Practical Session 1

Linear Regression

Alex Ferrer Ferre
Antonio Darder Bennassar

July 2022

1 Introduction

The aim of this session is to implement the basic concepts of Machine Learning in a Python code from scratch. The idea is to code a linear regression for a real example of data, and to see the phenomena of overfitting and underfitting.

Recalling from theory a standard linear regression looks like

$$\hat{y} = X \cdot \theta \tag{1}$$

where X is the matrix of data with m rows equal to the number of data points and n columns equal to the number of features. The term θ is a vector of length n which encloses the objective parameters that adjust the regression. Important, the first column of the matrix X is equal to all 1 making θ_0 the free term of the equation.

Equation 1 shows the model that tries to predict a certain phenomena, the other important characteristic of a linear regression are the metrics used to decide whether the model is accurate or not. In machine learning these equations are called cost functions, and the most used is L^2 but there are also L^1 and L^∞ norms which can be used to obtain various effects.

$$J(\theta) = \frac{1}{m}\sum_{i}^{m}(y_i - \hat{y}_i(\theta))^2 \tag{2}$$

$$J(\theta) = \frac{1}{m}\sum_{i}^{m}|y_i - \hat{y}_i(\theta)| \tag{3}$$

$$J(\theta) = max(|y_i - \hat{y}_i(\theta)|) \tag{4}$$

1.1 Optimization problems

The first part of the session will be focused on implementing the solution of the L^1, L^2 and L^∞ norms. The solution of the L^2 norm is the well known normal equation

$$\theta_2^* = (X^T \cdot X)^{-1} X^T Y \tag{5}$$

Although the L^1 and L^∞ norms have not a direct solution, the minimization problems can be rewritten as the following:

$$\begin{array}{ll} \underset{\text{x}}{\text{min}} & ||x||_1 \\ \text{subject to} & \text{Ax - b} = 0 \end{array} \quad \equiv \quad \begin{array}{ll} \underset{\text{x,t}}{\text{min}} & 1^T t \\ \text{subject to} & -t \leq Ax - b \leq t \end{array} \tag{6}$$

$$\begin{array}{ll} \underset{\text{x}}{\text{min}} & ||x||_\infty \\ \text{subject to} & \text{Ax - b} = 0 \end{array} \quad \equiv \quad \begin{array}{ll} \underset{\text{x,t}}{\text{min}} & t \\ \text{subject to} & -t1_m \leq Ax - b \leq t1_m \end{array} \tag{7}$$

To solve these problems a linear programming library can be used, these libraries demand the conditions to the minimization: c, A_{ub}, b_{ub}, A_{eq}, b_{eq}, l and u. It is important that these parameters are given in the same form as presented right below. That is to say, it has to be a minimization (not maximization) and the inequalities have to be "less or equal than".

$$\begin{array}{ll} \underset{\text{x}}{\text{min}} & c^T x \\ \text{such that} & A_{ub} x \leq b_{ub}, \\ & A_{eq} x = b_{eq}, \\ & l \leq x \leq b_{ub}, \end{array}$$

Consequently, equations (6) and (7) can be rewritten to equations (8) and (9) respectively:

$$\begin{array}{ll} \underset{\text{x}}{\text{min}} & [\ 0 \mid 1\ ]^T \begin{bmatrix} x \\ t \end{bmatrix} \\ \text{subject to} & \begin{bmatrix} A & -I \\ -A & -I \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \leq \begin{bmatrix} b \\ -b \end{bmatrix} \end{array} \tag{8}$$

$$\begin{array}{ll} \underset{\text{x}}{\text{min}} & [\ 0 \mid 1\ ]^T \begin{bmatrix} x \\ t \end{bmatrix} \\ \text{subject to} & \begin{bmatrix} A & -1 \\ -A & -1 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \leq \begin{bmatrix} b \\ -b \end{bmatrix} \end{array} \tag{9}$$

1.2 L^2 regularization

L^2 regularization adds another term to the cost function which "penalizes" high values of the parameter θ as

$$J(\theta) = L(\theta) + \lambda \frac{1}{2} \theta^T \theta \tag{10}$$

Where λ is the parameter which determines the influence of the regularization. The loss function $L(\theta)$ can be any norm already introduced. Note that the solution of the L^2 norm case is as follows

$$\theta_2^* = (X^T X + \lambda I)^{-1} X^T Y \tag{11}$$

1.3 The datasets

Anscombe's quartet First, to observe the behaviour of the different norms the datasets used will be the first and third sets from the anscombe's quartet. These datasets have 11 points and are shown in Figure 1. These 2 different particular datasets intentionally provide the same L^2 regression line. The third group of the anscombe's quartet is of special interest because it has a massive outlier and the difference between the norms will be clearly seen.

Third degree polynomial The last dataset that will be used is a third degree polynomial with some noise introduced to it. This data contains more points and it will be helpful to show the influence of the test and train ratio as well as the regularization.

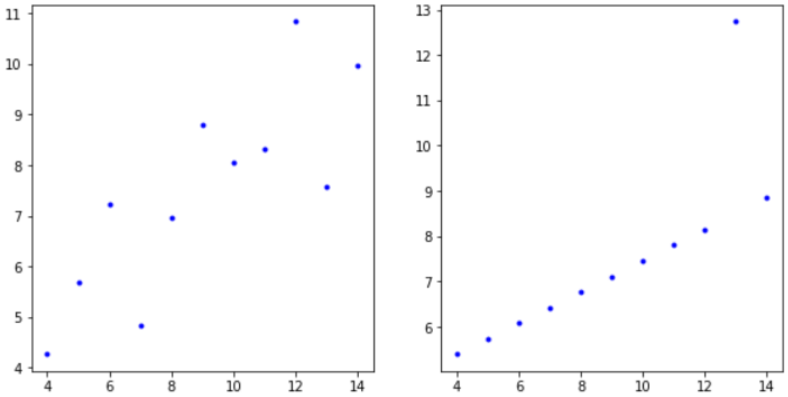

Figure 1 First and third sets of data from the anscombe's quartet