ISEP
IG.3510 Machine Learning
2022-2023

Tutorial course 1 : Introduction to Machine Learning

Patricia CONDE-CESPEDES

1 Machine Learning problems

Exercise 1. Explain whether each scenario is a classification or regression problem, describe the variables (input and output) as well as the type of variables. In the case of a qualitative variable, indicate the categories. Provide the number of observations n and the number of predictors p.

a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry (retail and wholesale) and the CEO salary. We are interested in understanding which factors affect CEO salary.

b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, the price charged for the product, marketing budget and the competition price.

c) We are interested in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Exercise 2. The Netflix prize
The *Netflix prize* is a competition started in October 2006. The training dataset consisted in ratings for 17,000 movies by 400,000 Netflix customers, each rating between 1 and 5.

A prize of $1,000,000 was offered for a contestant who could produce an algorithm that was 10% better than NetFlix's own algorithm at predicting movie ratings by users.

Netflix's original algorithm, called *CineMatch* achieved a root MSE of 0.953. The first team to achieve a 10% improvement would win one million dollars.

Finally, the prize was awarded in Sept., 2009, they improved the prediction in 10.06%.

Is this a supervised or unsupervised problem ? What if we are interested in detecting groups of customers that like similar kinds of movies ?

2 Application : the bias-variance trade-off

2.1 Machine Learning with Python

You can download python from the following website : `https://www.python.org/`. You will need to install version Python 2.7 or a later version. You have the following options to run Python :

1. Use a text editor to write your code and save it using file extension *.py*. Then, to run your file, you should use a Python interpreter. Open the command prompt and use the command `python` to invoke the Python interpreter. For example, if your file name is *TD1.py*, the command to type is :

```
$ python TD1.py
```

2. Alternatively, you can use IPython, an interactive prompt to the python interpreter. To start a new IPython session, run the following command :

```
$ ipython
```

3. Another option is to use a jupyter notebook. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. To open it you have to open a terminal and type :

```
$ jupyter notebook
```

If you want to know the current working directory or change it, you can type :

```
import os
os.getcwd()
os.chdir('/Users/path_to_directory')
```

Hint : if you need help about a function in python, you can type `help(name_of_function)`. You can also see the links given in the references.

2.2 Choice of model complexity and bias-variance trade-off

The file *train.txt* contains 10 observations of two variables X (the input) and Y (the output). We assume the relationship between X and Y is the following :

$$Y = f(X) + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{with } \sigma = 0.25.$$

We want to estimate the function $f(X)$ that provides accurate estimations of Y.

1. *Pandas* is a library providing data analysis tools in Python. The function `read_csv()` from the *Pandas* library allows to read external *.txt* and *.csv* files. For instance, run the following code to import the file *train.txt* which contains the training data :

```
import pandas
data_train=pandas.read_csv('train.txt',sep=';')
```

You can previously visualize the data to see how it looks like.

2. The library *Matplotlib* produces plots in python. You can use the function `plot()` in order to make a scatter plot of X and Y.

```
import matplotlib.pyplot as plt
plt.plot(data_train.X,data_train.Y,'bo')
plt.show()
```

3. We assume that $f(X)$ is a polynomial function of X :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + \epsilon \tag{1}$$

For p ranging from 1 to 9 follow the following steps :

(a) Use the function `polyfit` of the NumPy library to fit polynomial regression model in (1) using the training data. Set the option *full* to *True* in order to get the residual sum of squares. Do not forget to import *NumPy* !

For example, for $p = 2$ use the following code :

```
fit=np.polyfit(data_train.X, data_train.Y, deg=2, full=True)
```

The output of `polyfit()` returns the coefficient estimates `fit[0]`,the sum of squared residuals `fit[1]`, etc.

(b) Calculate the fitted values $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + ... + \hat{\beta}_p X^p$ for each observation in the training set. You can obtain this estimates using the function `polyval()` from the *NumPy* library.

(c) Plot the estimated functions $\hat{f}(X)$ and superpose to the scatter plot of question 3 for $p = \{1, 2, 5, 9\}$. Comment on the graphic. You can use the `linspace()` function from the NumPy library.

Hint In your plot, do not forget to set the x and y axis limits using `plt.ylim()` and `plt.xlim()`

(d) Evaluate the model accuracy by calculating the *training* RMSE (root mean square error) for each model fit.

(e) For the calculation of *test* RMSE you will need to import the data in the file *test.txt* and calculate the fitted values $\hat{Y}$ to the test data.

4. Finally, make a plot of both RMSE errors versus the complexity p. What model complexity would you choose to accurately estimate $f(X)$? Why ? What can you conclude about the bias and variance obtained with for the different fitted models ?

3 References

[1] Python Data Analysis Library (Pandas) `http://pandas.pydata.org`. Visited on August 21st 2018.

[2] Matplotlib `https://matplotlib.org`. Visited on August 21st 2018.

[3] Scientific computing tools for Python (SciPy) `https://scipy.org/docs.html`. Visited on August 7th 2020