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: \_\_G03\_\_\_\_ (e.g. G00)

Student 1: \_Divya Bhagavathiappan Shiva\_\_\_\_\_\_

Student 2: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student 3: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (if applicable)

Student 4: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (if applicable)

Date submitted: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*](https://arxiv.org/abs/1312.6199) (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*](https://authors.library.caltech.edu/94253/) (Dollar *et al*. 2010) is a fun title indicating that the goal is “pedestrian detection” and that the nature of the contribution is “speed.”

Human Activity Classification: A comparative study between different preprocessing approaches and classifiers

**(Divya) 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.

[Type your answer here. Aim for 3-6 sentences.]

This project aims to classify human activities using data obtained from accelerometer and gyroscope from phone and watch. The raw data will be preprocessed using two different approaches such as topological data analysis and statistical features extraction from time segments. The goal is to compare and evaluate the performance of different classifiers (Decision Tree, k Nearest Neighbors, Random Forest, SVM and NN) which are trained on the two sets of preprocessed data.

RoadMap

In this paper, “*Gary M. Weiss, Kenichi Yoneda, and Thaier Hayajneh. Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living. IEEE Access, 7:133190-133202, Sept. 2019*.”, Weiss et al has trained three models namely Decision Tree, k-Nearest Neighbors, and Random Forest for human activity classification by preprocessing the raw data using statistical features extraction from time segments. In this project we aim to do the following.

1. Train the same models - Decision Tree, k Nearest Neighbors, and Random Forest using the preprocessed data obtained from topological data analysis and compare and evaluate the performance between the two preprocessing approaches.
2. Train SVM and NN using the preprocessed data generated by Weiss et al and evaluate the performance against their Decision Tree, k Nearest Neighbors, and Random Forest models.

**(Reethu and Mohammad) 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?

[Type your answer here. Be concise. Aim for half a page.]

* *What are the ‘modalities’ that apply to the data? (images, video, speech, text, tabular, categorical, numerical, time series, experimental measurements, etc.)*

The dataset we will be working on is the WISDM (Wireless Sensor Data Mining) dataset, which is available from the UCI machine learning repository of the same name.

The dataset includes time-series data collected from the accelerometer and gyroscope sensors of a smartphone and smartwatch.

To collect this dataset, 51 subjects were asked to perform 18 different activities in 3 minutes. These activities include basic Non-hand-oriented activities, hand-based daily life activities, and various eating activities.

The following figure shows a graphical diagram of a smartphone's triaxial accelerometer data.



Time (seconds)

Acceleration (m/s2)

Image credit: GM Weiss, K Yoneda, T Hayajneh, “[Smartphone and smartwatch-based biometrics using activities of daily living](javascript:void(0))”

* *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)*

Each line of the time-series sensor file is considered as our input.

The format of this line is identical across both types of sensors and it would be as follows:

*Subject-id, Activity Code, Timestamp, x, y, z;*

* *Subject-id;* is a symbolic numeric identifier that represents the subject and is a single value between 1600 and 1650. (Each number belongs to one of the 51 subjects)
* *Activity Code;* is a symbolic single letter which identifies one of the 18 activities represented in dataset. (A-S except N value)
* *Timestamp;* is a Linux time displayed by an integer.
* *x, y, z*; are sensor values for x, y, z axes respectively.

1600, A, 252207666810782, -0.36476135, 8.793503, 1.0550842;

*Subject-id*

*x value*

*Activity Code*

*Timestamp*

*z value*

*y value*

*Ex:*

* *For an example input, what does a desired output look like? (show an example if possible, or at least try to describe)*

Considering that our goal is to classify human activities based on the time-series data obtained from the subject, our expected output will be the name of one of the 18 activities performed.

* *How many training and testing samples will there be?*

The dataset contains a total of 15,630,426 lines/sensor readings.

Each smartphone and smartwatch has its own accelerometer and gyroscope sensors.

Thus, there are four sensors and the number of lines of data per each sensor is as follows:

* Smartphone accelerometer: 4,804,403
* Smartphone gyroscope: 3,608,635
* Smartwatch accelerometer: 3,777,046
* Smartwatch gyroscope: 3,440,342

In addition to the raw time-series data, the dataset includes labeled examples.

The transformation process for generating the labeled examples has been carried out using a sliding window approach.

The raw time-series data for each sensor has been divided into 10-second non-overlapping segments, and then high-level features have been generated in each section.

As a result of the transformation process, 93 features have been extracted from the dataset and the transformed examples were placed into ARFF (Attribute-Relation File Format) files.

The number of number of examples per each category is as follows:

* Smartphone accelerometer: 23,173
* Smartphone gyroscope: 17,380
* Smartwatch accelerometer: 18,310
* Smartwatch gyroscope: 16,632
* *How are the training and testing data to be split? (randomly shuffled, by some grouping, by time period, etc.)*

**Topological approach**; first we need to preprocess the data. Then, we will split the data into two training and test sets. and …

**Statistical feature extraction approach**; given that we have the id of each subject, we consider to leave out a subject and ...

* *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)*

For the topological preprocessing approach, we need to …

* *Is the data small enough to train models on your computer, or is there a risk of scalability/engineering difficulties?*

[…]

**(Mahsa) 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.

[Type your answer here. Aim for 3-6 sentences.]

To evaluate our work, we will use two cross validation techniques.

Based on our aim #......, which is comparing our pre-processing approach with the one that researchers in [ …] applied, we must utilize the same validation approach that they used in order to be able to compare. Therefore, we will use 10-fold cross validation approach to evaluate our project.

However, there is an issue with 10-fold cross validation method in our case, it is prone to overfitting […]. Because features obtained from neighbor windows are not independent and identically distributed (i.i.d) and 10-fold cross validation shuffles the feature set and randomly choose some windows for training and others for testing, therefore, one window may appear in the training set and its neighbor may appear in the test set; then it seems like same window was used for both training and testing. That is the reason why we plan to also apply another cross-validation approach. We are planning to apply leave-one-subject-out (LOSO) cross validation to evaluate the performance of SVM and NN models. LOSO approach is less prone to overfitting.

Success in our project is getting higher evaluation metrics such as accuracy, when using topological data analysis as our preprocessing approach.

**(Later) 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!

[Type your answer here. High-level description only, like “Angela will train the neural networks, and Seyyed will preprocess the data and train the SVM. Both will write the report.”]

1. For Topological:

* process the raw data
* create persistence diagrams
* classify using k nearest neighbor
* classify using random forest
* classify using decision tree
* evaluating different models

1. For statistical features:

* classify using SVM
* training a recurrent neural network, or a convolutional network that convolves over time (very similar to RNN).
* classify using NN
* evaluate results

1. Writing final report

**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.)

[Type your answer here.]

* Scikit-Learn
* Pandas
* Matplotlib
* Numpy
* TensorFlow
* Keras
* Scikit-tda

**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.

1. Preprocess the data using the topological data analyze and feed them to different models. Is it appropriate to apply this preprocessing methodology on time series?

*I am not sure, I'm not experienced in topological data analysis and am not sure what the goal is in this scenario.*

2. Train the three models mentioned in the attached paper (Decision Tree, k Neighbors, and Random Forest) and compare our preprocessing approach with their approach.

*Sounds good. Please compare to the original results. If your results are better or worse, this is possible, but please try to provide hypotheses as to why that's the case.*

3. Train SVM and NN models using the pre-processed dataset mentioned in (A.1) and report the best performance.

*OK. Since the data is time-series where you need to provide a single output at the end, you might consider training a recurrent neural network, or a convolutional network that convolves over time (very similar to RNN). For example, you can run a convnet over a variable-length input sequence of (x,y,z) features (a 1-dimensional convolution) and then take the resulting variable-length feature maps and aggregate over time (e..g min, avg, max) to get classification output.*

4. Using leave one subject out cross validation approach instead of 10-fold CV, since the dataset is not i.i.d.

*Yes that makes sense. Do not train and validate/test on the same subject.*

1. Comparing different window sizes

*OK*

2. Using the transformed data and train NN and SVM models and then compare its results with their models.

*OK*