Elements of Statistical Learning.

- For several cases the goal is to use the inputs to predict the values of the outputs. (Supervised Learning)

Qualitative variables : Categorical or discrete variables.

- Distintichon in output type $\begin{cases} \text{Regression} & \text{Quantitative outputs} \\ \text{Classification} & \text{Qualitative outputs.} \end{cases}$

- Qualitative values : "Success" or "failure"
  "Survived" or "died"

Input variable : X
Quantitative output Y
Qualitative output G

Observed values are written in lowercase; ith of X is written as x_i

Linear models :

Given a vector of inputs $X^T = (x_1, x_2, \dots x_p)$, we predict the output Y via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{n} x_j \hat{\beta}_j \qquad (2.1)$$

Include $\hat{\beta}_0$ in the vector of coefficients $\hat{\beta}$ and then write the linear model in vector form:

$$\hat{Y} = X^T \beta \qquad (2.2)$$

How do we fit the linear model to a set of training data?

By far the most popular is the method of least squares. In this approach, we pick the coefficients β to minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \qquad (2.3)$$

y_i = Real values
$\hat{y}_i = x_i^T \beta$ Predicted values.

Cuadratic function of the parameters $\Rightarrow$ Minimum always exists, but may not be unique. We can write:

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \qquad (2.4)$$

$X \in \mathbb{R}^{N \times p}$ (Each row is an input vector)

$y \in \mathbb{R}^N$ (Vector of the outputs)

$$\alpha = x^T A x$$

$$\frac{\partial \alpha}{\partial x} = x^T (A + A^T)$$

$\alpha = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$ Taking $\frac{\partial \alpha}{\partial x_k} = \sum_{ij} a_{ij} x_i \delta_{jk} + \sum_{ij} a_{ij} \delta_{ik} x_j$

$$\frac{\partial \alpha}{\partial x_k} = \sum_i a_{ik} x_i + \sum_j a_{kj} x_j = x^T A^T + x^T A$$

$\frac{\partial \alpha}{\partial x} = x^T (A^T + A)$ if the matrix is symmetric:

$$\frac{\partial \alpha}{\partial x} = 2 x^T A.$$

$$\frac{\partial R}{\partial \beta} = -2X^T (y - X\beta) \Rightarrow X^T (y - X\beta) = 0$$

Organizing: $X^T y - X^T X \beta = 0$

$$\hat{\beta} = (X^T X) X^T y = 0 \qquad (2.6)$$

The fitted value at ith input is $\hat{y}_i = \hat{y}_i(x_i) = x_i^T \hat{\beta}$

- **Probabilistic approach:**

Regression problem with $p(y|x) = \mathcal{N}(y \mid f(x), \sigma^2)$

$x \in \mathbb{R}^D$ inputs
$y \in \mathbb{R}$ targets. $\quad > \quad \hat{y} = f(x) + \varepsilon.$

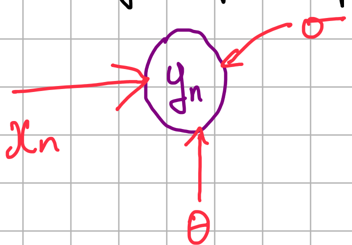
and $\varepsilon = \mathcal{N}(0, \sigma^2)$ Gaussian distribution with 0 mean and variance σ^2.

Model parameters θ (Learning this)

$$p(y \mid x, \theta) = \mathcal{N}(y \mid x^T\theta, \sigma^2) \iff y = x^T\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

⇒ **Parameter estimation:**

We are given a training set $D := \{(x_1, y_1) \ldots, (x_N, y_N)\}$ consisting of N inputs $\in \mathbb{R}^D$



y_i and y_j are conditionally independent given their respective inputs.

$$P(y \mid x, \theta) = p(y_1, \ldots, y_n \mid x_1, x_2, \ldots, x_n; \theta)$$

$$P(y \mid x, \theta) = \prod_{i=1}^{N} p(y_n \mid x_n, \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid x_n^T\theta, \sigma^2)$$

Maximum Likelihood estimation:

θ_{ML} : Maximizing likelihood (Maximize the predictive distribution of the training data given the model parameters:

$$\theta_{ML} = \underset{\theta}{\arg\max}\ p(y \mid x, \theta)$$

Using the properties of logarithms on products, we minimize the negative log-likelihood:

$$-\log p(y|x;\theta) = -\log \prod_{i=1}^{N} p(y_i|x_i;\theta) = -\sum_{i=1}^{N} \log(p(y_i|x_i;\theta))$$

Factorization due to independence.

$$\log p(y_n|x_n;\theta) = -\frac{1}{2\sigma^2}(y_n - x_n^T\theta)^2 + c.$$

$$\mathcal{L}(\theta) := \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x_n^T\theta)^2$$

$$\mathcal{L}(\theta) = \frac{1}{2\sigma^2}(y - X^T\theta)^T(y - X\theta) = \frac{1}{2\sigma^2}\|y - X\theta\|^2.$$

Computing the Gradient:

$$\frac{\partial\mathcal{L}}{\partial\theta} = \frac{\partial\mathcal{L}}{\partial\theta}\left(\frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta)\right)$$

$$= \frac{1}{2\sigma^2}\frac{d}{d\theta}\left(y^Ty - 2y^TX\theta + \theta^TX^TX\theta\right)$$

$$= \frac{1}{\sigma^2}\left(-y^TX + \theta^TX^TX\right)$$

Again $\hat{\theta}_{ML} = (X^TX)^{-1}X^Ty.$

Now we can use monomial to express different features.
For example for a second order polynomial:

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

$x_n \in \mathbb{R}^D$

$y_n \in \mathbb{R}$, $n = 1, 2, \ldots, N$

Number of inputs.

$$\theta_{ML} = (X^TX)^{-1}X^Ty$$