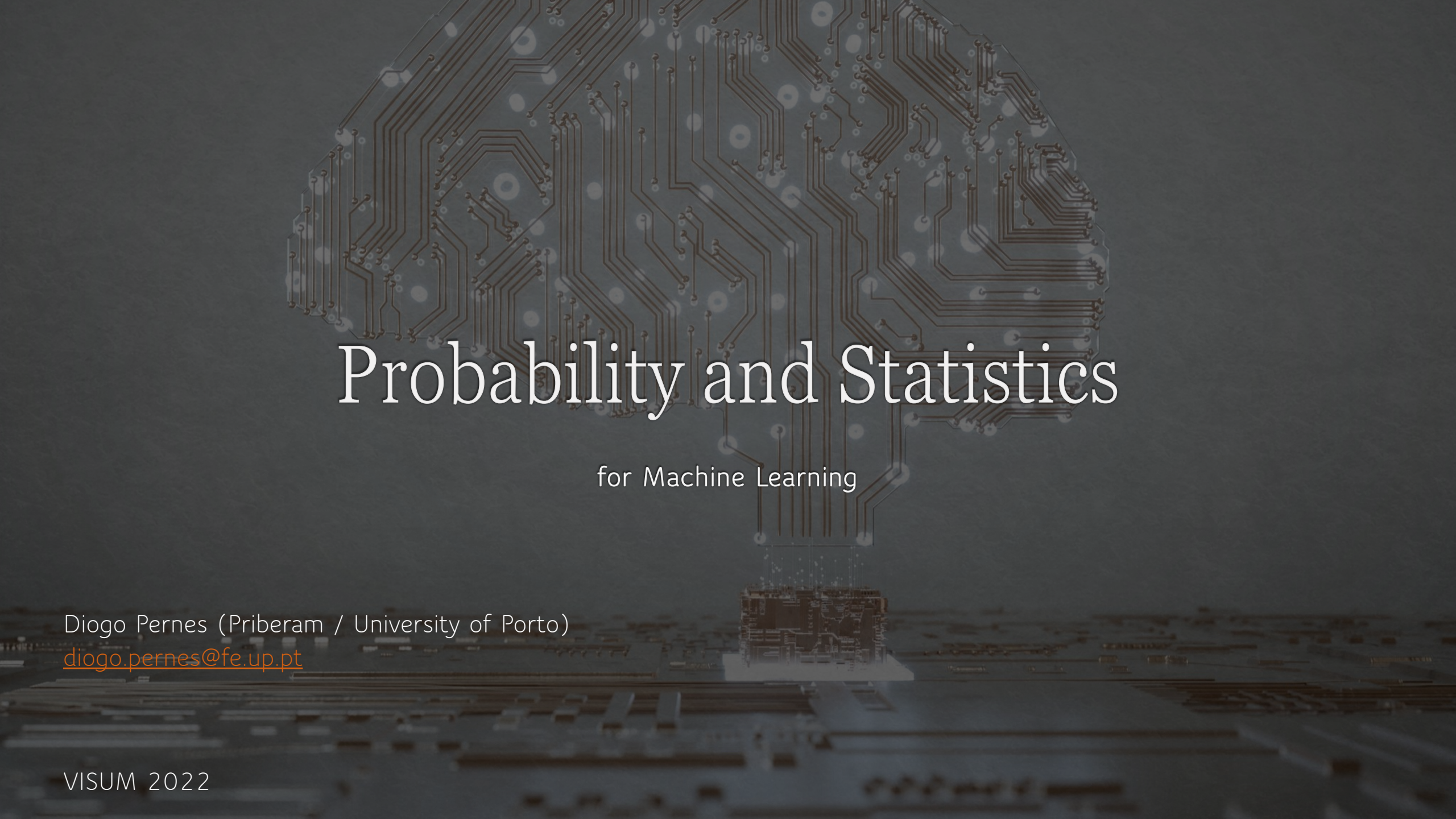


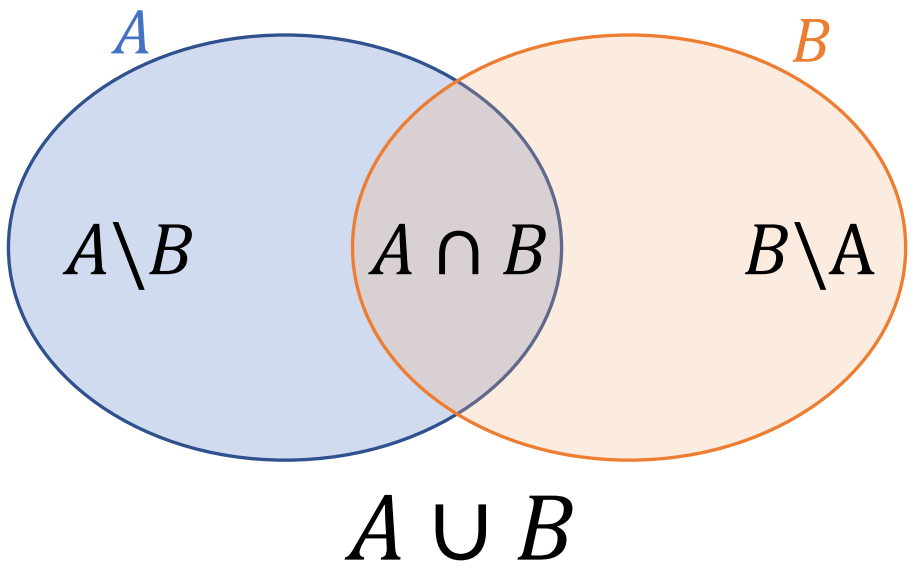
Probability and Statistics

for Machine Learning

Diogo Pernes (Priberam / University of Porto)
diogo.pernes@fe.up.pt

VISUM 2022

Sets: unions and intersections



What is probability?

Mathematically, it is essentially any function that obeys to some *axioms*:

Ω – sample space

1. $P(A) \geq 0, \ \forall A \subseteq \Omega$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots$ mutually exclusive,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

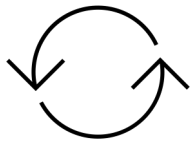
How to interpret probability?

Classical interpretation:

«The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.»

Laplace, *Théorie analytique des probabilités*

$$P(A) = \frac{\#\{\text{cases where } A \text{ occurs}\}}{\#\{\text{possible cases}\}}$$



How to interpret probability?

Frequentist interpretation:

Given a *repeatable* random experiment, the probability of an event A is the proportion of times that A occurs if we repeat the experiment many times.

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

How to interpret probability?

Bayesian interpretation:

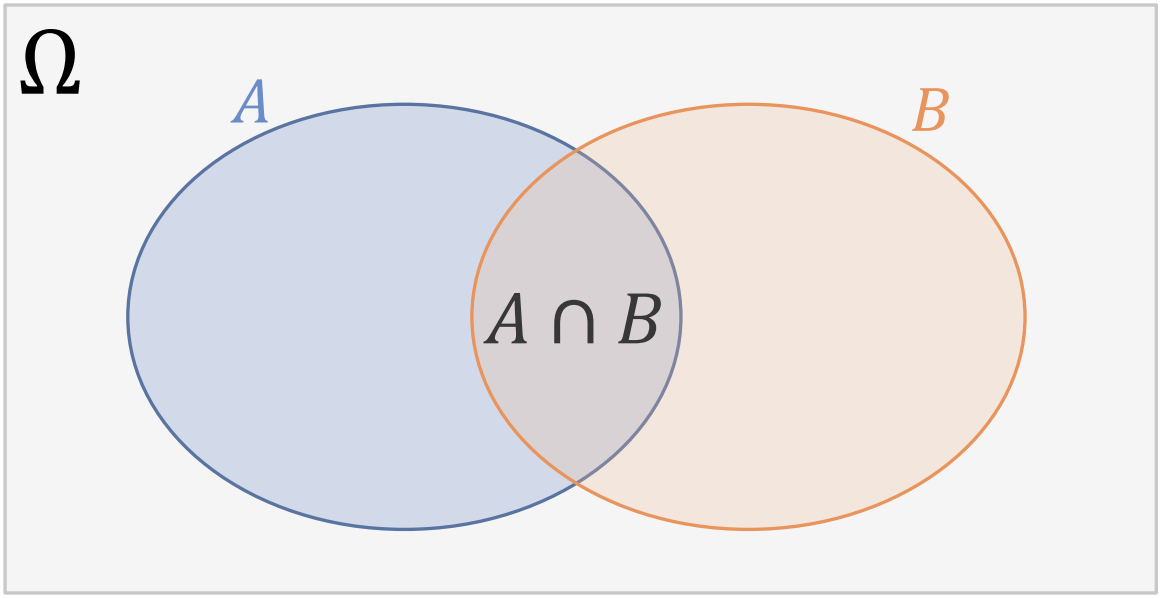
The probability of an event *A* measures the degree of belief about the occurrence of *A*.

Conditional probability

The probability of an event *A* when one knows/assumes that an event *B* has occurred is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

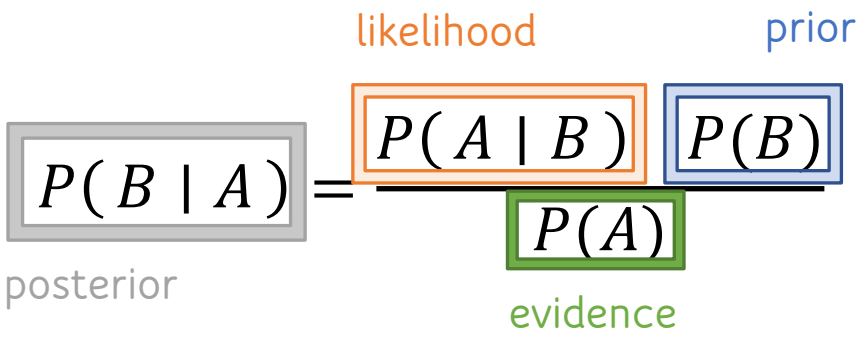
Conditional probability



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \mid \Omega) = \frac{P(A \cap \Omega)}{P(\Omega)} = \frac{P(A)}{1} = P(A)$$

Bayes rule


likelihood
prior
$P(B \mid A) = \frac{P(A \mid B)\,P(B)}{P(A)}$
posterior
evidence

Bayes rule

A disease *D* affects about 1 out of 10,000 people. There is a test to diagnose the disease that is quite accurate:

- If you have the disease, the test comes positive in 99% of the cases (true positive).
- If you do not have the disease, the test comes positive in 5% of the cases (false positive).

If you decide to the test and it comes positive, how likely is it that you have the disease?

$$P(D) = 0.01\% \qquad P(+ \mid D) = 99\% \qquad P(+ \mid D') = 5\%$$

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)} \approx 0.20\% \ \text{(!!!)}$$

Random variables

Take each event in the sample space Ω and assign it to a real number X, e.g.

- Flip a coin 10 times, count the number of "heads", $X \in \{0, 1, \dots, 10\}$
- Count the number of coin tosses until you get "heads", $X \in \{0, 1, \dots\}$
- Reveal an image which could be either of a cat or of a dog, $X \in \{0, 1\}$

discrete

- Choose a person from the audience randomly and measure the person's height, $X \in [0, \infty)$

continuous

Discrete random variables

Are characterized by a probability function $P(X)$ that assigns a probability value to every possible value of X.

$$P(X = 0) = P(X = 1) = 0.5$$

$$P(X = x) \geq 0 \quad \forall x$$

$$\sum_{\forall x} P(X = x) = 1$$

Continuous random variables

Are characterized by a probability density function p that assigns a probability to an interval of values for X.

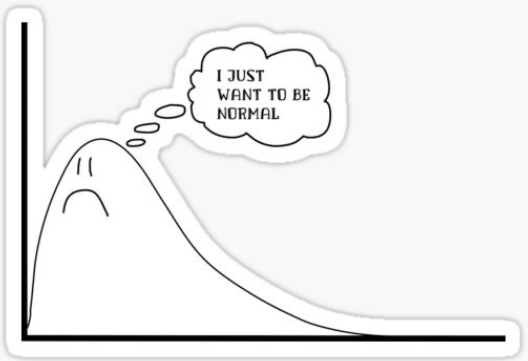
$$P(a \leq X \leq b) = \int_a^b p(x)\, dx$$

$$p(x) \geq 0 \;\; \forall x$$

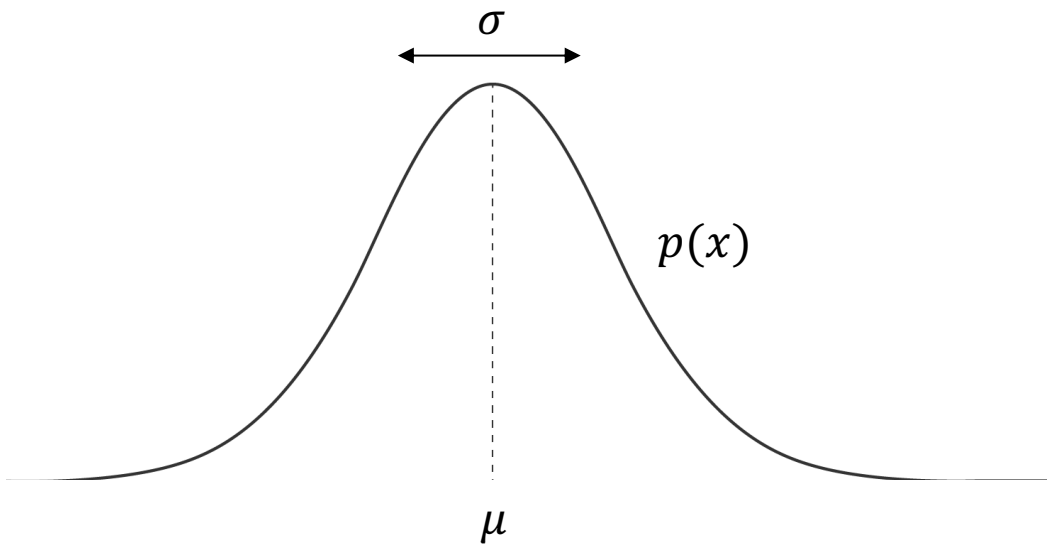
$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

$$P(X = a) = 0 \;\; \forall a$$

Normal/Gaussian distribution



$$X \sim N(\mu, \sigma^2)$$

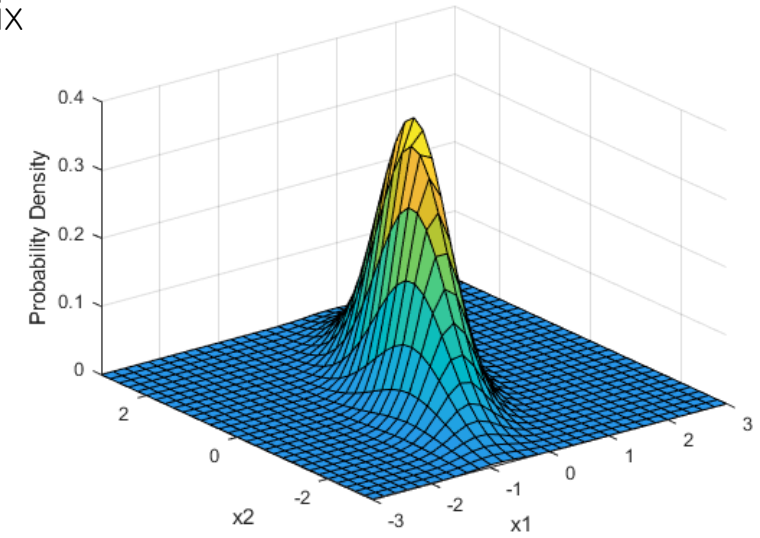


Normal/Gaussian distribution

(multivariate case)

$$(X_1, X_2, \ldots, X_d) \sim N(\mu, \Sigma)$$

μ is a d-dimensional vector
Σ is a $d \times d$ matrix



Expectation

$$\mathbb{E}_p[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)\,dx \qquad \text{(continuous } X\text{)}$$

$$\mathbb{E}_P[f(X)] = \sum_{\forall x} f(x)P(X = x) \qquad \text{(discrete } X\text{)}$$

Expectation

Intuition:

The expectation gives us the result we would get if we could observe the outcome of $f(X)$ many times and then average the results.

$$\mathbb{E}_p[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i), \quad x_i \sim p$$

Expectation

Mean: $\mu = \mathbb{E}[X]$

Variance: $\sigma^2 = \mathbb{E}[(X - \mu)^2]$

Entropy: $H(p) = \mathbb{E}[-\log(p(X))]$

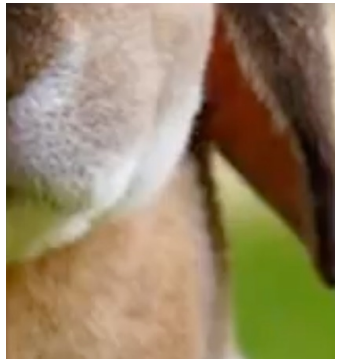
Conditional probability functions

Purpose:

Predicting one random variable Y when we observe a particular value x for another random variable X.

$$P(Y \mid X = x) = g(x)$$

E.g.: Image classification



x

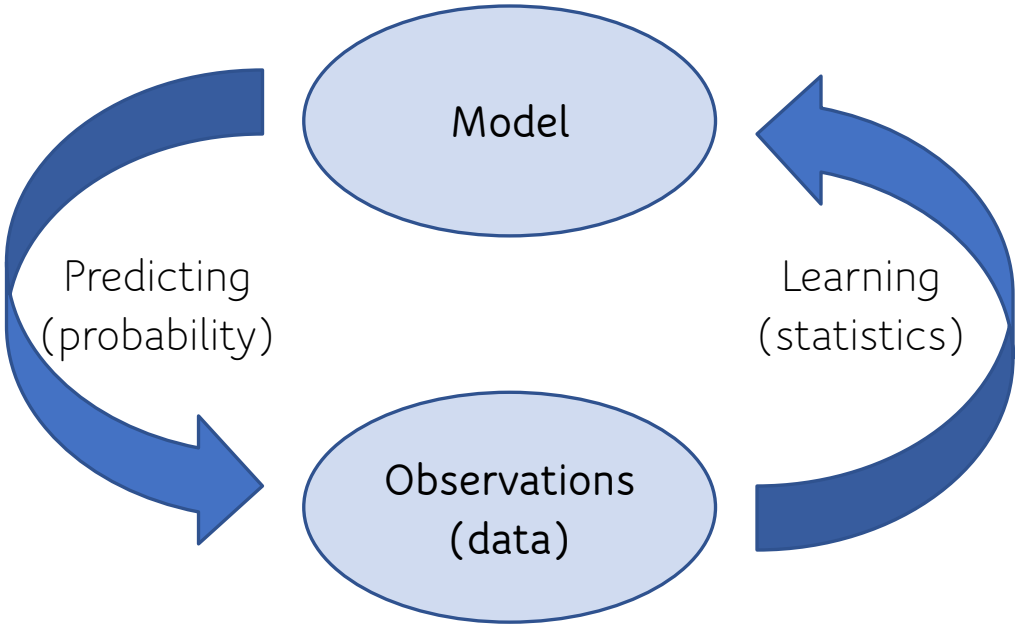
$$P(Y = \text{rabbit} \mid X = x) = 60\%$$

$$P(Y = \text{dog} \mid X = x) = 35\%$$

$$P(Y = \text{cat} \mid X = x) = 3\%$$

…

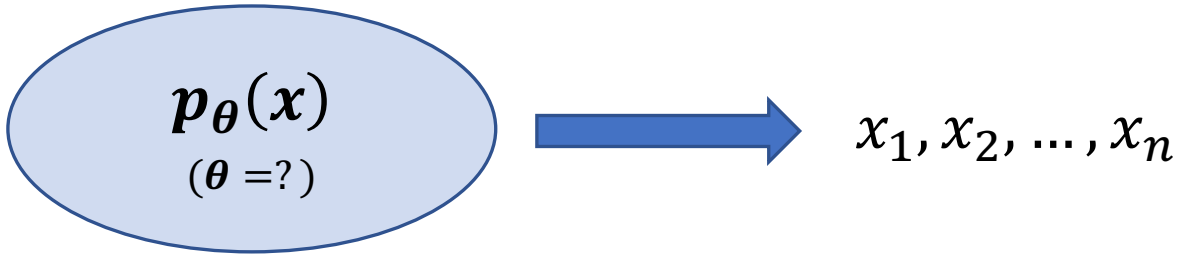
Probability vs. statistics



Parameter estimation

(learning problem)

You have a distribution (model) p_θ, with unknown parameters θ, and some data $x_1, x_2, \dots, x_n$ that was drawn from this distribution.



How to estimate θ from this data?

E.g.: Estimate μ and σ for $N(\mu, \sigma^2)$.

Maximum likelihood estimation

Idea:

Find the parameter θ^* that maximizes the likelihood (probability) of the observed data.

$$\theta^* = \operatorname{argmax}_\theta \prod_{i=1}^{n} p_\theta(x_i)$$

The function $\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$ is the likelihood function.

For numerical reasons, it is usually better to maximize its logarithm instead. And the maximizer is the same!

$$l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

Maximum likelihood estimation

Example:

You throw a (possibly) unfair coin 100 times and it turns heads 30 times. What is the maximum likelihood estimate of the probability of the coin turning heads?

$\theta \in [0,1] \quad P(H) = P(X = 1) = \theta \quad P(T) = P(X = 0) = 1 - \theta$

$\mathcal{L}(\theta) = P(X = 0)^{70} P(X = 1)^{30} = (1 - \theta)^{70} \theta^{30}$

$l(\theta) = 70 \log(1 - \theta) + 30 \log \theta$

$\frac{dl}{d\theta}(\theta^*) = 0 \Leftrightarrow \cdots \Leftrightarrow \theta^* = {}^{30}/_{100} = 0.30$

Maximum likelihood estimation

What if I don't know what distribution p_θ should be?

Just choose some very flexible model (e.g. a neural network).

How to solve the problem when θ is high-dimensional?

Then of course we can't solve it analytically, but we can apply numerical optimization methods to maximize the log-likelihood function (e.g. gradient ascent).

Information-theoretic perspective:

KL-divergence

Kullback-Leibler divergence (aka relative entropy):

Given two probability densities p and q, the KL divergence from q to p is defined as:

$$D_{KL}(p||q) = \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right] = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}\ dx$$

$$= \int_{-\infty}^{\infty} p(x)\log p(x)\,dx - \int_{-\infty}^{\infty} p(x)\log q(x)\,dx$$

$$= -H(p) - \mathbb{E}_p[\log(q)]$$

cross-entropy

$$= H(p,q) - H(p)$$

Information-theoretic perspective:

KL-divergence

Fact:

$D_{KL}(p||q)$ is always non-negative, being zero if and only if $p = q$.

So if we want p_θ to approximate the unknown data distribution p_{data}, then a possibility is to look for the θ that minimizes the KL divergence between these two distributions.

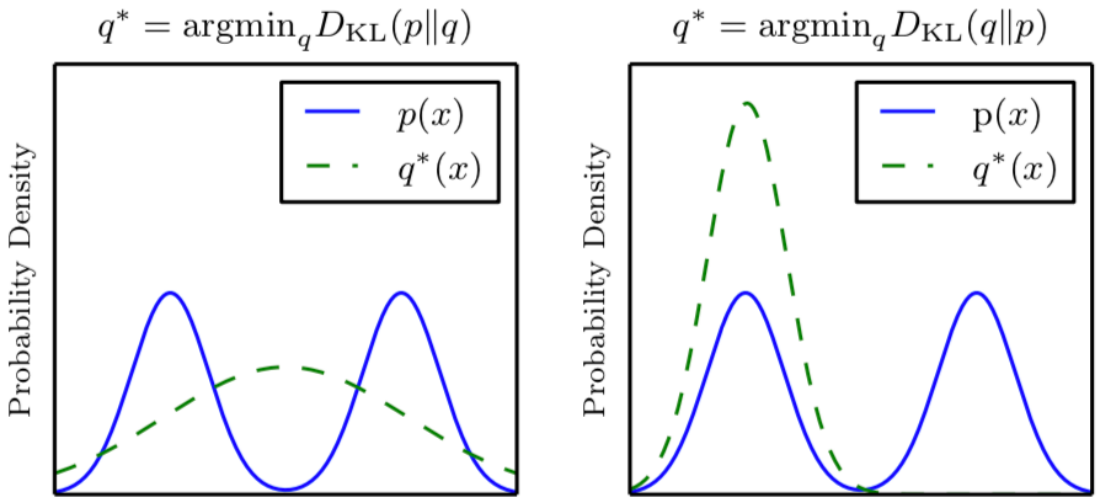
Constraint:

All we know about p_{data} are a few samples $x_1, x_2, \ldots, x_n$ drawn from it.

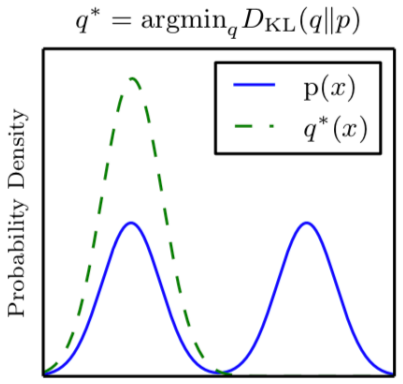
Problem:

$D_{KL}(p||q) \neq D_{KL}(q||p)$, so which one to choose?

KL-divergence asymmetry



Minimizing the KL-divergence (backward)



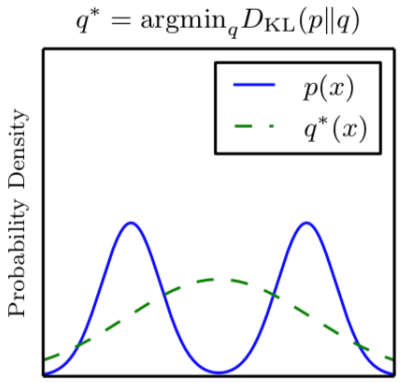
Let us try first minimizing $D_{KL}(p_\theta || p_{\text{data}})$:

$$D_{KL}(p_\theta || p_{\text{data}}) = H(p_\theta, p_{\text{data}}) - H(p_\theta)$$

$$= \int_{-\infty}^{\infty} p_\theta(x) \log p_{\text{data}}(x)\, dx - H(p_\theta)$$

???

Minimizing the KL-divergence (forward)



Fortunately, minimizing $D_{KL}(p_{\text{data}}||p_\theta)$ is easier:

$$D_{KL}(p_{\text{data}}||p_\theta) = H(p_{\text{data}}||p_\theta) - \cancel{H(p_{\text{data}})}$$

does not depend on $\boldsymbol{\theta}$

$$H(p_{\text{data}}, p_\theta) = -\mathbb{E}_{p_{\text{data}}}[\log p_\theta]$$

$$\approx -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i) = -l(\theta)$$

Minimizing $D_{KL}(p_{\text{data}}||p_\theta)$ $\Leftrightarrow$ Maximum likelihood estimation!

The Bayesian approach

Bayesians treat anything that you are uncertain about as a random variable. Since we don't know θ, we can treat it as random variable.

Then, we need to define:

- A prior over parameters, $p(\theta)$.
- A likelihood function, $p(x \mid \theta)$ (formerly denoted as $p_\theta(x)$).

The Bayesian approach

Then, given data $x_1, x_2, \ldots, x_n$, instead of finding a pointwise estimate for θ, we can use Bayes rule to update our belief about θ given the new evidence (i.e. the data):

$$p(\theta \mid x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n \mid \theta)p(\theta)}{p(x_1, x_2, \ldots, x_n)}$$

$$= \frac{\prod_{i=1}^{n} p(x_i \mid \theta)\, p(\theta)}{p(x_1, x_2, \ldots, x_n)}$$

$$= \frac{\mathcal{L}(\theta)p(\theta)}{p(x_1, x_2, \ldots, x_n)}$$

Maximum a-posteriori probability

If we really want a pointwise estimate θ^*, we can simply take the mode of this posterior distribution:

$$\theta^* = \text{argmax}_\theta \, p(\theta \mid x_1, x_2, \ldots, x_n)$$

$$= \text{argmax}_\theta \, \frac{\mathcal{L}(\theta)p(\theta)}{p(x_1, x_2, \ldots, x_n)}$$

$$= \text{argmax}_\theta \;\; \mathcal{L}(\theta)p(\theta)$$

This θ^* is called the maximum a-posteriori estimate of θ.

Maximum a-posteriori probability

E.g.: Model $p(x \mid \theta)$, Gaussian prior $\theta \sim N(0, \boxed{\sigma_0^2} I)$

fixed (hyperparameter)

$$p(\theta) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{\|\theta\|^2}{2{\sigma_0}^2}\right)$$

$$\theta^* = \text{argmax}_\theta \; \mathcal{L}(\theta) p(\theta)$$

$$\text{argmax}_\theta \; \log \mathcal{L}(\theta) + \log p(\theta)$$

$$\text{argmax}_\theta \; l(\theta) + \log p(\theta)$$

$$\text{argmax}_\theta \; l(\theta) + \log\left[\frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{\|\theta\|^2}{2{\sigma_0}^2}\right)\right]$$

$$\text{argmin}_\theta \; \boxed{-l(\theta)} + \frac{1}{2{\sigma_0}^2} \boxed{\|\theta\|^2}$$

cross-entropy loss L2 regularization

Thank you!

Questions?