Assignment 2: Machine Learning for Wind Energy Forecasting

General Consideration

Assignment 2 concentrates on wind energy forecasting by using different machine learning techniques. The work to undertake involves data pre-processing, training model building, algorithms implementation in your favorite programming language Python/R/Java/Matlab/etc., discussion of results, as well as presentation of the work in a report.

The expected outcome of Assignment 2 includes:

- a report of maximum 10 pages (excluding appendices)
- all **ForecastTemplate.csv** files
- code as supplementary material

Assignment 2 is to be performed in groups, where a group can consist of 4 students.
The evaluation of Assignment 2 will count for 15% of the grade. All students in the same group receive the same grade.

Description of the Assignment

Wind power forecasts are very critical for decision-making problems in electricity markets, e.g., energy load balance, electricity price, and power grid stability. It is therefore very important to understand how these forecasts can be generated.

You will generate wind power forecasts for a given period based on a long history of past cases (nearly 2 years) to learn from. The historical data includes the weather data, and the power observations that were eventually collected at the wind farm. Therefore, we can build a training model to learn the relationship between weather and the power eventually produced at the wind farm. You also have weather forecasts as input. Based on the relationship learned from the past, you can apply it to the weather forecasts input to forecast wind power generation over the evaluation period.

The data for the assignment includes real wind power data (normalized by the wind farm nominal capacity, to preserve anonymity) for a wind farm in Australia. The weather input information is from the European Centre for Medium-range Weather Forecasts (ECMWF - ecmwf.int), the world-leading research and operational weather forecasting center. In the wind energy forecasting, the input data consists of wind forecasts at 2 heights (10m and 100m above ground level). Wind forecasts are given in terms of their zonal and meridional components, which correspond to the projection of the wind vector on West-East and South-North axes, respectively. The data and the various files provided for the assignment are described below.

Description of data files

The files to be used for this assignment have all the necessary input weather forecasts, past wind power measurements, as well as templates for submitting the forecasts, for the assignment. The various files include:

TrainData.csv: This file gives the set of data that can be used to find a relationship between weather forecasts inputs (wind speed forecasts at 10m and 100m above ground level) and observed power generation. These data cover the period from 1.1.2012 to 31.10.2013.
Variables (columns):

- TIMESTAMP: time stamps giving date and time of the hourly wind power measurements in following columns. For instance, "20120708 13:00" is for the 8th of July 2012 at 13:00
- POWER: measured power values (normalized)
- U10: zonal component of the wind forecast (West-East projection) at 10m above ground level
- V10: meridional component of the wind forecast (South-North projection) at 10m above ground level
- WS10: wind speed at 10m above ground level
- U100: zonal component of the wind forecast (West-East projection) at 100m above ground level
- V100: meridional component of the wind forecast (South-North projection) at 100m above ground level
- WS100: wind speed at 100m above ground level

WeatherForecastInput.csv: This file gives the set of input weather that can be used as input to predict wind power generation for the whole month of 11.2013. These include wind speed forecasts at 10m and 100m above ground level. These data cover the period from 1.11.2013 to 30.11.2013.
Variables (columns):

- TIMESTAMP: time stamps for the wind forecasts
- U10: zonal component of the wind forecast (West-East projection) at 10m above ground level
- V10: meridional component of the wind forecast (South-North projection) at 10m above ground level
- WS10: wind speed at 10m above ground level
- U100: zonal component of the wind forecast (West-East projection) at 100m above ground level
- V100: meridional component of the wind forecast (South-North projection) at 100m above ground level
- WS100: wind speed at 100m above ground level

Solution.csv: This file gives the true wind power measurements for the whole month of 11.2013, which will be used to calculate the error between your forecasts and the true measured wind power.
Variables (columns):

- TIMESTAMP: time stamps for the wind power measurements, corresponding to the forecasts to be compiled in **ForecastTemplate.csv**
- POWER: true wind power measurement (normalized)

ForecastTemplate.csv: This file gives the template for submitting your forecasts for the whole month of 11.2013. For each question, this file should be accordingly renamed. Variables (columns):

- TIMESTAMP: time stamps for the wind power forecast values to be generated
- FORECAST: your forecast values

Tasks

1. We focus on the relationship between wind power generation and wind speed. Based on the training data from 1.1.2012 to 31.10.2013 in the file **TrainData.csv**, you apply machine learning techniques to find the relationship between wind power generation and wind speed. Here, we only use the wind speed at 10m above ground level. *Note that, through this project assignment, we only use weather data forecasting at 10m above ground level.* The machine learning techniques include: linear regression, k-nearest neighbor (kNN), supported vector regression (SVR), and artificial neural networks (ANN). Apparently, each machine learning technique has a different training model. Next, you can find the wind speed for the whole month of 11.2013 in the file **WeatherForecastInput.csv**. For each training model and the wind speed data, you predict the wind power generation in 11.2013 and save the predicted results in the files: **ForecastTemplate1-LR.csv** for the linear regression model**; ForecastTemplate1-kNN.csv** for the kNN model; **ForecastTemplate1-SVR.csv** for the SVR model and **ForecastTemplate1-NN.csv** for the neural networks model. Finally, you evaluate the prediction accuracy. You compare the predicted wind power and the true wind power measurements (in the file **Solution.csv**). Please use the error metric RMSE to evaluate and compare the prediction accuracy among the machine learning approaches.

2. Wind power generation may be not only dependent on wind speed, it may be also related to wind direction, temperature, and pressure. In this question, we focus on the relationship between wind power generation and two weather parameters (i.e., wind speed and wind direction). First, you may have noticed the zonal component U10 and the meridional component V10 of the wind forecast in the file **TrainData.csv**. You can calculate the wind direction based on the zonal component and meridional component. Then, you build Multiple Linear Regression (MLR) model between wind power generation and two weather parameters (i.e., wind speed and wind direction). Finally, you can predict the wind power production for the whole month 11.2013; based on the MLR model and weather forecasting data in the file **WeatherForecastInput.csv**. The predicted wind power production is saved in the file **ForecastTemplate2.csv**. You compare the predicted wind power and the true wind power measurements (in the file **Solution.csv**) by using the metric RMSE. You may also compare the prediction accuracy with the linear regression where only wind speed is considered.

3. In some situations, we may not always have weather data, e.g., wind speed data, at the wind farm location. In this question, we will make wind power production forecasting

when we only have wind power generation data; and we do not have other data. That is, in the training data file **TrainData.csv**, the following columns should be removed: U10, V10, WS10, U100, V100, WS100. In the new training data file, we only have two columns: TIMESTAMP and POWER, which is called as time-series data. We will apply the linear regression and recurrent neural network (RNN) techniques to predict wind power generation. You predict the wind power generation in 11.2013 and save in the files: **ForecastTemplate3-LR.csv** for the linear regression model and **ForecastTemplate3-RNN.csv** for the RNN model. Finally, you compare the predicted wind power and the true wind power measurements (in the file **Solution.csv**). You evaluate the prediction accuracy using RMSE. You may use a table to compare the prediction accuracy among the two machine learning approaches.

Structure and contents of the report to be delivered

The report for the assignment should include:

- For question 1, please use a table to compare the value of RMSE error metric among all four machine learning techniques. Please make comments why there are difference between the results;
- For question 1, for each machine learning technique, please plot a figure for the whole month 11.2013 to compare the true wind energy measurement and your predicted results. In each figure, there are two curves. One curve shows the true wind energy measurement and the other curve show the wind power forecasts results;
- For question 2, please plot a time-series figure for 11.2013 which has three curves. One curve shows the true wind energy measurement, the 2nd curve shows the wind power forecasts results by using linear regression, and the 3rd curve shows the wind power forecasts results by using multiple linear regression. In addition, please use a table to compare the prediction accuracy by using linear regression and multiple linear regression, and make comments if wind direction plays an important role in wind power prediction;
- For question 3, please explain the training data for the linear regression model and the RNN model. Please explain about how you encode the data as the input and the output in the linear regression mode and in the RNN mode. Please plot a time-series figure for the whole month 11.2013 which has three curves. One curve shows the true wind energy measurement, the 2nd curve shows the wind power forecasts results by using linear regression, and the 3rd curve shows the wind power forecasts results by using RNN. Then, please use a table to compare the forecasting accuracy.
- The code provided separately

Delivery of the Assignment

Assignment 2 is to be sent to the following email
Email: yanzhang@ifi.uio.no and hweiminc@ifi.uio.no
Submission form: the submission should be in a ZIP file with naming convention "INF5870-Assignment2 - GroupX.zip", where "X" is the group number.
Email subject: "[INF5870] Assignment 2 submission by Group X"
Firm deadline: 18 May 2018 (whole day included)

Questions? please contact HweiMing Chung. Email: hweiminc@ifi.uio.no; office: 4161

Reference

1. Example source codes in R for three machine learning techniques: linear regression, kNN and SVR. Two files: **LR-kNN-SVR-forHousePrice.R** builds the three different training models based on the data in the file **HousePriceData.csv.** Then, we can use the model to predict the prices of other houses with different size.
2. Examples source code in R for neural network. Two files: **NNforBostonHousePrice.R** builds the training model based on the data in the file **Boston_House.csv.** Then, we use the model to predict the prices of other houses in Boston.
3. Reference for questions 3: page 237-240 in the book "Deep Learning and Neural Networks" by Jeff Heaton. You can find these pages in the file ReferenceforQuestion3.pdf.