DATA SCIENCE FINAL PROJECT

Weather Data set

November 6, 2019

Please use these data sets and complete following tasks. You need to write a report where you explain the preprocessing/data exploration steps for the data, the method(s) you used, the result you received (plots and/or numbers). Remember, you need to return the report as a single PDF file (5-8 pages) and you do not need to include your code. You are allow to use different method but you need to be able to explain the method in a way that your peer students could understand. You may use all the material provided during the course, build-in function in different packages, and online sources. Just remember to cite the sources outside the course sources. The outline of the report is given in the slide "Final project and Peergrade".

1 Description of the data set

The data have been downloaded from the website `http://rp5.ru/`. The data are measurements collected by the weather station 2978 in Helsinki from September 2006 to May 2019 (the data have been downloaded on 22.05.2019). Complete explanations for the original labels can be found at `http://rp5.ru/archive.php?wmo_id=2978&lang=en`.

The original data have been resampled by day, and they have been simplified in the following way:

- Index:
  - "datetime" is the index, indicating the date in the format YYYY-MM-DD.
- Numerical attributes:
  - "T" is the air temperature, in degrees Celsius, 2 meters above the earth's. surface.
  - "Po" is the atmospheric pressure at weather station level, in millimeters of mercury.
  - "P" is the atmospheric pressure reduced to mean sea level, in millimeters of mercury.
  - "Ff" is the mean wind speed at a height of 10-12 meters above the earth's surface, in meters per second.
  - "Tn" is the minimum air temperature, in degrees Celsius, over the past day.
  - "Tx" is the maximum air temperature, in degrees Celsius, over the past day.
  - "W" is the horizontal visibility, in km.
  - "Td" is the dewpoint temperature at a height of 2 meters above the earth's surface, in degrees Celsius.
  - "U" is the relative humidity, in percentage, 2 meters above the earth's surface.
  - "OBSERVED" is a categorical variable, where 0 (not dry) indicates that the amount of precipitation was more than 0.3 millimeters, and 1 (dry) indicates that there was little or no precipitation.

The decimal separator used in the data is comma (','). For the numerical attributes, the mean and the variance values for each day are provided. For example, "T_mu" indicates the mean of the air temperature, and "T_var" indicates the variance of the air temperature.

The data have been split in train and test (approximately 70% of the data are the training set, and 30% are the test set), and are divided in 4 .csv files:

- Weather_data_train: contains 3140 observations, and 16 columns.
- Weather_data_train_labels: contains the labels 'U_mu' and 'OBSERVED' for these 3140 observations.
- weather_data_test: contains 1346 observations, and 16 columns
- weather_data_test_labels: contains the labels "U_mu" and "OBSERVED" for these 1346 observations

Here an example showing how to read the files in pandas:

```
In [222]: X_train = pd.read_csv('data/weather_data_train.csv', index_col='datetime',
                                sep=';', decimal=',', infer_datetime_format=True)
          print(X_train.shape)
          X_train.head(3)

          (3140, 16)
```

Out[222]:

datetime	T_mu	Po_mu	P_mu	Ff_mu	Tn_mu	Tx_mu	VV_mu	Td_mu	T_var	Po_var	P_var	Ff_var	Tn_var	Tx_var	VV_var	Td
2006-09-20	14.4875	751.3000	751.6375	3.500	13.30	15.95	11.425	12.550	0.926964	1.008571	0.979821	1.142857	0.320	4.205	155.590714	1.99
2006-09-21	14.1875	758.0625	758.3625	3.625	11.20	15.95	27.500	11.025	4.801250	7.965536	7.679821	0.267857	5.780	6.125	147.142857	1.94
2006-09-22	15.3000	762.1125	762.4375	3.000	13.15	16.70	12.875	12.875	3.754286	1.824107	1.742679	0.857143	1.445	10.580	23.553571	0.12

Please complete the following tasks and include your results and analysis in your report:

2 Data exploration

Before starting, explore your data. You are required to complete the followings;

- Histograms: Plot the histograms of Tn_mu and Tx_mu on the same plot: is it what you expect? Plot also other histograms that might be interesting for you and your analysis.
- Pair plots: Plot the pair plots for T_mu, P_mu, Td_mu, Ff_mu, VV_mu, U_mu. Plot also other pair plots that might be interesting for you and your analysis
- Correlation: Plot a correlation matrix of the features
- PCA: Prepare a PCA projection using the first 2 components.Plot cumulative explained variance.

3 Regression

The goal of this task is to predict the relative humidity based on the available measurements. You should complete the following; You can A

- Try to understand and explain which features are the strongest predictor for humidity and why.
- Try linear model. If you want, you can also try a quadratic model. Explain your result(s). If you try different models, remember to compare them and explain all of them.
- Adopt PCA to reduce the dimension of your data set. Select the right number of components by checking the variance explained by components. Use the projected train and test set to predict the relative humidity. compare your result obtained after adopting the PCA with without PCA and explain your results/observation.

4 Classification

The goal of this task is to classify the weather conditions either as “dry" or “not dry", based on the available measurements. Complete the following for this part:

- Try KNN-classification, describe how you choose the optimal number of neighbors and plot the confusion matrix. You might want to try other classification methods (for example, you can try to apply logistic regression, or some other method) and compare the outcomes.
- Since there many strongly related features, you might want to reduce the dimensionality (for example using PCA, or using some other method).Although it is not mandatory for this task. Explain how you choose the number of features (or components in the case of PCA. you may want to look at the plot of cumulative explained variance.)
- Compare your results from before and after feature selection (PCA), and try to explain your results and observations.

Tips: You may find the following commands and methods useful. From “sklearn.linear_model" use “LogisticRegression". See their documentations and online sources for more details information.

After finishing all the tasks, write a report based on the result you obtained in different part and upload it to the Peergrade platform.