COMP 6321: Machine Learning, Concordia University

PROJECT PROPOSAL FORM

This document serves as a template that you can either fill out and submit, *or* use as a guideline / checklist for writing your own project proposal document. Remember, this is just a proposal to demonstrate that you have put time into planning, but you are allowed to change your plans for the final project if you run into trouble or change your mind. The guidelines ("Aim for X sentences") are flexible, not hard constraints.

Group: _G16_____ (e.g. G00)

Student 1: _Pravesh Gupta (40152506)

Student 2: _Manjot Kaur Dherdi (40107905)

Student 3: _Konstantin Chemodanov (40049285)

Student 4: _Manish Yadav (40059711)

Date submitted: __10-October-2020

Propose a title for your project. If your project were written up as a research paper, what title would you give it? A good paper title will help each individual reader to know whether they should or should be interested in reading the paper. For example, the title *Intriguing properties of neural networks* (Szegedy *et al*. 2014) is a title that, although a little too vague, at least suggests that the nature of the work is an investigation, and that the focus was neural networks, and that the results are surprising. As another example, *The fastest pedestrian detector in the West* (Dollar *et al*. 2010) is a fun title indicating that the goal is "pedestrian detection" and that the nature of the contribution is "speed."

A comparative study of Classification Models in Machine Learning

Describe the goal of your project. What are you trying to achieve? What "main question" are you trying to answer, or at least to provide evidence for? Secondary goals are OK, but you should still have a clear "main goal" or "main question." From your description, it should also be clear whether your project is about: making better predictions for some application? speeding up training and/or predictions? simply comparing predictive performance and/or speed of several methods? assessing or comparing interpretability? understanding failure modes or sensitivities of some methods? Etc.

The project aims to achieve a comparative study of classification machine learning models to analyze confusion matrix for the accuracy, precision & recall, comparison using ROC curve, analyze model selection based on dataset variations, performance matrices

like speed/time taken for evaluation, feature transformation gains, generalization bound comparisons, conceptual simplicity, regularizations.

Describe the data you plan to use. One of the hardest steps for a good machine learning project is to find data that is truly suitable for your goals. Finding good data not the most fun part, but it's one of the most important—after all, for machine learning it is "garbage in, garbage out". Here are some things you should ideally know:

- What are the 'modalities' that apply to the data? (images, video, speech, text, tabular, categorical, numerical, time series, experimental measurements, etc.)
- What does an input look like? (show an example if possible, like an image, or a sound wave, or some features, or at least try to describe)
- For an example input, what does a desired output look like? (show an example if possible, or at least try to describe)
- How many training and testing samples will there be?
- How are the training and testing data to be split? (randomly shuffled, by some grouping, by time period, etc.)
- Will the data need preprocessing before you can feed it into a training algorithm? (It is OK if you are not sure, but try to guess)
- Is the data small enough to train models on your computer, or is there a risk of scalability/engineering difficulties?

For this comparative study, we have chosen 8 models and a variation of 8 different datasets. We aim to apply each model on each dataset. For now, we have analyzed 4 datasets and we aim to include similar 4 more dataset analysis with more or less similar feature engineering steps. Below is the analysis of 4 datasets that we have identified for now:

1. **Occupancy Detection Data Set-** represents time-series numerical data features captured from sensors (temperature, humidity, light and carbon dioxide percentage). https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

Input

date	Temperature	Humidity	Light	CO2	Humidity ratio	***Occupancy***
2015-02-02	14:19:00	23.7	26.272	585.2	0.004764163024164 14	1

Output: The output of the model is the one of occupancy classes (occupied or free) predicted based on features.

Samples: Dataset is already divided into training (8144 data instances) and test datasets (2666 and 9753 data instances), however it is possible to additionally split the data if needed based by time period.

Preprocessing: There will be small pre-processing needed because 2 files contain quotation symbols, and one does not.

Size: The data size should be good enough for training and testing.

2. **Bitcoin Heist Ransomware Address:** Dataset- Bitcoin Heist Ransomware Address Dataset represents multivariate timeseries bitcoin transaction data of 10 years from Jan,2009-Dec,2018.

(https://archive.ics.uci.edu/ml/datasets/BitcoinHeistRansomwareAddressDataset)

Input

Address	year	day	length	weight	count	looped	neighbors	income	***label***
19XjYC7yXsKkCxUrseWJiPc6ZNRdUG2TqN	2017	11	18	1	1	0	2	00050000	princetonCerber

Output: Output is a predicted label (white or ransom) for any input transaction.

Samples: There are 1048575 data instances which can be split into training set and test set in the ratio of 80:20.

Preprocessing: Some features may need preprocessing. Labels need to fit to two classes.

Size: Dataset is considerably large, and some reduction technique need to be employed to make it workable.

3. **Activity recognition with healthy older people using a batteryless wearable sensor Data Set:** Represents a text-based time sequenced classification data having real number based continuous signal values collected by sensors to label activity based on sensor information. Data is sparse, noisy and big for training purpose.
(https://archive.ics.uci.edu/ml/datasets/Activity+recognition+with+healthy+older+people+using+a+batteryless+wearable+sensor)

Input:

seconds	acceleration_frontal	acceleration_	acceleration_lateral	antenna_id	signal_strength	phase	frequency	***activity_label***	gender_id	person_id

		vertical,								
0	*0.27203*	*1.0082*	*-0.082102*	*1*	*-63.5*	*2.4252*	*924.25*	*1*	0	1

Output: The desired output is activity label which is a nominal category number for different type of activities (1: sit on bed, 2: sit on chair, 3: lying, 4: ambulating)

Sampling: For cross-validation purpose, K-fold splitting would be done with shuffling and 80:20 split on 90% of the data. Rest 10% will be used for final evaluations.

Preprocessing: Data does contain duplicates and those will be removed first. Columns 'seconds' may be dropped out as it does not contribute to finding the actions as of now. Columns acceleration_frontal, acceleration_lateral contain outliners which need to be regularized using scaling methods. Nominal categorical data in columns antenna_id, gender_id, person_id can be converted into more features using one-hot encoding to minimize risk of irrelevant information feeding. Binning can be employed for continuous real numbered data columns like acceleration_*, signal_strength, phase, frequency. Labels are analyzed to be slightly skewed and hence may need preprocessing there as well.

Size: After removing duplicates, data has shape of (34000, 10). More columns will be introduced by nominal data and hence we can say that data is sufficiently large for small scale comparisons.

4. **Bank Marketing Data Set-** The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Input

age	Job	Marital	Edu	Default	Balance	Housing	Loan	Contact	Day	Month	duration	Pdays	outcome
30	None	M	Primary	No	1787	No	no	Cell	19	Oct	79	Unknown	no

Output- has the client subscribed a term deposit? (binary: 'yes','no').

Samples- Total 41188 samples. randomly shuffled with 80:20 split.

Preprocessing- Remove missing values, remove duplicate values.

Size: Data is of medium size and should work on personal computer to train model.

Describe how you will measure "success." You should explain how you will know whether you have achieved the goal(s) that you described earlier. What does "success" look like? What does "failure" look like? Keep in mind that your project can still succeed (in the sense of a good grade!) even if the experimental results are bad—what is important is that your experimental results are *conclusive*! A bad project is one in which you cannot even tell whether the goal was achieved or not.

We will measure success using cross-validation strategy and find out the confusion matrix to draw ROC curve. Model having higher accuracy on test-set will be taken as more robust. Also, whether the model can contain outliers or not will also be part of the study. Same study will be done without feature transformations and accuracy will be compared with the feature transformations to find out best case implementation. After successful completion of the project we expect to know the speed of different models based on their training and evaluation time, we are planning to successfully compare models based on their performance and determine if some models work better with specific datasets.

Describe how work will be divided. It is very important for everyone to have a meaningful role in the project. If one person (the most experienced person) does all the programming or writing, then everyone else in the group loses this important chance to gain experience. For example, if there is no way to "happily divide" the work because two group members want to work on the same part, that is totally OK and no one should feel guilty for wanting that; both group members can do their own version of that part of the project, and then the final report can say "two group members each implemented did this part, and their results {matched, didn't match}" When two people attempt and come to different conclusions, that is interesting and a chance for everyone to learn!

We are a team of 4. Work will be divided equally among all.

We have 8 models and 8 datasets.

Initially each team member will train 2 models and output metrices for all datasets. At this step, we will record changes in predictions and accuracy with dataset variations for specific model.

We will be having 8 different model doing analysis on each dataset. At this point, we aim to assign 2 datasets to each member and do a model comparison analysis for each dataset and record changes in predictions and accuracy with model variations for specific dataset.

For final report, we aim to divide sections of report and do the report collaboratively.

List the main Python packages you expect to use. PyTorch? TensorFlow? Scikit-learn? Special packages for working with your data? (It is OK if this list is incomplete or changes for the final project.)

- Numpy
- Matplotlib
- Scikit-learn
- Pandas

Frequently asked questions. Below are some questions students have asked.

Q: Can we use a pre-trained model, or do we have to train a model ourselves?

A: This question applies mainly to computer vision or natural language processing. Your project must involve training one or more models. However, you are definitely allowed to incorporate pre-trained models in that effort. For example, if you download a pre-trained model and then use it to convert your raw training data into more a more useful representation, then that is OK, but (a) you should still train new models on top of that representation for the task you are trying to solve and (b) you should consider training simple baselines (e.g. a linear model, a random forest) on your raw data, to demonstrate that the pre-trained model was important for performance.

Q: Are bigger groups expected to do more ambitious projects?

A: No. But in bigger groups there is a higher chance that the least-experienced group member will be "left out" of important programming or writing activities. So please be conscious of this and give everyone a chance to learn.

Q: Can we try machine learning algorithms beyond what we learn in the course?

A: Yes absolutely. If you want to try reinforcement learning, that is OK. But you should still try your best to apply some of the methods we've learned about—even if they are not a natural fit to your "task" and you expect them to perform poorly, you should try. This can also be useful for dividing work among group members: some can try to apply fancier methods outside the course, where other group members can try to apply the basic methods even if the results are not expected to be "state of the art."

Q: How do we find data?

A: Here are some thoughts:

- Google something you are interested about, like "climate change datasets for machine learning" and you may get lucky.
- You can try to look at machine learning data set repositories, such as the UCI machine learning repository or the OpenML repository.
- You can use or create synthetic data, generated by simulators or other software that you are capable of running on your computer. For example, if the main question of your project was "can we use ML to approximate the output of <insert

software here>” then you may pursue this approach. (However, notice that basic machine learning algorithms tend to produce fixed-dimensional outputs, whereas most software produces variable-length outputs, so defining a good project along these lines takes time!).

- You can look for “challenges” or “competition tracks” that have been hosted as part of machine learning conferences, or as part of a Kaggle challenge.

Be careful about the size of the data sets. If you choose to train a model directly on a huge data set (a huge text corpus, or 3D medical images, or the ImageNet dataset), then it is very likely that some group members cannot participate in the training because they do not have GPUs, and so most group members do not have a good experience or are stuck waiting.