Machine Learning

Machine Learning Framework

Souhaib Ben Taieb

March 8, 2021

University of Mons

Table of contents

Table of contents

Learning from data

"Machine learning is a **scientific discipline** that explores the **construction and study of algorithms** that can **learn from data**."

- The essence of machine learning
    - A pattern exists
    - We cannot pin it down mathematically
    - We have data on it
- Learning examples
    - Spam Detection
    - Product Recommendation
    - Credit Card Fraud Detection
    - Medical Diagnosis

Table of contents

Components of supervised learning

The **input** variables[1] are typically denoted using the symbol X. If we observe p different variables, we write $X = (X_1, X_2, \ldots, X_p)$. The inputs belong to an *input space* $\mathcal{X} \subseteq \mathbb{R}^p$.

The **output** variable[2] is typically denoted using the symbol Y. The output belongs to an *output space* $\mathcal{Y}$.

- Regression: $\mathcal{Y} \subseteq \mathbb{R}$
- Binary classification: $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$
- Multi-class classification (with K categories): $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$

[1] also called *predictors*, *independent variables*, *features*, *variables* or just *inputs*.
[2] also called the *response* or *dependent variable*.

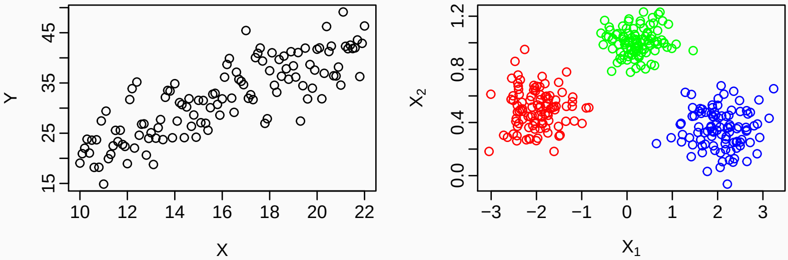
Components of supervised learning

The **data**, also called *training set*, is a set of n input-output pairs

$$\begin{aligned}\mathcal{D} &= \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \\ &= \{(x_i, y_i)\}_{i=1}^{n},\end{aligned}$$

where $x_i = (x_{i1}, \ldots, x_{ip})$. Each pair, also called an *example* or a *data point*, belongs to the *data space* $\mathcal{X} \times \mathcal{Y}$.

Components of supervised learning



- Left figure: $\mathcal{X} = \mathbb{R}$ (one-dimensional input) and $\mathcal{Y} \subseteq \mathbb{R}$
- Right figure: $\mathcal{X} = \mathbb{R}^2$ (two-dimensional input) and $\mathcal{Y} = \{\mathit{RED}, \mathit{GREEN}, \mathit{BLUE}\}$

Table of contents

Probabilistic data model

We assume the data points are identically and independently distributed (i.i.d.) realizations from a fixed <u>*unknown*</u> **data distribution** $p_{X,Y}(x, y)$, which represents different sources of uncertainty.

The probability distribution $p_{X,Y}(x, y)$ can be factorized as

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y|x)$$

where

- the marginal distribution $p_X(x)$ models uncertainty in the sampling of the inputs.
- the conditional distribution $p_{Y|X}(y|x)$ describes a stochastic (non-deterministic) relation between inputs and output.

Probabilistic data model

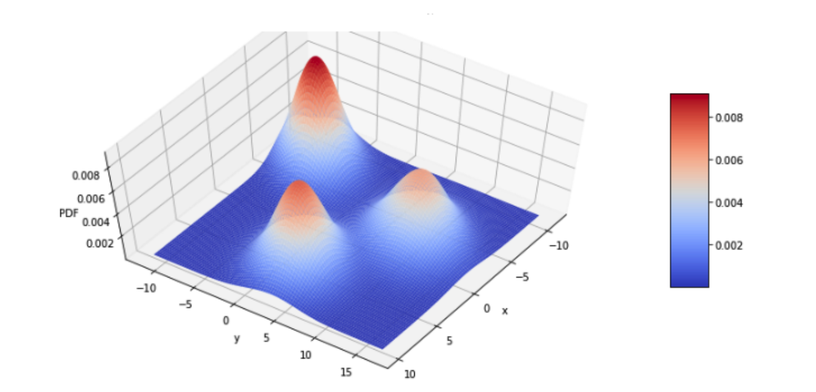
In other words, for $i = 1, 2, \ldots, n$, we have

$$(x_i, y_i) \sim p_{X,Y},$$

or, equivalently,

$$x_i \sim p_X \text{ and } y_i | x_i \sim p_{Y|X}(\cdot | x_i).$$

Probabilistic data model



Source: `https://tinyurl.com/19bdt531`

Probabilistic data model - Regression

The data distribution is often implicitly specified, i.e. $p_{X,Y}$ is not given explicitly.

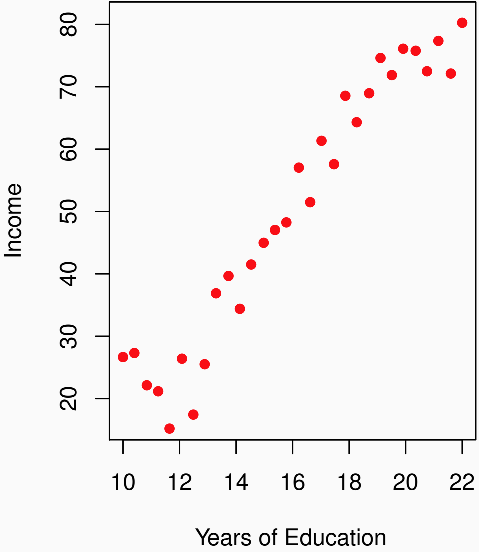
In regression, the following (additive error) model is often considered:

$$y = f(x) + \varepsilon, \tag{1}$$

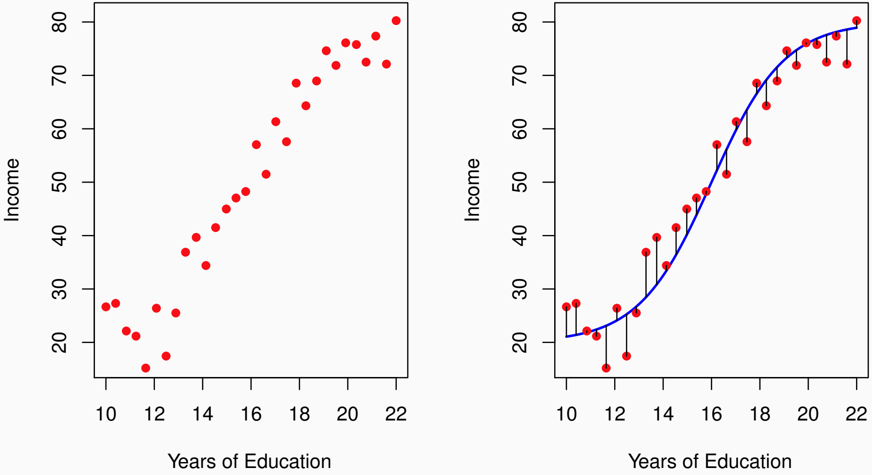
where

- $x \sim P_X$
- f is a fixed unknown function (e.g. $x \in \mathbb{R}$ and $f(x) = x^2$)
- ε is random noise, where
    - $\mathbb{E}[\varepsilon|x] = 0$
    - $\mathrm{Var}(\varepsilon|x) = \sigma^2$, with $\sigma \in [0, \infty)$.

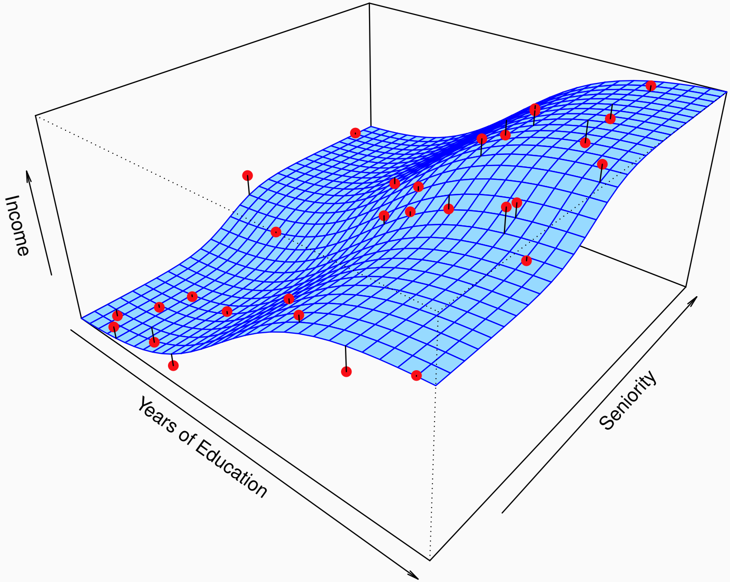
Probabilistic data model - Regression



Probabilistic data model - Regression



Probabilistic data model - Regression



Probabilistic data model - Regression

If $x \in \mathbb{R}^p$, and

- $f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ ($\beta_j \in \mathbb{R}$)
- $\varepsilon | x \sim \mathcal{N}(0, \sigma^2)$,

what is $y|x$?

Probabilistic data model - Regression

If $x \in \mathbb{R}^p$, and

- $f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ ($\beta_j \in \mathbb{R}$)
- $\varepsilon | x \sim \mathcal{N}(0, \sigma^2)$,

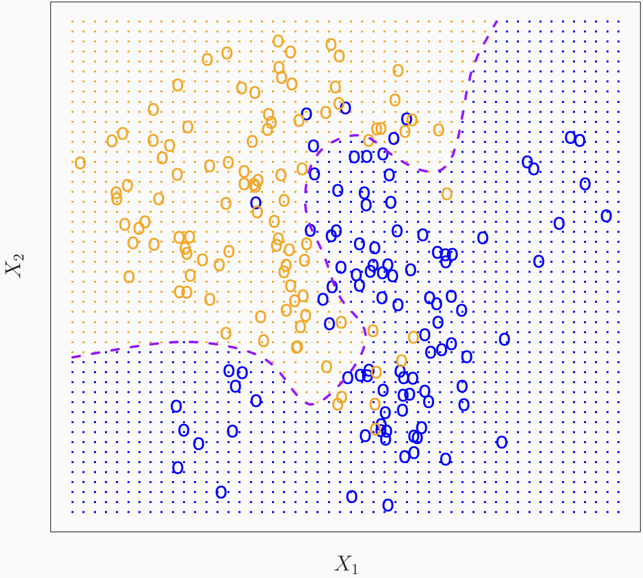
what is $y|x$?

$$y|x \sim \mathcal{N}(f(x), \sigma^2)$$

The data model in (1) implies that the conditional distribution $y|x$ depends on x only through the conditional mean. In fact, we have

- $\mathbb{E}[y|x] = f(x)$
- $\text{Var}[y|x] = \sigma^2$

Probabilistic data model - Classification



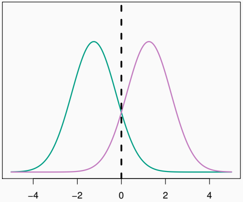
Probabilistic data model - Classification

In classification, y is a discrete random variable ($p(y|x)$ is a conditional pmf). We cannot use the previous additive error model. The notion of "noise" is different.

Using Bayes' rule, we can write

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y) \overset{y \text{ uniform}}{\propto} p(x|y)$$

Let us consider $K = 2$ and $p = 1$.



$p(x|y = -1)$ $p(x|y = +1)$

Table of contents

Common goals and models in supervised learning

- **Prediction**: predict the output for new inputs.
- **Inference (or explanation)**: which predictors are associated with the response? what is the relationship between the response and each predictor? ...

- **Discriminative models**: learn/estimate the underlying conditional distribution $p(y|x)$ (or conditional expectation $\mathbb{E}_{y|x}[y|x]$).
- **Generative models**: learn/estimate the underlying joint distribution of output and inputs, $p(x, y)$. In other words, learn both $p(y|x)$ and $p(x)$.

Components of supervised learning (continued)

In prediction problems, we often want to produce the "best" prediction (value) for a given input x. Suppose the conditional distribution $p(y|x)$ is **known**, which *value* should we use as output prediction for an input x?

Components of supervised learning (continued)

In prediction problems, we often want to produce the "best" prediction (value) for a given input x. Suppose the conditional distribution $p(y|x)$ is **known**, which *value* should we use as output prediction for an input x?

To do so, we need to define "best" by specifying a **loss function**

$$\mathsf{L} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty),$$

which is a (pointwise) measure of the error $L(y, \hat{y})$ we incur in when predicting $\hat{y}$ in place of y. Some examples are

- $L(y, \hat{y}) = (y - \hat{y})^2$ (square error loss)
- $L(y, \hat{y}) = |y - \hat{y}|$ (absolute error loss)
- $L(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$ (zero-one loss)
- ...

Components of supervised learning (continued)

Given a **loss function** $L(\cdot, \cdot)$ and a **hypothesis ("prediction function")**

$$h : \mathcal{X} \to \mathcal{Y},$$

the **expected error** of h on a single data point x (or expected conditional risk) is given by

$$\begin{aligned} E(h, x) &= \mathbb{E}_{y|x}[L(y, h(x))|x] && (2) \\ &= \mathbb{E}_{y \sim p(y|x)}[L(y, h(x))] && (3) \end{aligned}$$

Components of supervised learning (continued)

Given a **loss function** $L(\cdot,\cdot)$ and a **hypothesis ("prediction function")**

$$h : \mathcal{X} \to \mathcal{Y},$$

the **expected error** of h on a single data point x (or expected conditional risk) is given by

$$\begin{aligned} E(h,x) &= \mathbb{E}_{y|x}[L(y,h(x))|x] && (2) \\ &= \mathbb{E}_{y\sim p(y|x)}[L(y,h(x))] && (3) \end{aligned}$$

The (global) **expected error** (or expected risk) of h is given by

$$\begin{aligned} E(h) &= \mathbb{E}_{x,y}[L(y,h(x))] && (4) \\ &= \mathbb{E}_{x}\left[\mathbb{E}_{y|x}[L(y,h(x))|x]\right] && (5) \\ &= \mathbb{E}_{x\sim p(x)}\left[\mathbb{E}_{y\sim p(y|x)}\left[L(y,h(x))\right]\right] && (6) \\ &= \mathbb{E}_{x\sim p(x)}\left[E(h,x)\right] && (7) \end{aligned}$$

Components of supervised learning (continued)

We want to pick the "best" hypothesis $h : \mathcal{X} \to \mathcal{Y}$, where "best" is defined by the **loss function** L. In other words, we want to solve the following optimization problem:

$$h^* = \operatorname*{argmin}_{h:\mathcal{X}\to\mathcal{Y}} E(h),$$

where $E(h)$ is defined in (4).

Since $E(h) = \mathbb{E}_x[E(h, x)]$, it suffices to minimize the error pointwise, i.e. solve

$$h^*(x) = \operatorname*{argmin}_{h:\mathcal{X}\to\mathcal{Y}} E(h, x),$$

for all $x \in \mathcal{X}$.

Table of contents

Optimal predictions in regression

In regression (with squared error loss), we want to solve

$$h^*(x) = \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ E(h,x) \tag{8}$$

$$= \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}_{y|x}[(y-h(x))^2|x] \tag{9}$$

for all $x \in \mathcal{X}$.

Let $z = h(x)$., and let us solve (9) for a single input x. We want to minimize

$$E(h,x) = \mathbb{E}_{y|x}[(y-z)^2|x = x].$$

Optimal predictions in regression

The necessary condition for optimality is given by

$$\begin{aligned}
\frac{\mathrm{d}E(h,x)}{\mathrm{d}z} = 0 \iff & \frac{\mathrm{d}\mathbb{E}_{y|x}[(y-z)^2|x=x]}{\mathrm{d}z} && = 0 \\
\iff & \mathbb{E}_{y|x}\left[\frac{\mathrm{d}(y-z)^2}{\mathrm{d}z}|x=x\right] && = 0 \\
\iff & \mathbb{E}_{y|x}\left[-2(y-z)|x=x\right] && = 0 \\
\iff & z = \mathbb{E}_{y|x}\left[y|x=x\right] \\
\iff & h^*(x) = \mathbb{E}_{y|x}\left[y|x=x\right]
\end{aligned}$$

The sufficient condition for optimality (a minimum) is given by

$$\frac{\mathrm{d}^2E(h,x)}{\mathrm{d}z^2} > 0 \iff 2 > 0.$$

Optimal predictions in regression

In regression, the optimal hypothesis is

$$h^*(x) = \mathbb{E}_{y|x}[y|x],$$

the conditional expectation, also known as the **regression function**.

In other words, when *best is measured by expected squared error*, the best prediction of y at any point x is the conditional expectation at x.

If $y = f(x) + \varepsilon$ with $\mathbb{E}[\varepsilon|x] = 0$ and $\text{Var}(\varepsilon|x) = \sigma^2$, then we have

$$h^*(x) = f(x),$$

and

$$E(h, x) = \mathbb{E}_{y|x}[(y - h^*(x))^2|x] = \mathbb{E}_{y|x}[(y - f(x))^2|x] = \sigma^2,$$

which is the smallest conditional risk in regression.

Optimal predictions in classification

Let us consider multi-class classification with K categories where $y \in \mathcal{C} = \{C_1, \ldots, C_K\}$.

With the zero-one loss, we want to solve

$$h^*(x) = \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ E(h, x) \tag{10}$$

$$= \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}_{y|x}[\mathbb{1}\{y \neq h(x)\}|x] \tag{11}$$

for all $x \in \mathcal{X}$.

Note that

$$\mathbb{E}_{y|x}[\mathbb{1}\{y \neq h(x)\}|x] = P(y \neq h(x)|x).$$

Let $z = h(x)$., and let us solve (11) for a single input x. We want to minimize

$$E(h, x) = \mathbb{E}_{y|x}[\mathbb{1}\{y \neq h(x)\}|x = x].$$

Optimal predictions in classification

We have

$$
\begin{aligned}
&\mathbb{E}_{y|x}[\mathbb{1}\{y \neq z\}|x = x] \\
&= \sum_{k=1}^{K} \mathbb{1}\{C_k \neq z\} P(y = C_k|x = x) \\
&= \sum_{k: C_k \neq z} 1 \times P(y = C_k|x = x) + 0 \times P(y = z|x = x) \\
&= \sum_{k: C_k \neq z} P(y = C_k|x = x) \\
&= \sum_{k: C_k \neq z} P(y = C_k|x = x) + P(y = z|x = x) - P(y = z|x = x) \\
&= \sum_{k=1}^{K} P(y = C_k|x = x) - P(y = z|x = x) \\
&= 1 - P(y = z|x = x).
\end{aligned}
$$

Optimal predictions in classification

This implies that

$$\begin{aligned} h^*(x) &= \operatorname*{argmin}_{z \in \mathcal{C}} \mathbb{E}_{y|x}[\mathbb{1}\{y \neq z\} | x = x] \\ &= \operatorname*{argmin}_{z \in \mathcal{C}} 1 - P(y = z | x = x) \\ &= \operatorname*{argmax}_{z \in \mathcal{C}} P(y = z | x = x). \end{aligned}$$

Equivalently, we have that

$$h^*(x) = C_k \text{ if } P(y = C_k | x = x) = \max_{z \in \mathcal{C}} P(y = z | x = x).$$

The optimal classifier is called the **Bayes classifier**, which has the following error rate at x:

$$1 - \max_{k=1,\dots,K} P(y = C_k | x = x),$$

also called the **Bayes error rate**, which gives the lowest possible error rate that could be achieved <u>if we knew</u> $P(y|x)$.