CONCORDIA UNIVERSITY

Department of Computer Science and Software Engineering

Machine Learning

COMP 432

Project Proposal

Zhaoyang Li 27838824

Weichen Wan 40072743

Yufei Li 40165883

Concordia University

November 5, 2022

1 Title

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." [Type your answer here.]

Weather Prediction with Machine Learning

2 Goal

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.]

Weather is the most prevalent topic in small talk across the full world. This project will attempt to answer these questions for the city of Montreal. The first being whether it'll rain/snow and the second being the amount of rain/snow that will fall from the sky in a given day.

A common approach to this type of questiony is to build two separate models. One for classification to answer the first question and one for regression to answer the second. After fitting the two models the prediction process, is as follows:

$$\mathbb{E}[\text{Precipitation}|X] = \begin{cases} 0 & , \mathbb{P}[\text{Precipitation}|X] < \text{Threshold} \\ \text{Regression Prediction} & , \text{Otherwise} \end{cases}$$

In words, first the classification model will be used to probability of precipitation for a given day. If the probability of precipitation is above a threshold (usually 0.4 or 0.5), then the prediction of the amount of precipitation from the regression will be generated. If it is below the threshold, the value of zero will be the prediction.

This approach is chosen for the majority of the modelling techniques learned in this course, as well as machine learning classification techniques.

3 Dataset Description and Manipulation

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 the desired output look like?
- How many training and testing samples will there be? Can some be realistically trained on a laptop, or is something more powerful needed?
- Anything special about how the training and testing data should be split?
- For an example input, what does the desired output look like?
- Might the data need preprocessing before you can use it?

[Type your answer here. Be concise.]

3.1 Description

We concern ourselves with giving day predictions for the weather station located at the Montreal International Airport. Data was chosen between 2013-2018 due to the quality of daily and hourly data being high, when compared to other weather stations.
Here is a list of relevant terminology. For further reference, see weather data source.

Dew Point Temperature (C) The dew point temperature in degrees Celsius (C), a measure of the humidity of the air, is the temperature to which the air would have to be cooled to reach saturation with respect to liquid water. Saturation occurs when the air is holding the maximum water vapmy possible at that temperature and atmospheric pressure.

Heating degree-days Heating degree-days for a given day are the number of degrees Celsius that the mean temperature is below 18 C. If the temperature is equal to or greater than 18 C, then the number will be zero. For example, a day with a mean temperature of 15.5 C has 2.5 heating degree-days; a day with a mean temperature of 20.5 C has zero heating degree-days. Heating degree-days are used primarily to estimate the heating requirements of buildings.

Cooling degree-days Cooling degree-days for a given day are the number of degrees Celsius that the mean temperature is above 18 C. If the temperature is equal to or less than 18 C, then the number will be zero. For example, a day with a mean temperature of 20.5 C has 2.5 cooling degree-days; a day with a mean temperature of 15.5 C has zero cooling degree-days. Cooling degree-days are used primarily to estimate the air-conditioning requirements of buildings.

Gusts Gusts are sudden, rapid and brief changes in the wind speed. They are characterized by more or less continual fluctuations between the high (peak) and low (lull) speed. The extreme gust speed is the instantaneous peak wind observed from the anemometer dials, abstracted from a continuous chart recording, or from a data logger.

Pressure (kPa) The atmospheric pressure in kilopascals (kPa) at the station elevation. Atmospheric pressure is the force per unit area exerted by the atmosphere as a consequence of the mass of air in a vertical column from the elevation of the observing station to the top of the atmosphere.

Description of variables that were included on a daily basis (all temperatures are in Celsius):

- `max_tmp` - the maximum temperature
- `min_tmp` - the minimum temperature
- `mean_tmp` - the average temperature
- `mean_RH` - the average relative humidity expressed as a percentage.
- `mean_DP` - the average dew point temperature
- `mean_press` - the average station pressure
- `mean_Wind_speed` - the average wind speed (km/h) usually 10 kilometres above the ground

- `Heat_Deg_Days` - See Definition of Heating Degree-Days
- `Cooling_Deg_Days` - See Definition of Cooling Degree-Days
- `spd_max_gust` - the maximum gust speed (km/h) recorded in a day
- `Precipitation` - the total precipitation (mm) recorded in a day

3.2 Dataset Manipulation

To be able to perform cross-validation, the dataset has been divided into two parts. The training set contains 1421 observations, the testing set contains 391 observations.

3.2.1 Features Selection

Correlation Before fitting any models, it would be ideal to check the relation between all features, there are two main reasons,

- If two features are highly linear correlated, the model may have multi-collinearity effect.
- Since we are going to fit some of the models like a neural network which contains a lot of hyper parameters, by reducing the number of input, we may obtain less parameters to be tuning when fitting model and consequently reduce the computational load.

The set up in this analysis of features is to rule out those features that have a correlation coefficient of more than 0.7 in absolute value.

3.2.2 Standardization and Scaling

Since our features are expressed in different measurement scales, we standardize or scale the features based on the following:

- If the feature can take either positive or negative values, then we standardize it.
- If the feature can only take positive values, then we scale it by dividing the range.

3.2.3 Artificial Feature

We are going to add one feature called `rain` base on the existed feature `precipitation` which serves as a label for the precipitation. If for a given day, there is positive precipitation, we can consider that day as raining/snowing, and label `rain = 1`, otherwise we label `rain = 0`.

3.2.4 Merging

Since we are trying to predict the o given day (either rain or precipitation) by considering all features of the previous 7 days, it would convenient to rebuild a bigger dataset by stacking all the features in the past 7 days as the features for predicting the 8 th day. Note that we also included the raining and the precipitation record of 7 days as a set of new features of predictors, and we believe that this is a valid inclusion.

By rebuilding the dataset, we have $6 \times 7 + 2 \times 7 = 56$ features for predicting the 8th day.

4 Analyses

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.]

In this section, we will provide several methods to perform the analysis. For comparison purpose, we also include the traditional times series model. For the machine learning models, we will use some common evaluation metric and cross validation to determine some models.

4.1 Evaluation Metric and Cross Validation

4.1.1 Evaluation Metric

We use the common evaluation metric described as follows,

- For classification, We use *Accuracy*[1] plus *Confusion Matrix*
- For regression, We use *Mean Square Error*.

4.1.2 Cross Validation

Classification My plan is: We use a sequential cross validation based on month. We split the training data by month, then we train the model with the first month and use the trained model to predict the next month, we record the number of errors we made. Again, we use all previous months to train the model and make prediction for the new month. As the procedure goes on when running all the training data, we obtain total number of out-of-sample error count, and compute the error rate accordingly.
We choose the model with smallest error rate and use the setting of the model to train all data in the training set and use it to predict the data in the testing set and obtain the error rate.

Regression The regression applies similar strategy of sequential cross validation described above for classification but slightly different when training the model. Since we are trying to

[1] Accuracy = 1 - Error Rate

predict the precipitation, we only use the observations with non-zero precipitation, i.e. the days with rain or snow. Then we make prediction for every single day of the next month. Of course, we are going to set negative precipitation to zero, then we compute the sum of square error. Similarly, we choose the model with least sum of square error and train the whole subset of training data with only raining or snowing days and predict the precipitation in the testing set and obtain MSE.

In summary, we use a subset of all previous months to train the model to predict the next month.

4.2 Regression and Neural Network

In this section, we are going to use the linear regression model and neural network to preform two kind of predictions, the classification and the regression.

4.2.1 Logistic Regression

The logistic regression is to perform the first step of the prediction, i.e. whether a given day will rain (snow) or not. For such a simple logistic regression, my learning strategy is designed as follows,

- We train logistic regression model with the train data and obtain parameters.
- We obtain predictions of training data from the model, and choose the classification threshold with minimum error rate.
- We use the trained model and the optimized threshold to predict the test data and evaluate the error rate.

4.2.2 Linear Regression

The linear regression to predict the precipitation is similar We have to sub-setting the days with precipitation and use the subset to train the model.

- We train linear regression model with the train data and obtain parameters.
- We obtain predictions of training data from the model.
- We use the trained model to predict the precipitation in the test data.
- To avoid negative prediction values, take the maximum of 0 and the predicted precipitation.

- We evaluate the prediction by MSE.

Note that, in this case, misclassification of raining day and non-raining day have, practically, no difference, so we do not assign cost of misclassification.
The training MSE and testing MSE are summarized below,

4.2.3 Classic Neural Network

Single Layer Neural Network For single layer Neural Network, with consideration of the computational load, we only try hidden unit in the set $S = \{10, 20, \ldots, 150\}$. For each hidden unit in the set, to obtain the error rate, we should run approximately 48 times of a neural net with same level of hidden unit by the cross validation strategy mentioned above.

4.3 Other Analyses

We will also perform some of the traditional analyses for the classification, such as *Random Forest, Adaptive Boosting* and *K-Nearest Neighbor*

5 Work load

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."]

Zhaoyang will fetch and preprocess the date.
Weichen will train regression model.
Yufei will train NN model.
All of us will write the report.

6 Python package

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.]

Matplotlib, NLTK. Pandas, Seaborn, Numpy, Keras, Scipy, PyTorch, TensorFlow, Scikit-learn should be good for now, we will add more packages in the future if necessary.