or visualization of the results.
- Estimate of the average asymmetric loss L. Optionally, an interval [Lmin, Lmax] where you believe L will be, and how you estimated these.
- Optional: references
- Total length: no more than 5 pages of contents, with extra pages containing references or figures, up to no more than 10 pages total.

In writing the report, assume that the readers (=instructor and TA) are very familiar with all the predictors and with machine learning terminology; there is no need to reproduce textbook like defintions (and there would be no space for it). What the reader need to know are the specifics of what you did with these predictors. What parameters you used for learning, what inputs, and if there were any variations from the standard methods. For example, if you use a Random Forests package, although we know what a RF is, assume we don't know what variant of RF the package implements, or what the parameters mean. You need to specify these in your report.

How to submit your test set results

- for the ~10,000 examples in the test set, you will produce ~10,000 labels with values from 0 or 1 by your Predictor. Create the file **y_out.txt** with the following format:
  predicted_false_positive_rate
  predicted_false_negative_rate
  yhat_example_1
  yhat_example_2
  ....
  (10001 lines)
  E.g. y.out
  For the value of predicted_false_positive/negative_rate you should input your best guess of how your method will perform on the training set. The values predicted_false_positive/negative_rate entered should be the averages on the test set (hence both will be in [0,1]).
  We will use a script to download and evaluate your results, so please do not vary from these format or file name, lest your prediction error be distorted. There will be two files to submit, one for your assigned predictor, and one for your choice of predictor.
- Go to the Canvas dropbox. Upload the files.
- T.B. set upEnter also: the names of the method used by your Predictor.

Marina Meila
Last modified: Sat Nov 30 15:43:52 PST 2013 >>>>>>> 9f548f6c949b352d0afee16f9376ac0a08d62fb4