NTUST, CSIE

Machine Learning (CS5087701), Fall 2017

Final Project

Guidelines

A. Each project should be done by a group of 1 to 3 students. Let me know if you have any requests to form a group of more than 3 students.

B. A project should include an oral presentation (20%), which will be scheduled soon and a written report (20%) due near end of this semester. A team needs to have each member in the group on the stage for the oral presentation. The written report should have fewer than six pages, with references included.

C. Project title due as soon as possible.

D. You can either use Matlab, R, Julia or other languages to write your own code or use software/package to help you to do experiments and evaluations. That means writing codes is not a must, but you get some credits by writing codes by yourselves.

E. Any further questions can be discussed with me.

Candidate Topics: Data oriented

You can choose from one of the following datasets and test a few machine learning algorithms on it.

I. *T*he *Bank Marketing* dataset (homework 2)
(http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

II. *T*he Spooky Author Identification (homework 2)
(https://www.kaggle.com/c/spooky-author-identification)

III. *The MNIST Database of Handwritten Digits* (homework 3)
http://yann.lecun.com/exdb/mnist/

IV. *ECML/PKDD 15: Taxi Trip Time Prediction (II)* (homework 3)
https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii

V. Your own choice of dataset & problem.
(send me one-page summary and/or discuss with me!)

Some suggestions on how to use these datasets: (you do not need to cover all, but a significant portion should be great!)

- Because the first four are the old datasets. You should extend what you already have from your homework and continue improving your result. Especially, I would like to see you consider a few items to make your analysis more complete:
  - A brief summary about what your result was and what you expect to improve (or to focus on).

- Anything you learned from your classmates' presentation, so that you can borrow and make your prediction better than before.
- Comparing to state-of-the-art approach: compare your result to others, or the state-of-the-art approach in particular.
- Feature engineering: finding any possible set of features so that you may have an improved result. You can either extend the feature set or shrink the feature set (dimensionality reduction).
- Any strategy that you can make your new model more efficiently than the one you obtained in your homework.
- Comparing to end-to-end models such as deep learners to see if the deep leaners really work better than those we have learned in class.
- Try your best to deliver a final statement about the dataset. The statement should provide some general guidelines when novices begin to analyze the dataset.

● After all, the larger the dataset, the more complicated dataset you work on, the more credits you can receive for your project.

Candidate Topics: Model oriented

On the other hand, you can study some machine learning techniques by testing the techniques on a set of datasets and draw conclusions based on the result.

I. Increasing the model complexity: How to design a model for prediction, an easy model or a complex model? One strategy could be using an easy model at beginning and gradually increasing the model complexity when we acquire more data? Can we avoid overfitting in this case? Some example can be increasing the tree depth (or tree nodes) in a decision tree model, or adding nodes to a hidden layer in a neural network model.

II. Combining generative and discriminative modeling techniques: Pick a pair of generative model and discriminative model and suggest an approach on combining the results from both parts together and create a good result that cannot be obtained by using only a single type of technique.

III. Active data/attribute learning: If we acquire the data and the attributes incrementally, how to obtain the best performance given various strategies to add data and attributes? That is, we may have a subset of data and a subset of attributes at beginning and we continuous to acquire more data and decide to collect more attributes for each of the data. How to design a good strategy so that we can add data and attributes one by one and obtain good prediction result in the end.

IV. Various learning algorithms: Perceptron, incremental or batch-mode, with or

without thresholding, logistic regression to name a few in your comparison candidates. The effectiveness may be different given different data sets. In your study, you should point out what the algorithm we should adopt given different data sets.

V. Various deep networks: regular deep networks, convolutional networks, recurrent networks, etc. It is also interesting to try the network with different depths.

The general issues to study and discussion

You should emphasize the following items in your oral and written presentations.

I. *Model Effectiveness*

The prediction accuracy is not the only way to judge how effective a model is; however, it is the most common one. A typical approach to estimate the prediction accuracy is based on a cross-validation procedure, e.g., 10-fold; with several repeats, e.g., 5 different partitions of the original training set, for a total of 50 trials. Is there any other way to measure the model effectiveness? Clearly state the evaluation measure in your experiments and explain why this measure should be used in your case.

II. *Model Complexity*

How many attributes do you use in your model? You may try to use as few attributes as possible given similar performance from your model. On the other hand, you should care about the model size. State clearly for your own definition of model complexity. Generally speaking, more complex models may have a better chance to overfit the data, or may introduce more local optima. Make sure that you do not run into the above problems in your modeling procedure.

III. *The Data Size*

In general, the more data the better for the model effectiveness. Can you see this from your experiments? At the same time, can you use a smaller subset of the whole set to make your model perform as good as you using the whole dataset?

IV. *Comparison to the baselines*

Some candidates of baseline benchmark learning methods include *k*NN and naïve Bayes (if applicable). It is good that your model is at least more effective than those models. Moreover, you can compare your model to the state-of-the-art models. On the other hand, you can also compare your model to other models in terms of model complexity. We can say a simple model is preferable than a complex model based on the above discussion.

V. *Creative ideas*

How your model is designed? Any creative preprocessing, design of experiment process is used? You may also share the ideas why you choose particular modeling techniques.

VI. *How much domain knowledge is involved?*

Usually domain knowledge helps us to improve the learning model. You can show how you use some domain knowledge to help you achieve better performance in your evaluation.

Datasets & Software

- **WEKA** http://www.cs.waikato.ac.nz/ml/weka/
- **Julia** https://julialang.org
- **OpenCV**
  - Open Source Computer Vision
  - http://opencv.org/
- **MALLET**
  - MAchine Learning for LanguagE Toolkit
  - http://mallet.cs.umass.edu/
- **MLC++**
  - Machine learning library in C++ (http://www.sgi.com/tech/mlc/)
- **Stalib**
  - Data, software and news from the statistics community
  - http://lib.stat.cmu.edu
- **GALIB** MIT GALib in C++ (http://lancet.mit.edu/ga)
- **UCI**
  - Machine Learning Data Repository UC Irvine
  - http://www.ics.uci.edu/~mlearn/MLRepository.html
- **UCI KDD Archive** http://kdd.ics.uci.edu/summary.data.application.html
- **Delve**
  - Data for Evaluating Learning in Valid Experiments
  - http://www.cs.utoronto.ca/~delve

Major journals & conferences

- PAMI (IEEE T. on Pattern Analysis and Machine Intelligence (PAMI))
- JMLR (Journal of Machine Learning Research)
- NIPS (Neural Information Processing Systems)
- ICML (International Conference on Machine Learning)
- ECML (European Conference on Machine Learning)
- UAI (Uncertainty in Artificial Intelligence)
- COLT (Computational Learning Theory)
- IJCAI (International Joint Conference on Artificial Intelligence)