# Theory of statistical learning

## Exercises - Linear regression

*Rémi Flamary*

### Exercise 1 Simple linear regression

Simple linear regression consists in estimating the parameters of the following linear model.

$$y = ax + b \tag{1}$$

This is done by minimizing the mean squared errors of the model on the training dataset $\{x_i, y_i\} \in \mathbb{R}^2, \forall i \in 1, \dots, n$:

$$\min_{a,b} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - ax_i - b)^2 \tag{2}$$

The problem is convex and differentiable, the minimum can be obtained by computaion the derivatives wrt $a$ and $b$ and setting them to zero, yielding two equations with two unknown.

1. Compute the derivative of the loss function wrt $a$ and $b$.
2. Express the linear equation system when those derivatives are null.
3. Express $\hat{b}$ as a function of $\bar{x} = \frac{1}{n}\sum_i x_i$ and $\bar{y} = \frac{1}{n}\sum_i y_i$ and $\hat{a}$.
4. Express the solution $\hat{a}$ as a function of the training data and $\bar{x}, \bar{y}$.
5. We can now study the linear relation between the number of employees ($x$) of a company and its annual sales ($y$). The data is as follows:

| *Year* | *Number of employees* | *Annual sales* |
|---|---|---|
| 1957 | 294 | 634 |
| 1959 | 314 | 728 |
| 1961 | 383 | 819 |
| 1963 | 402 | 938 |
| 1965 | 475 | 1136 |
| 1967 | 786 | 1317 |

a) Represent the point cloud for this data (between employees and sales). Use the solutions above to estimate the parameters $a$ and $b$, and plot the model on the point cloud.
b) What is the quality of the model ? Compute its correlation coefficient on the data.
c) Predict the annual sales of the company if next year it has 800 employees.
d) Discuss the training examples, are there any outliers?
e) Does time has any effect on the model?

### Exercise 2 Weighted least squares

When the samples come from different sensors that can have different properties such as noise levels, it is interesting to take into account this information during the training. This can be done by weighting differently the training samples in the loss (more weight on less noisy examples for instance).

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\sum_{i=1}^{n} p_i \left(y_i - \mathbf{w}^\top \mathbf{x}_i - b\right)^2$$

1. Show that if $p_i = 1, \forall i$ then the problem boils down to classical least squares.
2. Express the minimization problem in its matrix form. To this end you can use vector $\mathbf{p}$ containing the weights $p_i$ and its corresponding diagonal matrix $\mathbf{P} = diag(\mathbf{p})$.
3. Compute the gradient of the cost function and deduce the solution of the weighted least squares.
4. Following the same approach, express the solution for the weighted least squares with ridge regularization.

## Exercise 3 Polynomial regression

We investigate in this exercise the estimation of linear regression on a polynomial basis, which leads to a non-linear regression.

We want to find the polynomial relation between the input variable $x$ and the output variable $y$ such that:

$$y = \sum_{k=0}^{p} c_k x^k \quad \text{with } \mathbf{c} = [c_p, \ldots, c_k, \ldots, c_0] \in \mathbb{R}^{p+1} \tag{3}$$

where the parameters are stored in the column vector $\mathbf{c}$. In order to estimate coefficients $\{c_k\}$, we have $N$ training samples $(x_i, y_i)$ that can be stored in columns vectors $X$ and $Y$.

1. Express the column vector $\mathbf{x}$ as a function of $x$ such that $y = \mathbf{x}^T\mathbf{c}$. Deduce the matrix form equivalent to 3 between vector $Y$ et and matrix $Z$ with

$$Z = \begin{bmatrix} x_1^p & \ldots & x_1 & 1 \\ \vdots & \ddots & & \vdots \\ x_N^p & \ldots & x_N & 1 \end{bmatrix} = [X^p \; X^{p-1} \; \ldots X \;\; 1] \tag{4}$$

   constructed from the training data $(x_i, y_i)$. we admit in the following that the power is computed component-wise in the vectors.
2. Apply the least square principle to this problem and solve the resulting optimization problem. Express the optimal $\mathbf{c}$ as a function of $Y$ and $Z$.
3. What happens if $p > N$ ?

## Exercise 4 Non-linear regression

In the following models $y$ is the value to predict $w_i$ for $i = 1, \ldots, d'$ are the model parameters, $x_i$ for $i = 1, \ldots, d$ correspond to variables (features) and $\epsilon$ is a noise.

1. $y = w_1x_1 + w_2x_1^2 + \epsilon$, $d' = 2, d = 1$
2. $log(y) = w_1x_1 + w_2x_2 + w_3 + \epsilon$, $d' = 3, d = 2$
3. $log(y + \epsilon) = w_1x_1 + w_2x_2 + w_3$, $d' = 3, d = 2$
4. $log(y + w_4) = w_1x_1 + w_2x_2 + w_3 + \epsilon$, $d' = 4, d = 2$
5. $y = (x_1)^{w_1}(x_2)^{w_2}(10)^{w_3}\epsilon$, $d' = 3, d = 2$

For all models above, answer the following questions:

- Is it possible to reformulate the problem with a linear formulation?
- If yes, describe the linearization procedure (transformation $y \to \tilde{y}$ et $x \to \tilde{x}$).
- Discuss the impact of this linearization o the noise. Is the additive noise assumtion still valid? valable?