Statistical Learning II

Lecture 2 - Supervised learning

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
18.09.2024

Recap of Probability

The butter of statistical learning

Random variable

A random variable X mathematically formalises the notion of a “measurement” or “random event”.

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Discrete: when the possible outcomes are countable.

Examples:

- the outcome of tossing a coin $X \in \{\text{head}, \text{tail}\}$
- rolling a dice $X \in \{1,\ldots,6\}$
- The number of people in France $X \in \mathbb{N}$

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Discrete: when the possible outcomes are countable.

Examples:

- the outcome of tossing a coin $X \in \{\text{head}, \text{tail}\}$
- rolling a dice $X \in \{1, \ldots, 6\}$
- The number of people in France $X \in \mathbb{N}$

Discrete r.v.s are described by their probability distribution

$$\boxed{\mathbb{P}(X = k)}$$

A positive "function" that sums to one. $\displaystyle\sum_{k \in \mathrm{supp(X)}} \mathbb{P}(X = k) = 1$

Random variable

A random variable X mathematically formalises the notion of a “measurement” or “random event”. It can be:

- Continuous: when the possible outcomes are uncountable.

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Examples:
- The temperature in the room $X \in \mathbb{R}$
- The GDP of France next year $X \in \mathbb{R}$

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Examples:
- The temperature in the room $X \in \mathbb{R}$
- The GDP of France next year $X \in \mathbb{R}$

Continuous r.v.s are described by their probability density function (p.d.f.), which integrates to probabilities:

$$\mathbb{P}(X \in [a, b]) = \int_a^b \mathrm{d}x \; p_X(x)$$

A "function" that integrates to one: $\int_{\mathrm{supp}(X)} \mathrm{d}x \; p_X(x) = 1$

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Examples:
- The temperature in the room $X \in \mathbb{R}$
- The GDP of France next year $X \in \mathbb{R}$

Continuous r.v.s are described by their probability density function (p.d.f.), which integrates to probabilities:

$$\mathbb{P}(X \in [a, b]) = \int_a^b \mathrm{d}x \; p_X(x)$$

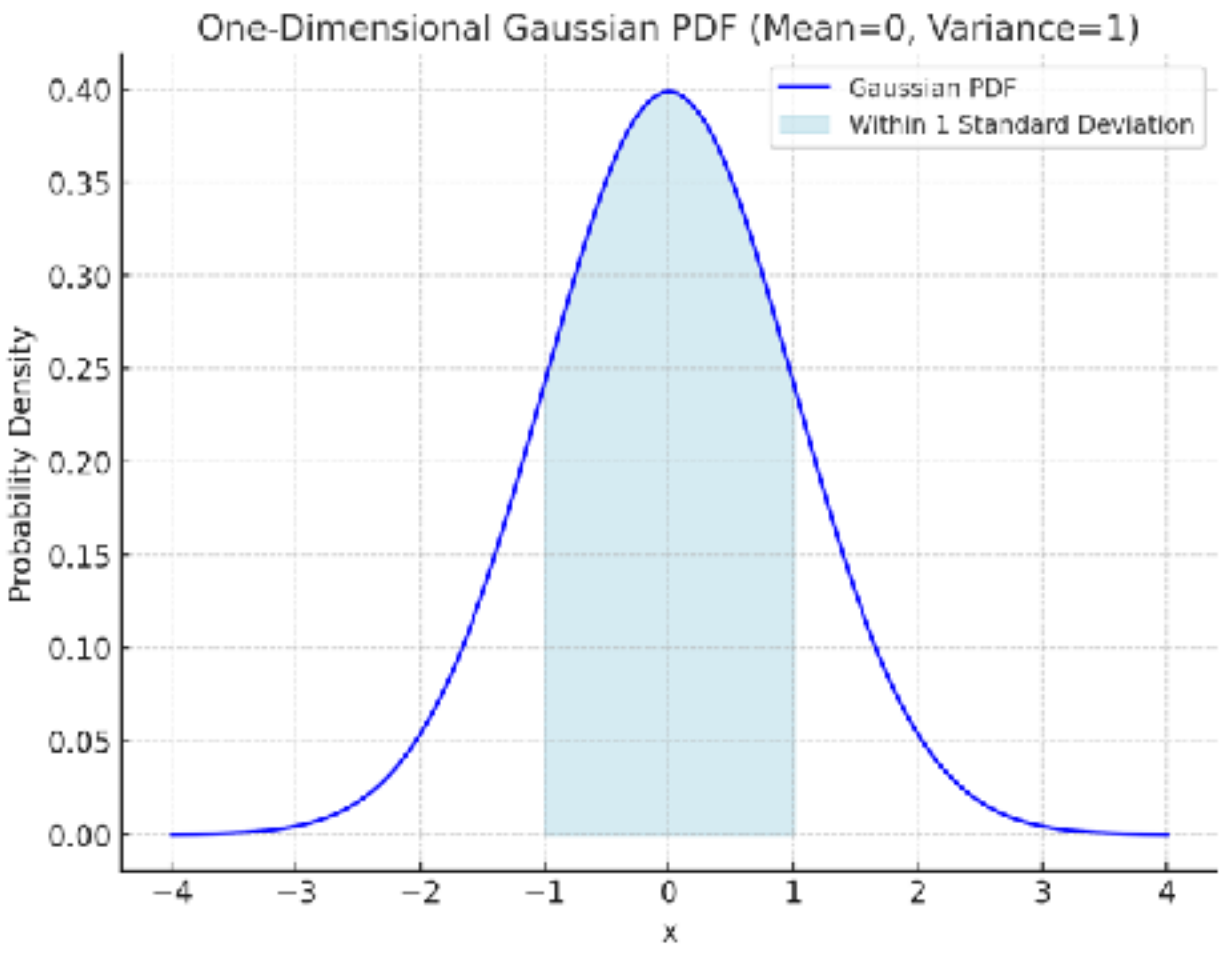
A "function" that integrates to one: $$\int_{\mathrm{supp}(X)} \mathrm{d}x \; p_X(x) = 1$$

The p.d.f. is NOT a probability. It can be negative.

Normal distribution

A Gaussian r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ has the following p.d.f.:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



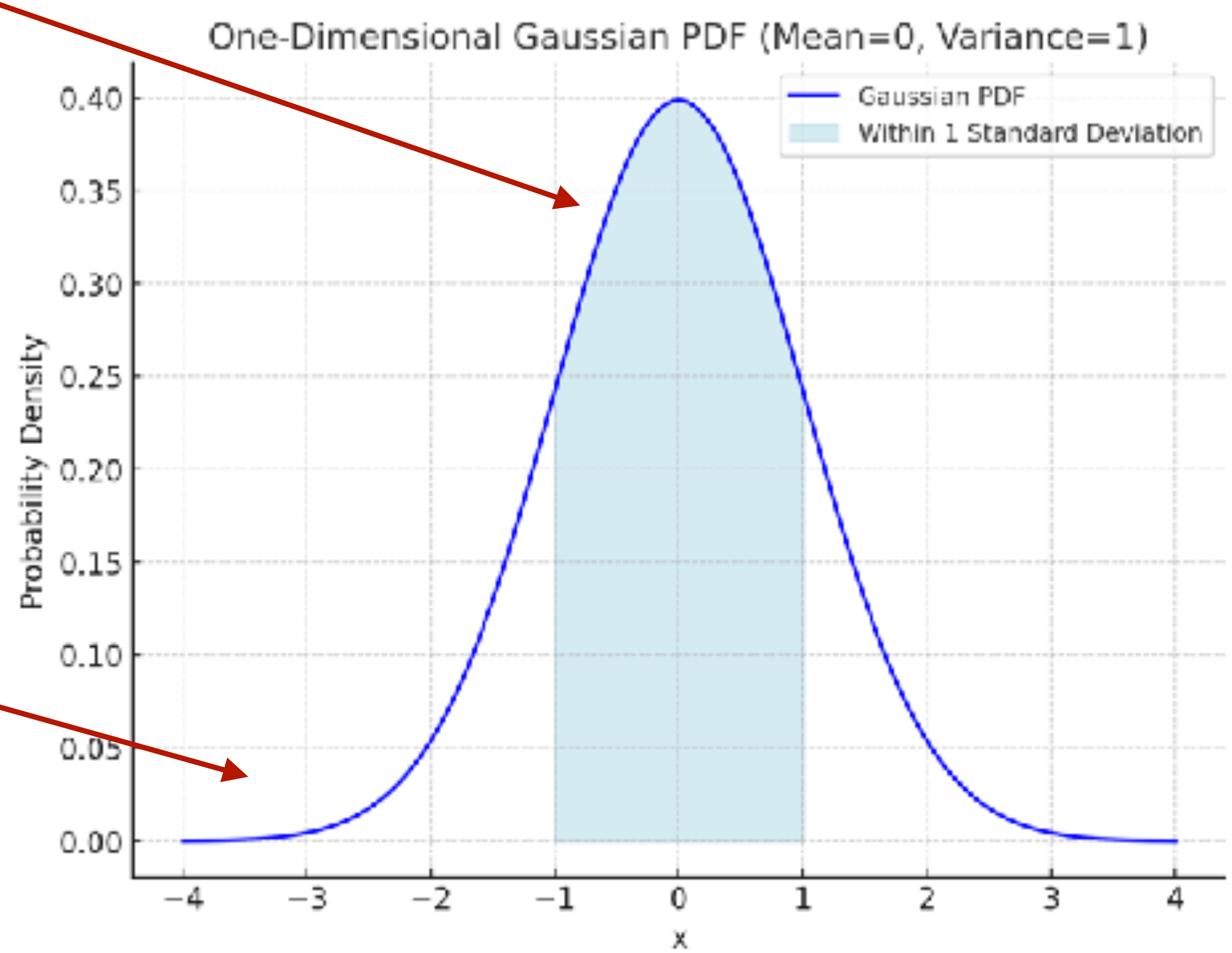
Normal distribution

A Gaussian r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ has the following p.d.f.:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

High-probability

Low-probability



Expectation and variance

Let $X \sim p_X$ denote a continuous r.v.

- The expectation (or mean) of X is defined as

$$\mathbb{E}[X] = \int \mathrm{d}x \; p_X(x)x$$

For example, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}[X] = \mu$

Expectation and variance

Let $X \sim p_X$ denote a continuous r.v.

- The expectation (or mean) of X is defined as

$$\mathbb{E}[X] = \int \mathrm{d}x \; p_X(x) x$$

For example, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}[X] = \mu$

- The variance of X is defined as:

$$\mathrm{Var}[\mathrm{X}] = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2] = \mathbb{E}[\mathrm{X}^2] - \mathbb{E}[\mathrm{X}]^2$$

For example, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathrm{Var}[\mathrm{X}] = \sigma^2$

Change of variables

Let $X \sim p_X$ denote a continuous r.v. and $f : \mathbb{R} \to \mathbb{R}$

Change of variables

Let $X \sim p_X$ denote a continuous r.v. and $f : \mathbb{R} \to \mathbb{R}$

Then, $Y = f(X)$ is also a random variable, with p.d.f. given by

$$p_Y(y) = \int \mathrm{d}x \; p_X(x)\delta(y - f(x))$$

Where $\delta(x)$ is the "Dirac delta function":

$$\int_{\mathbb{R}} \mathrm{d}x \; \delta(x - y)f(x) = f(y)$$

Joint distribution

Two random variables X, Y that concern the same random experiment are characterised by their joint p.d.f.

$$p_{X,Y}(x, y)$$

Joint distribution

Two random variables X, Y that concern the same random experiment are characterised by their joint p.d.f.

$$p_{X,Y}(x,y)$$

The correlation between X, Y is defined by

$$\mathbb{E}[XY] = \int \mathrm{d}x \int \mathrm{d}y \; p_{X,Y}(x,y) xy$$

Joint distribution

Two random variables X, Y that concern the same random experiment are characterised by their joint p.d.f.

$$p_{X,Y}(x, y)$$

The correlation between X, Y is defined by

$$\mathbb{E}[XY] = \int \mathrm{d}x \int \mathrm{d}y \; p_{X,Y}(x, y) xy$$

We say X, Y are uncorrelated if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Independence

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the marginal distributions

$$p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y) \qquad\qquad p_Y(y) = \int \mathrm{d}x \; p_{X,Y}(x, y)$$

Independence

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the marginal distributions

$$p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y) \qquad p_Y(y) = \int \mathrm{d}x \; p_{X,Y}(x, y)$$

- We say the r.v.s. X, Y are independent if

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

Independence

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the marginal distributions

$$p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y) \qquad\qquad p_Y(y) = \int \mathrm{d}x \; p_{X,Y}(x, y)$$

- We say the r.v.s. X, Y are independent if

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

Note that independence implies uncorrelated,
but not the converse!

Exercise: Construct a counter-example

Conditional distribution

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the conditional p.d.f.

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Conditional distribution

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the conditional p.d.f.

$$p_{X|Y}(x\,|\,y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Note that $X, Y \sim p_{X,Y}$ are independent if and only if:

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

Conditional distribution

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the conditional p.d.f.

$$p_{X|Y}(x\,|\,y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Note that $X, Y \sim p_{X,Y}$ are independent if and only if:

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

Theorem (Bayes theorem)

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

Law of large numbers

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$

Define the sample mean (note this is itself a r.v.)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Law of large numbers

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$

Define the sample mean (note this is itself a r.v.)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Theorem (Weak LLN)

$$\bar{X}_n \xrightarrow{P} \mu \qquad \text{as} \qquad n \to \infty$$

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| < \epsilon) = 1$$

Be aware there are many variations of the LLN.

Central limit theorem

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Again, consider the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Central limit theorem

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Again, consider the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Theorem (Lindeberg CLT)

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$$

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(\bar{X}_n - \mu) \le z) = \mathbb{P}(Z \le z/\sigma) \qquad Z \sim \mathcal{N}(0,1)$$

Be aware there are many variations of the CLT.

Supervised learning

Supervised Learning

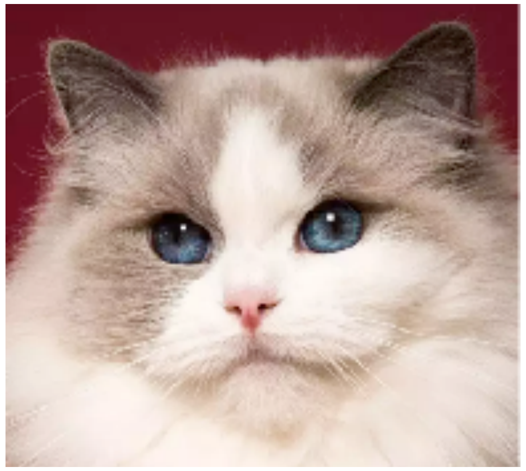
Not grumpy

Grumpy

Supervised Learning

Not grumpy

Grumpy



???

Supervised Learning

Inputs / covariates $x \in \mathcal{X}$

x_1

x_2

x_3

y_1
Not grumpy

y_2
Grumpy

y_3
???

Outputs / Labels / response $y \in \mathcal{Y}$

Supervised Learning

Inputs / covariates $x \in \mathcal{X}$

x_1

x_2

x_3

y_1
Not grumpy

y_2
Grumpy

y_3
???

Outputs / Labels / response $y \in \mathcal{Y}$

Supervised Learning

Inputs / covariates $x \in \mathcal{X}$

x_1

x_2

x_3

$f : \mathcal{X} \to \mathcal{Y}$

y_1
Not grumpy

y_2
Grumpy

y_3
???

Outputs / Labels / response $y \in \mathcal{Y}$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1, \ldots, n\}$ denote the training data.

- The input space $\mathscr{X}$ is often assumed to be a vector space $\mathscr{X} \subset \mathbb{R}^d$. But keep in mind in real life it can be any data structure (e.g. a pandas.DataFrame)

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

- The input space $\mathscr{X}$ is often assumed to be a vector space $\mathscr{X} \subset \mathbb{R}^d$. But keep in mind in real life it can be any data structure (e.g. a `pandas.DataFrame`)

- The output space $\mathscr{Y}$ is often assumed to be a subset $\mathscr{Y} \subset \mathbb{R}$.

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

- The input space $\mathscr{X}$ is often assumed to be a vector space $\mathscr{X} \subset \mathbb{R}^d$. But keep in mind in real life it can be any data structure (e.g. a `pandas.DataFrame`)
- The output space $\mathscr{Y}$ is often assumed to be a subset $\mathscr{Y} \subset \mathbb{R}$.
- In particular, if $|\mathscr{Y}| = k$ is a discrete set, we say we have a classification problem.

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

- The input space $\mathscr{X}$ is often assumed to be a vector space $\mathscr{X} \subset \mathbb{R}^d$. But keep in mind in real life it can be any data structure (e.g. a pandas.DataFrame)
- The output space $\mathscr{Y}$ is often assumed to be a subset $\mathscr{Y} \subset \mathbb{R}$.
- In particular, if $|\mathscr{Y}| = k$ is a discrete set, we say we have a classification problem.
- If $\mathscr{Y}$ is a continuous set, we say we have a regression problem

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

- The input space $\mathscr{X}$ is often assumed to be a vector space $\mathscr{X} \subset \mathbb{R}^d$. But keep in mind in real life it can be any data structure (e.g. a pandas.DataFrame)
- The output space $\mathscr{Y}$ is often assumed to be a subset $\mathscr{Y} \subset \mathbb{R}$.
- In particular, if $|\mathscr{Y}| = k$ is a discrete set, we say we have a classification problem.
- If $\mathscr{Y}$ is a continuous set, we say we have a regression problem

It is very common to consider a one-hot encoding $\mathscr{Y} = \{0,1\}^k$ in classification.

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

Examples of classification:

- Grumpy vs. Non-grumpy cats

$\mathscr{X} = \{\text{photos of cats}\}$, $\mathscr{Y} = \{\text{grumpy}, \text{not grumpy}\}$

- E-mail spam detection

$\mathscr{X} = \{\text{your inbox}\}$, $\mathscr{Y} = \{\text{spam}, \text{not spam}\}$

- Medical diagnosis

$\mathscr{X} = \{\text{medical data}\}$, $\mathscr{Y} = \{\text{diseases}\}$

- Sentiment analysis

$\mathscr{X} = \{\text{text}\}$, $\mathscr{Y} = \{\text{positive}, \text{negative}, \text{neutral}\}$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

Examples of regression:

- Temperature prediction

  $\mathscr{X} = \mathbb{R}^3$, $\mathscr{Y} = \mathbb{R}$

- Stock price prediction

  $\mathscr{X} = \{\text{list of stocks}\}$, $\mathscr{Y} = \mathbb{R}_+$

- Life expectancy

  $\mathscr{X} = \{\text{medical data}\}$, $\mathscr{Y} = \mathbb{R}_+$

- Any price, cost, income, etc. prediction.

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1, \ldots, n\}$ denote the training data.

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

In supervised learning, our goal is to use the data to learn a function that correctly assigns the labels to the responses.

$$f : \mathscr{X} \to \mathscr{Y}$$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1, \ldots, n\}$ denote the training data.

In supervised learning, our goal is to use the data to learn a function that correctly assigns the labels to the responses.

$$\boxed{f : \mathscr{X} \to \mathscr{Y}}$$

For classification, it is common to define instead:

$$\boxed{f : \mathscr{X} \to [0,1]^{|\mathscr{Y}|}}$$

Where $f(x)$ is a vector of class probabilities. In this case, final prediction is given by:

$$\boxed{\hat{y} = \underset{k \in |\mathscr{Y}|}{\operatorname{argmax}} f(x)}$$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

In supervised learning, our goal is to use the data to learn a function that correctly assigns the labels to the responses.

$$f : \mathscr{X} \to \mathscr{Y}$$

Two key words: correctly and learn.

Loss function

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

In supervised learning, our goal is to use the data to learn a function that correctly assigns the labels to the responses.

$$f : \mathscr{X} \to \mathscr{Y}$$

Two key words: correctly and learn. To quantify the first, it is common to introduce a loss function:

$$\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$$

Loss function

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathscr{Y} : i = 1,\ldots,n\}$ denote the training data.

In supervised learning, our goal is to use the data to learn a function that correctly assigns the labels to the responses.

$$\boxed{f : \mathscr{X} \to \mathscr{Y}}$$

Two key words: correctly and learn. To quantify the first, it is common to introduce a loss function:

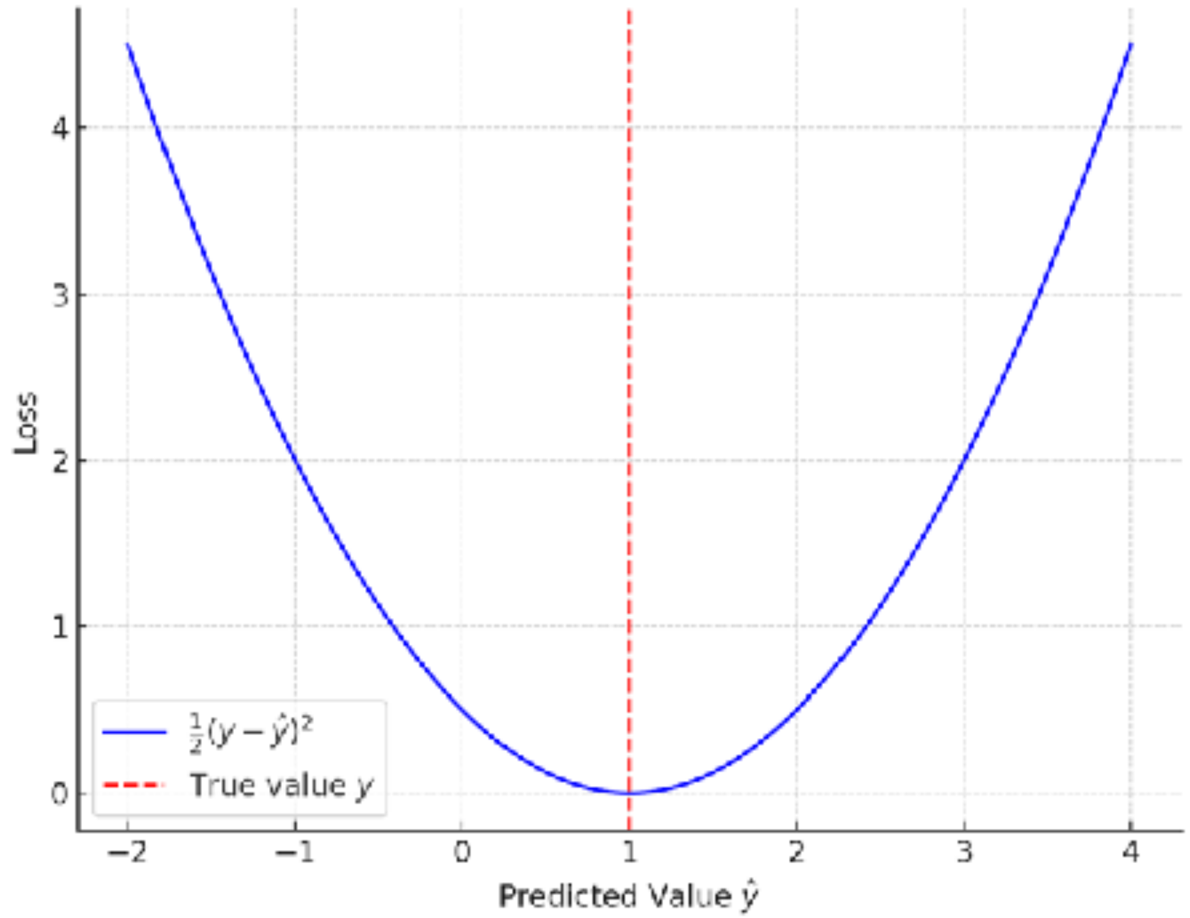
$$\boxed{\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+}$$

For classification this will also depend on the encoding.

Regression losses

Examples in regression:

- Square loss: $\ell(y, z) = \frac{1}{2}(y - z)^2$



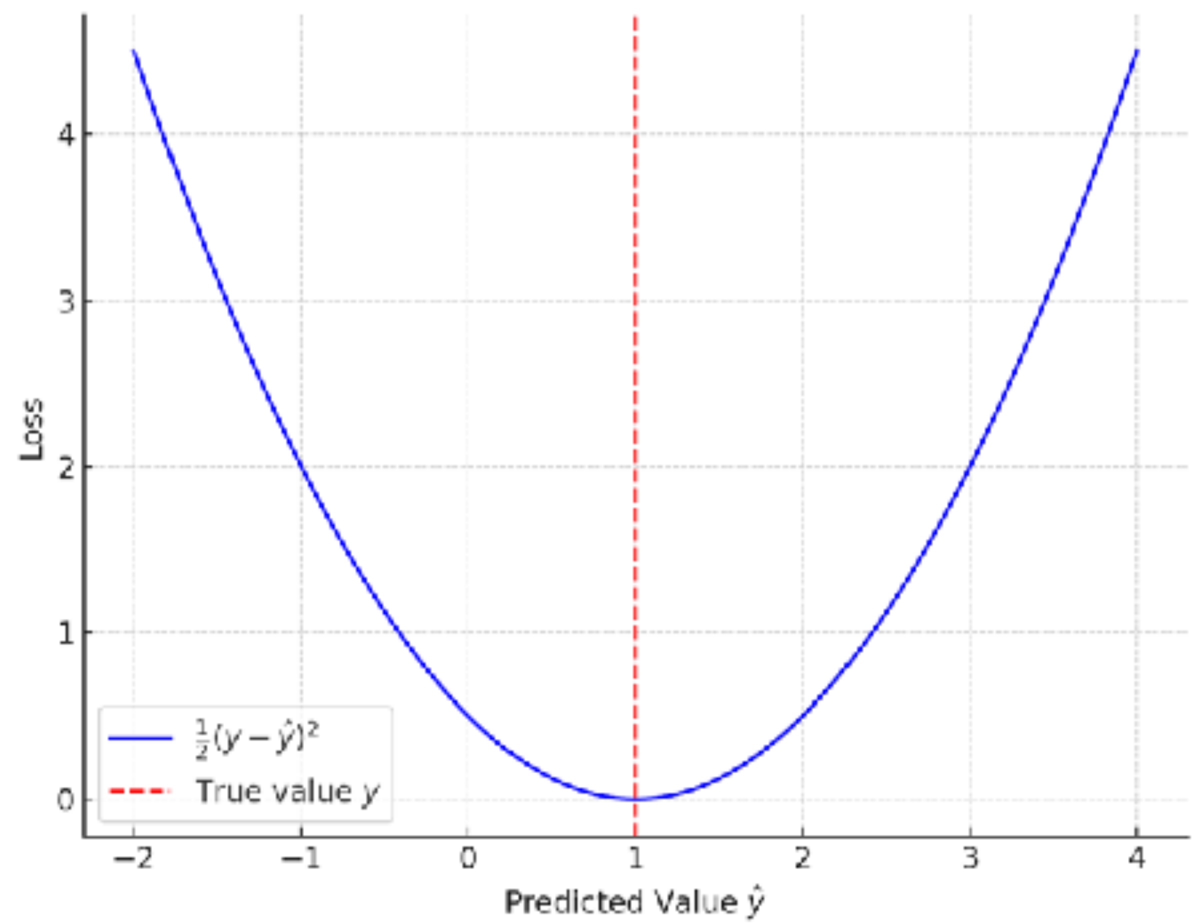
Regression losses

Examples in regression:

- Square loss: $\ell(y, z) = \frac{1}{2}(y - z)^2$

The square loss is sensitive to outliers

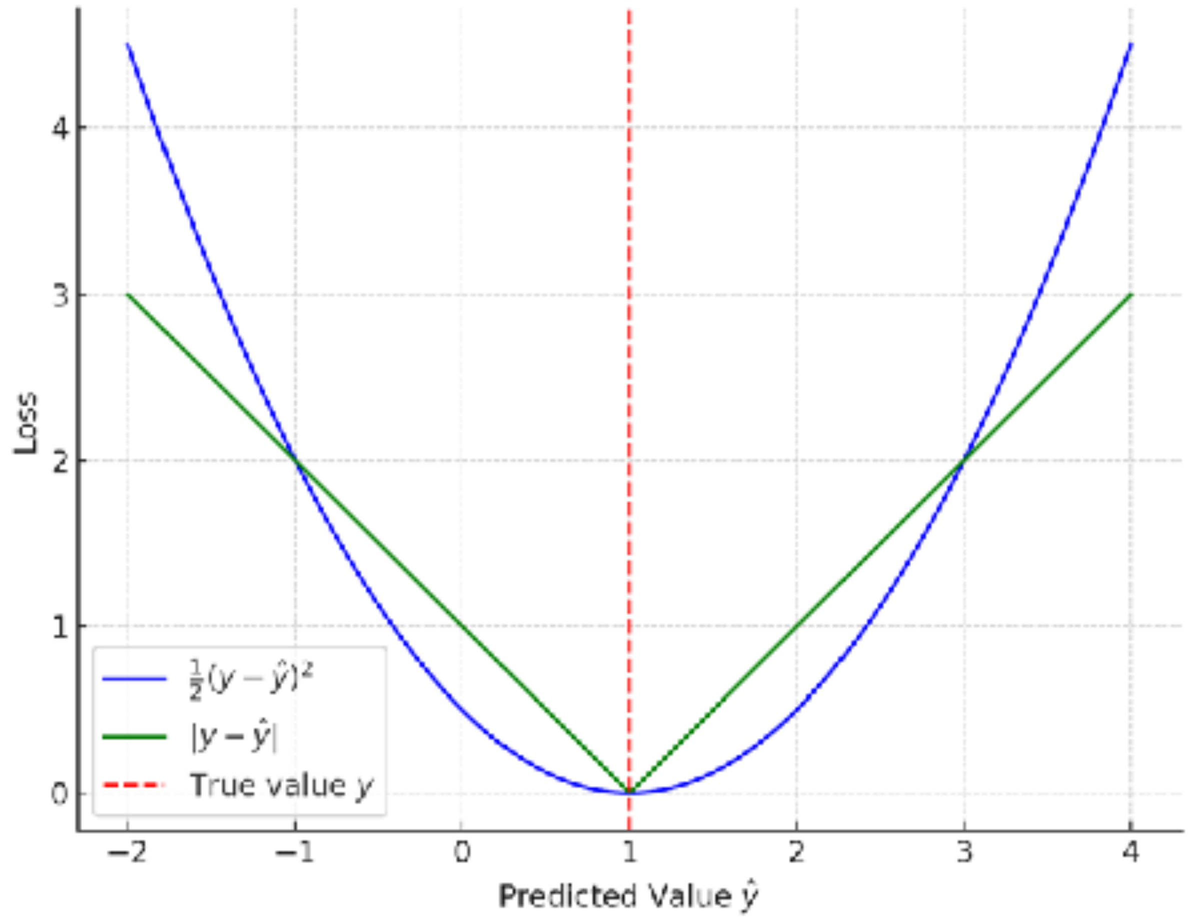
Exercise: show this.



Regression losses

Examples in regression:

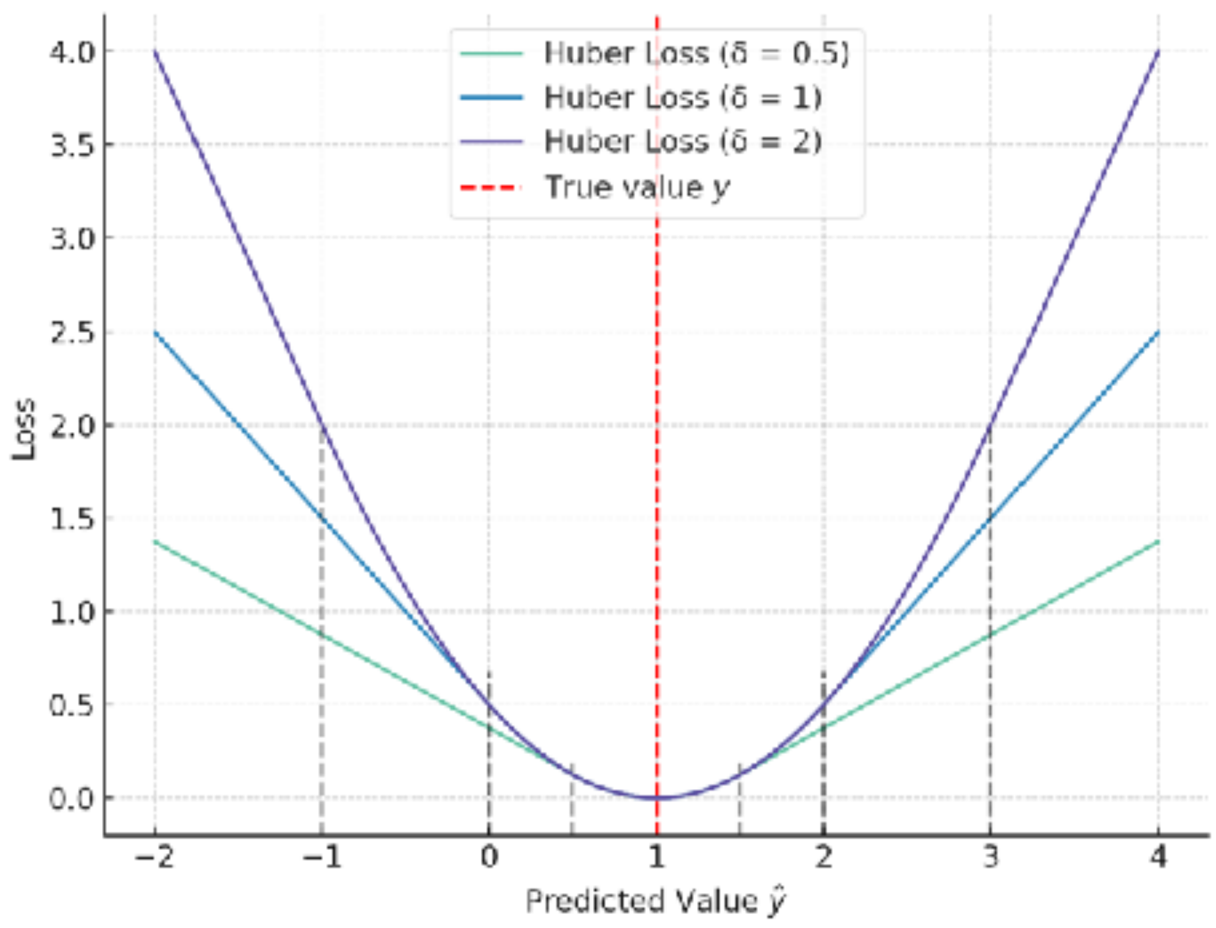
- Square loss: $\ell(y, z) = \dfrac{1}{2}(y - z)^2$
- Absolute loss: $\ell(y, z) = |y - z|$



Regression losses

Examples in regression:

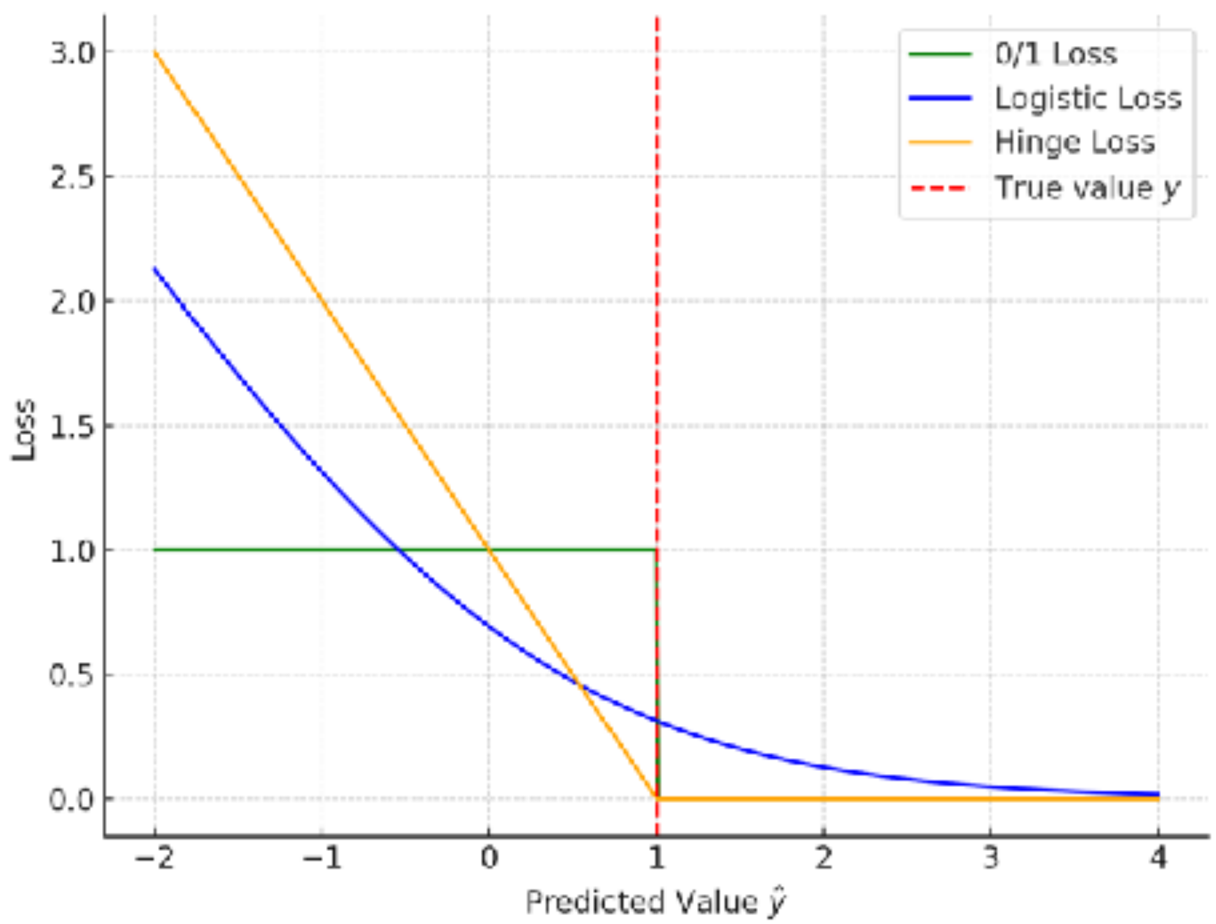
- Huber loss: $\ell_\delta(y,z) = \begin{cases} \frac{1}{2}(y-z)^2 & \text{if } ||y-z| \le \delta \\ \delta(|y-z| - \frac{1}{2}\delta) & \text{if } |y-z| > \delta \end{cases}$



Classification losses

Examples in binary classification $\mathcal{Y} = \{-1, +1\}$:

- 0/1 loss: $\ell(y, z) = \delta_{yz}$ (or $\ell(y, z) = \theta(y - z) = \begin{cases} 1 & \text{if } y - z \leq 0 \\ 0 & \text{otherwise} \end{cases}$)
- Logistic loss: $\ell(y, z) = \log(1 + e^{-yz})$
- Hinge loss: $\ell(y, z) = \max(0, 1 - yz)$



Empirical risk

Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \ldots, n\}$ denote the training data.

Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and a predictor $f : \mathcal{X} \to \mathcal{Y}$ define the empirical risk:

$$\hat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

Also known as the training loss.

Empirical risk

Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1,\ldots,n\}$ denote the training data.

Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and a predictor $f : \mathcal{X} \to \mathcal{Y}$ define the empirical risk:

$$\hat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

Also known as the training loss. This quantifies how well we fit the data. But is this a good notion of learning?

$$f(x) = \begin{cases} y_i & \text{if } x \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad \hat{\mathcal{R}}_n = 0$$

Probabilistic framework

Instead, we would like predictors that do well on unseen data.

Probabilistic framework

Instead, we would like predictors that do well on unseen data.

Assume there is an underlying data distribution p over $\mathcal{X} \times \mathcal{Y}$:

$$(x_i, y_i) \sim p \qquad \text{i.i.d.}$$

Probabilistic framework

Instead, we would like predictors that do well on unseen data.

Assume there is an underlying data distribution p over $\mathcal{X} \times \mathcal{Y}$:

$$(x_i, y_i) \sim p \qquad \text{i.i.d.}$$

- The "i.i.d." assumption might not always hold. (Sampling bias, distribution shift, etc.)

- Under this assumption, $\hat{\mathcal{R}}_n$ is a random function.

Population risk

Instead, we would like predictors that do well on unseen data.

Assume there is an underlying data distribution p over $\mathcal{X} \times \mathcal{Y}$:

$$(x_i, y_i) \sim p \qquad \text{i.i.d.}$$

Define the notion of population risk of a predictor $f : \mathcal{X} \to \mathcal{Y}$:

$$\mathcal{R}(f) = \mathbb{E}\left[\ell(y, f(x))\right]$$

Also known as the generalisation or test error.

Population risk

Instead, we would like predictors that do well on unseen data.

Assume there is an underlying data distribution p over $\mathcal{X} \times \mathcal{Y}$:

$$(x_i, y_i) \sim p \qquad \text{i.i.d.}$$

Define the notion of population risk of a predictor $f : \mathcal{X} \to \mathcal{Y}$:

$$\mathcal{R}(f) = \mathbb{E}\left[\ell(y, f(x))\right]$$

Also known as the generalisation or test error.

$\mathcal{R}$ is a deterministic function of the predictor f