Statistical Machine Learning

Lecture 1: Introduction

2022-23

Course Structure

1. Introduction, definitions and problems.
2. Regression techniques.
3. Classification techniques.
4. Theoretical arguments for complexity control and dimensionality reduction.

Definition

1. Artificial Intelligence (AI) is the area of knowledge that endeavors towards constructing systems (hardware or software) that behave *intelligently.* (Observe that I have not defined what I mean by intelligent just yet).
2. Machine Learning (ML) is a sub-area of AI that tackles problems by extracting the patterns that link questions with correct answers (provided to the AI system during the *training phase*). The ML paradigm differs from the traditional manner to solve problems on the fact that, in the traditional way the rules (patterns) that produce an answer given a question are supposed to be known, whereas in the ML paradigm such rules are unknown and need to be discovered.

Definition

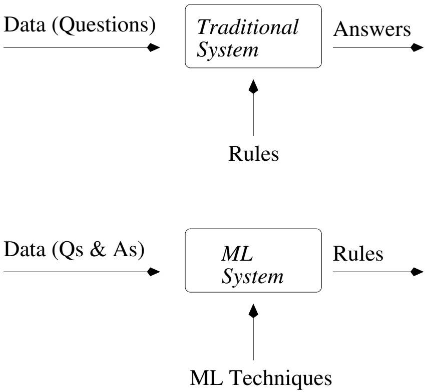


Figure: Different paradigms for solving problems

Definition

1. Statistical ML: The adjective *statistical* acknowledges the nature of the elements in the Data Set to be used $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^{M}$, which can be described through a probability distribution $\mathcal{P}(\boldsymbol{x})$.
2. Most of the work in SML involves making models of the $\mathcal{P}(\boldsymbol{x})$ and produce results based on inferences using such a model (the ML techniques in the figure above).

Problems to be Tackled

1. In the most general terms: Modeling stationary processes, i.e. probabilistic processes where the density distribution describing the observations does not change with time.
2. In more restrictive terms: Pattern recognition (speech, face, hand-written characters, etc.)
3. All these tasks require an approach based on statistics to help extract the relevant features (patterns) out of a big volume of information.

More on Statistics

1. Both characters are drawn in a 256 x 256 pixels figure.
2. The total number of possible figures (in black and white) is $2^{256 \times 256} \sim 10^{20000}$. The size of the set with all the possible figures is huge.
3. No all the possible figures are meaningful. There are many (many indeed) figures that wouldn't carry any meaning at all.

Figure: Hand-written characters a and b.

Pattern extraction

1. We can define x_1 and x_2 as the horizontal and vertical dimensions of the characters.
2. a's and b's are of similar width, $x_1(a) \simeq x_2(b)$.
3. a's are expected to be shorter than b's, therefore $x_2(a) < x_2(b)$.
4. Both attributes are distributed variables $x_1(a) \sim \mathcal{P}_{1,a}$, $x_2(a) \sim \mathcal{P}_{2,a}$, $x_1(b) \sim \mathcal{P}_{1,b}$, $x_2(b) \sim \mathcal{P}_{2,b}$.

.

Pattern extraction

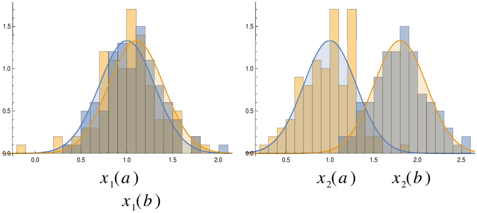


Figure: Histogram and correspondent distribution for x_1 (left) and x_2 (right). Observe that, even for the clear separation between x_2 distributions, there steel exists an *overlapping* range of values where no decision can be taken.

First Problem: Linear Regression

1. The Data Set that characterizes this problem has the following structure: $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{M}$, where $\boldsymbol{x}_n \in \mathbb{R}^d$ (independent variable or the *features*) and $\boldsymbol{y}_n \in \mathbb{R}^s$ (the dependent variables or the *labels* or the *targets*).
2. We assume there exists a map $\mathscr{M} \ni h : \mathbb{R}^d \to \mathbb{R}^s$ such that $\boldsymbol{y} = h(\boldsymbol{x}) + \varepsilon$, (we will discuss in the near future the fundamentals for this assumption) where $\varepsilon \in \mathbb{R}^s$ represents all the variables we have no control over (for instance, these variables account for the variability in style shown by different people while writing characters a and b).
3. Neither the map h or the variable ε are known or given.

First Problem: Linear Regression

1. We assume the existence of a basis set $\{\varphi_m\}_{m=0}^{\infty} \subset \mathscr{M}$ such that for all $g \in \mathscr{M}$ there exist a collection of real numbers $c_m \in \mathbb{R}$ such that

$$g(\boldsymbol{x}) = \sum_m c_m \varphi_m(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^d$. The linear regression problem consist of finding the coefficients $\{c_m\}$, given a particular set $\{\varphi_m\}$, that minimize the Sum Of Squares error for a particular data set $\mathcal{D}$:

$$E\left(\{c_m\} | \mathcal{D}, \{\varphi_m\}\right) = \frac{1}{M} \sum_{n=1}^{M} \left\| \boldsymbol{y}_n - \sum_m c_m \varphi_m(\boldsymbol{x}_n) \right\|^2$$

$$\|z_n\|^2 = \sum_{j=1}^{s} z_j^2$$

Solution

1. The sum-of-square error admits a unique minimum (it is a convex optimisation problem) and the solution is obtained by solving the set of equations:

$$\frac{\partial E}{\partial c_{m'}} = \frac{1}{M}\sum_{n=1}^{M}\left[\sum_{j=1}^{s}[\boldsymbol{y}_n]_j[\varphi_{m'}(\boldsymbol{x}_n)]_j - \sum_{j=1}^{s}\sum_{m} c_m[\varphi_m(\boldsymbol{x}_n)]_j[\varphi_{m'}(\boldsymbol{x}_n)]_j\right]$$
$$= 0$$

or

$$\sum_{m} c_m \left[\frac{1}{M}\sum_{n=1}^{M}\sum_{j=1}^{s}[\varphi_m(\boldsymbol{x}_n)]_j[\varphi_{m'}(\boldsymbol{x}_n)]_j\right] = \frac{1}{M}\sum_{n=1}^{M}\sum_{j=1}^{s}[\boldsymbol{y}_n]_j[\varphi_{m'}(\boldsymbol{x}_n)]_j.$$

Solution

1. By defining the matrix $[\boldsymbol{A}]_{m',m} = \frac{1}{M}\sum_{n=1}^{M}\sum_{j=1}^{s}[\varphi_{m'}(\boldsymbol{x}_n)]_j[\varphi_m(\boldsymbol{x}_n)]_j$, the vector $[\boldsymbol{t}]_{m'} = \frac{1}{M}\sum_{n=1}^{M}\sum_{j=1}^{s}[\boldsymbol{y}_n]_j[\varphi_{m'}(\boldsymbol{x}_n)]_j$ and the vector $\boldsymbol{c}^T = (c_0, c_1, \dots, c_m \dots)$ we have that:

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{t},$$

   which, if $\boldsymbol{A}^{-1}$ exists, admits the solution

$$\boldsymbol{c} = \boldsymbol{A}^{-1}\boldsymbol{t}.$$

2. Observe that the matrix $\boldsymbol{A}$ and the vector $\boldsymbol{t}$ depend on the data set $\mathcal{D}$.
3. The problem is linear in $\boldsymbol{c}$.

Initial Stages in Model Selection

1. Consider the following data, that has been approximated by a polynomial of order 2 and by a polynomial of order 11

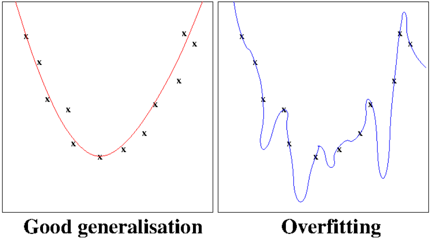


Figure: Two different models applied to the same data. Which one is the best?

Particular Case

1. Observe that I have not indicated the number of elements of the basis $\{\varphi_m\}$ I am using.
2. To illustrate how to proceed with a typical case we consider the following data set

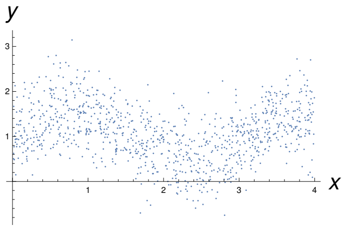


Figure: Scatter plot of the data set.

Particular Case

1. Observe that $d = s = 1$ thus $x, y \in \mathbb{R}$ and suppose that $\{x^m\}$ are the elements of the basis set (polynomial regression).
2. The error becomes

$$E\left(\{c_m\}|\mathcal{D}, \{x^m\}_{m=0}^{L}\right) = \frac{1}{M}\sum_{n=1}^{M}\left(y_n - \sum_{m=0}^{L} c_m x_n^m\right)^2.$$

Particular Case

1. The entries of the matrix $\boldsymbol{A}$ become $[\boldsymbol{A}]_{m',m} = \frac{1}{M}\sum_{n=1}^{M} x_n^{m+m'} = \overline{x^{m+m'}}$, and the vector $[\boldsymbol{t}]_{m'} = \frac{1}{M}\sum_{n=1}^{M} y_n x_n^{m'} = \overline{yx^{m'}}$, thus

$$\boldsymbol{A} = \begin{pmatrix} 1 & \overline{x} & \overline{x^2} & & \dots & \overline{x^L} \\ \overline{x} & \overline{x^2} & \overline{x^3} & & \dots & \overline{x^{L+1}} \\ \vdots & & & \ddots & & \vdots \\ \overline{x^k} & \overline{x^{k+1}} & & & \dots & \overline{x^{L+k}} \\ \vdots & & & & & \vdots \\ \overline{x^L} & \overline{x^{L+1}} & & & \dots & \overline{x^{2L}} \end{pmatrix} \qquad \boldsymbol{t} = \begin{pmatrix} \overline{y} \\ \overline{yx} \\ \\ \overline{yx^k} \\ \\ \overline{yx^L} \end{pmatrix}$$

Particular Case

1. Training error and validation error:

$$e_t(C) = \frac{1}{L_t} \sum_{n=1}^{L_t} (y_n - p_C(x_n))^2 \qquad e_v(C) = \frac{1}{L - L_t} \sum_{n=1+L_t}^{L} (y_n - p_C(x_n))^2$$

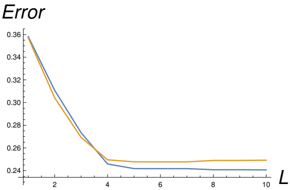
2. 

Figure: Training (blue) and validation (orange) errors as functions of the order of the polynomial.

Generalization

1. The error landscape gets determined by the mathematical form of the error function.
2. Consider the expression

$$E_p\left(\{c_m\}|\mathcal{D},\{x^m\}_{m=0}^{L}\right)=\frac{1}{M}\sum_{n=1}^{M}\left|y_n-\sum_{m=0}^{L}c_m x_n^m\right|^p$$

   which for $p=2$ is the one we have considered.
3. Different values of p may produce different results, particularly in the presence of *outlayers* in the dataset.

Data Linearization

1. Suppose we have model the data $\mathcal{D} = \{(x_n, y_n)\},\ x, y \in \mathbb{R}$, with the expression

$$y = f\left(\sum_{k=0}^{L} c_k x^k\right)$$

   with $f : \mathbb{R} \to \mathbb{R}$ bijective.

2. Therefore, there exists $f^{-1} : \mathbb{R} \to \mathbb{R}$ such that $f^{-1}(f(x)) = x$. Thus

$$\begin{aligned} Y &= f^{-1}(y) \\ &= \sum_{k=0}^{L} c_k x^k \end{aligned}$$

Nonlinear Regression and Gradient Descent

1. Suppose we have model the data $\mathcal{D} = \{(x_n, y_n)\}$, $x, y \in \mathbb{R}$, with the expression

$$E_2(\boldsymbol{c}|\mathcal{D}) = \frac{1}{N}\sum_n (y_n - f(x_n, \boldsymbol{c}))^2$$

2. Assume f is a *smooth* function on $\boldsymbol{c} \in \mathbb{R}^m$, $m < N$, thus

$$E_2(c_0 + \delta c|\mathcal{D}) = \frac{1}{N}\sum_n (y_n - f(x_n, c_0))^2 + \frac{1}{N}\sum_n (y_n - f(x_n, c_0))\,\nabla f(x_n, c_0)\cdot \delta c + O(\delta c^2)$$

$$\frac{\partial f}{\partial c_k} = 0.$$

3. This set of equations may have 1, many, none or infinite number of solutions and there is no general method to solve them.
4. Most methods need a very good initial approximation to the global minimum.

Propagation of Errors; Errors When Changing Variables

- Within the subject of Statistical Machine Learning we will explore quantities x that are distributed, i.e. there exists $\mathcal{P} : \mathbb{D} \to \mathbb{R}^+ \cup \{0\}$ and $\sum_{x\in\mathbb{D}} \mathcal{P}(x) = 1$ that describes the statistics of x.
- Every measurement process produces an estimate $\overline{x}$ (for a quantity of interest x) and an error $\overline{\sigma}_x$ which provides an estimate of the statistical dispersion around the estimate.
- Both quantities are estimates. They are not the true values of the $x_0 = \sum_{x\in\mathbb{D}} \mathcal{P}(x)x = \mathbb{E}[x]$ and $\sigma_x^2 = \sum_{x\in\mathbb{D}} \mathcal{P}(x)(x - x_0)^2 = \mathbb{E}[(x - \mathbb{E}[x])^2]$. These values, x_0 and σ_x, are, in general, not accessible but can be estimated.

Propagation of Errors; Errors When Changing Variables

- Suppose we measure x which has a mean value x_0 and a variance σ_x^2 and y with a mean y_0 and variance σ_y^2.
- Suppose we have a function $G(x, y)$ and wish to determine the variance of $G(x, y)$,i.e., propagate the errors in x and y to G. Thus

$$
\begin{aligned}
G(x, y) &= G(x_0 + x - x_0, y_0 + y - y_0) \\
&= G(x_0, y_0) + \left.\frac{\partial G}{\partial x}\right|_{x_0, y_0} (x - x_0) + \left.\frac{\partial G}{\partial y}\right|_{x_0, y_0} (y - y_0) + O(\Delta^2) \\
G_0 &= G(x_0, y_0) + O(\Delta^2) \\
\sigma_G^2 &= \left(\left.\frac{\partial G}{\partial x}\right|_{x_0, y_0} \right)^2 \sigma_x^2 + \left(\left.\frac{\partial G}{\partial y}\right|_{x_0, y_0} \right)^2 \sigma_y^2 + O(\Delta^4).
\end{aligned}
$$

Worked Problem

- Suppose we take n independent measurements of the same quantity x. Suppose each measurement x_i has the same mean $x_0$and variance σ_x^2.
- Given the following definition

$$G(\{x_i\}) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

find the mean and variance of G.

Solution

- The mean is given by (up to corrections of $O(\Delta^2)$)

$$G_0 = \frac{1}{n}\sum_{i=1}^{n} x_0 = x_0.$$

- The variance is given by (up to corrections of $O(\Delta^4)$)

$$\sigma_G^2 = \sum_{i=1}^{n}\left(\left.\frac{\partial G}{\partial x_i}\right|_{x_0}\right)^2 \sigma_x^2 = \sum_{i=1}^{n}\frac{1}{n^2}\sigma_x^2 = \frac{\sigma_x^2}{n}.$$

Interpretation

- The mean $x_0$is in general inaccessible. We usually substitute $x_0$with the arithmetic mean $\frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x}$.
- Let us define the experimental variances

$$\overline{\sigma^2_{\mathrm{exp}}} \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\sigma^2_{\mathrm{exp}} \equiv \mathbb{E}[(x - \overline{x})^2].$$

- The first expression can be measured, the second is, in general, inaccessible (we do not know the distribution $\mathcal{P}(x)$).

.

Interpretation

- The experimental variance is linked to the true variance in the following way:

$$
\begin{aligned}
\sigma^2_{\mathrm{exp}} &= \mathbb{E}[(x-\overline{x})^2] = \mathbb{E}[(x_i-\overline{x})^2] \\
&= \mathbb{E}\left[\left(\frac{n-1}{n}x_i - \frac{1}{n}\sum_{j\neq i} x_j\right)^2\right] \\
&= \left(\frac{n-1}{n}\right)^2 \mathbb{E}[x_i^2] - 2\frac{n-1}{n^2}\mathbb{E}[x_i]\sum_{j\neq i}\mathbb{E}[x_j] + \frac{1}{n^2}\mathbb{E}\left[\sum_{j\neq i} x_j^2 + 2\sum_{j\neq i}\sum_{k\neq i,j} x_j x_k\right] \\
&= \left(\frac{n-1}{n}\right)^2 \mathbb{E}[x_i^2] - 2\frac{n-1}{n^2}\mathbb{E}[x_i]\sum_{j\neq i}\mathbb{E}[x_j] + \frac{1}{n^2}\sum_{j\neq i}\mathbb{E}\left[x_j^2\right] + \frac{2}{n^2}\sum_{j\neq i}\sum_{k\neq i,j}\mathbb{E}[x_j]\mathbb{E}[x_k] \\
&= \left(\frac{n-1}{n}\right)^2 \mathbb{E}[x^2] - 2\left(\frac{n-1}{n}\right)^2\mathbb{E}[x]^2 + \frac{n-1}{n^2}\mathbb{E}[x^2] + \frac{2}{n^2}\frac{(n-1)^2-(n-1)}{2}\mathbb{E}[x]^2 \\
&= \frac{n-1}{n^2}(n-1+1)\mathbb{E}[x^2] - 2\frac{n-1}{n^2}\left(n-1-\frac{n-1-1}{2}\right)\mathbb{E}[x]^2 = \frac{n-1}{n}\left(\mathbb{E}[x^2]-\mathbb{E}[x]^2\right) \\
&= \frac{n-1}{n}\sigma_x^2
\end{aligned}
$$

Interpretation

- If we estimate σ^2_{exp} using $\overline{\sigma^2_{\mathrm{exp}}}$ we can estimate the true variance by

$$\sigma_x^2 \approx \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Therefore

$$G_0 \approx \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\sigma_G^2 \approx \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$