EE 559 **Course Project Assignment** Posted: Thur, 4/5/2018
Jenkins Due: Tues., 5/1/2018, 11:59 AM

Introduction

The goal of this project is to develop a pattern recognition system that operates on a given real-world dataset that is nontrivial to solve. In doing so, you will apply tools and techniques that we have covered in class. You will also confront and solve issues that occur in practice and that have not been covered in class. Discussion sessions and past homework problems will provide you with some tips and pointers for this; also piazza discussions will likely be helpful.

Projects can be individual (one student working on the project) **or team** (2 students working together on one project). There are two datasets to choose from (described below). Each individual project will use one dataset (your choice of which dataset); team projects can use the Bank Marketing dataset as the only dataset, or can use both datasets (your choice).

In your final report you will describe the work that you have done, and where you describe the work of others, it must be cited and referenced as such. For team projects, the final report must also describe which parts were done by which team member. Plagiarism (copying information from elsewhere without crediting the source) of text, figures, or code will cause substantial penalty.

You will have significant freedom in designing what you will do for your project. You must cover various topics that are listed below (as "Required Elements"); the methods you use, and degree of depth you go into each topic, are up to you. And, you are encouraged to do more than just the required elements.

Everyone will choose their dataset(s) from the two datasets listed below. Collaboration and comparing notes on piazza may be helpful, and may make it more fun and engaging.

Datasets:

Choose from the following two datasets:

1. Online News Popularity

Dataset is in the D2L project folder: **`OnlineNewsPopularityReduced.csv`**

Brief description: the dataset provides several attributes from different articles published by Mashable. Your goal is to predict if an article is popular or not. The labels (popular/ not popular) are based on the number of shares that article received.

Original link (for information about the data):

https://archive.ics.uci.edu/ml/datasets/online+news+popularity#

For a 2-class problem, we suggest you define "popular" as shares > 1600; "not popular" if shares ≤ 1600; this gives a dataset that is roughly balanced. You can set a different threshold if you prefer; be sure to state so in your final report.

For a multiclass problem, we suggest using 5 classes as defined below (which gives an approximately balanced dataset); but you can use different thresholds if you clearly state so in your report.

Class	Popularity	s = # of shares
S1	1	$s \leq 900$
S2	2	$900 < s \leq 1200$
S3	3	$1200 < s \leq 1600$
S4	4	$1600 < s \leq 3400$
S5	5	$3400 < s$

Obviously, you cannot use the feature "shares" during training.

Difficulty: 3 out of 5.

2. Bank Marketing Data Set

Dataset is in the D2L project folder: **`bank-additional.csv`**

Brief description: the dataset contains information on people targeted by a bank marketing campaign. Your goal is to predict if a client will subscribe to a long term deposit at a certain bank after receiving a marketing call.

Original link (for information about the data):

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

Note that the original feature 11 (call duration) has already been deleted from the csv data file provided on D2L, because it would not be a known input in the envisioned predictive scenario.

This is inherently a 2-class problem.

Difficulty: 4 out of 5 – this dataset needs substantial preprocessing (for example, some features have 'unknown' values - how to treat this?) and it is unbalanced.

Note: the links above are only meant to provide you more information on the datasets. Please use the reduced versions of the datasets provided on the D2L site.

Computer languages and available code

You may use Matlab, Python, or C/C++. (To use another language, please ask the TA or instructor first.)

Matlab and Python are recommended, and supported.

You may use any toolbox or libraries you prefer.

For Matlab users, some common choices include PRTools, LIBSVM and PMTK.

For Python users, scikit-learn and LIBSVM are common choices.

Be sure to state in your project report what languages and toolboxes/libraries you used; what you coded up yourself specifically for this class project; and any code from other sources.

Required elements

- **Preprocessing**
  - Tip: you might find it easier to let Matlab (using the "import" button) or Python (using pandas) handle csv parsing.
  - Consider the feature types: numerical, ordered categorical, or unordered categorical. Re-cast the representation of data as appropriate. [Hint: see Discussion 11]
  - Missing data. If there is any missing data, decide how you will deal with it. [Hint: see Discussion 11]
  - Normalization. Decide whether, and how, you will normalize the data, and which features will be normalized. [Hint: see Discussion 6 and 7.]
- **Feature-space dimensionality adjustment.**
  - Use a method to try reducing and/or expanding the dimensionality, and to choose a good dimensionality.
- **Cross validation**
  - Use for choosing parameters and/or for dimensionality adjustment.
- **Training and classification.**
  - Try at least 3 different classification techniques that we have covered in class; include both distribution-free and statistical classification. Beyond this, feel free to optionally try other methods (either from those we covered in class or other pattern recognition/machine learning methods).
- **Proper dataset (and subset) usage.**
  - Final test set, training set, validation sets, cross validation.
- **Interpretation.**
  - Interpret intermediate results and final results. Can you explain (or hypothesize) reasons for) what you observe?

Evaluation of performance.

- Randomly set aside some percentage (at least 10% is recommended) of the dataset as the test set (preserving percent representation of each class) before using the data; or else use cross-validation for final error estimation. Be sure to describe in

your final report the method you used, including how test set(s) were generated and used.

- For Online News Popularity dataset, Use classification accuracy (percent) and mean F1 score as your main performance measures. Also show the confusion matrix. [See Discussion 11.]
- For Bank Telemarketing dataset, use F1 score and AUC (Area Under Curve) as your main performance measures. [See Discussion 11 and 12.]

Tips

1. Be careful to keep your final test set uncorrupted, by setting it aside at the beginning, or by using cross-validation procedures appropriately.
2. For unbalanced datasets, for many statistical classifiers you can use minimum risk as a criterion instead of minimum error [Discussion 12].
3. If possible, it can be helpful to consider degrees of freedom (d.o.f.), and number of constraints, as discussed in class. However, this is easier to evaluate for some classifiers than for others; and yet for others, such as SVM, it doesn't directly apply.
4. It's good to start out with a baseline system and result to compare with. The choice of baseline is up to you. For example, it could be assignment based on priors only (if dataset is approximately balanced), or based on a simple classifier (like Naïve Bayes) using the given feature set, or based on purely random classification.

Grading criteria

Please note that your project will not be graded like a homework assignment. There is no set of problems with pre-defined completion points. The project is open-ended, and the work you do is largely up to you. Grading criteria will include: inclusion of required elements; understanding and interpretation (of approach, algorithms used, and results); technical soundness and final performance; quantity and quality of effort; difficulty of the problem you are solving; and report write-up (clarity, conciseness, and completeness).

Final report

More detailed guidelines for the written report will be posted later.