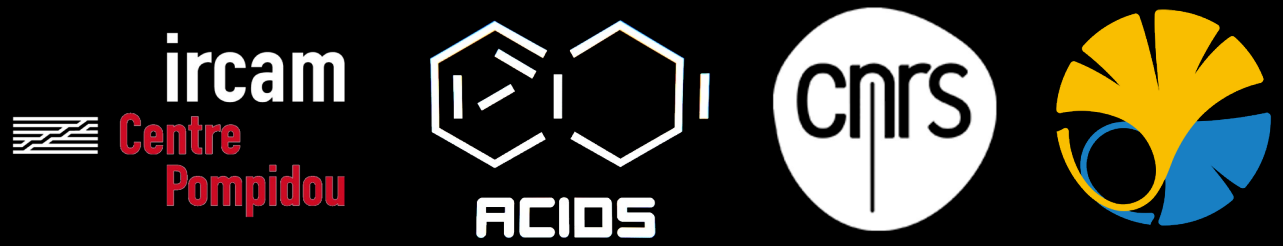


Probabilistic and Bayesian models

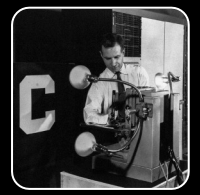
Creative Machine Learning - Course 05

Pr. Philippe Esling
esling@ircam.fr

Brief history of AI

1943 - Neuron

First model by McCulloch & Pitts (purely theoretical)



1957 - Perceptron

Actual **learning machine** built by Frank Rosenblatt

Learns character recognition analogically

1986 - Backpropagation

First to learn neural networks efficiently (*G. Hinton*)

1989 - Convolutional NN

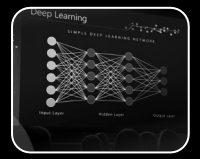
Mimicking the vision system in cats (*Y. LeCun*)

This lesson

2012 - Deep learning

Layerwise training to have deeper architectures

Swoop all state-of-art in classification competitions



Lesson #6

2015 - Generative model

First wave of interest in generating data

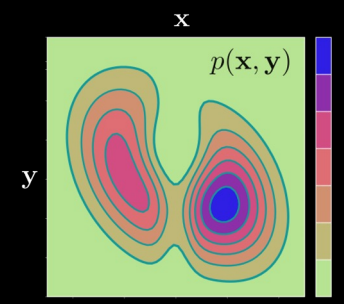
Led to current model craze (VAEs, GANs, Diffusion)

2012 onwards Deep learning era

Brief history of AI

Pre-requisites for understanding generative models

1

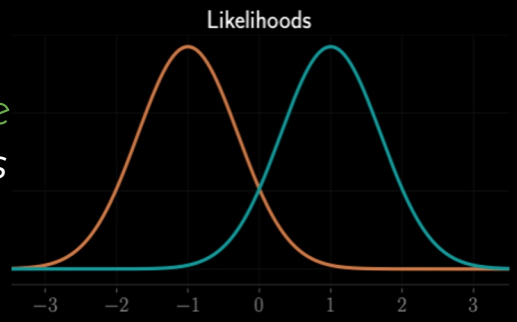


Probability theory

Random variables, distributions, independence

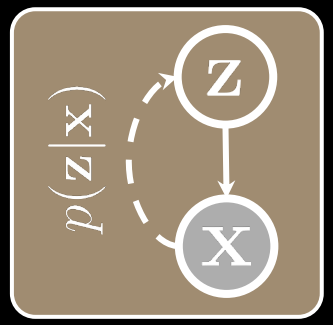
Bayesian inference

Bayes' theorem, likelihood, conjugate priors



2

3

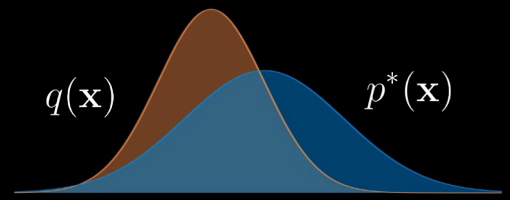


Latent models

Latent variables, probability graphs

Approximate inference

Latent variables, probability graphs



4

Lesson #6

2015 - *Generative model*

First wave of interest in generating data

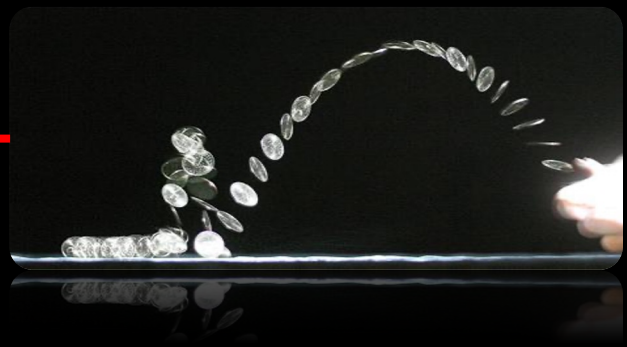
Led to current model craze (VAEs, GANs, Diffusion)

2012 onwards Deep learning era

Probability - intuition

We can witness events that appear uncertain or random

- Typical example of *tossing a coin* or *rolling a die*
- *Events* that appear random (too complex)
- If we had access to all variables
- We might predict exactly the outcome



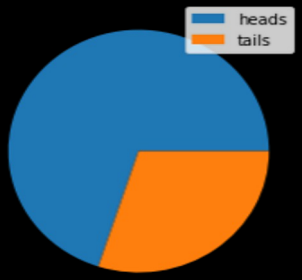
Say we *observe* a set of *samples* $\mathcal{X} = \{\mathbf{x}_i\}_{i \in [1,n]}$

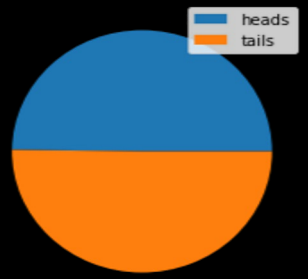
For instance, we can throw our coin a very large number of times

$\mathcal{X} =$ 

- Each throw is *independent* (appears random)
- But we can *expect* some behavior on the *distribution* of all throws

5 throws





10000 throws

Probability - terminology

Experiment (or *trial*)
Action with an uncertain outcome

$\Omega = \{\text{Heads}, \text{Tails}\}$

Sample space Ω
set of all possible outcomes of an experiment

$\omega_0 = Heads$
$\omega_1 = Tails$

Sample point $\omega \in \Omega$
a single possible outcome in the set

Event (or *sample*) $X : \omega \in \Omega \rightarrow X(\omega) \in \mathcal{E}$
outcome of single instance of an experiment

$p(\{\}) = 0$
$p(\Omega) = 1$

Probability value $p(\omega) \in [0, 1]$
indicates the likelihood of a particular event

$p(\omega) = 0$: event is impossible
$p(\omega) = 1$: event is inevitable

Allows to define the three axioms of probability

Probability - terminology

The three axioms of probability

1. For any event $\omega \in \Omega \quad 0 \leq p(\omega) \leq 1$
2. Probability of the sample space $p(\Omega) = 1$
3. For a set of disjoint events $\omega_1, \cdots, \omega_n$, we have $p(\bigcup_{i=1}^{n} \omega_i) = \sum_{i=1}^{n} p(\omega_i)$

Practical example (coin tossing)

Going back to our example of coin tossing, we have

$$p(\{Heads\}) = \frac{1}{2} = 0.5 \qquad p(\{Tails\}) = \frac{1}{2} = 0.5$$

The probability of having heads **or** tails is $p(\{Tails\} \cup \{Heads\}) = 0.5 + 0.5 = 1$

Note also that $p(\{Tails\} \cup \{Heads\}) = p(\{Tails, Heads\}) = p(\Omega) = 1$

The probability of having heads **and** tails is $p(\{Tails\} \cap \{Heads\}) = 0$

These events are said to be *disjoint*

Basic probabilities

Given two random events a and b, the probability of having one **or** the other is

$$p(a \cup b) = p(a) + p(b) - p(a \cap b)$$

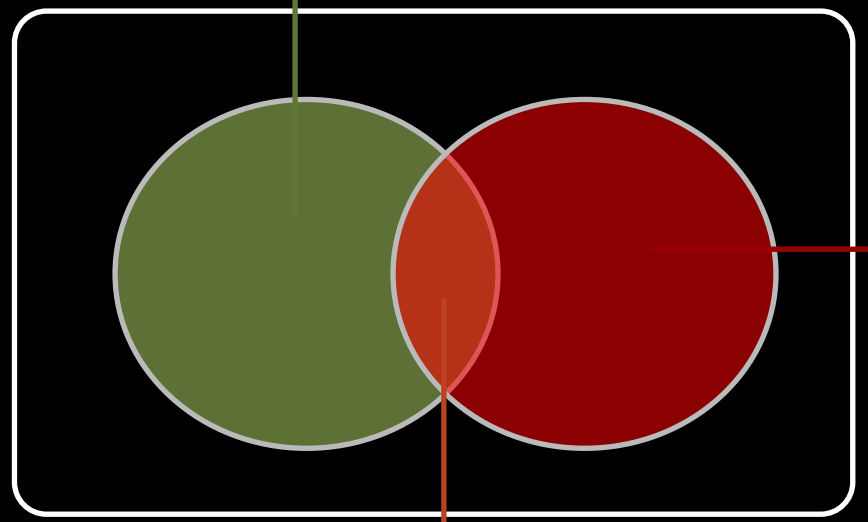
Intuition: notion of space

Space of possibilities

Probability of landing anywhere in the space $p(\Omega) = 1$

Probability of landing on infinitesimal point $p(\{\}) = 0$

Space of our coin toss



Probability of landing in the circle
area of the circle divided by area of space

If we add the two areas, this intersection is counted twice

Conditional probability

Given a dice, we can define two random events

a : Getting a 6

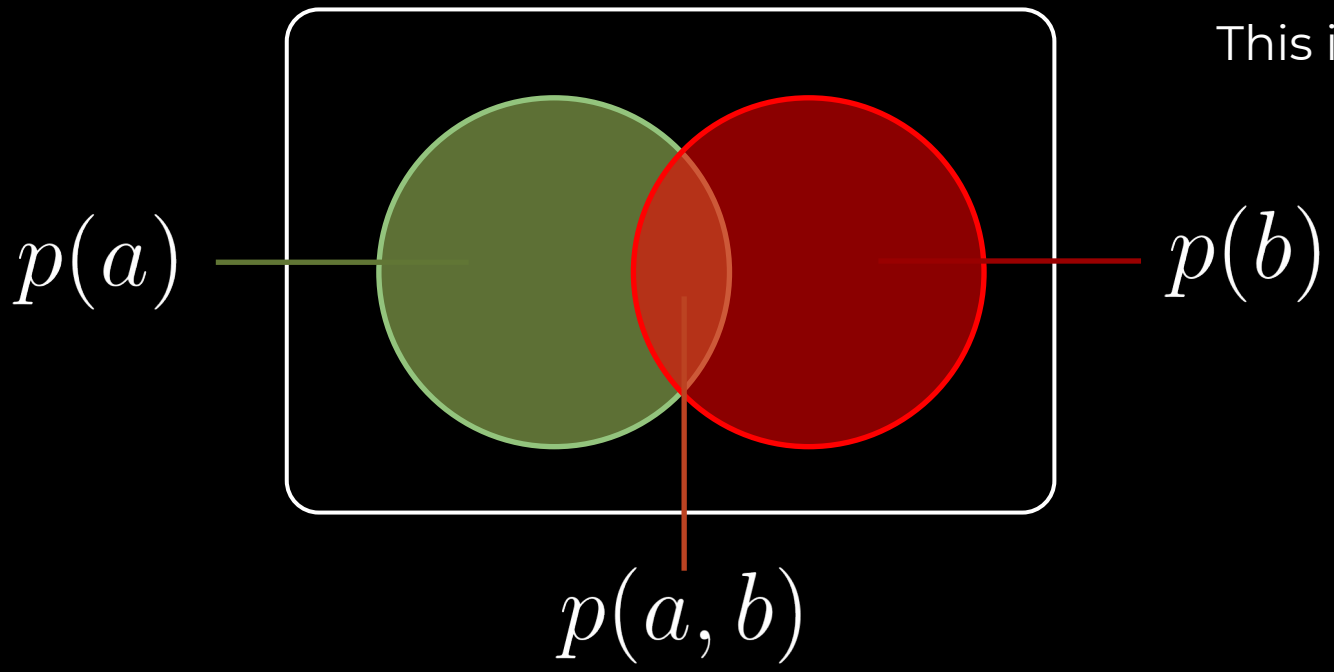
b : Getting an even number

Now if we observe b, it changes the probability of a

Joint (*and*) notation

$$p(a \cap b) = p(a, b)$$

Conditional probability



This is expressed by the *conditional probability*

$$p(a|b) = \frac{p(a, b)}{p(b)}$$

So it is the probability of a *If we restrict the world of possibilities to* b

Conditional probability

Definition $p(a|b) = \frac{p(a,b)}{p(b)}$

$$p(a,b) = p(a|b)p(b)$$

$p(a,b,c) = ?$ $\quad y = b,c$

= p(a,b|c) p(c)
= p(a|b,c) p(b|c) p(c)

$$\begin{aligned} &= p(a,y) \\ &= p(a|y)p(y) \\ &= p(a|b,c)p(b,c) \\ &= p(a|b,c)p(b|c)p(c) \end{aligned}$$

$$p\left(x_1, \cdots, x_n\right) = \prod_{i=n}^{1} p\left(x_i | x_{i-1}, \cdots, x_1\right)$$

Chain rule

Independence

We now define two random events over a dice and a coin

a : Getting a 6 on the dice

b : Getting *heads* with the coin

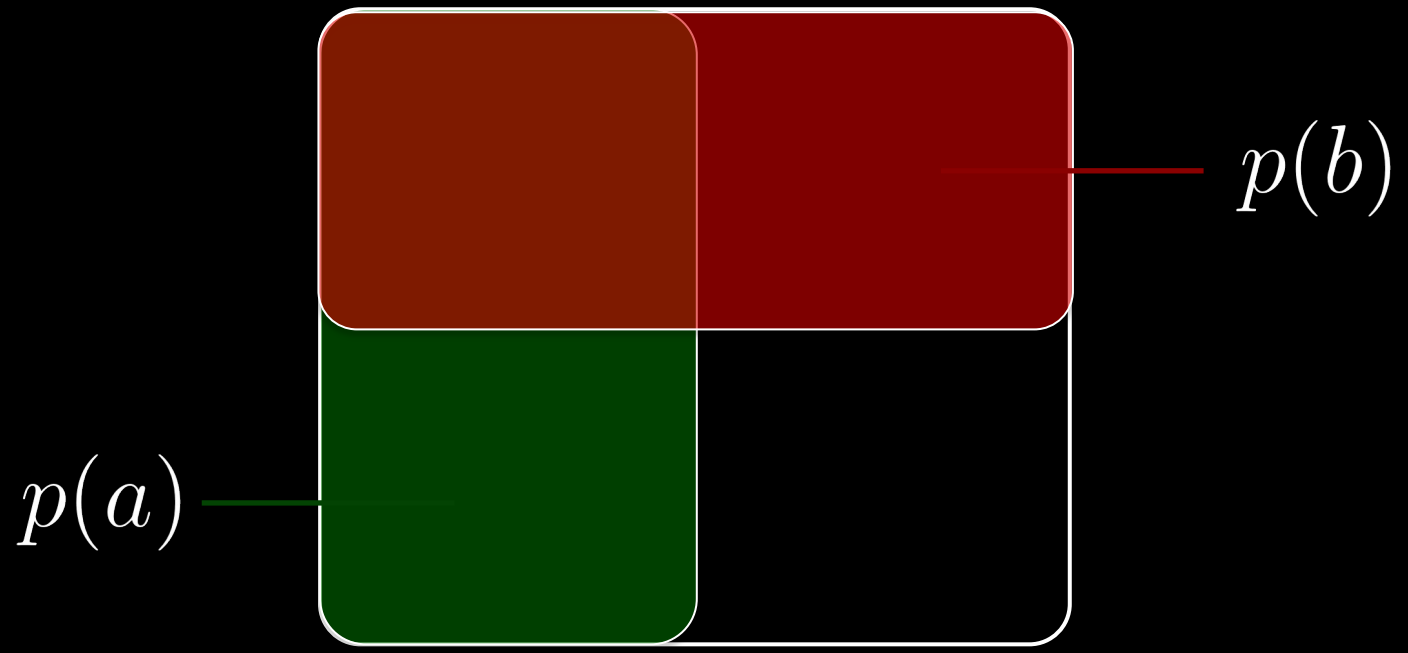
Observing a will **supposedly not influence** b

Definition

a independent of b

$$p(a|b) = p(a)$$

a and b are disjointe

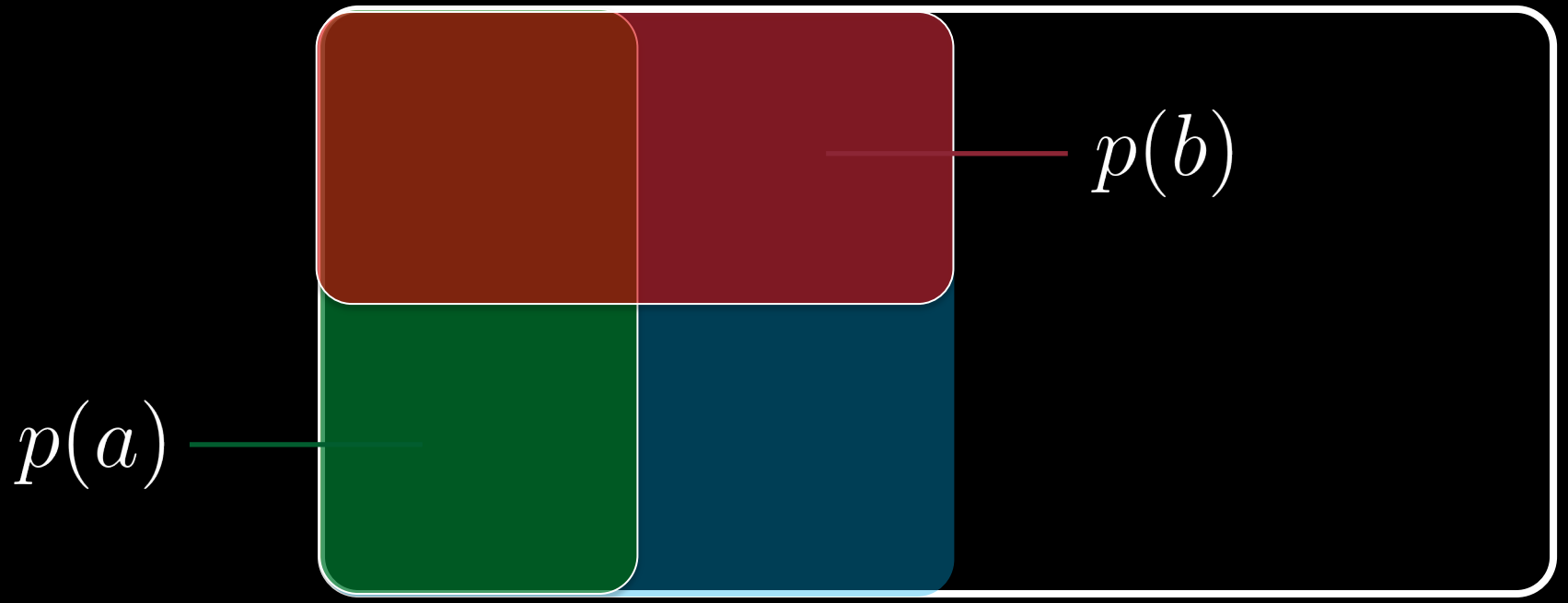


Conditional independence

= p(a|b,z) p(b|z)

Definition $p(a, b|z) = p(a|z)p(b|z)$

$p(a|b, z) = p(a|z)$



Note that **independence does not imply conditional independence**

(and conversely)

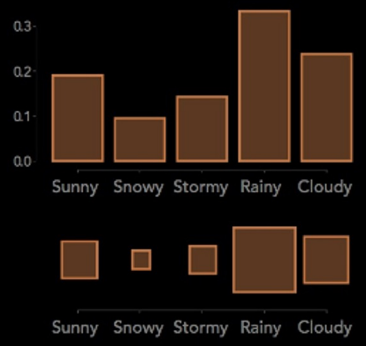
Random variables

Definition

A random variable X is a function that maps events from the sample space Ω to a measurable space $\mathcal{E}$

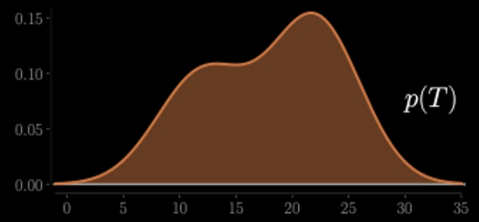
$$X : \omega \in \Omega \to X(\omega) \in \mathcal{E}$$

Notion of distribution



Discrete random variable

Defined on a countable set of *discrete* values
Follows a *probability mass function (pmf)*



Continuous random variable

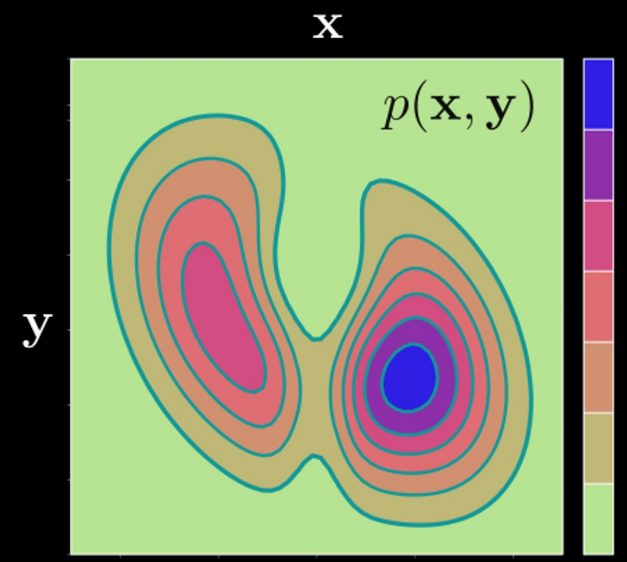
Defined on a continuous domain
Follows a *probability density function (pdf)*

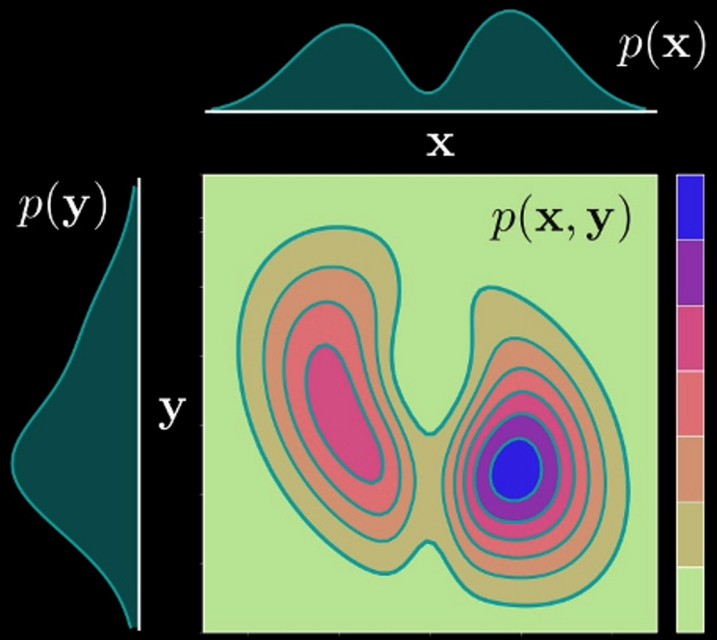
Joint probability and marginalization

Joint distribution

If we consider two random variables $\mathbf{x}$ and $\mathbf{y}$.
We can observe their *paired* outcomes.
This defines the ***joint* probability distribution**

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$$





Marginalization

Given the joint distribution $p(\mathbf{x}, \mathbf{y})$
We can retrieve $p(\mathbf{x})$ through ***marginalization***

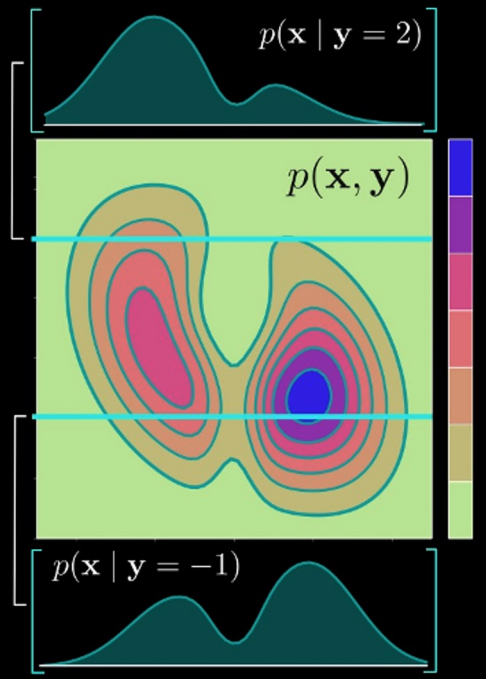
$$p(\mathbf{x}) = \int_{\mathcal{R}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$$

$p(\mathbf{x})$ $p(\mathbf{y})$

Marginal distributions

$$p(\mathbf{x}, \mathbf{y}) = \int_{\mathcal{R}} \int_{\mathcal{R}} p(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{w} d\mathbf{z}$$

Conditional probability and independence



Conditional probability

The conditional probability of $\mathbf{x}$ given $\mathbf{y}$ $p(\mathbf{x}|\mathbf{y} = y_i)$

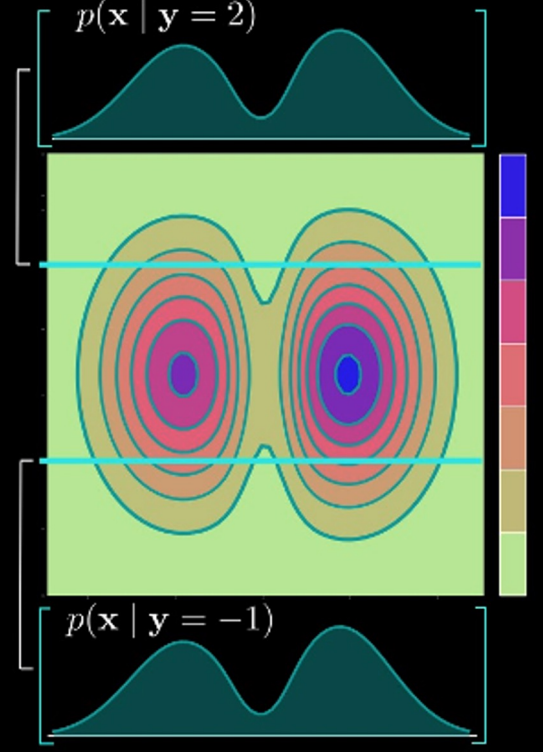
Explains how $\mathbf{x}$ behaves if we observe $\mathbf{y}$

As discussed before, we have

$$p(\mathbf{x}|\mathbf{y} = y_i) = \frac{p(\mathbf{x}, \mathbf{y} = y_i)}{\int p(\mathbf{x}, \mathbf{y} = y_i) d\mathbf{x}} = \frac{p(\mathbf{x}, \mathbf{y} = y_i)}{p(y_i)}$$

Note that this becomes a function of $\mathbf{x}$

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})}$$



Independence

Two random variables $\mathbf{x}$ and $\mathbf{y}$ are said to be independent if

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

This also implies that $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$ ------

Expectation and inequalities

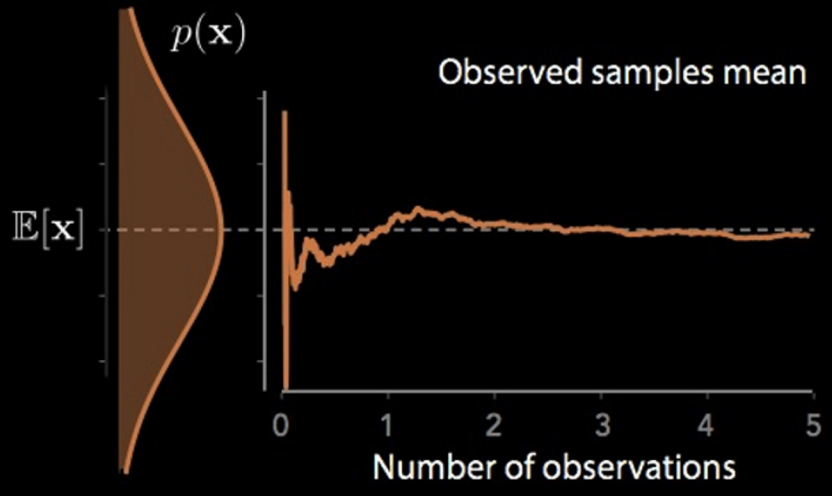
Expectation

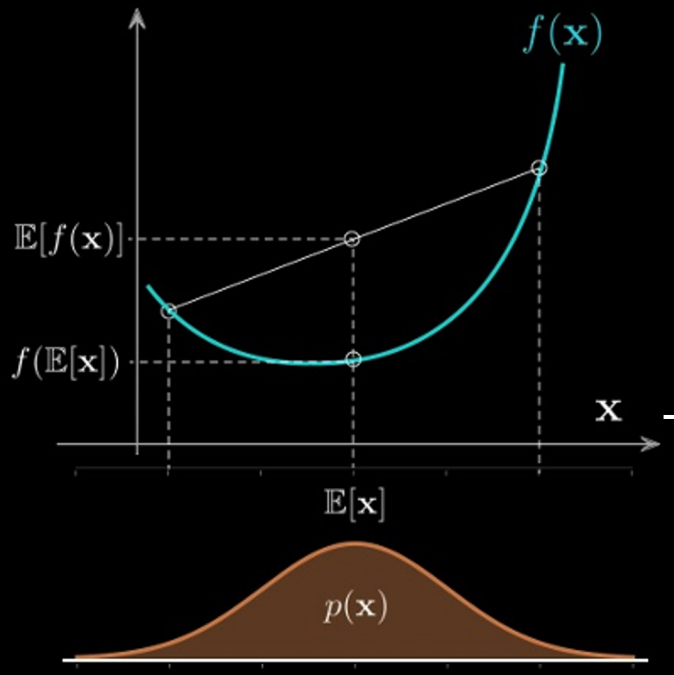
$\mathbf{x}$ random variable with distribution $p(\mathbf{x})$

Then its *expectation* is given by $\mathbb{E}[\mathbf{x}] = \int_{\mathbb{R}} \mathbf{x} p(\mathbf{x})\, d\mathbf{x}$

The discrete equivalent is $\mathbb{E}[\mathbf{x}] = \sum_i \mathbf{x}_i p(\mathbf{x} = \mathbf{x}_i)$

This defines the *expected value (central limit theorem)*





Useful properties of the expectation

Linearity $\mathbb{E}[\alpha\mathbf{x} + \beta\mathbf{y}] = \alpha\mathbb{E}[\mathbf{x}] + \beta\mathbb{E}[\mathbf{y}]$

Cauchy-Schwarz $(\mathbb{E}[\mathbf{xy}])^2 \leq \mathbb{E}[\mathbf{x}^2]\mathbb{E}[\mathbf{y}^2]$

Jensen inequality

If $\varphi : \mathbb{R} \to \mathbb{R}$ is a *convex* function

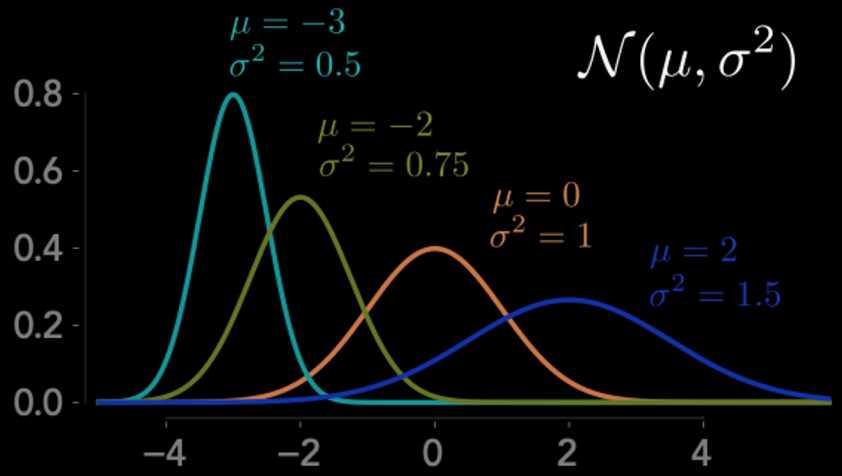
$$\varphi(\mathbb{E}[\mathbf{x}]) \leq \mathbb{E}[\varphi(\mathbf{x})]$$

Continuous distributions: Normal (Gaussian)

A continuous random variable x *follows a Normal distribution of mean* μ *and variance* σ^2 *noted* $x \sim \mathcal{N}(\mu, \sigma^2)$ *if we have*

$$p(x) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Different parameters for the Gaussian



Attractive properties of the Gaussian (Normal) distribution

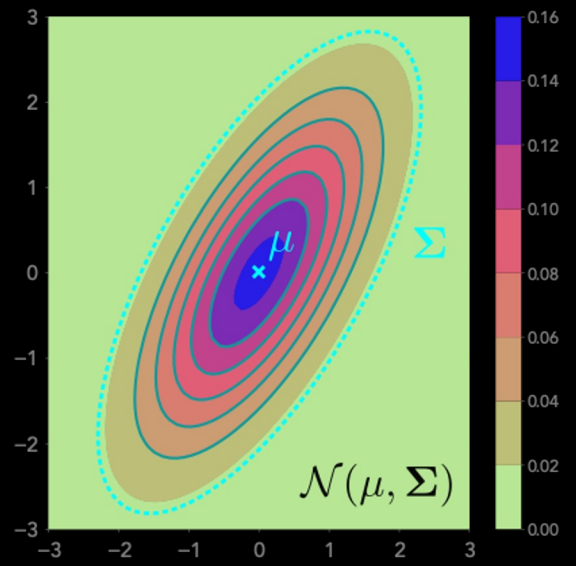
- *Unimodal* and *symmetric* distribution
- *Stable by linearity : if* $x \sim \mathcal{N}(\mu, \sigma^2)$ *then* $\alpha x + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$
- Density is *infinitely differentiable*
- *Mathematical simplicity*

Continuous distributions: Multivariate Normal

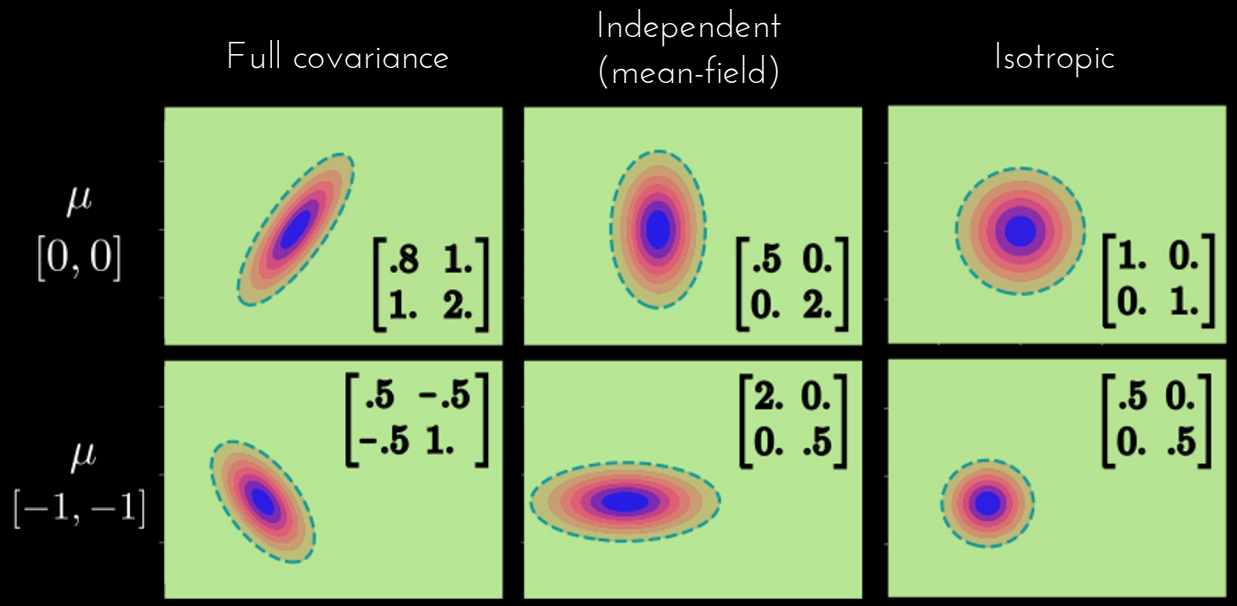
A continuous random variable $\mathbf{x}$ *follows a multivariate Normal distribution of mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$, *noted* $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *if we have*

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Understanding parameters

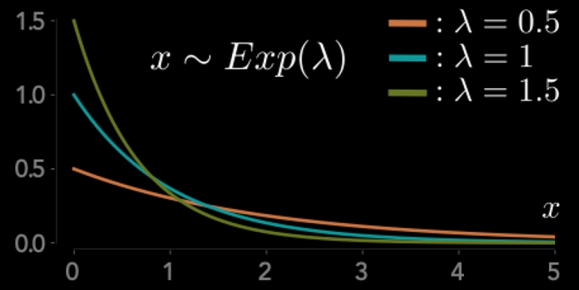


Different types of covariances



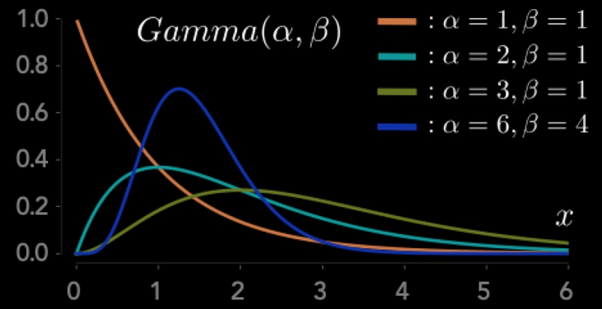
Also a whole set of attractive mathematical properties

Continuous distributions bestiary



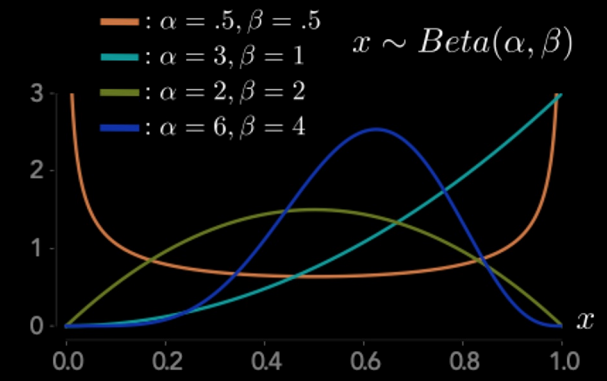
Exponential

$$p(x) = Exp_x(\lambda) = \lambda e^{-\lambda x}$$



Gamma

$$p(x) = \Gamma(\alpha, \beta) = Gamma(\alpha, \beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$



Beta

$$p(x) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Learning

Introducing probabilities in machine learning

- **Supervised learning** is inferring a function from labeled training data
- **Unsupervised learning** is trying to find hidden structure in unlabeled data

> We defined machine learning problems as $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$
>
> First need to define what *function* could *approximate* this process
>
> $$f_\theta \in \mathcal{F} \quad \theta \in \Theta$$
>
> Then evaluate the *quality* of this approximation $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}|\theta, f_\theta)$

Introducing probability distributions as models

Probability distributions are *parametric functions* $f_\theta \in \mathcal{F}$

Hence, can be used as models in our learning frameworks

$$p_\theta(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

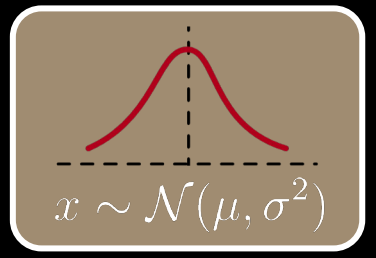
Gaussian distribution

$$\theta = \{\mu, \sigma\}$$

Learnable parameters

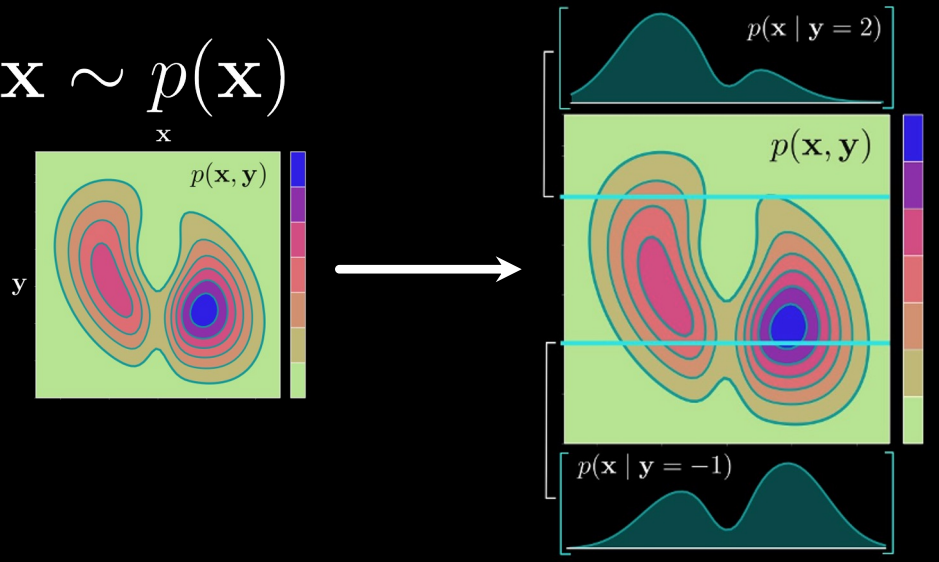
Probabilistic machine learning

Consider **data** $\mathbf{x} \in \mathbb{R}^n$ following a **distribution** $\mathbf{x} \sim p(\mathbf{x})$



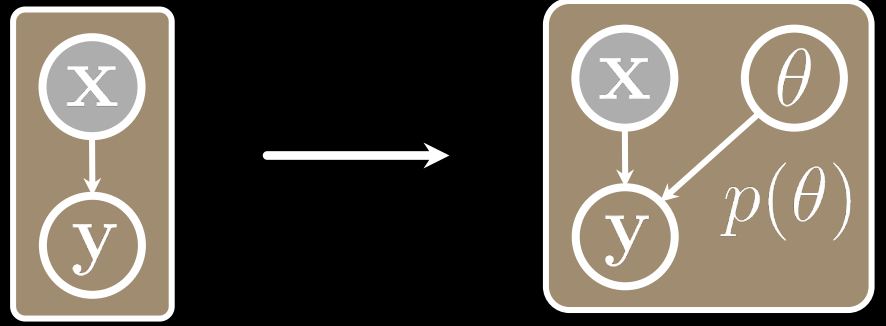
Sampling new individuals

$$x_1 \rightarrow 0.12$$
$$x_2 \rightarrow -1.3$$
$$x_3 \rightarrow 0.76$$



Now how to define (*discriminative*) learning ?

Looking for the answer $\mathbf{y}$ *given* the input $\mathbf{x}$

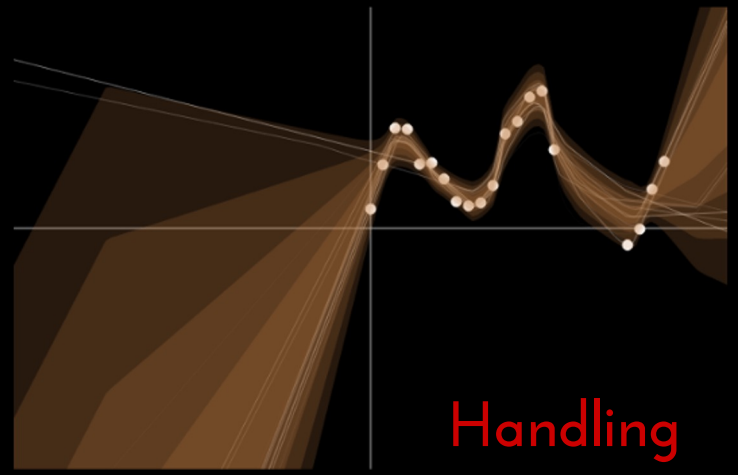


$$p(\mathbf{y}|\mathbf{x})$$

Probability of a class given a datapoint

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x}, \theta) d\theta$$

Incorporating parameters



Handling uncertainty

Bayes' theorem

Conditional probability $$p(\mathbf{a}|\mathbf{b}) = \frac{p(\mathbf{a},\mathbf{b})}{p(\mathbf{b})}$$

As probability is commutative, we rewrite it as

$$p(\mathbf{a}|\mathbf{b})p(\mathbf{b}) = p(\mathbf{a},\mathbf{b}) = p(\mathbf{b},\mathbf{a}) = p(\mathbf{b}|\mathbf{a})p(\mathbf{a}) \qquad p(\mathbf{a}|\mathbf{b}) = \frac{p(\mathbf{b}|\mathbf{a})p(\mathbf{a})}{p(\mathbf{b})}$$

Say we have a classification problem with $\mathbf{x}$ = features $\mathbf{y}$ = class

Bayes' rule

Easy

Hard $$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$$

- Decide the **class given some features**
- Using **probability of the feature given each class**
- And the *a priori probability* of each class

(denominator is same for all classes, discussed later)

Usually more than one feature $$p(y_i|\mathbf{x}) = \frac{p(\mathbf{x}_1,\cdots,\mathbf{x}_n|y_i)p(y_i)}{p(\mathbf{x})}$$

If features are **independent** $$p(y_i|\mathbf{x}) = \frac{p(\mathbf{x}_1|y_i)p(\mathbf{x}_2|y_i)\cdots p(\mathbf{x}_n|y_i)p(y_i)}{p(\mathbf{x})}$$

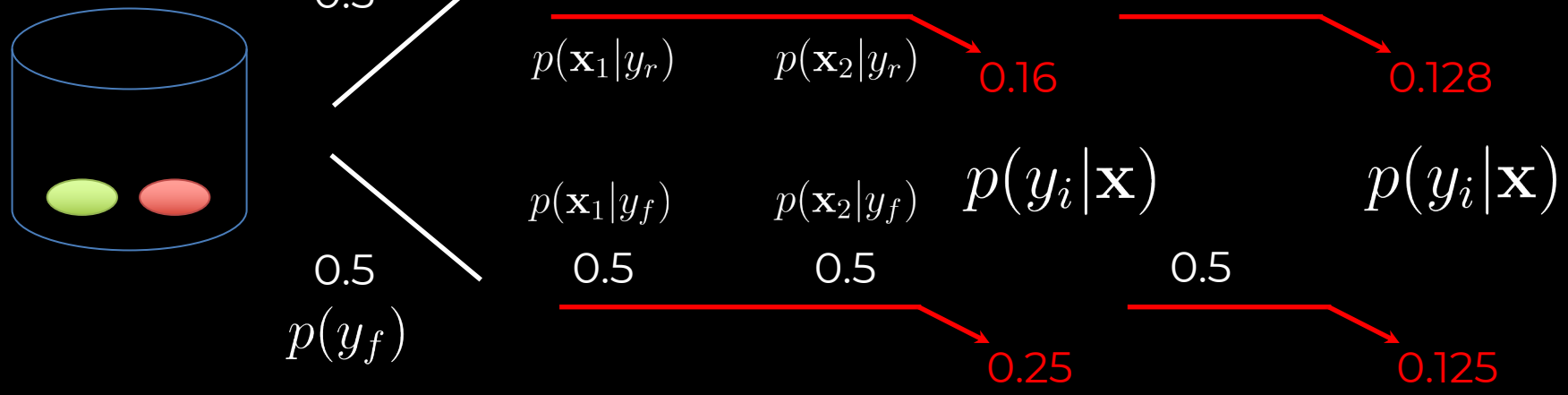
Bayesian classification

Deciding if our coin is fair or rigged

Let's say I have two coins, in my pocket
One is fair and one is rigged

Rigged
$p(tails) = 0.2$
$p(heads) = 0.8$

Fair
$p(tails) = 0.5$
$p(heads) = 0.5$



$$p(y_i|\mathbf{x}) = \frac{p(\mathbf{x}_1|y_i)p(\mathbf{x}_2|y_i)\cdots p(\mathbf{x}_n|y_i)p(y_i)}{p(\mathbf{x})}$$

Bayesian inference

Goal | Method for *updating our* **beliefs** about unknown parameters
Based on **observed data** and **prior knowledge.**

Update **prior** beliefs with observations (**likelihood**) to obtain new view (**posterior**)

Likelihood

Posterior $p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$ Prior

Evidence

Obtained by **marginalization**

$$p(\mathbf{x}) = \sum_j p(\mathbf{x}|y_j)p(y_j)$$

Posterior ∝ Likelihood × Prior

Bayesian Inference aims to model the **entire posterior probability**

Priors usually unknown (can be removed by assuming equal belief)
Posterior is prior influenced by the likelihood according to new observations

Posterior becomes prior of the next iteration

Bayesian inference

Goal

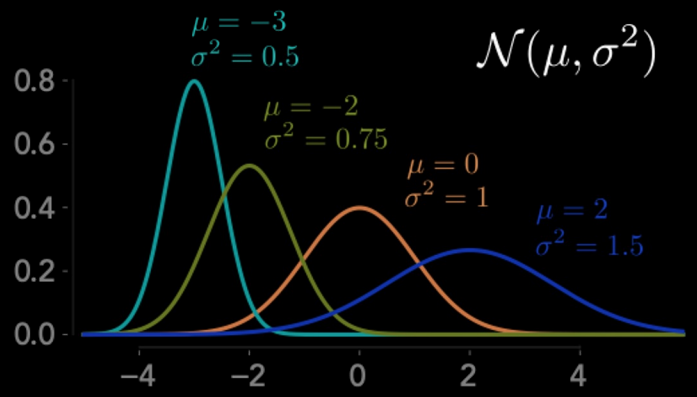
Method for *updating our* **beliefs** about unknown parameters
Based on **observed data** and **prior knowledge.**

Update **prior** beliefs with observations (**likelihood**) to obtain new view (**posterior**)

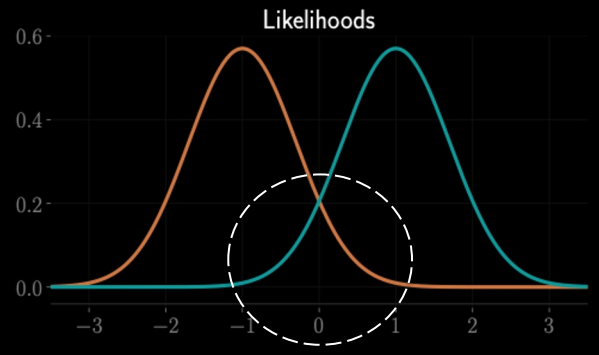
Updating beliefs

Prior: Start with initial prior distribution (based on prior knowledge).
Likelihood: Compute likelihood function for the observed data.
Posterior: Update the prior using Bayes' to obtain posterior
New data: Iteratively update the posterior using new data.

$$x \sim \mathcal{N}(\mu, \sigma^2)$$
$$\theta = \{\mu, \sigma\}$$



Cf. notebook



Allows class uncertainty (membership)

Bayesian inference

Goal

Method for *updating our* **beliefs** about unknown parameters
Based on **observed data** and **prior knowledge.**

Update **prior** beliefs with observations (**likelihood**) to obtain new view (**posterior**)

Advantages

Incorporates prior knowledge
About parameters, can lead to more accurate and robust estimates.

Uncertainty quantification
Full description of the uncertainty in parameter estimates (useful for decision).

Online learning
Naturally handles online learning, where data is collected incrementally over time.

Problems

Computational complexity
Can be computationally expensive (high-dimensional spaces or complex models).

Choice of prior
Can have significant impact on results, challenging choice of appropriate prior

How to model our distributions or optimize parameters ?

Maximum A Posteriori (MAP)

Question | How to choose the best class given the model (**inference**) ?

Choose the Maximum A Posteriori (MAP) class $\hat{y} = \underset{y_i}{\operatorname{argmax}}\ p(y_i|\mathbf{x})$

Intuition Choose the most probable class given the observation(s)

> We do not have direct access to $p(y_i|\mathbf{x})$
> However, simpler to know $p(\mathbf{x}|y_i)$

Apply Bayes' theorem inside $\hat{y} = \underset{y_i}{\operatorname{argmax}}\ p(y_i|\mathbf{x})$

We obtain $\hat{y} = \underset{y_i}{\operatorname{argmax}}\ \frac{p(\mathbf{x}|y_i)p(y_i)}{\sum_j p(\mathbf{x}|y_j)p(y_j)}$ <— for all y_i, this term is the same

We have a similar denominator ($p(\mathbf{x})$) for all y_i

Need to solve for $\hat{y} = \underset{y_i}{\operatorname{argmax}}\ p(\mathbf{x}|y_i)p(y_i)$

Bayesian inference

Applying Bayesian ideas to our coin toss

Now how to estimate the parameters of a (coin toss) distribution ?

Bernoulli $\phi(x) = p^x(1-p)^{(1-x)}$

We will observe a set of trials x_i leading to the likelihood

start from equiv prob of p(y_i)
then replace it by posterior

$$\mathcal{L}(p|\mathbf{x}) = \prod_{i=1}^{n} p_i^x(1-p)^{(1-x_i)}$$

$$\hat{y} = \underset{y_i}{\operatorname{argmax}}\ p(\mathbf{x}|y_i)p(y_i)$$

Need to compute $\frac{\partial \mathcal{L}(p|\mathbf{x})}{\partial p}$ and find the maximum $\frac{\partial \mathcal{L}(p|\mathbf{x})}{\partial p} = 0$

Note that finding the maximum of a function
is equivalent to finding the maximum of the log of this function

Exercise in notebook

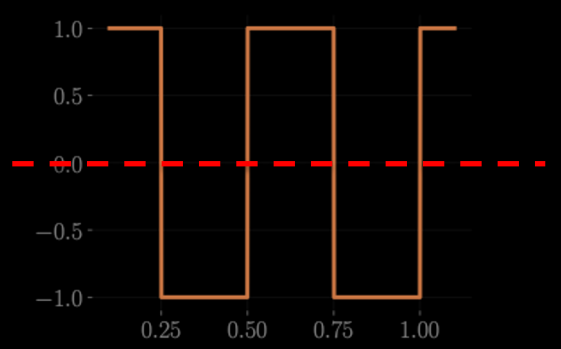
Bayesian inference

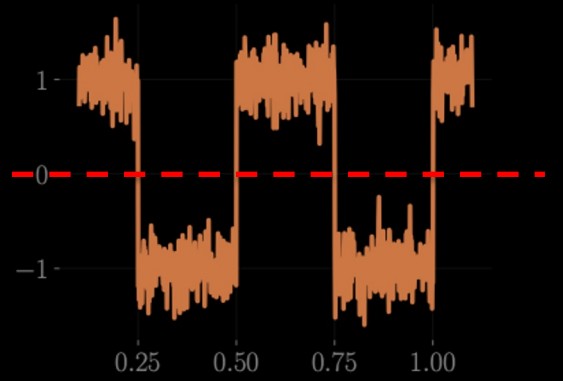
Defining classification

Transmit class data through a communication channel

Class 1

Class 0





Class 1

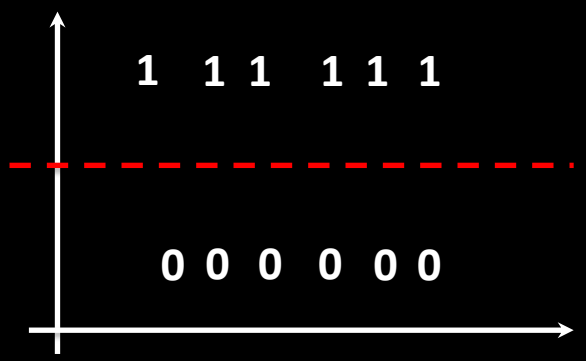
Class 0

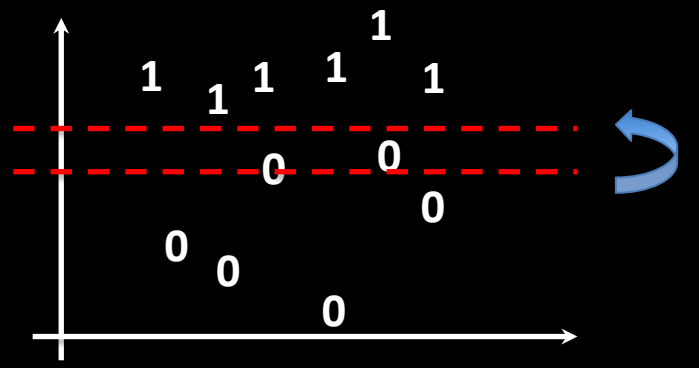
$\bar{x}$ → Channel → x

Communication channel adds noise to data $x = \bar{x} + \epsilon \quad \epsilon \sim \mathcal{N}(\mu, \sigma^2)$

What is the optimal decision threshold in *evolving situations*

T



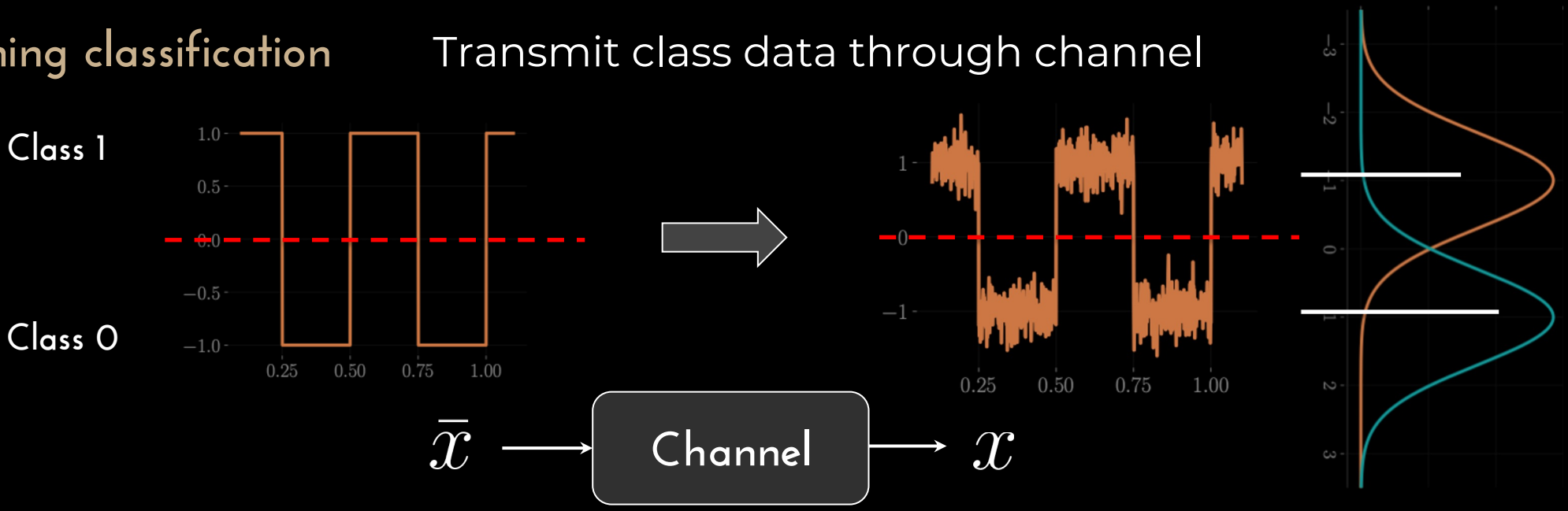


From a priori
To a posteriori

Bayesian inference

Defining classification

Transmit class data through channel



Laying out a Bayesian inference problem

- **(A priori) class probabilities**
  Without other knowledge an equal belief

$$p_y(0) = p_y(1) = 1/2$$

- **Class-conditional probabilities**
  Class depends on the noise added by the channel

$$p_\theta(\mathbf{y}|\mathbf{x}) = \mathcal{N}_x(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Central limit theorem: we can assume noise is Gaussian $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$

Bayesian inference

Gaussian probability density function $p_\theta(\mathbf{y}|\mathbf{x}) = \mathcal{N}_x(\mu, \sigma^2)$

Since we assumed that the noise is Gaussian and additive

$$x = \bar{x} + \epsilon \quad \epsilon \sim \mathcal{N}(\mu, \sigma^2)$$

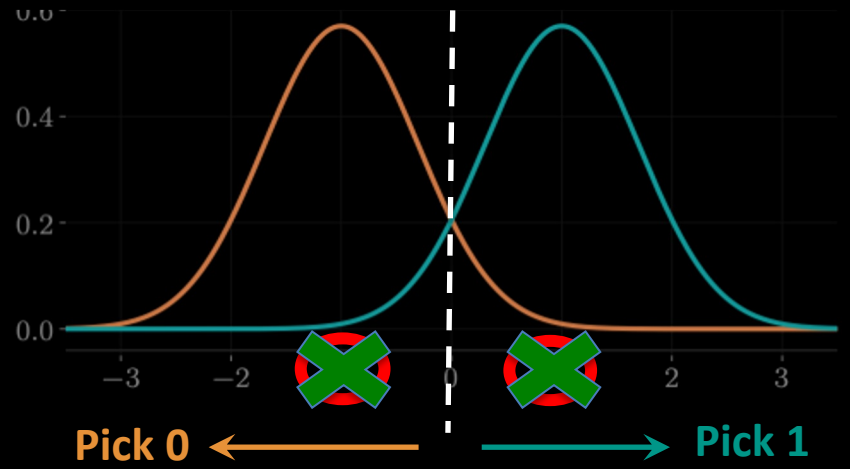
How to select the correct class ?

Probability of input given classes

Class 0 $p_{x|y}(x|0) = \mathcal{N}_x(\mu_0, \sigma^2)$

Class 1 $p_{x|y}(x|1) = \mathcal{N}_x(\mu_1, \sigma^2)$

$$p_y(0) = p_y(1) = 1/2$$



Real-world

$$p_{x|y}(x|0) = \mathcal{N}_x(\mu_0, \sigma^2)$$
$$p_{x|y}(x|1) = \mathcal{N}_x(\mu_1, \sigma^2)$$

Bayesian inference

What happens for the general case $p_{x|y}(x|0) = \mathcal{N}_x(\mu_0, \sigma^2)$ $p_{x|y}(x|1) = \mathcal{N}_x(\mu_1, \sigma^2)$

To compute the Bayesian Decision Rule (BDR), we can use log probabilities

$$i^* = \underset{i}{\text{argmax}} \left[\log p_{x|y}(x|i) + \log p_y(i)\right]$$

And note that the priors are equal for everybody so $i^* = \underset{i}{\text{argmax}} \log p_{x|y}(x|i)$

If we develop this equation

$$\begin{aligned} i^* &= \underset{i}{\text{argmax}} \log p_{x|y}(x|i) \\ &= \underset{i}{\text{argmax}} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}\right] \\ &= \underset{i}{\text{argmax}} \left[-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{(x-\mu_i)^2}{2\sigma^2}\right] \\ &= \underset{i}{\text{argmin}} \left[\frac{(x-\mu_i)^2}{2\sigma^2}\right] \end{aligned}$$

Bayesian decision rule

Consider both distributions have the same variance

$$
\begin{aligned}
i^* &= \underset{i}{\operatorname{argmin}} \left[\frac{(x-\mu_i)^2}{2\sigma^2}\right] \\
&= \underset{i}{\operatorname{argmin}} \left[x^2 - 2x\mu_i + \mu_i^2\right] \\
&= \underset{i}{\operatorname{argmin}} \left[-2x\mu_i + \mu_i^2\right]
\end{aligned}
$$

Optimal decision

$$
\begin{aligned}
-2x\mu_0 + \mu_0^2 &< -2x\mu_1 + \mu_1^2 \\
-2x(\mu_1 - \mu_0) + &< \mu_1^2 - \mu_0^2
\end{aligned}
$$

$$
x < \frac{\mu_1 + \mu_0}{2}
$$

Seems like too much work to find a very intuitive rule ?

But we had to make lots of assumptions

- Uniform class probabilities, additive noise, gaussianity

In other problems, we can remove these simplifications

Bayesian decision rule (BDR)

Deriving the BDR for multivariate Gaussian

$$i^* = \underset{i}{\operatorname{argmax}} \; \left[\log p_{\mathbf{x}|y}(\mathbf{x}|i) + \log p_y(i)\right]$$

$$p_{\mathbf{x}|y}(\mathbf{x}|i) = \frac{1}{\sqrt{(2\pi)^d \, \boldsymbol{\Sigma}_i}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)\right]$$

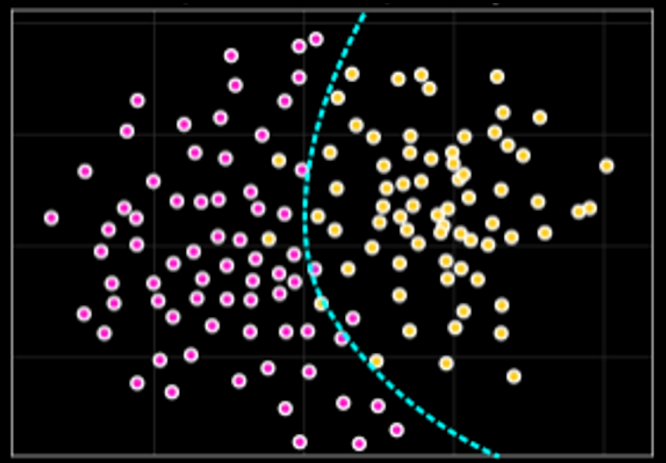
By using the *log* the BDR becomes

$$i^* = \underset{i}{\operatorname{argmax}} \; \left[-\frac{1}{2}\underbrace{(\mathbf{x}-\mu_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)}_{d_i(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{y})} - \frac{1}{2}\underbrace{\log (2\pi)^d \, \boldsymbol{\Sigma}_i + \log p_y(i)}_{\alpha_i = \frac{1}{2}\log (2\pi)^d \, \boldsymbol{\Sigma}_i + \log p_y(i)}\right]$$

Final BDR

$$i^* = \underset{i}{\operatorname{argmin}} \; \left[d_i(\mathbf{x}, \mu_i) + \alpha_i\right]$$

Optimal rule is to assign **x** to the closest class
Mesured with the Mahalanobis distance (**d**)
Constant is added to account for class prior



Maximum Likelihood

The BDR provides the optimal (and geometric) solution for **class selection**

$$i^* = \underset{i}{\operatorname{argmax}} \left[-\frac{1}{2}(\mathbf{x} - \mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x} - \mu_i) - \frac{1}{2}\log(2\pi)^d \mathbf{\Sigma}_i + \log p_y(i) \right]$$

Question | How to obtain the **parameters of class distributions**

$$\mu_i \quad \mathbf{\Sigma}_i \quad \log p_y(i)$$

We have to estimate these values from a training set

Maximum Likelihood Estimation (MLE)

Principle for parameter estimation (of class distribution)

1. Choose **parametric** probabilistic model
2. Assemble our training classification dataset
3. Find the parameters that maximize the **likelihood**

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p_\theta(\mathbf{x}|\mathbf{y})$$

Maximum Likelihood

Deriving the MLE solution

1. Choose a parametric model for all probabilities

   We usually denote parameters by $\boldsymbol{\Theta}$ and class-conditional distributions by

   $$p_{\mathbf{x}|y}(\mathbf{x}|i;\boldsymbol{\Theta})$$

   not a random variable but a parameter (probabilities are function of it)

2. Assemble a collection of datasets

   Set of examples independently drawn from class i (cf. sampling schemes)

3. Select the parameters that **maximize the likelihood** of the data

$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\ p_{\mathbf{x}|y}(\mathbf{x}|i;\boldsymbol{\Theta}) = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\ \log p_{\mathbf{x}|y}(\mathbf{x}|i;\boldsymbol{\Theta})$$

log-likelihood

(mathematical simplicity)

How do we solve this ?

Learning parameters

How to find parameters that maximize likelihood or posterior

Analytical solutions

For simple models (linear regression, Gaussian distributions)
Sometimes we have analytical solution to the optimization

Maximum Likelihood Estimation

$$\theta_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Gradient-based optimization

For more complex models, usually no analytical solution is available
We can use gradient-based optimization methods

$$\left(\frac{\partial \log p_{\mathbf{x}|y}(\mathbf{x}|i;\mathbf{\Theta_1})}{\partial \mathbf{\Theta}_1}, \cdots, \frac{\partial \log p_{\mathbf{x}|y}(\mathbf{x}|i;\mathbf{\Theta_n})}{\partial \mathbf{\Theta_n}}\right)$$

Bayesian - MAP - MLE

Bayesian inference, MLE, and MAP have different purposes and properties

Bayesian

Goal: Calculate entire posterior distribution of parameters given observations

✓ Provides full uncertainty in estimates for more robust decision.

x Can be computationally expensive (high-dimensional spaces).

Uncertainty quantification, Bayesian networks, nonparametric models

MLE

Goal: Estimate single value for parameters by maximizing the likelihood.

✓ Simple, computationally efficient with asymptotically efficient estimates.

x Does not incorporate prior knowledge, sensitive to outliers

Classical statistics and parameter estimation for range of models

MAP

Goal: Estimate single value for parameters by maximizing posterior (with prior).

✓ Combines likelihood with prior for more robust estimates.

x Choice of prior distribution may introduce biases

Applications with prior knowledge such ridge regression, LASSO, and Bayesian neural networks.

- Bayesian inference provides more comprehensive *parameter uncertainty*
- MLE and MAP are point estimation methods (single value for parameters).
- MLE and MAP computationally more efficient than Bayesian inference
- MAP is a compromise between MLE and Bayesian inference (prior knowledge)

Conjugate priors

Complete Bayesian inference development

We should sweat over the mathematics ... but note that

- The prior is Gaussian $p_\mu(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right)$
- The posterior is Gaussian $p_x(x|\mu) = \mathcal{N}\left(x|\mu, \sigma_x^2\right)$

Whenever the **posterior is in the same family as the prior**

$p_\mu(\mu)$ is a conjugate prior for the likelihood $p_x(x|\mu)$

Posterior $p_\mu(\mu|x)$ is the reproducing density

A number of likelihoods have conjugate priors

Likelihood	Conjugate prior
Bernoulli	Beta
Poisson	Gamma
Exponential	Gamma
Normal	Normal