Bayesian Machine Learning

03/05/21 - François HU

Outline

1 Bayesian statistics
- define a probabilistic model
- apply bayesian inference
- conjugate priors

2 Latent variable models
- define latent variable and apply them to simplify probabilistic model
- cluster data with latent models like GMM
- train probabilistic models with EM-algorithm

3 Variational Inference
- apply variational inference for probabilistic models
- understand variational interpretation of LDA
- application of LDA to text mining

4 Markov Chain Monte Carlo
- train / do inference almost any probabilistic model with MCMC
- pros and cons of MCMC / VI

5 Extensions and oral presentations

PREREQUISITE

THEORY

1. Notions of **probability & statistics**
2. **Statistical Learning :** supervised **&** unsupervised learning
3. **Information theory :** Entropy, KL-divergence, …
4. **Monte Carlo & Markov Chain**

APPLICATION

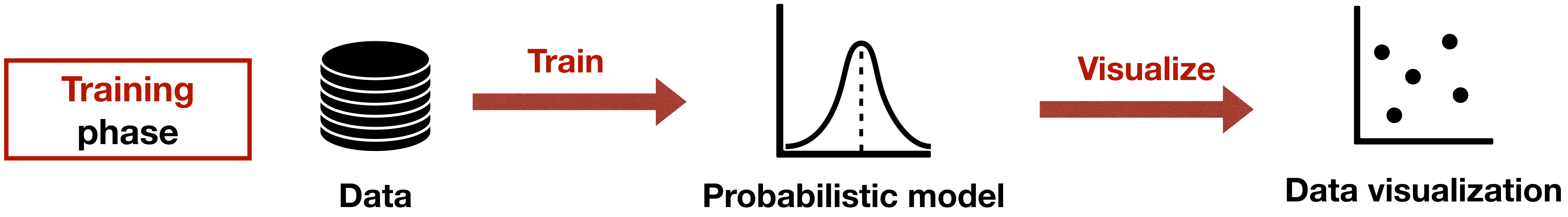
Python (or at least R)

ALGORITHM

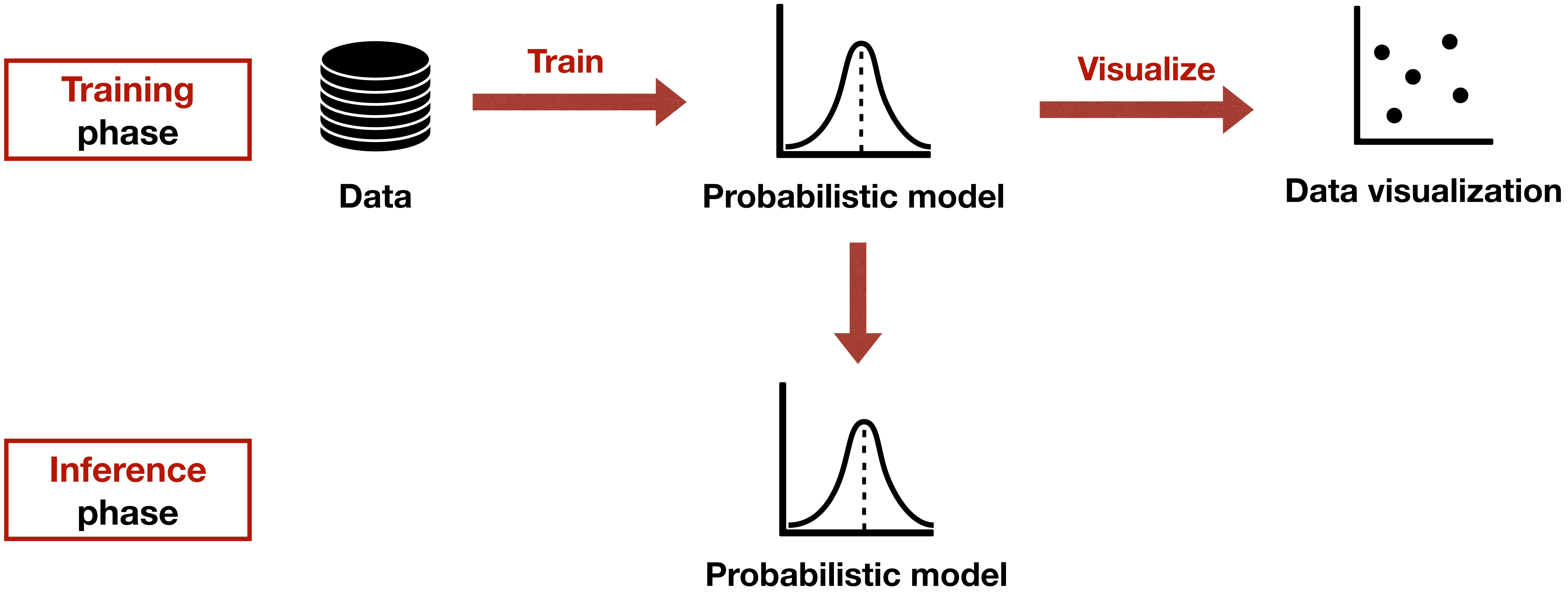
Some « classical » supervised & unsupervised models

0 Gentle introduction to statistical learning

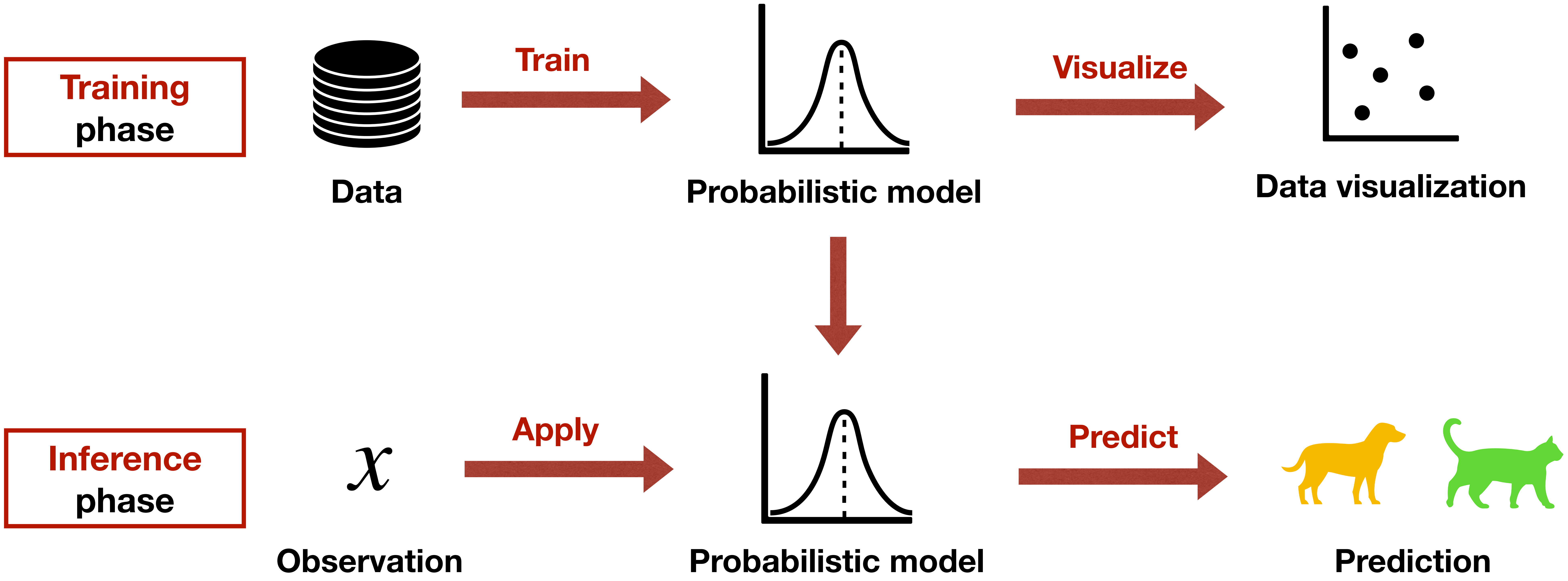
Simplified statistical learning process



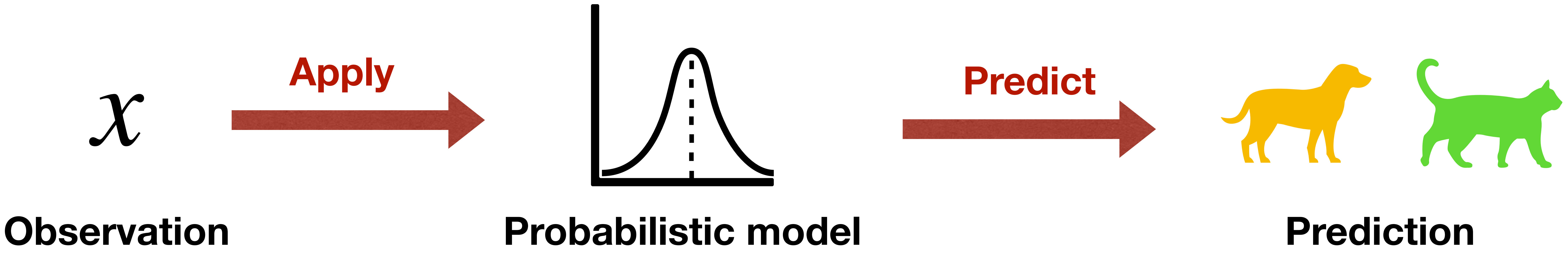
Simplified statistical learning process



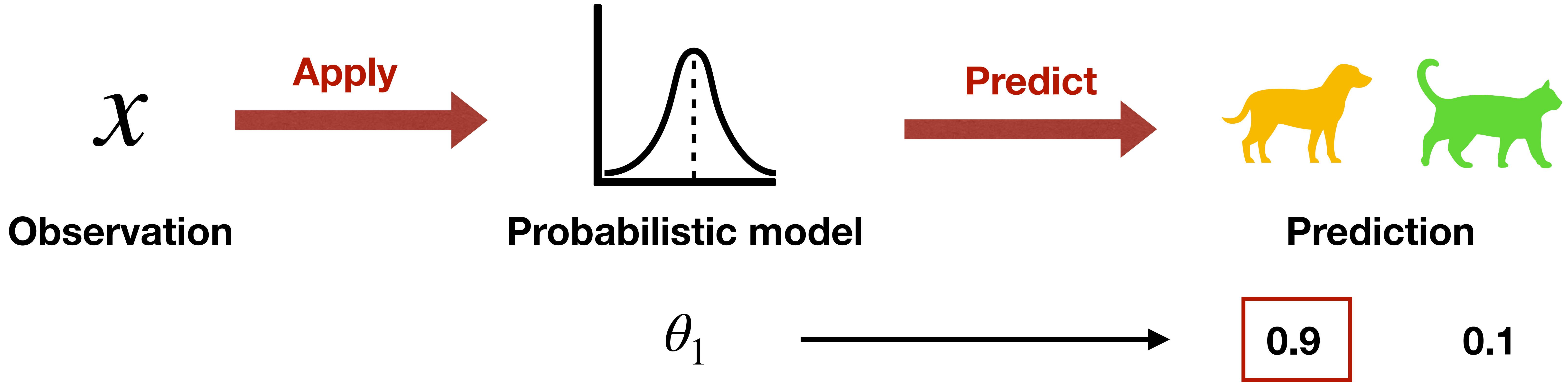
Simplified statistical learning process



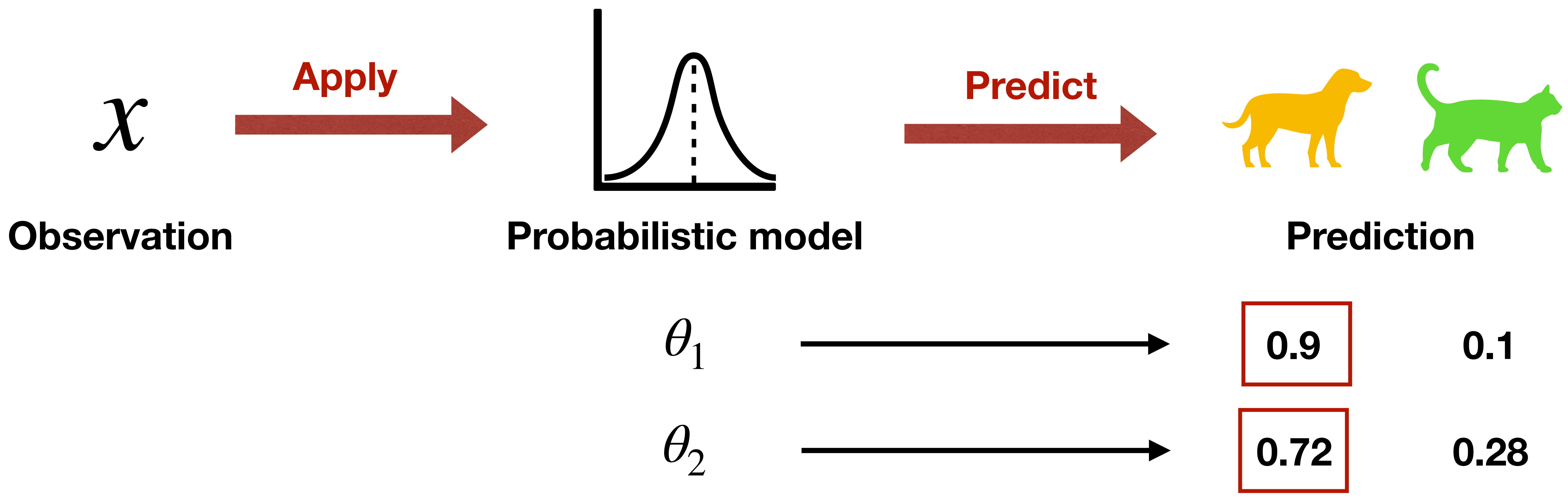
Inference phase



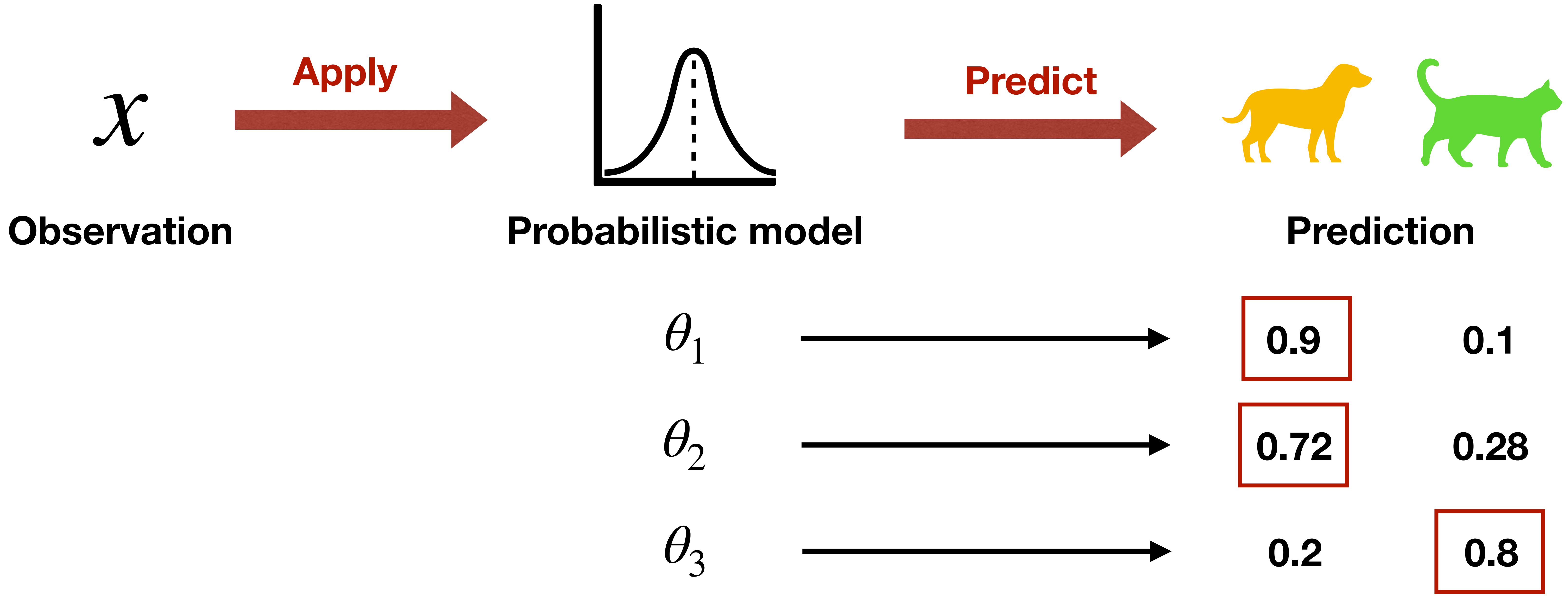
Inference phase



Inference phase



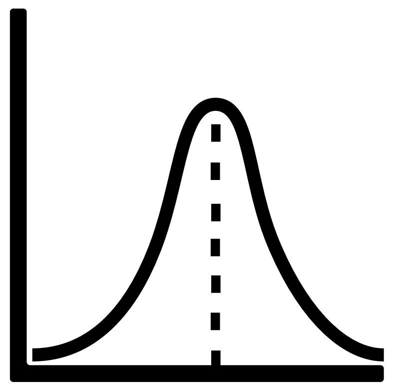
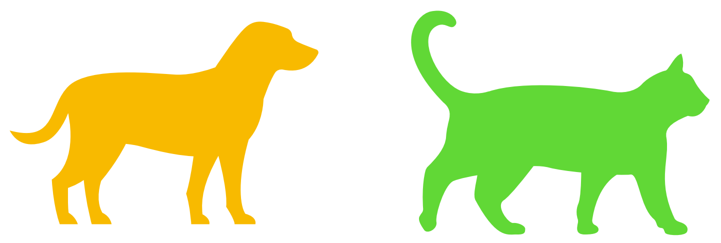
Inference phase



Inference phase

Aim during training phase : find

θ **such that** $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

x $\xrightarrow{\text{Apply}}$  $\xrightarrow{\text{Predict}}$ 

Observation **Probabilistic model** **Prediction**

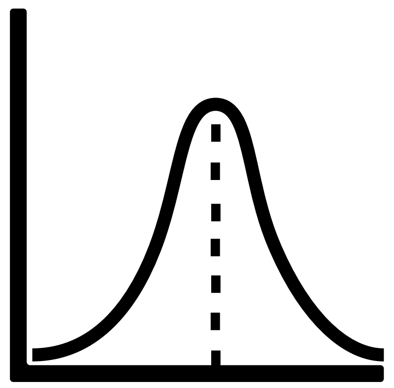
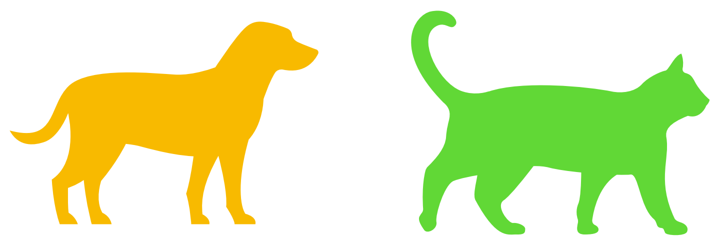
$\theta_1 \longrightarrow$	$\boxed{0.9}$	0.1
$\theta_2 \longrightarrow$	$\boxed{0.72}$	0.28
$\theta_3 \longrightarrow$	0.2	$\boxed{0.8}$

$\dots$

Inference phase

Aim during training phase : find

θ such that $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

x $\xrightarrow{\text{Apply}}$  $\xrightarrow{\text{Predict}}$ 

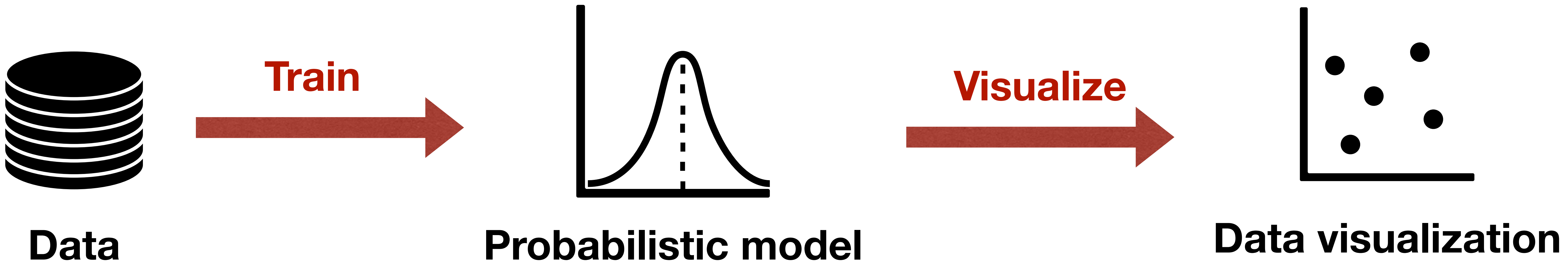
Observation **Probabilistic model** **Prediction**

$\theta_1 \longrightarrow$	$\boxed{0.9}$	0.1
$\theta_2 \longrightarrow$	$\boxed{0.72}$	0.28
$\theta_3 \longrightarrow$	0.2	$\boxed{0.8}$

...

- « Classical » frequentist point of view
- Parameter θ is fixed and the data X is random

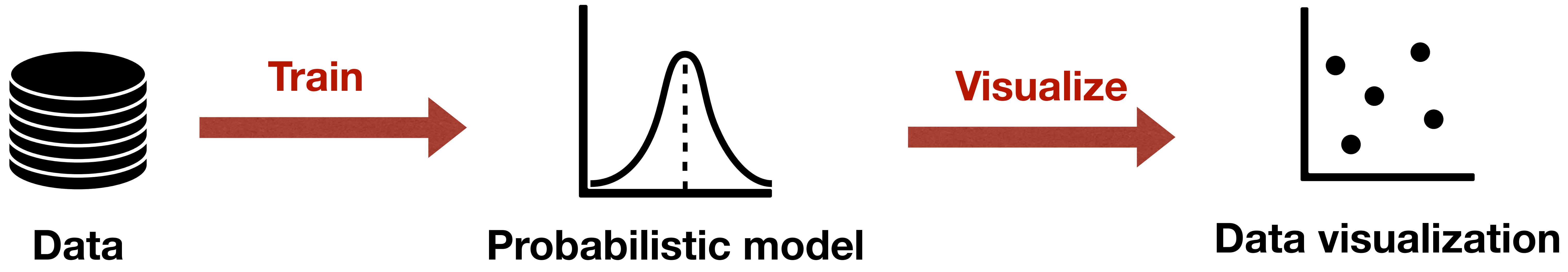
Training phase



Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Usually in frequentist statistics we use the MLE : Maximum Likelihood Estimation $\hat{\theta} = \arg\max_\theta P(X \mid \theta)$

Training phase



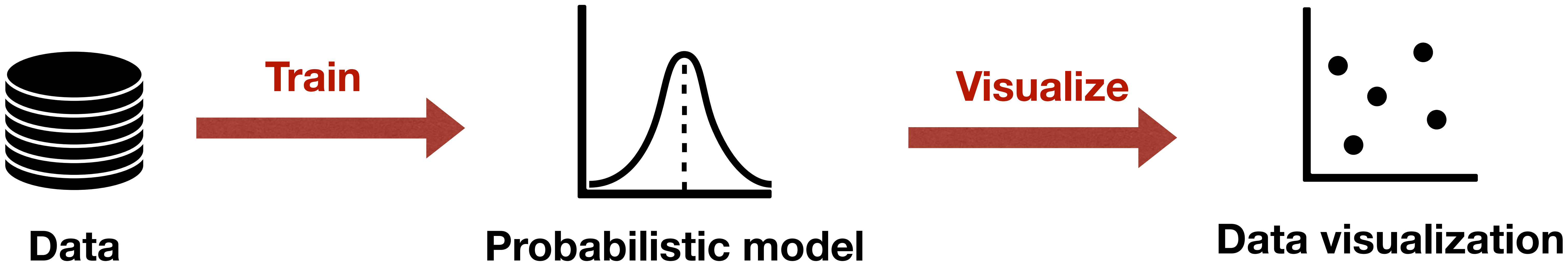
Aim during training phase : find θ such that $\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Usually in frequentist statistics we use the MLE : Maximum Likelihood Estimation $\hat{\theta} = \underset{\theta}{\arg\max}\, P(X|\theta)$

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start with a « belief » hence not practical nor flexible
- Cannot express uncertainty of estimated model parameters and predictions

Training phase



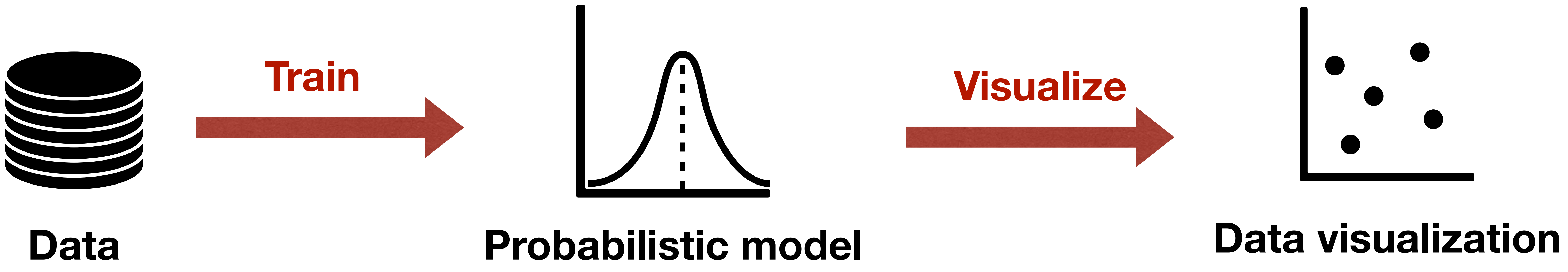
Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Usually in frequentist statistics we use the MLE : Maximum Likelihood Estimation $\hat{\theta} = \arg\max_\theta P(X|\theta)$

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start with a « belief » hence not practical nor flexible
- Cannot express uncertainty of estimated model parameters and predictions

Training phase



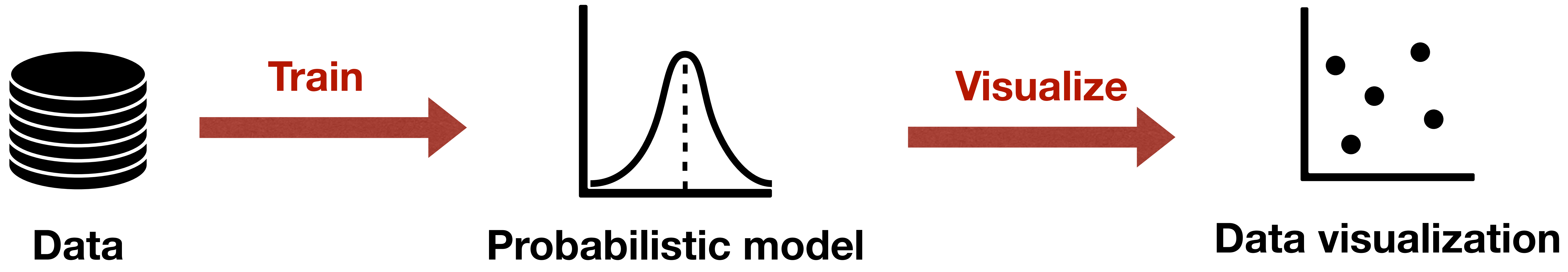
Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Usually in frequentist statistics we use the MLE : Maximum Likelihood Estimation $\hat{\theta} = \underset{\theta}{\arg\max}\, P(X \mid \theta)$

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start with a « belief » hence not practical nor flexible
- Cannot express uncertainty of estimated model parameters and predictions

Training phase



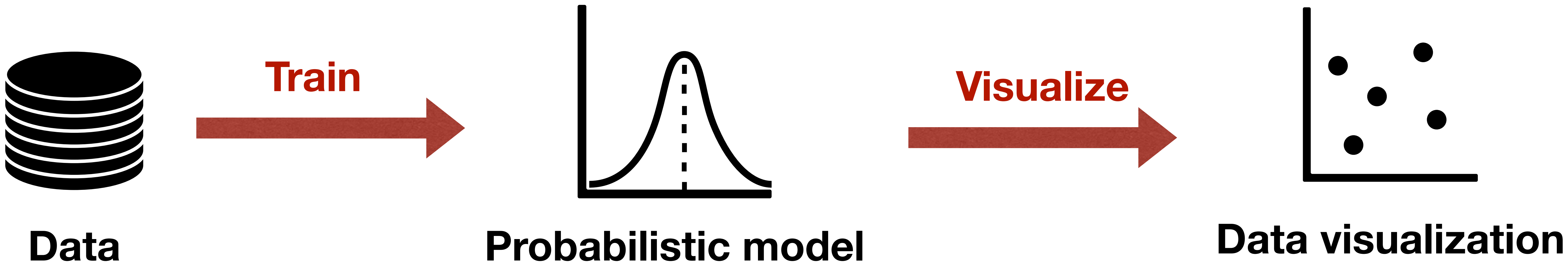
Aim during **training phase** **: find** θ **such that** $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Usually in frequentist statistics we use the MLE : **Maximum Likelihood Estimation** $\hat{\theta} = \arg\max_\theta P(X|\theta)$

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start with a « belief » hence not practical nor flexible
- Cannot express uncertainty of estimated model parameters and predictions

Training phase



Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Usually in frequentist statistics we use the MLE : Maximum Likelihood Estimation $\hat{\theta} = \arg\max_\theta P(X|\theta)$

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start with a « belief » hence not practical nor flexible
- Cannot express uncertainty of estimated model parameters and predictions

Bayesian statistics

1 Introduction to bayesian statistics

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Example : a deck of 52 playing cards

$P($  $) = \frac{1}{52}$ $\qquad$ $P(\spadesuit) = \frac{1}{4}$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable

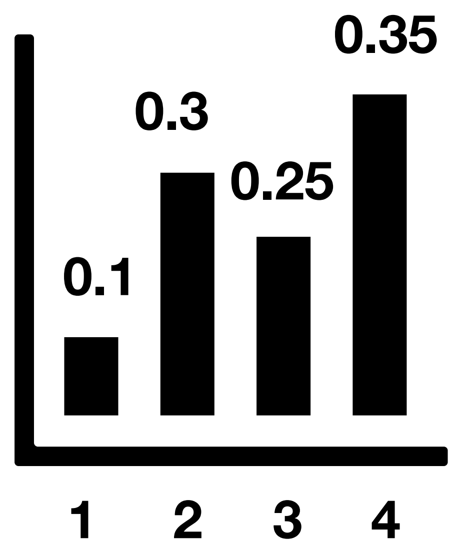
1. Introduction to bayesian statistics

Probability & statistics : small review

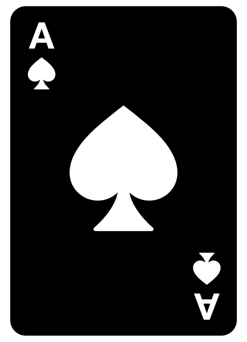
Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable

- **Discrete** variable



Example



Probability Mass Function (PMF)

$$P(X) = \begin{cases} 0.1 & \text{if } X = 1 \\ 0.3 & \text{if } X = 2 \\ 0.25 & \text{if } X = 3 \\ 0.35 & \text{if } X = 4 \end{cases}$$

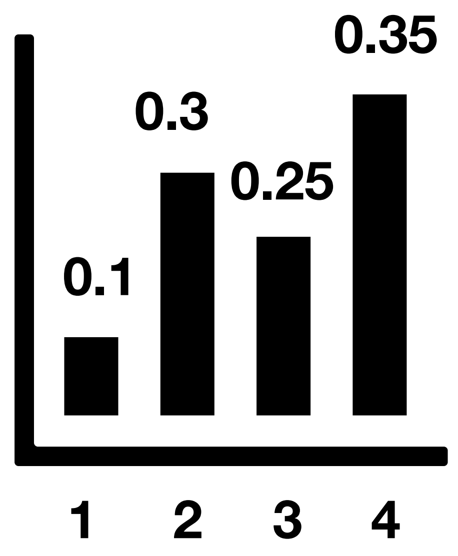
1. Introduction to bayesian statistics

Probability & statistics : small review

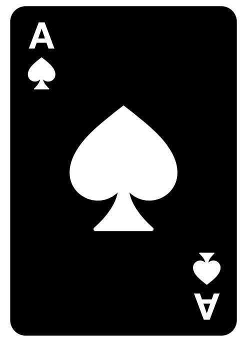
Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable

- **Discrete** variable



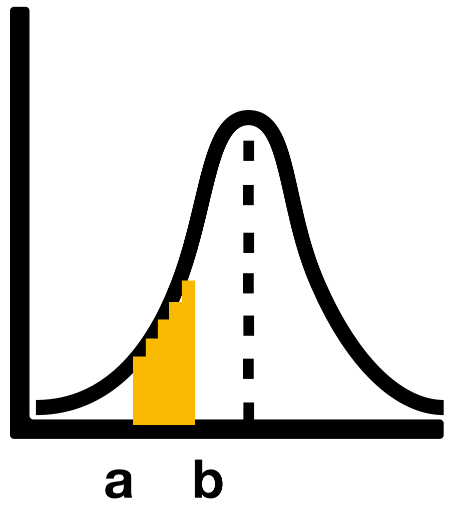
Example



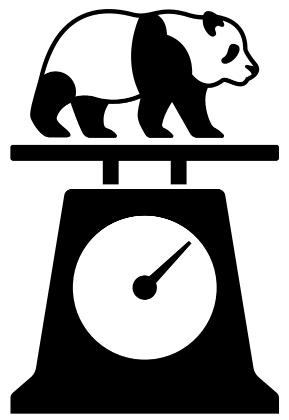
Probability Mass Function (PMF)

$$P(X) = \begin{cases} 0.1 & \text{if } X = 1 \\ 0.3 & \text{if } X = 2 \\ 0.25 & \text{if } X = 3 \\ 0.35 & \text{if } X = 4 \end{cases}$$

- **Continuous** variable



Example



Probability Density function (PDF)

$$P(X \in [a, b]) = \int_a^b p(s)ds$$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$\underbrace{P(X,Y)}_{\textbf{joint probability}} = \underbrace{P(X)P(Y)}_{\textbf{marginals}}$$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

joint probability **marginals**

Example :

dependency : **one** deck of 52 playing cards

$$P(X_1 = A\heartsuit, X_2 = A\heartsuit) = 0$$

$$P(X_1 = A\heartsuit) \cdot P(X_2 = A\heartsuit) = \frac{1}{52^2}$$

independency : **two** dices

$$P(X_1 = \text{⚃}, X_2 = \text{⚃}) = \frac{1}{6^2}$$

$$P(X_1 = \text{⚃}) \cdot P(X_2 = \text{⚃}) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{6^2}$$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if $P(X, Y) = P(X)P(Y)$

Conditional probability

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

joint probability

$$\underbrace{P(X|Y)}_{\textbf{conditional}} = \frac{P(X, Y)}{P(Y)}$$

marginal

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

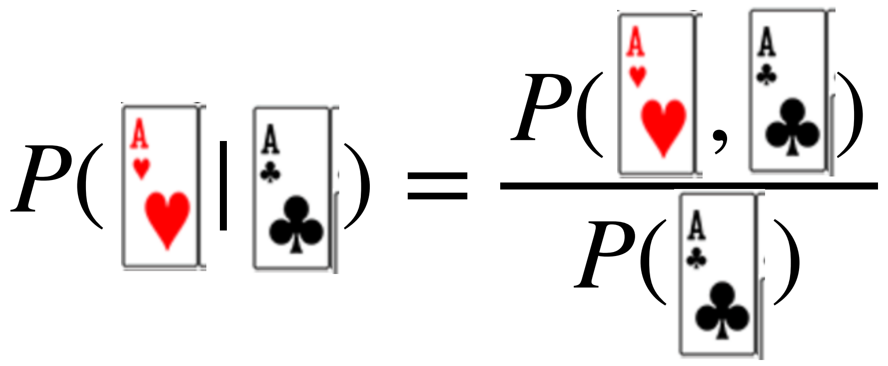
conditional

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

joint probability

marginal

Example : **one** deck of 52 playing cards



1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

Rules :

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

Rules :

Chain rule $P(X_1, X_2) = P(X_1|X_2) \times P(X_2)$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

Rules :

Chain rule $P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3) \times P(X_2 \mid X_3) \times P(X_3)$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

Rules :

Chain rule $$P(X_1, \ldots, X_n) = \prod_{k=1,\ldots,n} P(X_k | X_1, \ldots, X_{k-1})$$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

Rules :

Chain rule

$$P(X_1, \ldots, X_n) = \prod_{k=1,\ldots,n} P(X_k \mid X_1, \ldots, X_{k-1})$$

Sum rule

$$P(X) = \sum_{Y \in \mathcal{Y}} P(X, Y) \qquad P(X) = \int_{Y \in \mathcal{Y}} P(X, Y) \cdot dY$$

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

Rules : **Chain rule** **Sum rule**

Bayes theorem :

1. Introduction to bayesian statistics

Probability & statistics : small review

Probability (non axiomatic definition) of an event : relative **frequency** of an event in an infinite trials

Random variable - **Discrete** variable - **Continuous** variable

Independence : two random variables X and Y are **independent** if

$$P(X, Y) = P(X)P(Y)$$

Conditional probability : probability of X **given that Y happened**

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

Rules : **Chain rule** **Sum rule**

Bayes theorem :

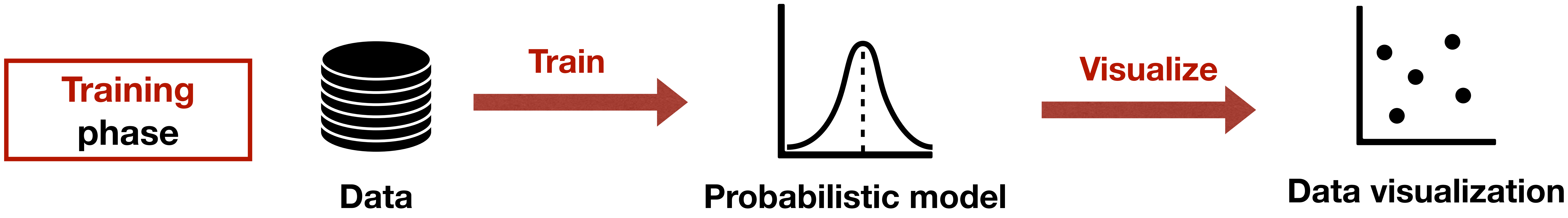
θ Parameters

X Observations (data)

$$\underbrace{P(\theta \mid X)}_{\text{Posterior}} = \frac{P(X, \theta)}{P(X)} = \frac{\overbrace{P(X \mid \theta)}^{\text{Likelihood}} \times \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(X)}_{\text{Evidence}}}$$

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view

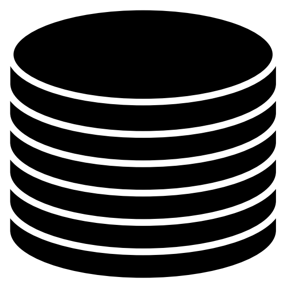


find θ **such that** $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

1. Introduction to bayesian statistics

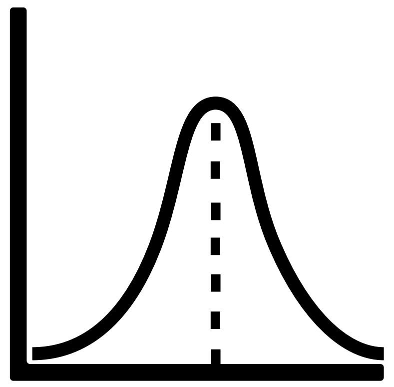
Frequentist VS Bayesian point of view

Training phase



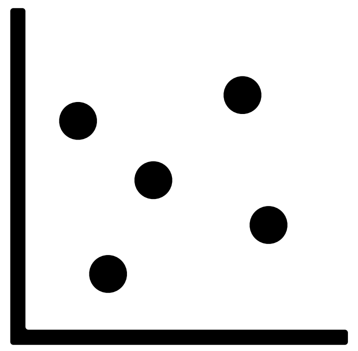
Data

Train



Probabilistic model

Visualize



Data visualization

Frequentist

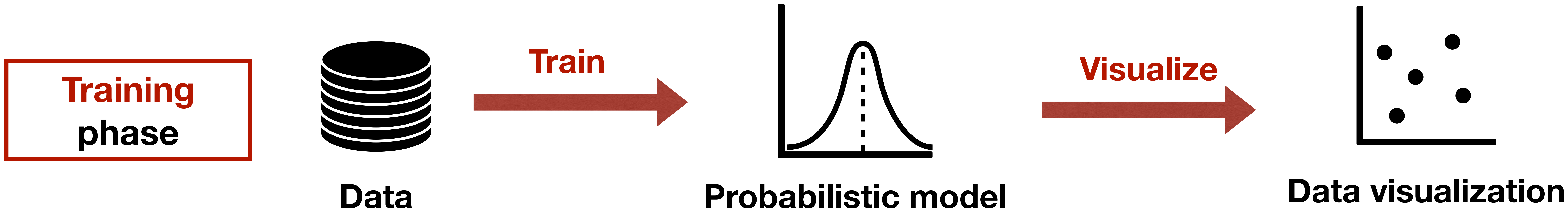
θ is fixed
X is random

Maximum Likelihood Estimation

$$\hat{\theta} = \arg\max_{\theta} P(X \mid \theta)$$

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view



Frequentist

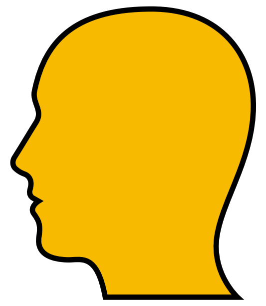
θ is fixed
X is random

Maximum Likelihood Estimation

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)$$

Bayes theorem

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$



Bayesian

θ is random
X is fixed

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view

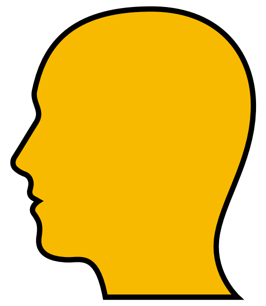
Frequentist

Maximum Likelihood Estimation

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)$$

Bayes theorem

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$



Bayesian

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view

Frequentist

Maximum Likelihood Estimation

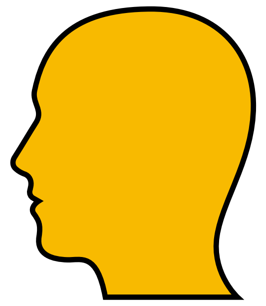
$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)$$

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start with a « belief » hence not practical nor flexible
- Cannot express uncertainty of estimated model parameters and predictions

Bayes theorem

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$



Bayesian

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view

Frequentist

Maximum Likelihood Estimation

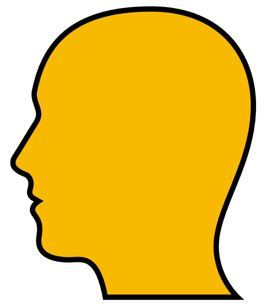
$$\hat{\theta} = \arg\max_{\theta} P(X \mid \theta)$$

Problems :

- ~~Only works well if we have big data : $|X| \gg |\theta|$~~
- ~~Cannot start with a « belief » hence not practical nor flexible~~
- ~~Cannot express uncertainty of estimated model parameters and predictions~~

Bayes theorem

$$P(\theta \mid X) = \frac{P(X \mid \theta) \times P(\theta)}{P(X)}$$



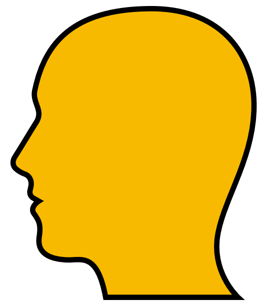
Bayesian

1. Introduction to bayesian statistics

Bayesian point of view : classification

Bayes theorem

$$P(\theta | X) = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$



Bayesian

Training **phase**

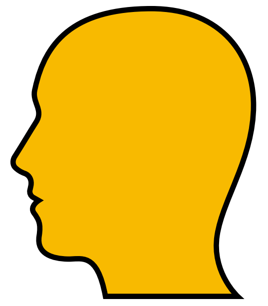
$$P(\theta | X_{train}, y_{train}) = \frac{P(y_{train} | X_{train}, \theta) \times P(\theta)}{P(y_{train} | X_{train})}$$

1. Introduction to bayesian statistics

Bayesian point of view : training

Bayes theorem

$$P(\theta | X) = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$



Bayesian

Training phase

$$P(\theta | X_{train}, y_{train}) = \frac{P(y_{train} | X_{train}, \theta) \times \boxed{P(\theta)}}{P(y_{train} | X_{train})}$$

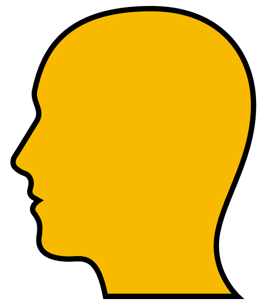
Can regularize your model when training on your data

1. Introduction to bayesian statistics

Bayesian point of view : inference

Bayes theorem

$$P(\theta | X) = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$



Bayesian

Training phase

$$P(\theta | X_{train}, y_{train}) = \frac{P(y_{train} | X_{train}, \theta) \times \boxed{P(\theta)}}{P(y_{train} | X_{train})}$$

Can regularize your model when training on your data

Inference phase

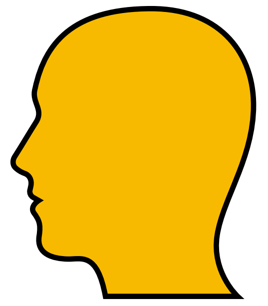
$$P(y_{new} | X_{new}, X_{train}, y_{train}) = \int P(y_{new} | X_{train}, \theta) \times P(\theta | X_{train}, y_{train}) d\theta$$

1. Introduction to bayesian statistics

Bayesian point of view : online learning

Bayes theorem

$$P(\theta | X) = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$



Bayesian

Training phase

$$P(\theta | X_{train}, y_{train}) = \frac{P(y_{train} | X_{train}, \theta) \times \boxed{P(\theta)}}{P(y_{train} | X_{train})}$$

Can regularize your model when training on your data

Inference phase

$$P(y_{new} | X_{new}, X_{train}, y_{train}) = \int P(y_{new} | X_{train}, \theta) \times P(\theta | X_{train}, y_{train}) d\theta$$

Online learning

$$P_{new}(\theta) = P(\theta | x_{new}) = \frac{P(x_{new} | \theta) \times P_{old}(\theta)}{P(x_{new})}$$

New prior

Posterior

2 Probabilistic models

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

- **Nodes :** random variables
- **Links :** probabilistic relationships

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Bayesian networks
(**Directed** graphical models)

Nodes : random variables

Links : probabilistic relationships

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

- **Nodes :** random variables
- **Links :** probabilistic relationships

Bayesian networks
(**Directed** graphical models)

Markov random fields
(**Undirected** graphical models)

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Nodes : random variables

Links : probabilistic relationships

Bayesian networks
(**Directed** graphical models)

The focus of our course !

Markov random fields
(**Undirected** graphical models)

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Nodes : random variables

Links : probabilistic relationships

Bayesian networks
(**Directed** graphical models)

The focus of our course !

Markov random fields
(**Undirected** graphical models)

Model : joint probability over all variables $P(X_1, \cdots, X_N) = \ldots$

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Nodes : random variables

Links : probabilistic relationships

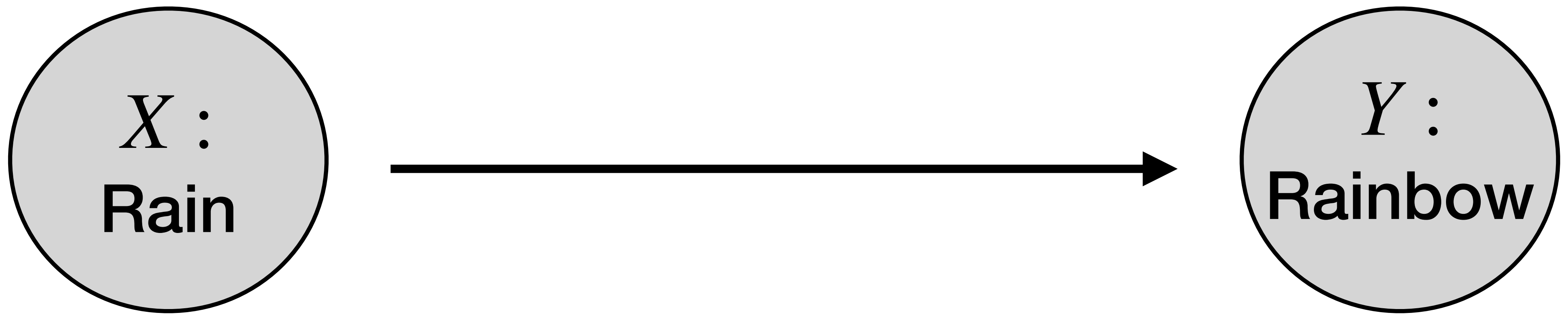
Bayesian networks
(**Directed** graphical models)

Markov random fields
(**Undirected** graphical models)

The focus of our course !

Model : joint probability over all variables $P(X_1, \cdots, X_N) = \ldots$

Example :



$$P(X, Y) = \ldots$$

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

- **Nodes :** random variables
- **Links :** probabilistic relationships

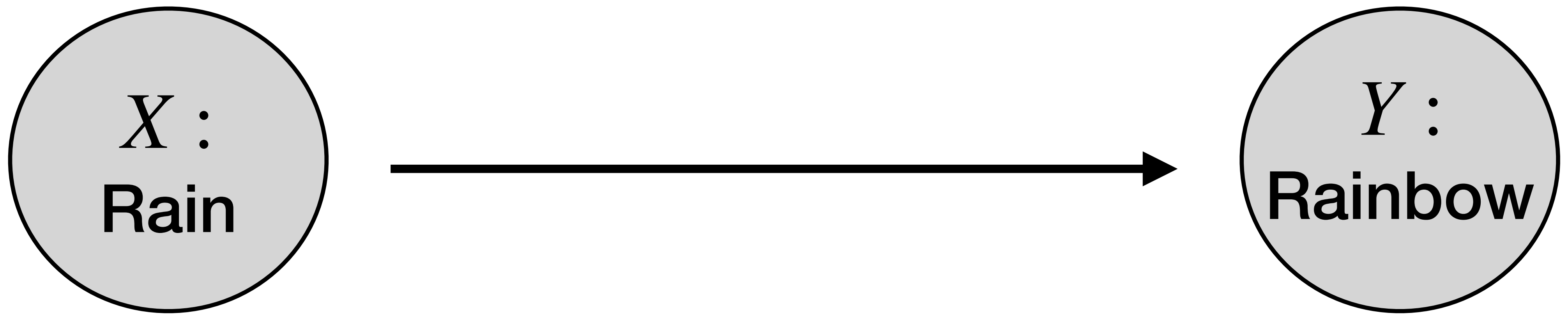
Bayesian networks
(**Directed** graphical models)

The focus of our course !

Markov random fields
(**Undirected** graphical models)

Model : joint probability over all variables $P(X_1, \cdots, X_N) = \ldots$

Example :



$$P(X, Y) = P(Y|X) \times P(X)$$

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

- **Nodes :** random variables
- **Links :** probabilistic relationships

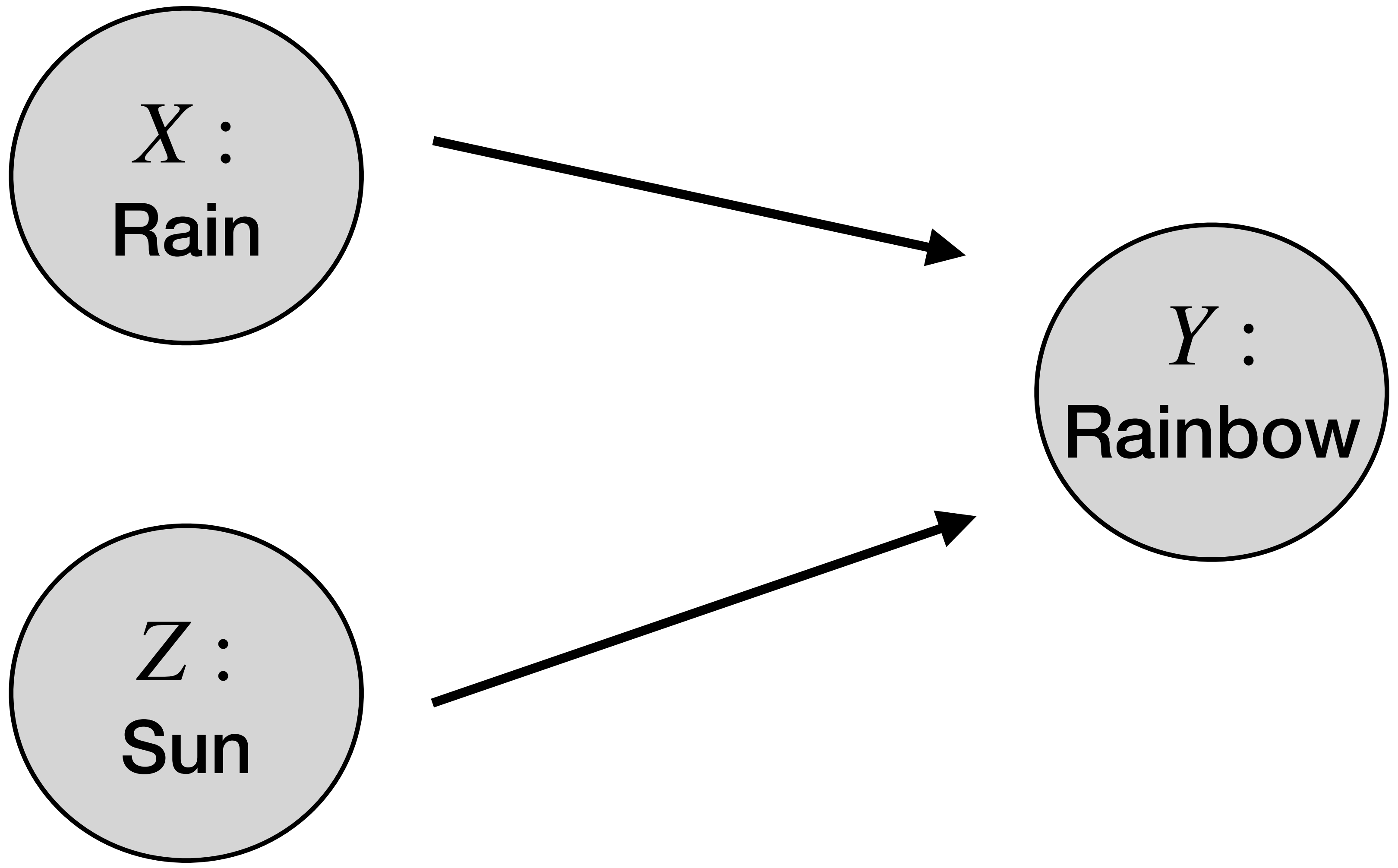
Bayesian networks
(**Directed** graphical models)

The focus of our course !

Markov random fields
(**Undirected** graphical models)

Model : joint probability over all variables $P(X_1, \cdots, X_N) = \ldots$

Example :



$$P(X, Y, Z) = \ldots$$

Features

X , Z conditionally independent given Y

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Bayesian networks
(**Directed** graphical models)

The focus of our course !

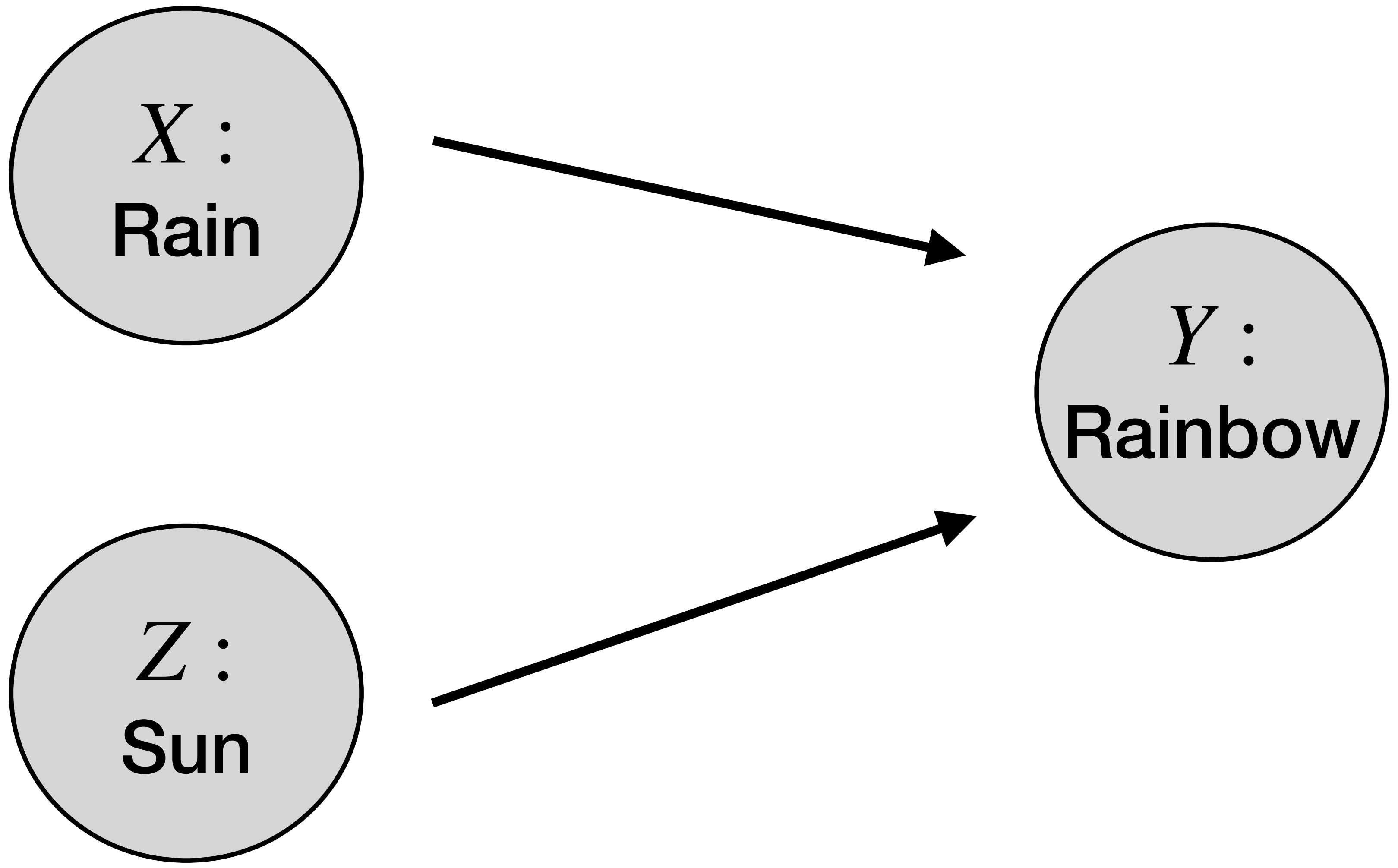
Markov random fields
(**Undirected** graphical models)

Nodes : random variables

Links : probabilistic relationships

Model : joint probability over all variables $P(X_1, \cdots, X_N) = \dots$

Example :



$$P(X, Y, Z) = P(Y|X, Z) \times P(X, Z)$$
$$= P(Y|X, Z) \times P(X) \times P(Z)$$

Features

X , Z conditionally independent given Y

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Nodes : random variables

Links : probabilistic relationships

Bayesian networks
(**Directed** graphical models)

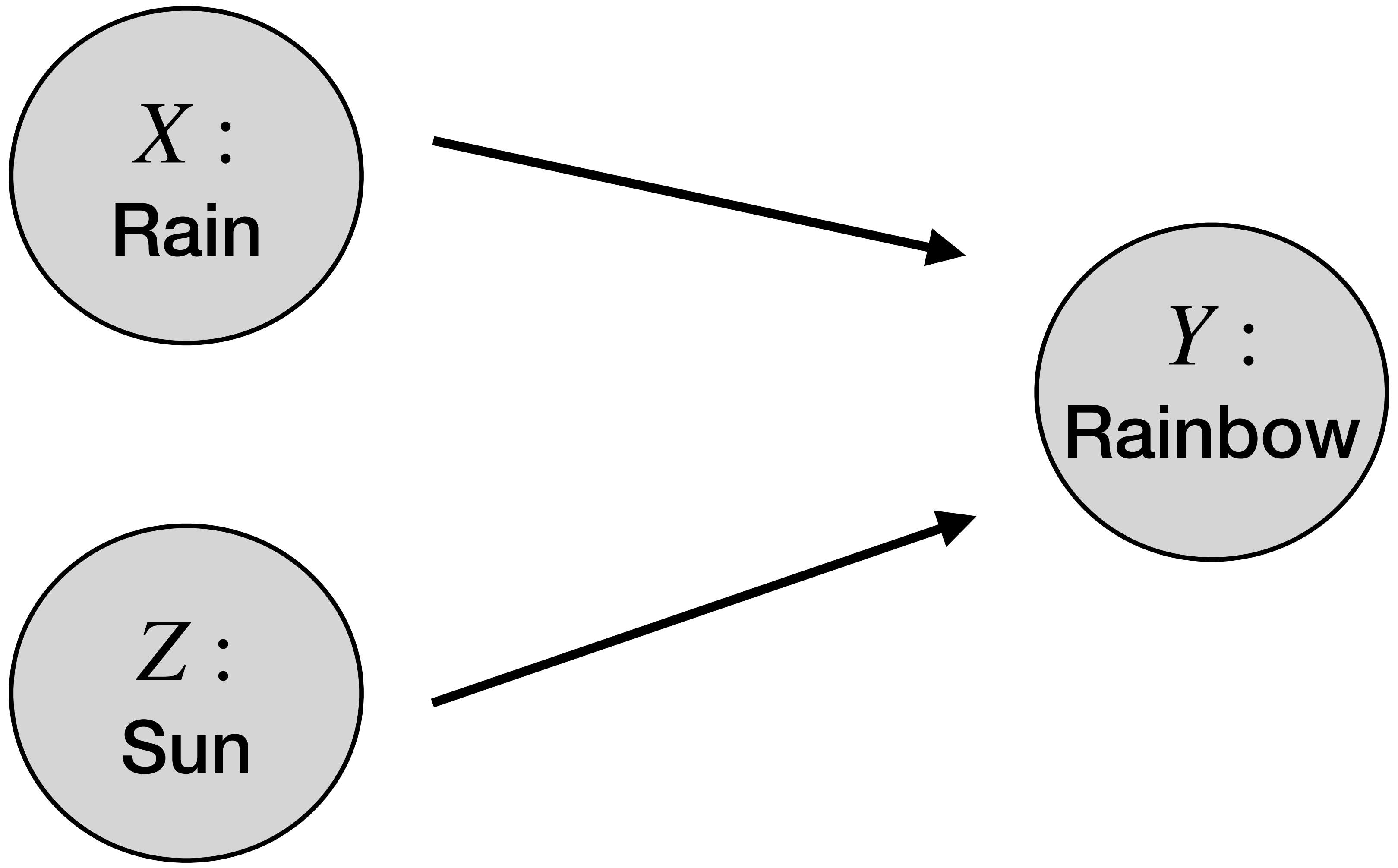
The focus of our course !

Markov random fields
(**Undirected** graphical models)

Model : joint probability over all variables

$$P(X_1, \cdots, X_N) = \prod_{i=1,\cdots,N} P(\ X_i \mid parents(X_i)\ )$$

Example :



$$P(X, Y, Z) = P(Y | X, Z) \times P(X, Z)$$
$$= P(Y | X, Z) \times P(X) \times P(Z)$$

Features

X , Z conditionally independent given Y

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Nodes : random variables

Links : probabilistic relationships

Bayesian networks
(**Directed** graphical models)

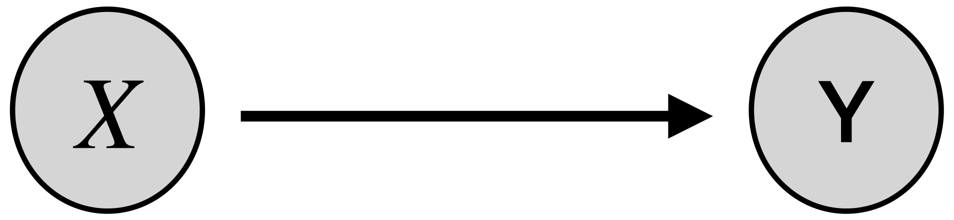
Markov random fields
(**Undirected** graphical models)

The focus of our course !

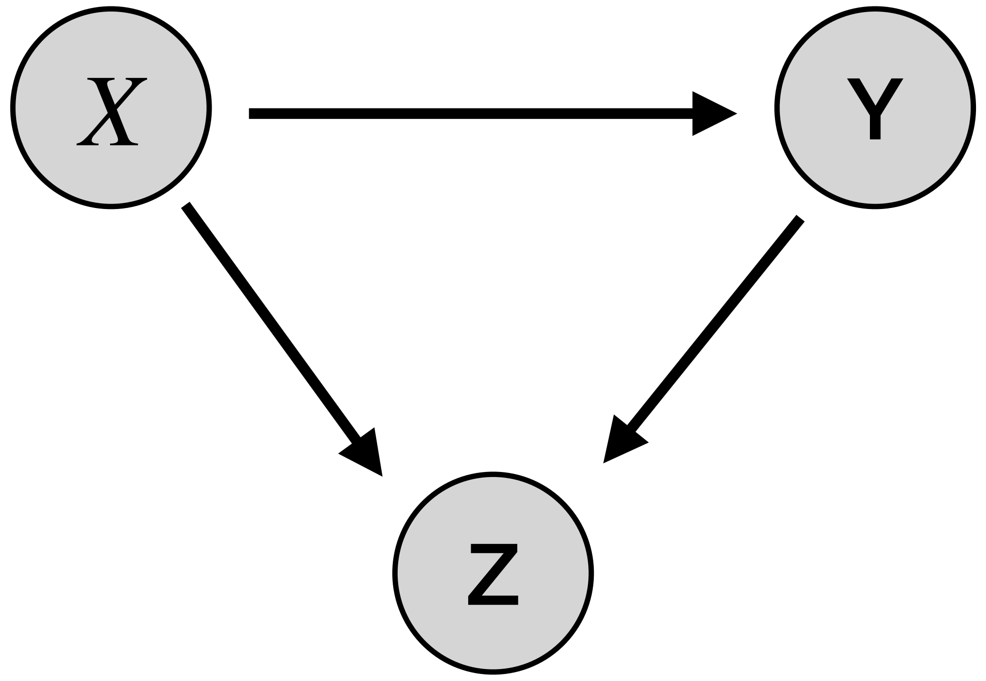
Model : joint probability over all variables

$$P(X_1, \cdots, X_N) = \prod_{i=1,\cdots,N} P(\ X_i \mid parents(X_i)\ )$$

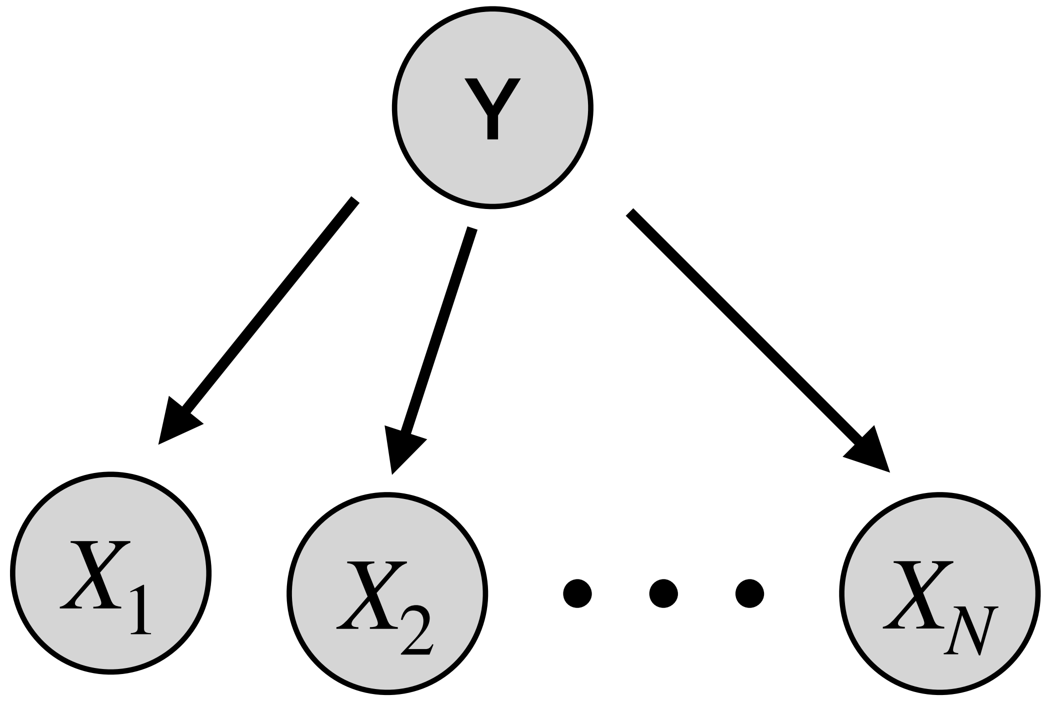
Example :



$P(X, Y) = P(Y | X) \times P(X)$



$P(X, Y, Z) = \ldots$



$P(Y, X_1, \cdots, X_N) =$

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using **diagrammatic representations** of probability distributions

Nodes : random variables

Links : probabilistic relationships

Bayesian networks
(**Directed** graphical models)

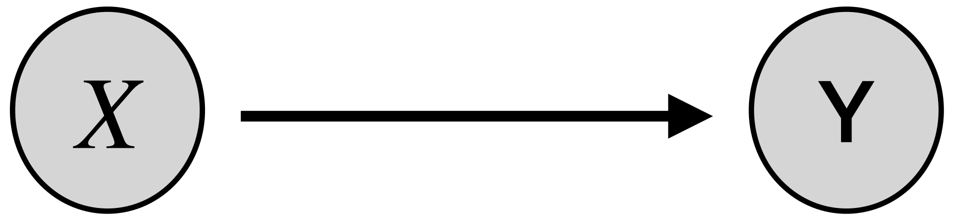
The focus of our course !

Markov random fields
(**Undirected** graphical models)

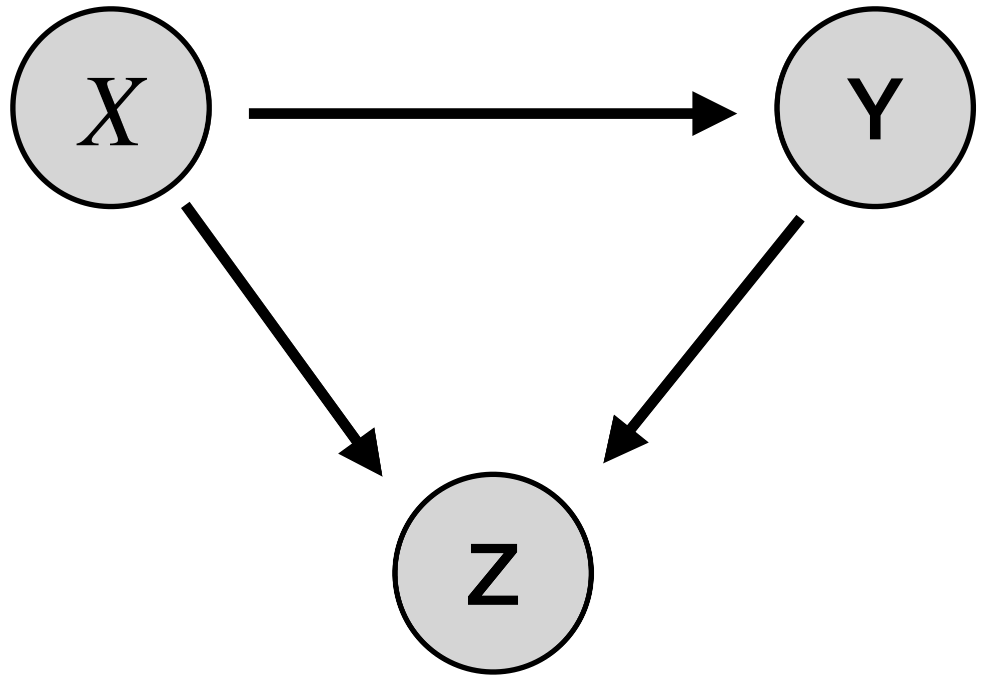
Model : joint probability over all variables

$$P(X_1, \cdots, X_N) = \prod_{i=1,\cdots,N} P(\ X_i \mid parents(X_i)\ )$$

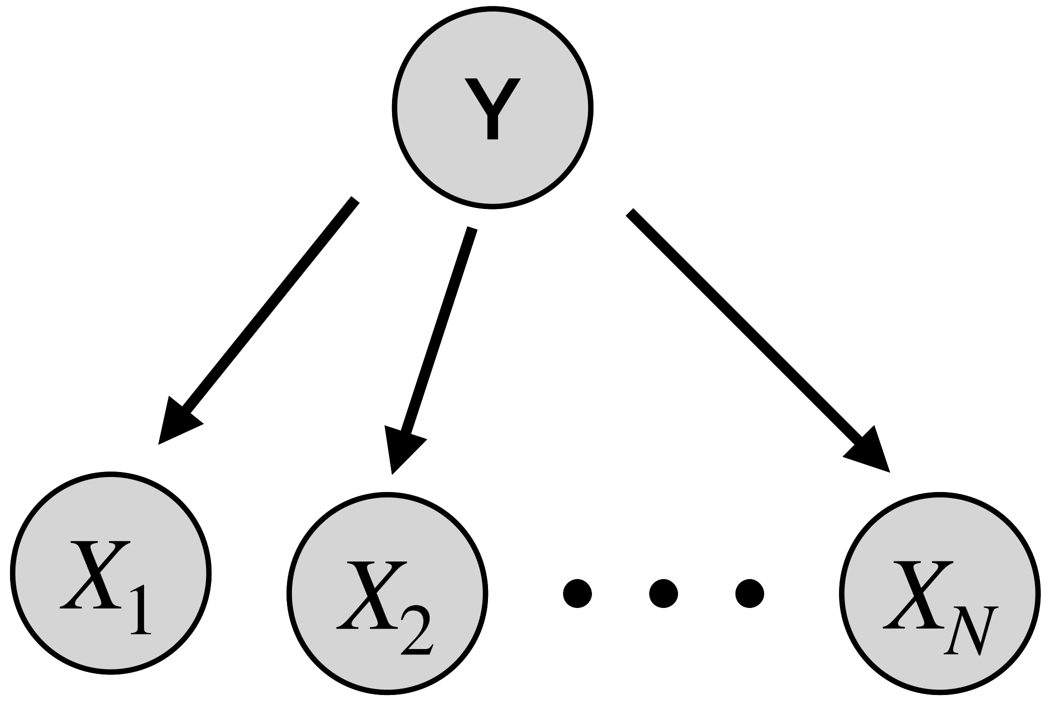
Example :



$$P(X, Y) = P(Y|X) \times P(X)$$



$$P(X, Y, Z) = P(Z|X, Y) \times P(X, Y)$$
$$= P(Z|X, Y) \times P(Y|X) \times P(X)$$



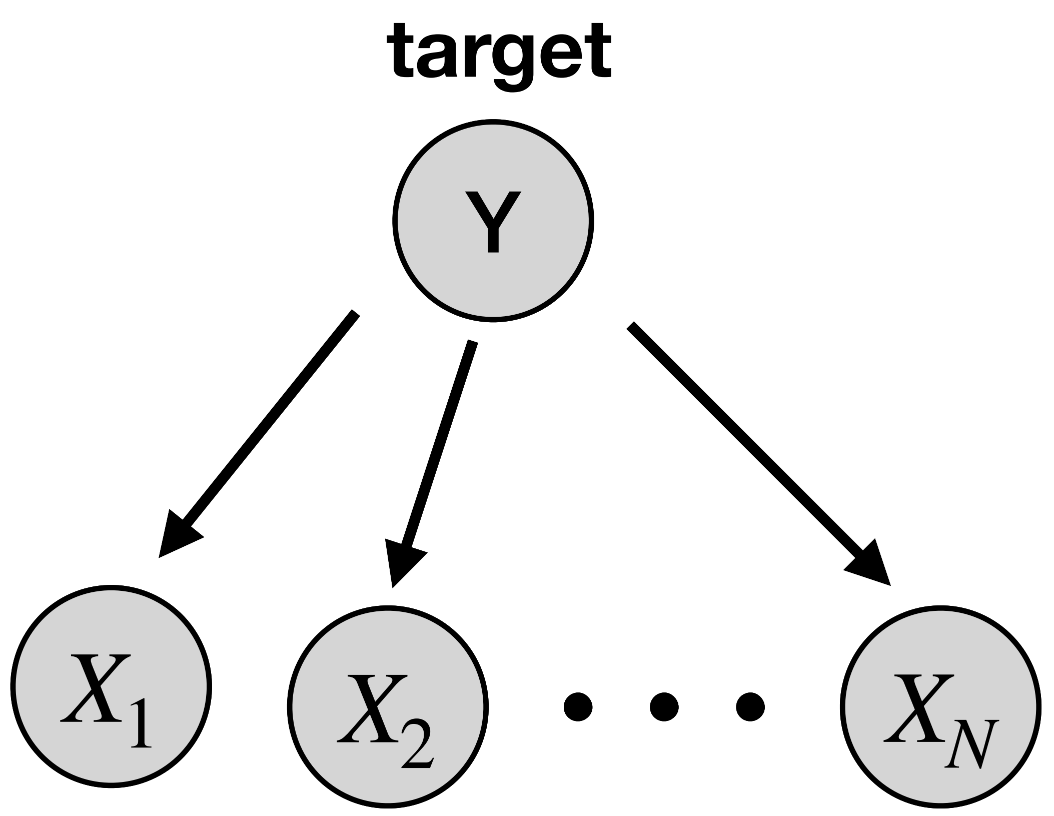
$$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i | Y)$$

2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier

$$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$$



Features
X_i , X_j conditionally independent given Y

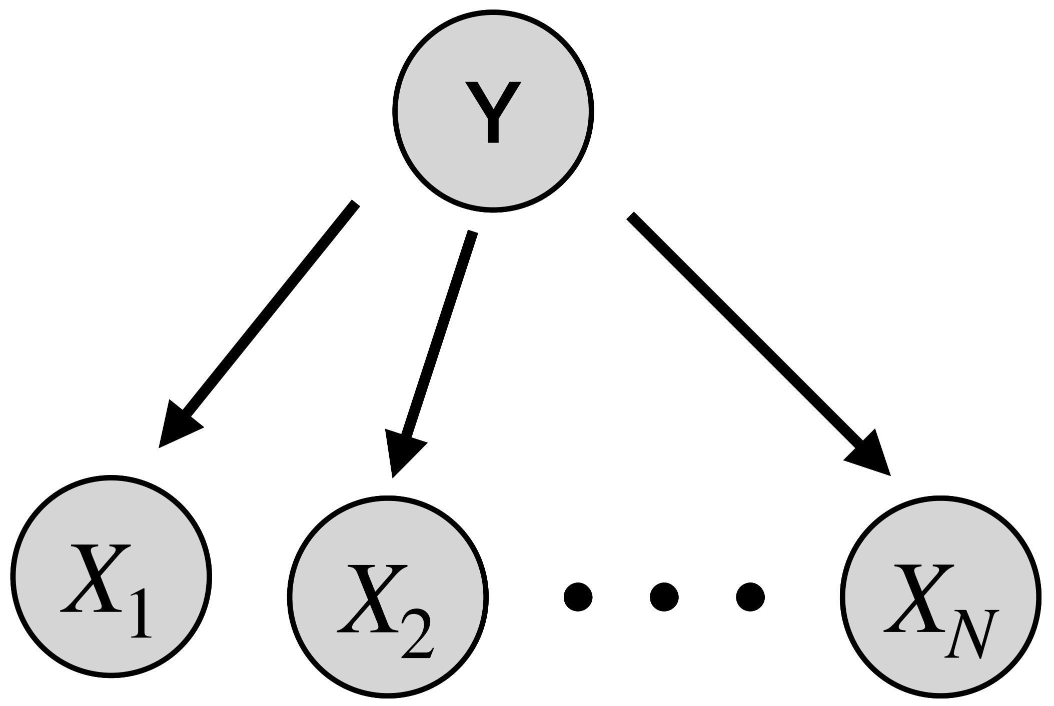
2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier

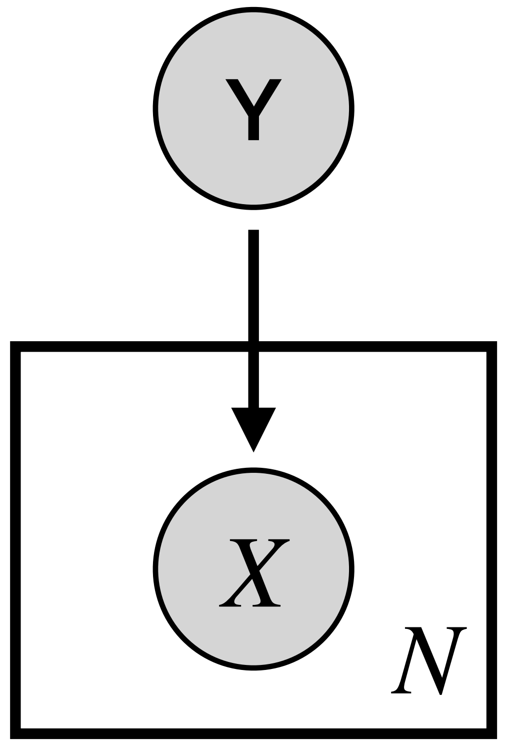
$$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$$

target



Features
X_i , X_j conditionally independent given Y

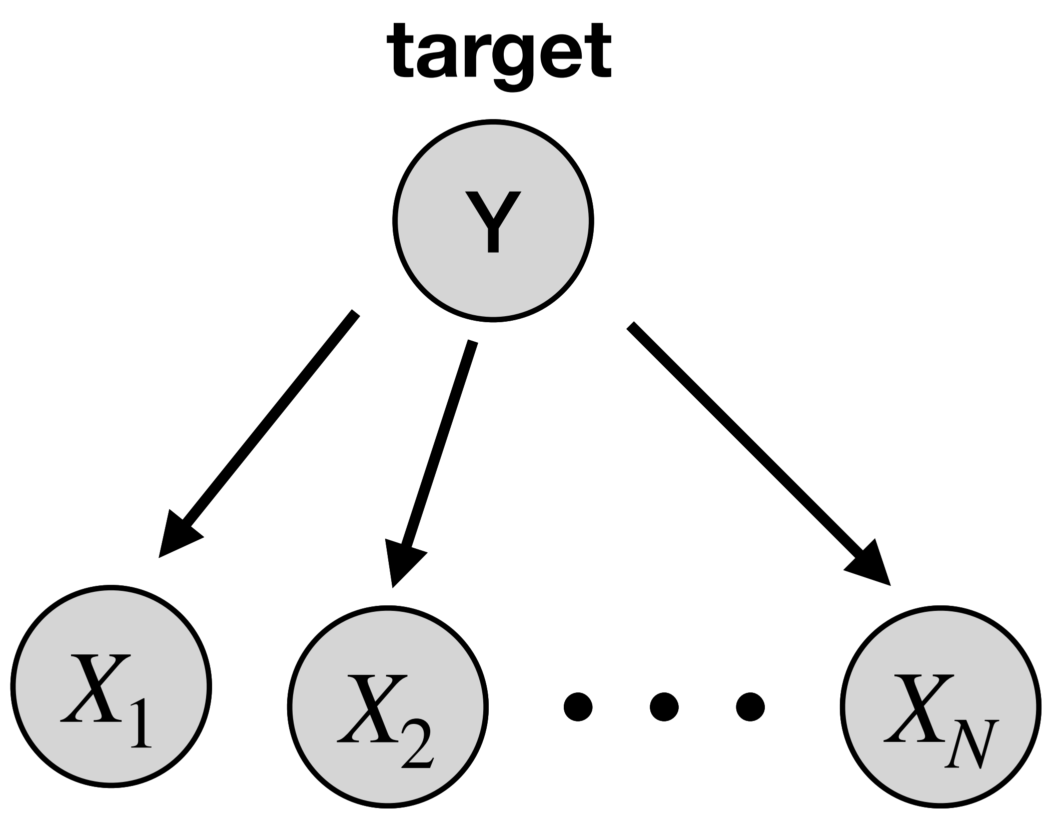
Plate notation $\longrightarrow$



2. Probabilistic model

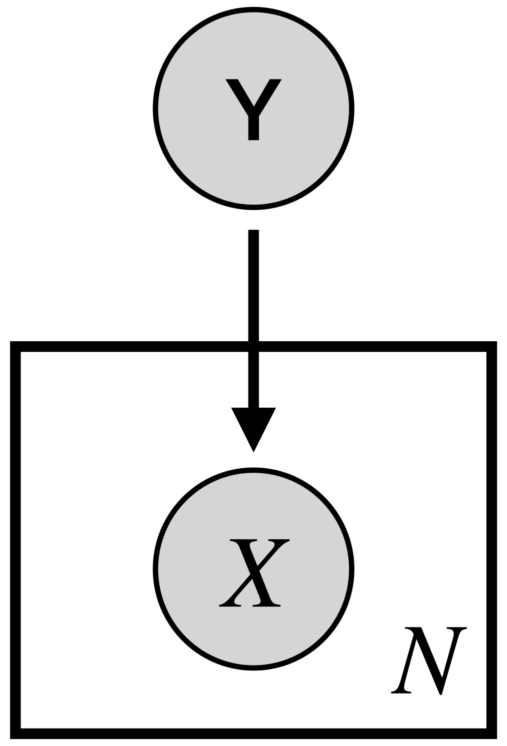
Plates and examples of probabilistic model

Naive Bayes Classifier $$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$$



Features
X_i , X_j conditionally independent given Y

Plate notation →



Reminder : Frequentist linear regression $x_i \in \mathbb{R}^d \quad y_1 \in \mathbb{R}$

Scalar notation :

$$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$$y_i = x_i^T \theta + \epsilon_i$$

Matrix notation :

$\mathbf{X} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$

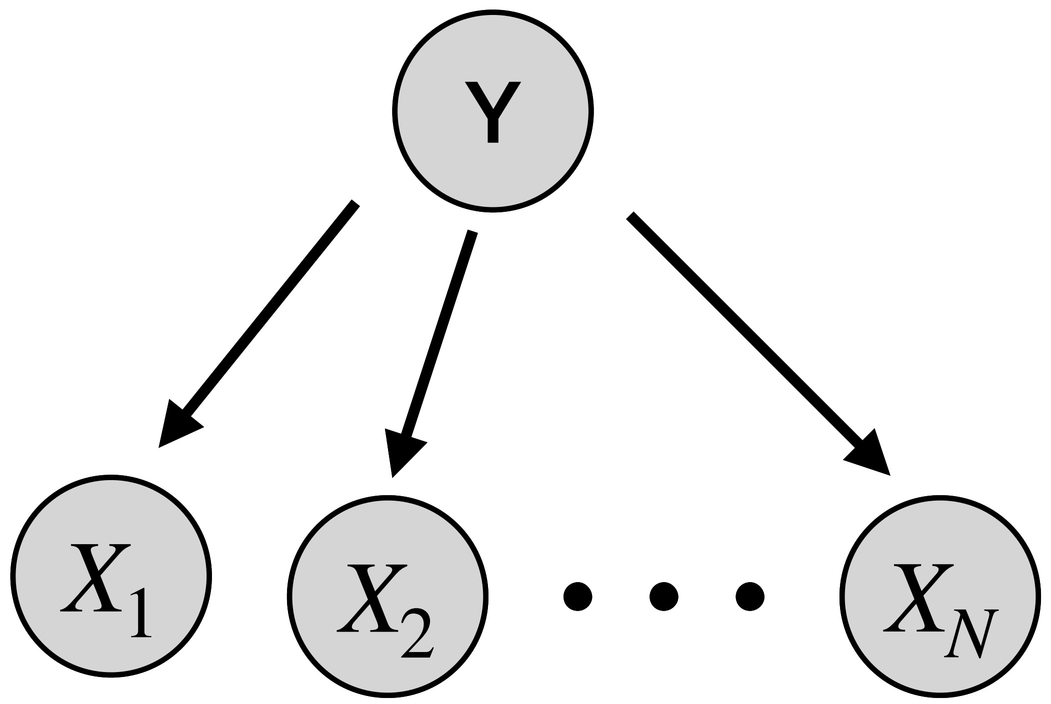
$$y = \mathbf{X}^T \theta + \epsilon$$

2. Probabilistic model

Plates and examples of probabilistic model

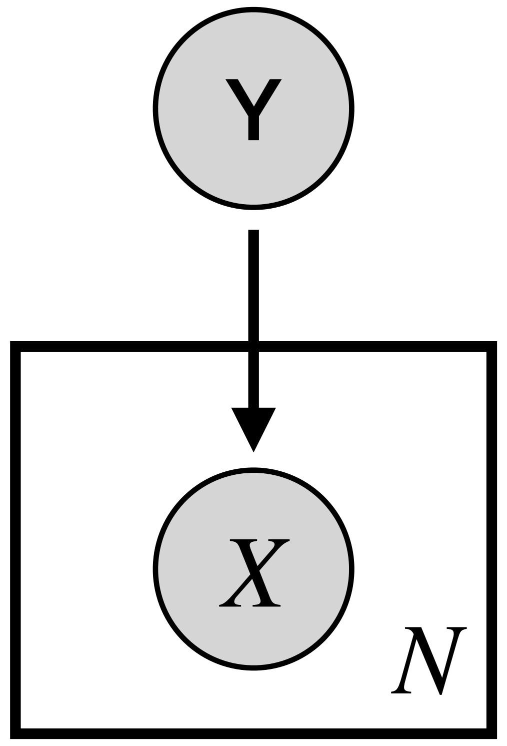
Naive Bayes Classifier $P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$

target



Features
X_i , X_j conditionally
independent given Y

Plate notation



Reminder : Frequentist linear regression $x_i \in \mathbb{R}^d$ $y_1 \in \mathbb{R}$

Scalar notation :

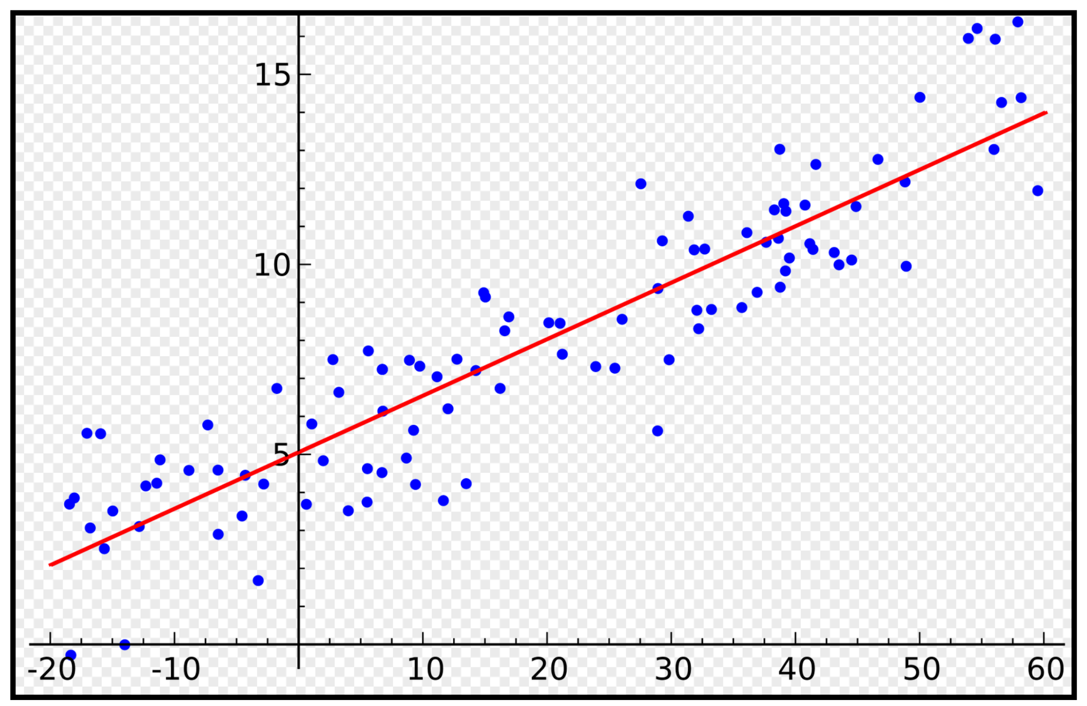
$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

$$y_i = x_i^T \theta + \epsilon_i$$

Matrix notation :

$\mathbf{X} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$

$$y = \mathbf{X}^T \theta + \epsilon$$



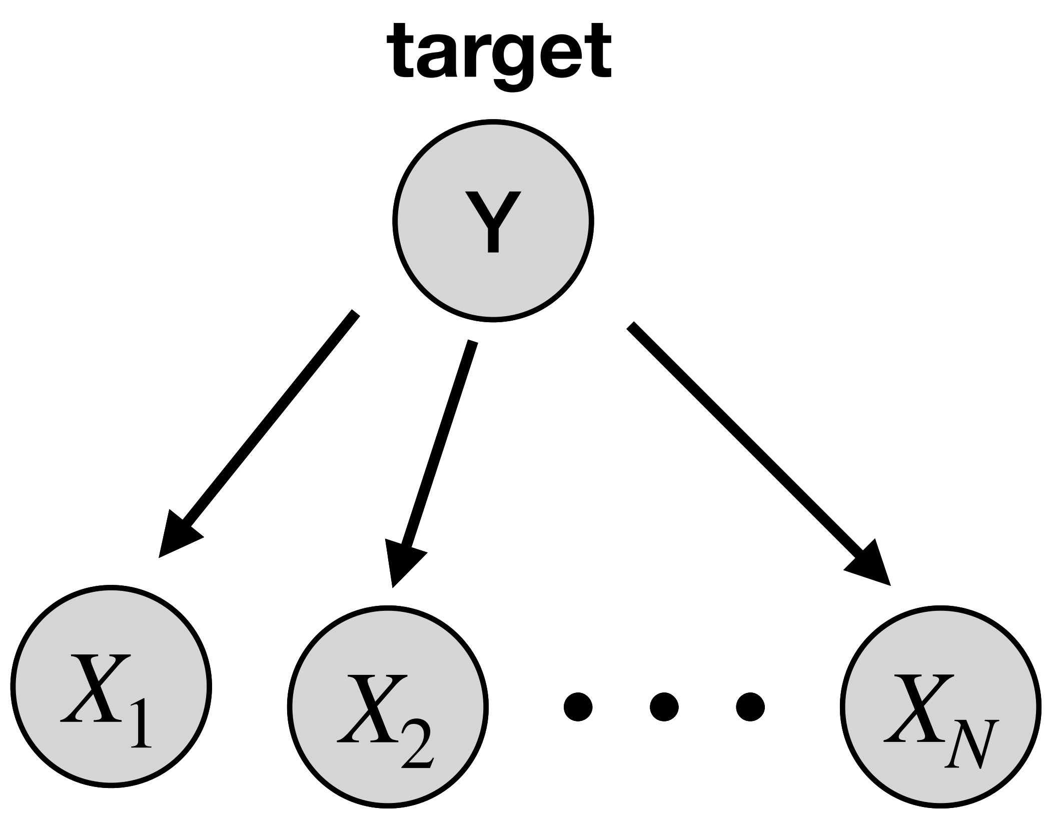
$$\min_{\theta} \|\theta^t X - y\|^2$$

2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier

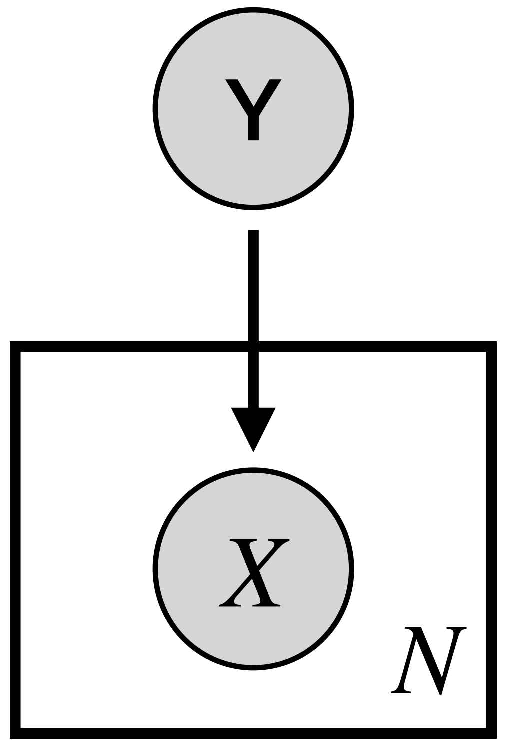
$$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$$



Features

X_i , X_j conditionally independent given Y

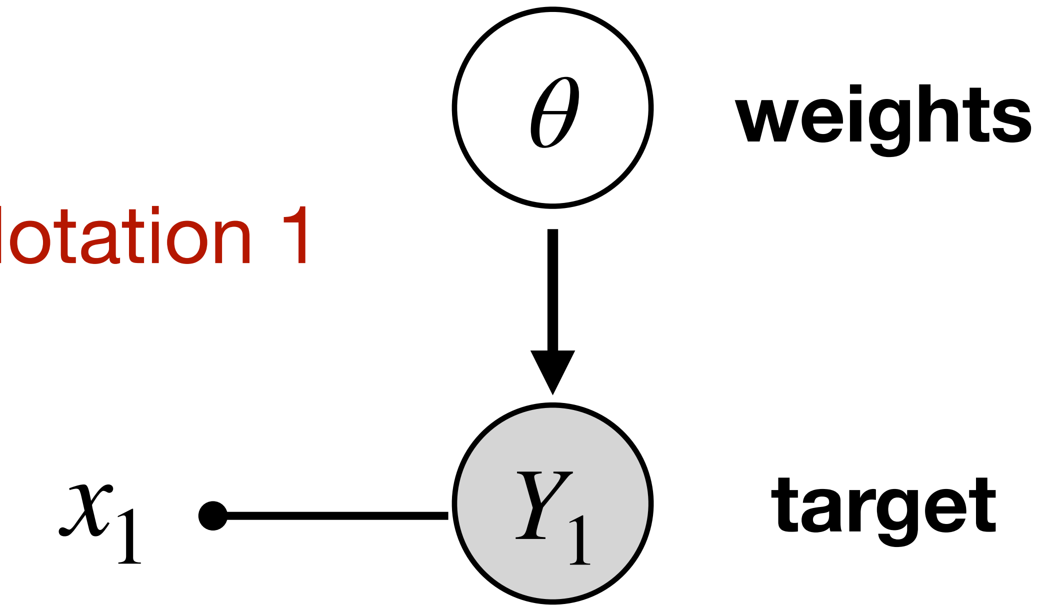
Plate notation



Bayesian Linear regression

(x_1, y_1) with $x_1 \in \mathbb{R}^d$, $y_1 \in \mathbb{R}$

Notation 1



$$P(\theta, y_1 \mid x_1) = P(y_1 \mid \theta, x_1) \times P(\theta)$$

$$P(y_1 \mid \theta, x_1) = \ldots$$

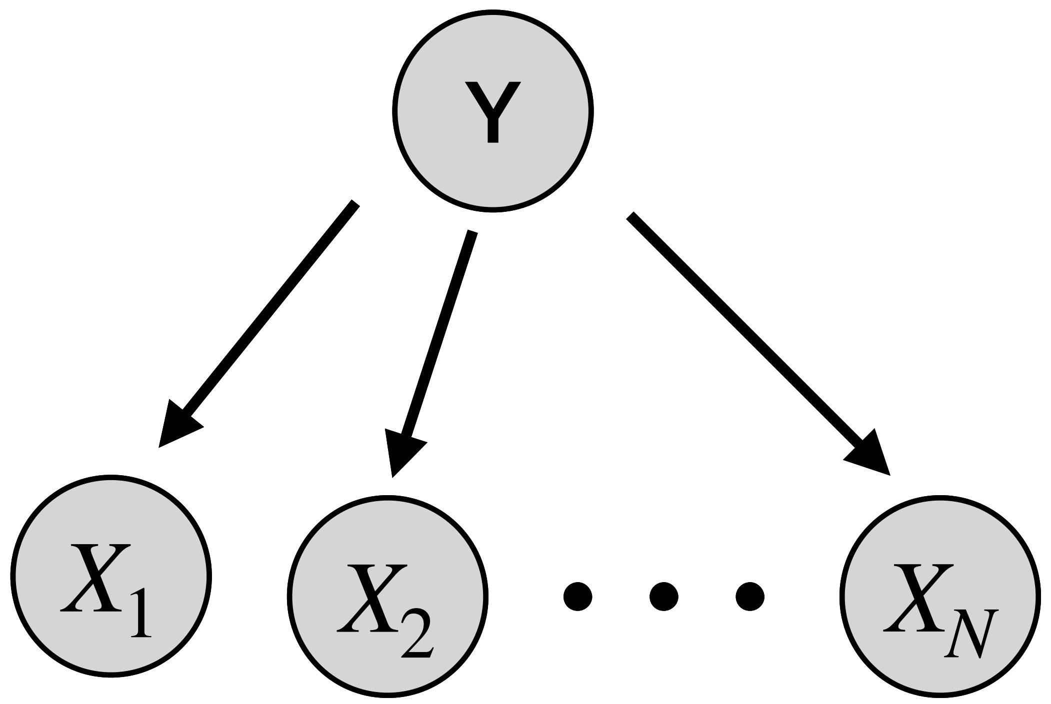
$$P(\theta) = \ldots$$

2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier $P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$

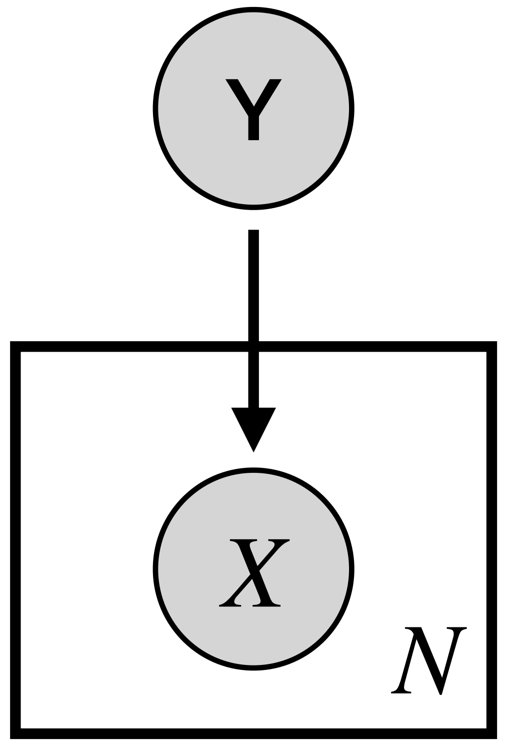
target



Features

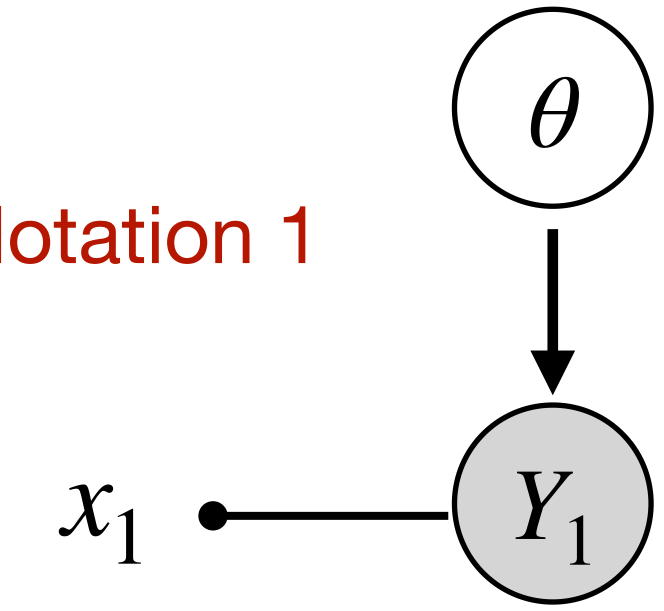
X_i , X_j conditionally independent given Y

Plate notation



Bayesian Linear regression (x_1, y_1) with $x_1 \in \mathbb{R}^d$, $y_1 \in \mathbb{R}$

Notation 1



weights

target

$$P(\theta, y_1 \mid x_1) = P(y_1 \mid \theta, x_1) \times P(\theta)$$

$$P(y_1 \mid \theta, x_1) = \mathcal{N}(y_1 \mid \theta^T x_1,\ \sigma^2)$$

$$P(\theta) = \mathcal{N}(\theta \mid 0,\ \gamma^2)$$

Legend :

Fixed variable

Hidden (latent) r.v.

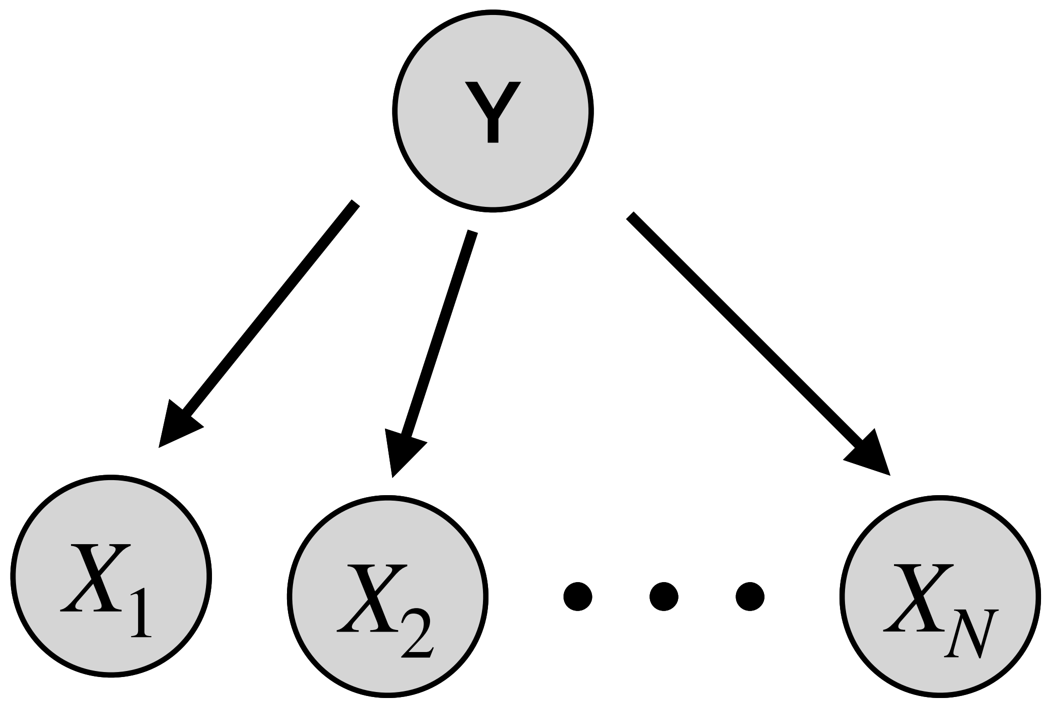
Observed r.v.

2. Probabilistic model

Plates and examples of probabilistic model

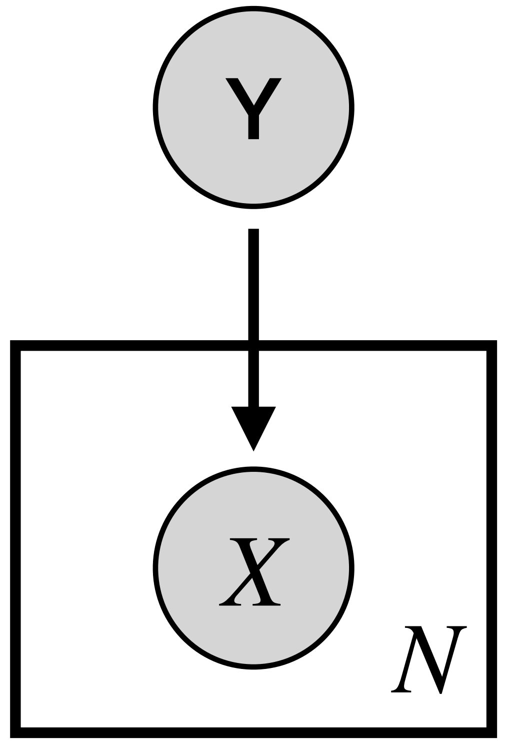
Naive Bayes Classifier $P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$

target



Features
X_i , X_j conditionally
independent given Y

Plate notation



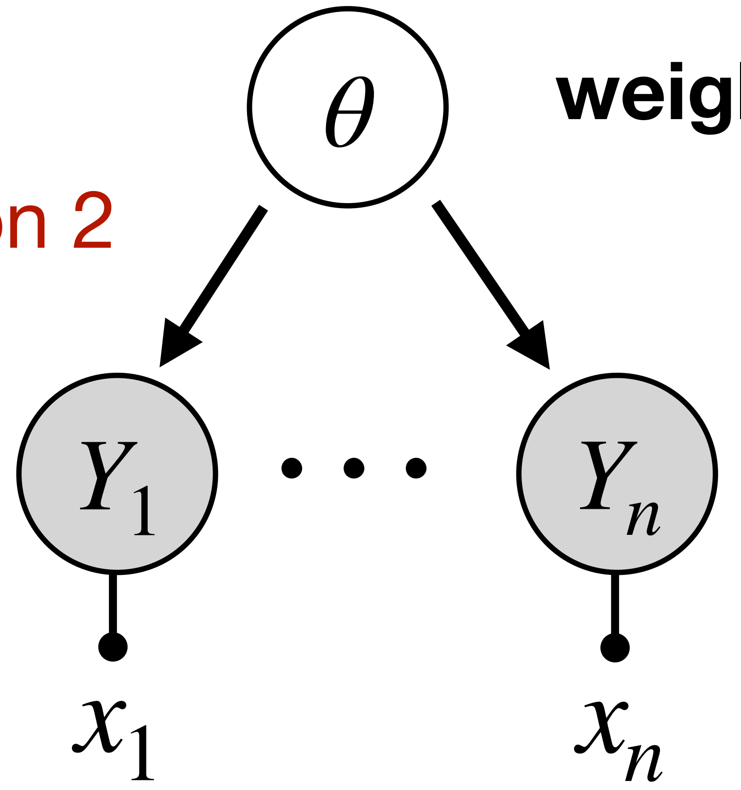
Bayesian Linear regression $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$$P(\theta, y_i \mid x_i) = P(y_i \mid \theta, x_i) \times P(\theta)$$

$$P(y_i \mid \theta, x_i) = \mathcal{N}(y_i \mid \theta^T x_i, \sigma^2)$$

$$P(\theta) = \mathcal{N}(\theta \mid 0, \gamma^2)$$

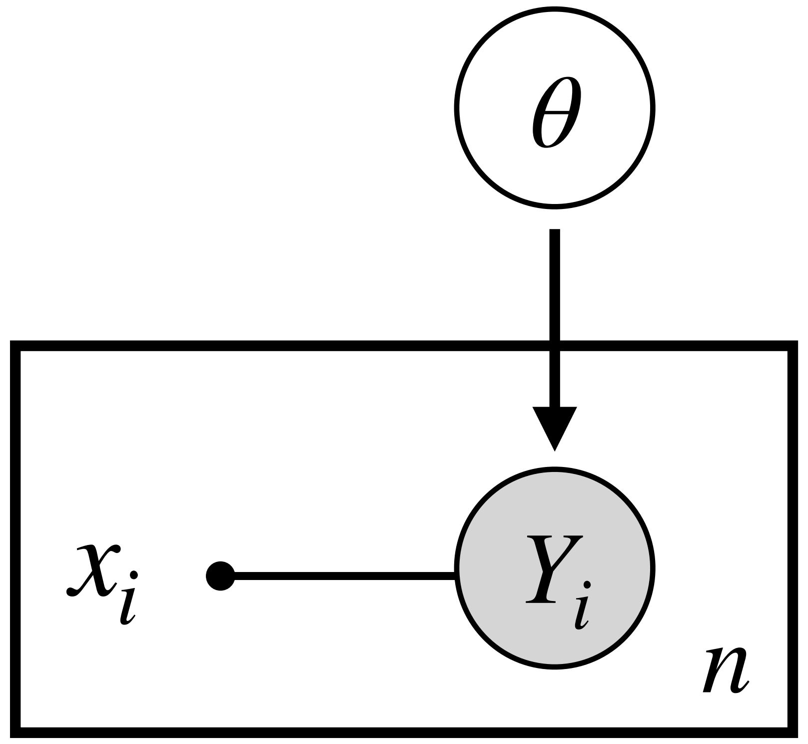
Notation 2



weights

target

Plate notation



Legend :

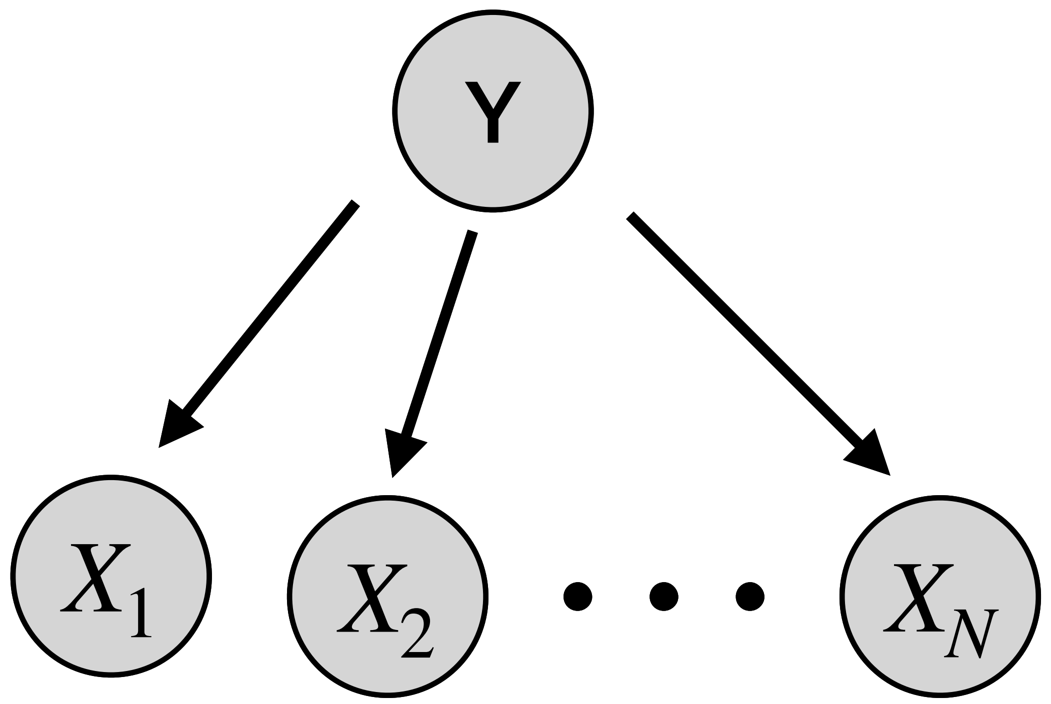
2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier

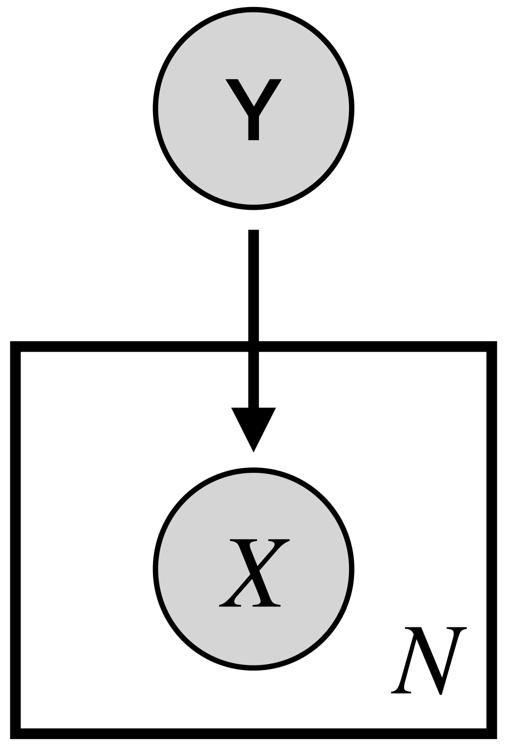
$$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$$

target



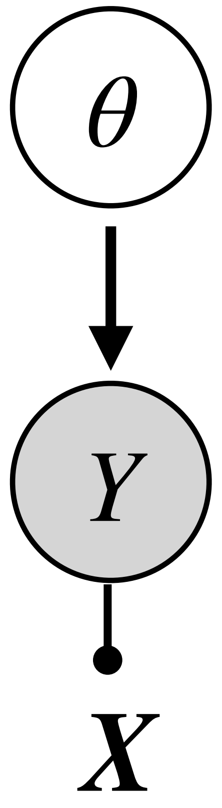
Features
X_i , X_j conditionally
independent given Y

Plate notation



Bayesian Linear regression $\mathbf{X} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Notation 3



weights

target

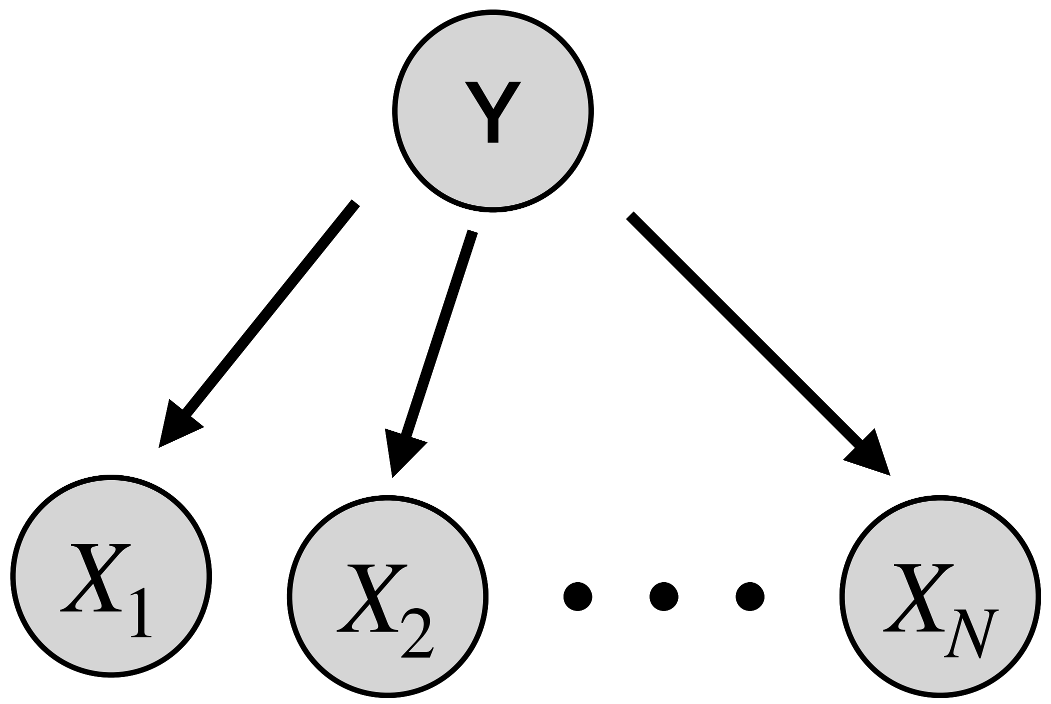
$$P(\theta, y \mid x) = P(y \mid \theta, x) \times P(\theta)$$
$$P(y \mid \theta, x) = \ldots$$
$$P(\theta) = \ldots$$

2. Probabilistic model

Plates and examples of probabilistic model

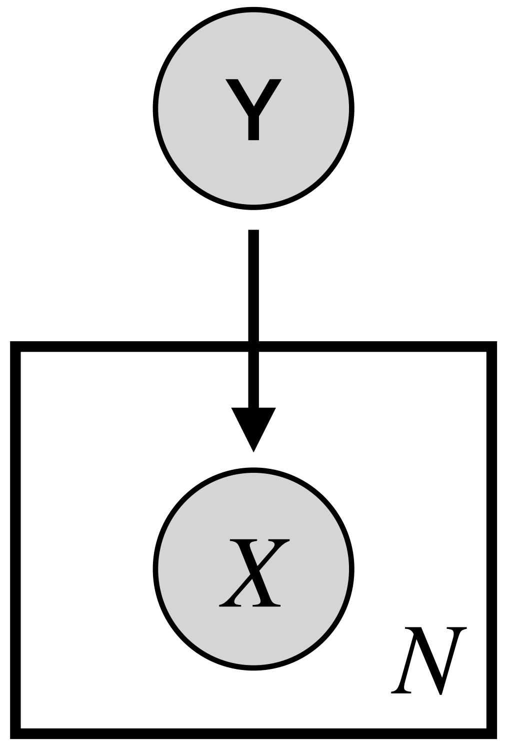
Naive Bayes Classifier $P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$

target



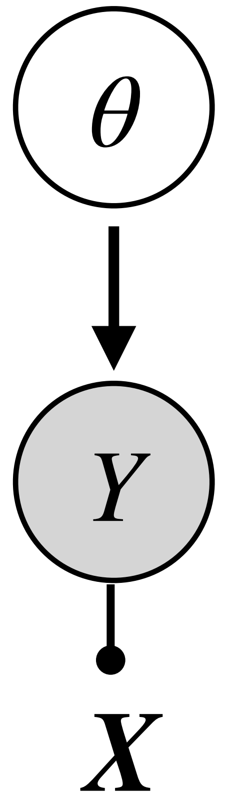
Features
X_i , X_j conditionally
independent given Y

Plate notation



Bayesian Linear regression $\mathbf{X} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Notation 3



$$P(\theta, y \mid x) = P(y \mid \theta, x) \times P(\theta)$$

$$P(y \mid \theta, x) = \mathcal{N}(y \mid \theta^T X, \sigma^2 I_n)$$

$$P(\theta) = \mathcal{N}(\theta \mid 0, \gamma^2 I_n)$$

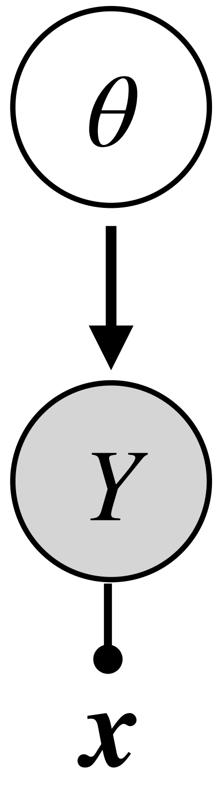
Legend :

Fixed variable
Hidden (latent) r.v.
Observed r.v.

2. Probabilistic model

Linear regression

Bayesian Linear regression



$$P(\theta, y \,|\, X) = P(y \,|\, \theta, X) \times P(\theta)$$

$$P(y \,|\, \theta, X) = \mathcal{N}(y \,|\, \theta^T X,\, \sigma^2 I_n)$$

$$P(\theta) = \mathcal{N}(\theta \,|\, 0,\, \gamma^2 I_n)$$

Objective :

Frequentist linear regression

Objective : $\min\limits_{\theta} \mathcal{L}(\theta) = \min\limits_{\theta} \|\theta^t X - y\|^2$

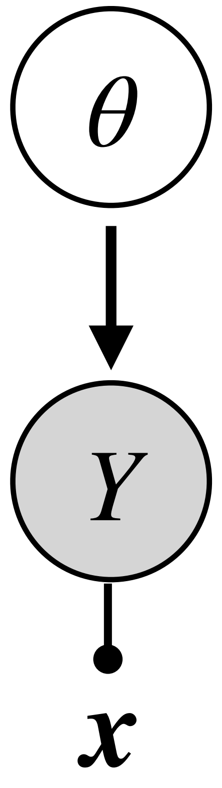
$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta)$$

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

2. Probabilistic model

Linear regression

Bayesian Linear regression



$$P(\theta, y \mid X) = P(y \mid \theta, X) \times P(\theta)$$

$$P(y \mid \theta, X) = \mathcal{N}(y \mid \theta^T X,\ \sigma^2 I_n)$$

$$P(\theta) = \mathcal{N}(\theta \mid 0,\ \gamma^2 I_n)$$

Objective : $\arg\max_{\theta} P(\theta \mid X, y) = \arg\max_{\theta} P(\theta, y \mid X)$

Frequentist linear regression

Objective : $\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \|\theta^t X - y\|^2$

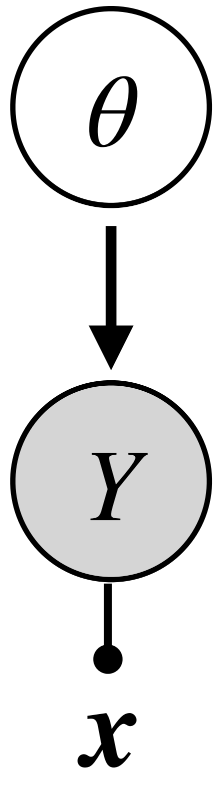
$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta)$$

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

2. Probabilistic model

Linear regression

Bayesian Linear regression



$$P(\theta, y \mid X) = P(y \mid \theta, X) \times P(\theta)$$

$$P(y \mid \theta, X) = \mathcal{N}(y \mid \theta^T X, \sigma^2 I_n)$$

$$P(\theta) = \mathcal{N}(\theta \mid 0, \gamma^2 I_n)$$

Objective : $\arg\max_{\theta} P(\theta \mid X, y) = \arg\max_{\theta} P(\theta, y \mid X)$

Frequentist linear regression

Objective : $\min_{\theta} \mathscr{L}(\theta) = \min_{\theta} \|\theta^t X - y\|^2$

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}(\theta)$$

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Theorem :

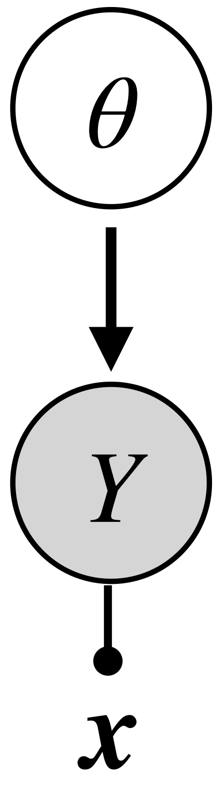
There exists $\lambda \in \mathbb{R}$ such that : $\arg\max_{\theta} P(\theta \mid X, y) = \arg\min_{\theta} \left\{ \|\theta^t X - y\|^2 + \lambda \|\theta\|^2 \right\}$

So by adding a normal prior on the weight we turned this problem into a L_2 **regularised problem**

2. Probabilistic model

Linear regression

Bayesian Linear regression



$$P(\theta, y \,|\, X) = P(y \,|\, \theta, X) \times P(\theta)$$

$$P(y \,|\, \theta, X) = \mathcal{N}(y \,|\, \theta^T X,\, \sigma^2 I_n)$$

$$P(\theta) = \mathcal{N}(\theta \,|\, 0,\, \gamma^2 I_n)$$

Objective : $\arg\max_\theta P(\theta \,|\, X, y) = \arg\max_\theta P(\theta, y \,|\, X)$

Frequentist linear regression

Objective : $\min_\theta \mathcal{L}(\theta) = \min_\theta \|\theta^t X - y\|^2$

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta)$$

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Theorem :

There exists $\lambda \in \mathbb{R}$ such that : $\arg\max_\theta P(\theta \,|\, X, y) = \arg\min_\theta \left\{ \|\theta^t X - y\|^2 + \lambda \|\theta\|^2 \right\}$

So by adding a normal prior on the weight we turned this problem into a L_2 **regularised problem**

Proof : see the whiteboard in class <u>**or**</u> left as an exercice

3 Analytical Inference

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

Likelihood **Prior**

$$P(\theta \mid X) = \frac{P(X, \theta)}{P(X)} = \frac{P(X \mid \theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

Fixed by **model**

Fixed by **us**

Likelihood   Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

Fixed by **data**

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

Fixed by model — **Likelihood**

Prior — **Fixed by us**

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

HARD TO COMPUTE

Fixed by data

$$P(X) = \int_\theta P(X|\theta) \cdot P(\theta) \cdot d\theta$$

3. Analytical Inference

Maximum a posteriori (MAP) : definition & remarks

Posterior distribution

Fixed by model

Fixed by us

Likelihood Prior

$$\underbrace{P(\theta|X)}_{\text{Posterior}} = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Evidence

HARD TO COMPUTE

Fixed by data

$$P(X) = \int_\theta P(X|\theta) \cdot P(\theta) \cdot d\theta$$

Remarks

- We have to avoid computing the evidence
- Naive approach : maximum a posteriori ,

$$\hat{\theta}_{MAP} = \arg\max_\theta \left\{ \frac{P(\theta|X) \cdot P(\theta)}{P(X)} \right\}$$

$$= \arg\max_\theta P(X|\theta) \cdot P(\theta)$$

- This maximization can be done with numerical optimization problem

3. Analytical Inference

Maximum a posteriori (MAP) : limitations

Posterior distribution

Fixed by model

Fixed by us

Likelihood Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

HARD TO COMPUTE

Fixed by data

$$P(X) = \int_\theta P(X|\theta) \cdot P(\theta) \cdot d\theta$$

Remarks

- We have to avoid computing the evidence
- Naive approach : maximum a posteriori ,

$$\hat{\theta}_{MAP} = \arg\max_\theta \left\{ \frac{P(\theta|X) \cdot P(\theta)}{P(X)} \right\}$$

$$= \arg\max_\theta P(X|\theta) \cdot P(\theta)$$

- This maximization can be done with numerical optimization problem

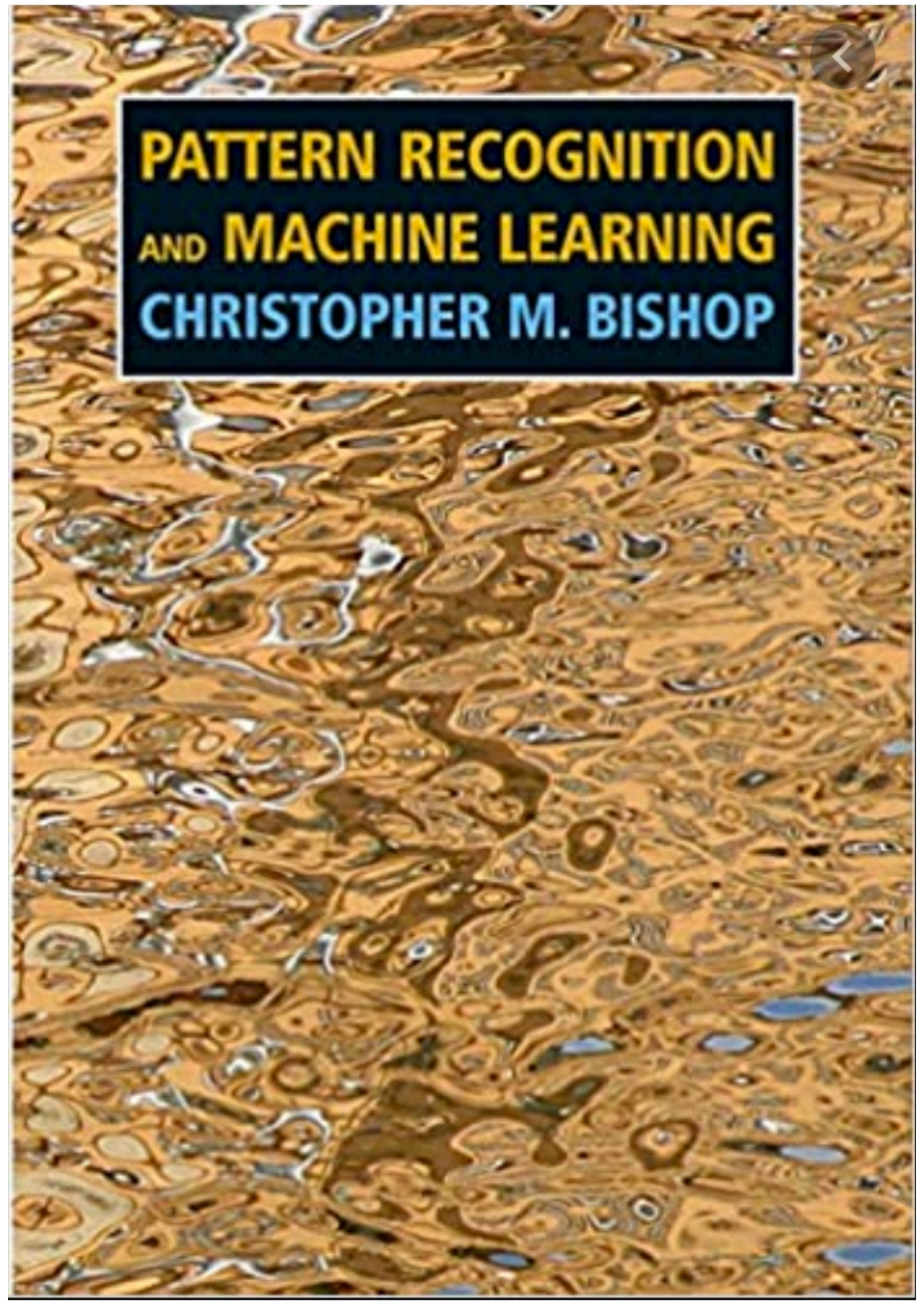
Limitations (among many others)

1. in general, not representative of bayesian methods : $\hat{\theta}_{MAP}$ is a point estimate like $\hat{\theta}_{MLE}$
   - can't compute credible intervals because it doesn't return a pdf/pmf (not a bayesian inference)
2. can't use online learning : the prior is not well updated

3. Analytical Inference

Maximum a posteriori (MAP) : book

For more theoretical details (and example on analytical inference) :



4 Conjugate distributions

4. Conjugate distributions

Conjugate distributions : avoid computing evidence

Posterior distribution

Fixed by model

Fixed by us

Likelihood

Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

Fixed by data

4. Conjugate distributions

Conjugate distributions : avoid computing evidence

Posterior distribution

Fixed by **model**

Fixed by **us**

Likelihood **Prior**

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

Fixed by **data**

Remarks

- We have to avoid computing the evidence
- We can choose a convenient prior which enable us to compute the **posterior :**

  Conjugate prior

4. Conjugate distributions

Conjugate distributions : avoid computing evidence

Posterior distribution

Fixed by model — **Likelihood**

Prior — **Fixed by us**

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence — **Fixed by data**

Remarks

- We have to avoid computing the evidence
- We can choose a convenient prior which enable us to compute the **posterior :**

  Conjugate prior

Conjugate prior

$P(\theta)$ is **conjugate** to $P(X|\theta)$ if the $P(\theta)$ and $P(X|\theta)$ lie in the same family of distributions (gaussian for example)

4. Conjugate distributions

Conjugate distributions : avoid computing evidence

Posterior distribution

Fixed by **model**

Fixed by **us**

Likelihood Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

Fixed by **data**

Remarks

- We have to avoid computing the evidence
- We can choose a convenient prior which enable us to compute the **posterior :**

  Conjugate prior

Conjugate prior

$P(\theta)$ is **conjugate** to $P(X|\theta)$ if the $P(\theta)$ and $P(X|\theta)$ lie in the same family of distributions (gaussian for example)

Example

$$P(\theta|X) = \frac{\mathcal{N}(X|\theta,\sigma^2) \times P(\theta)}{P(X)}$$

$\mathcal{N}(\theta|\mu_{prior}, \sigma^2_{prior})$

$\mathcal{N}(\theta|\mu_{posterior}, \sigma^2_{posterior})$

In the context of a gaussian, the prior for the **mean** is a gaussian !

4. Conjugate distributions

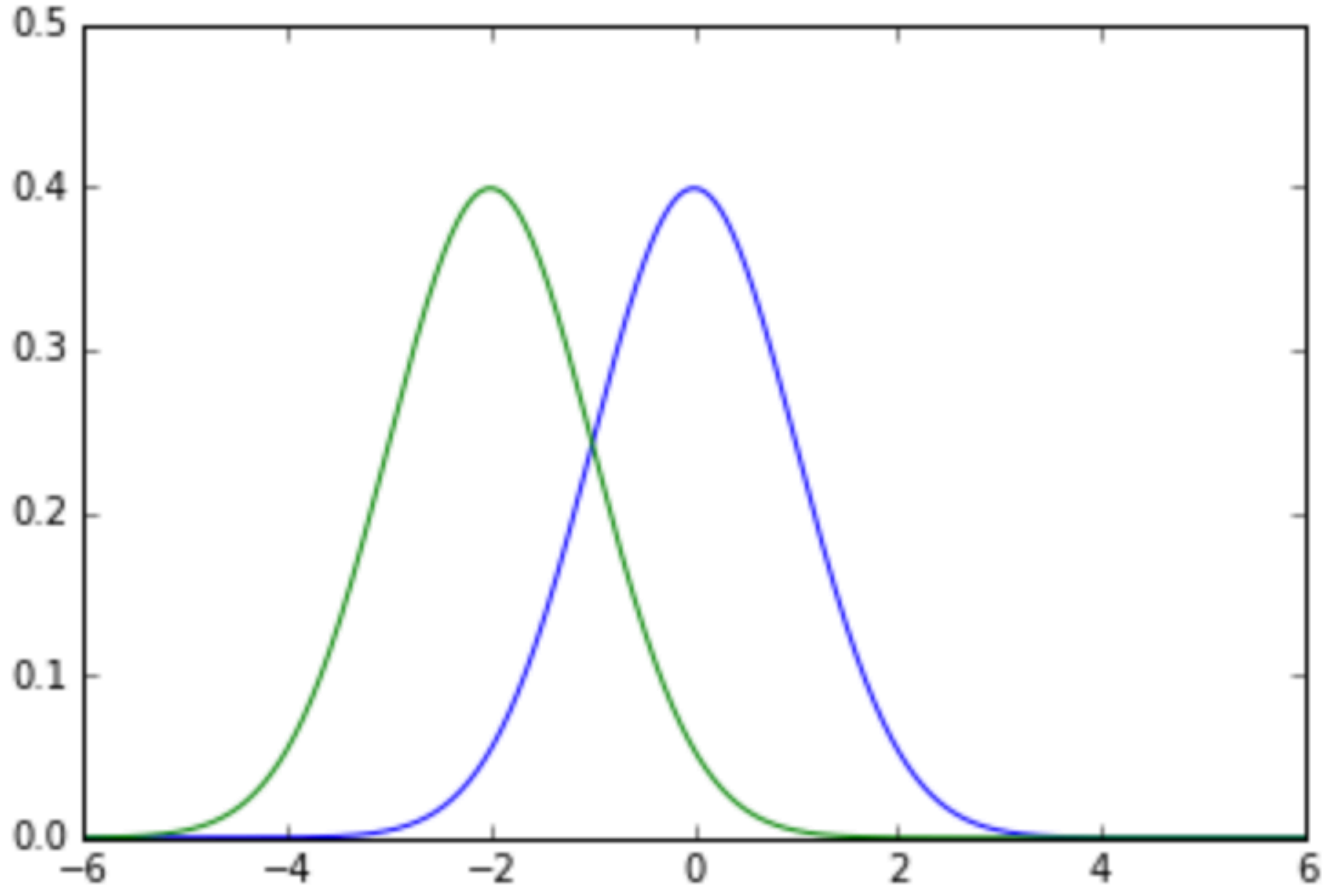
Conjugate distributions : example

Example

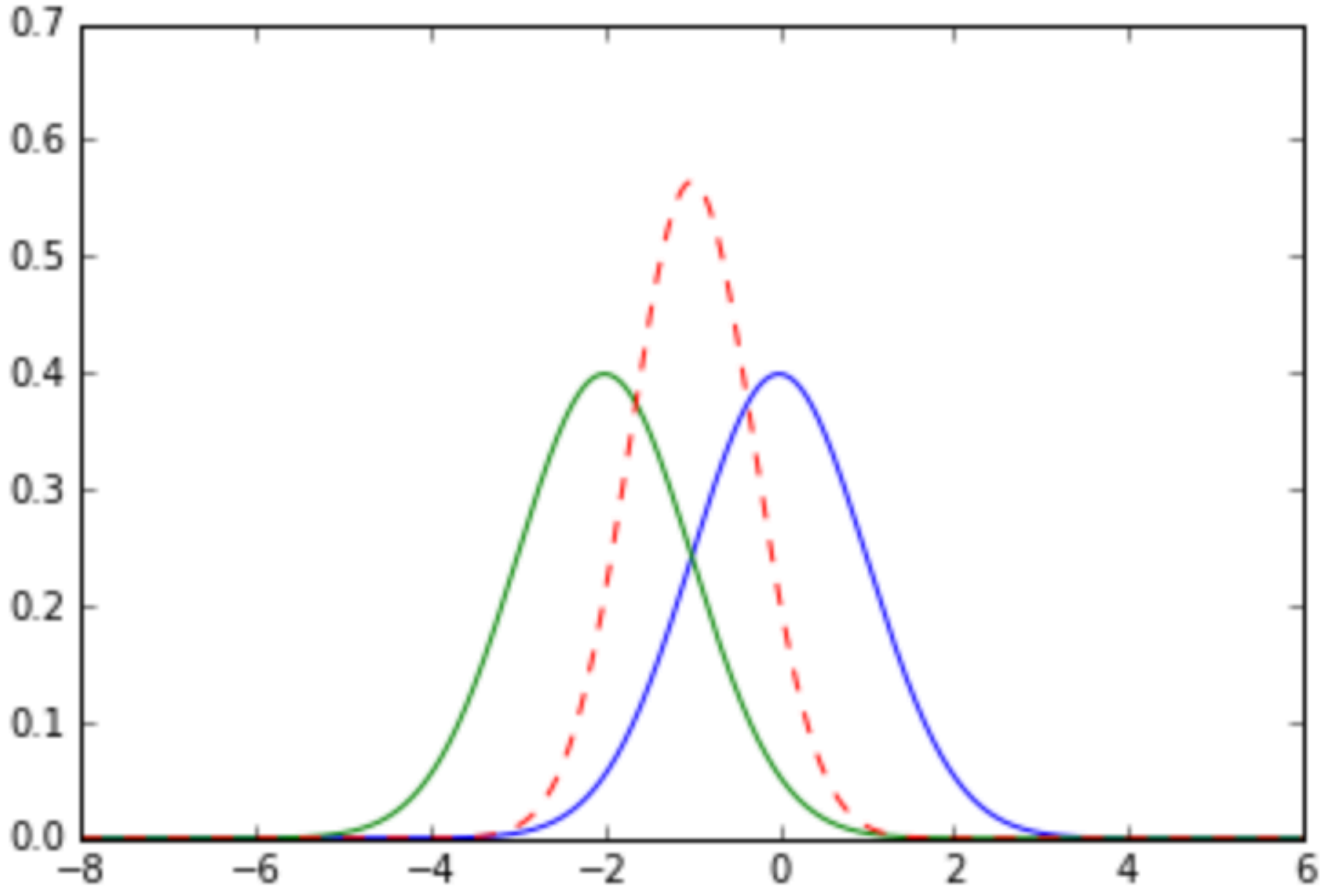
$$P(\theta|X) = \frac{\mathcal{N}(X|\theta,\sigma^2) \times P(\theta)}{P(X)}$$

$\mathcal{N}(\theta|\mu_{posterior}, \sigma^2_{posterior})$

$\mathcal{N}(\theta|\mu_{prior}, \sigma^2_{prior})$



pointwise product



Exercice (left as an exercice, correction in the next lecture)

Show that $\mathcal{N}(\theta|x/2,1/2) = \dfrac{\mathcal{N}(x|\theta,1) \times \mathcal{N}(\theta|0,1)}{P(x)}$

4. Conjugate distributions

Usual distributions : Gamma distribution

Gamma distribution

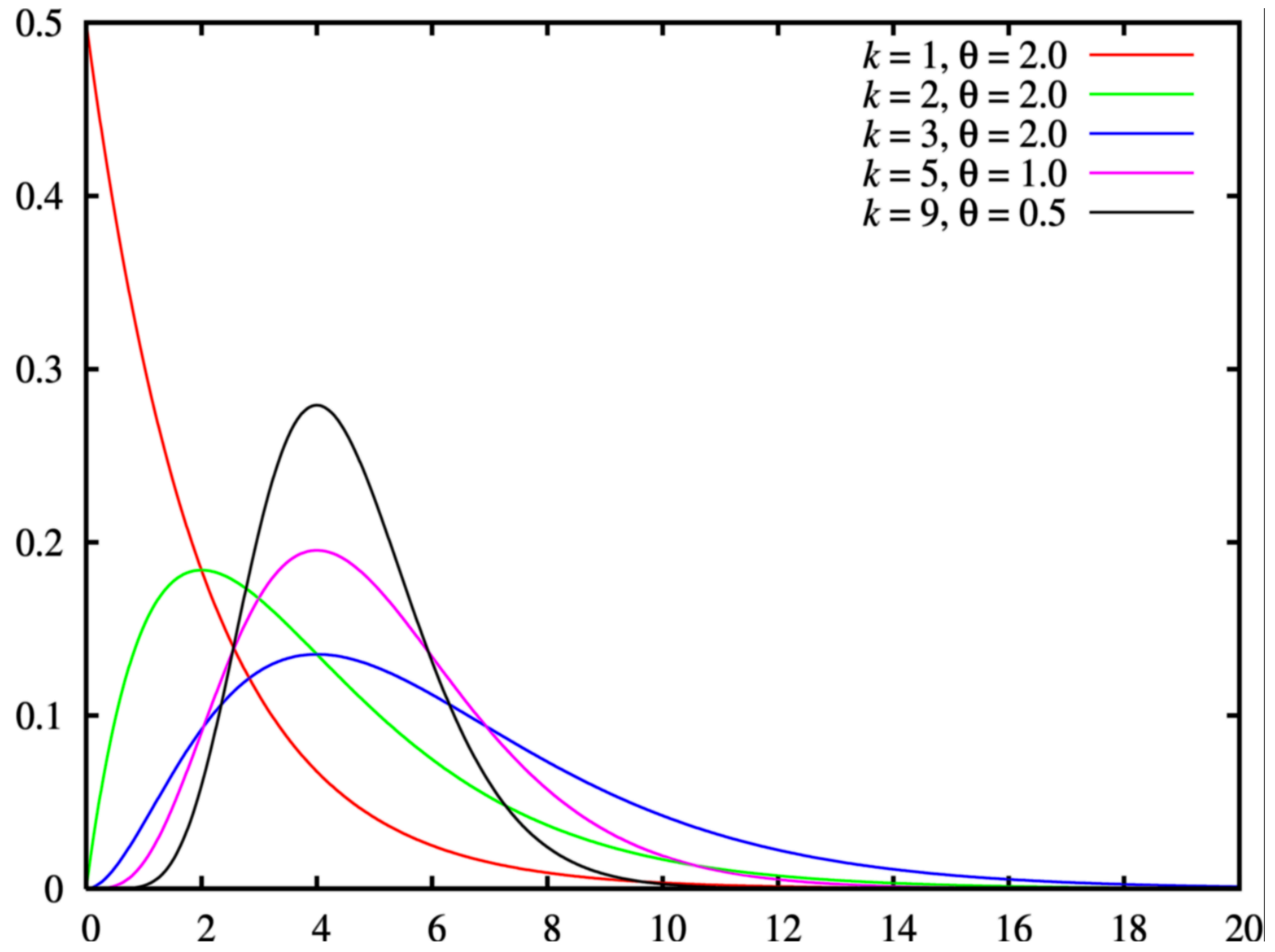
PDF : $\Gamma(x\,|\,\alpha,\beta) = \dfrac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ with $x, \alpha, \beta > 0$

$\Gamma(\alpha) = (\alpha - 1)!$

mean : $\mathbb{E}[x] = \dfrac{\alpha}{\beta}$

variance : $V(x) = \dfrac{\alpha}{\beta^2}$

mode : $Mode\,[x] = \dfrac{\alpha - 1}{\beta}$



Example

$$\Gamma(\gamma\,|\,\alpha_{posterior}, \beta_{posterior}) \quad P(\gamma\,|\,x) = \frac{\mathcal{N}(x\,|\,\mu, \gamma^{-1}) \times P(\gamma)}{P(x)} \quad \Gamma(\gamma\,|\,\alpha_{prior}, \beta_{prior})$$

In the context of a gaussian, the prior for the **precision** is a gamma !

4. Conjugate distributions

Usual distributions : Gamma distribution

Gamma distribution

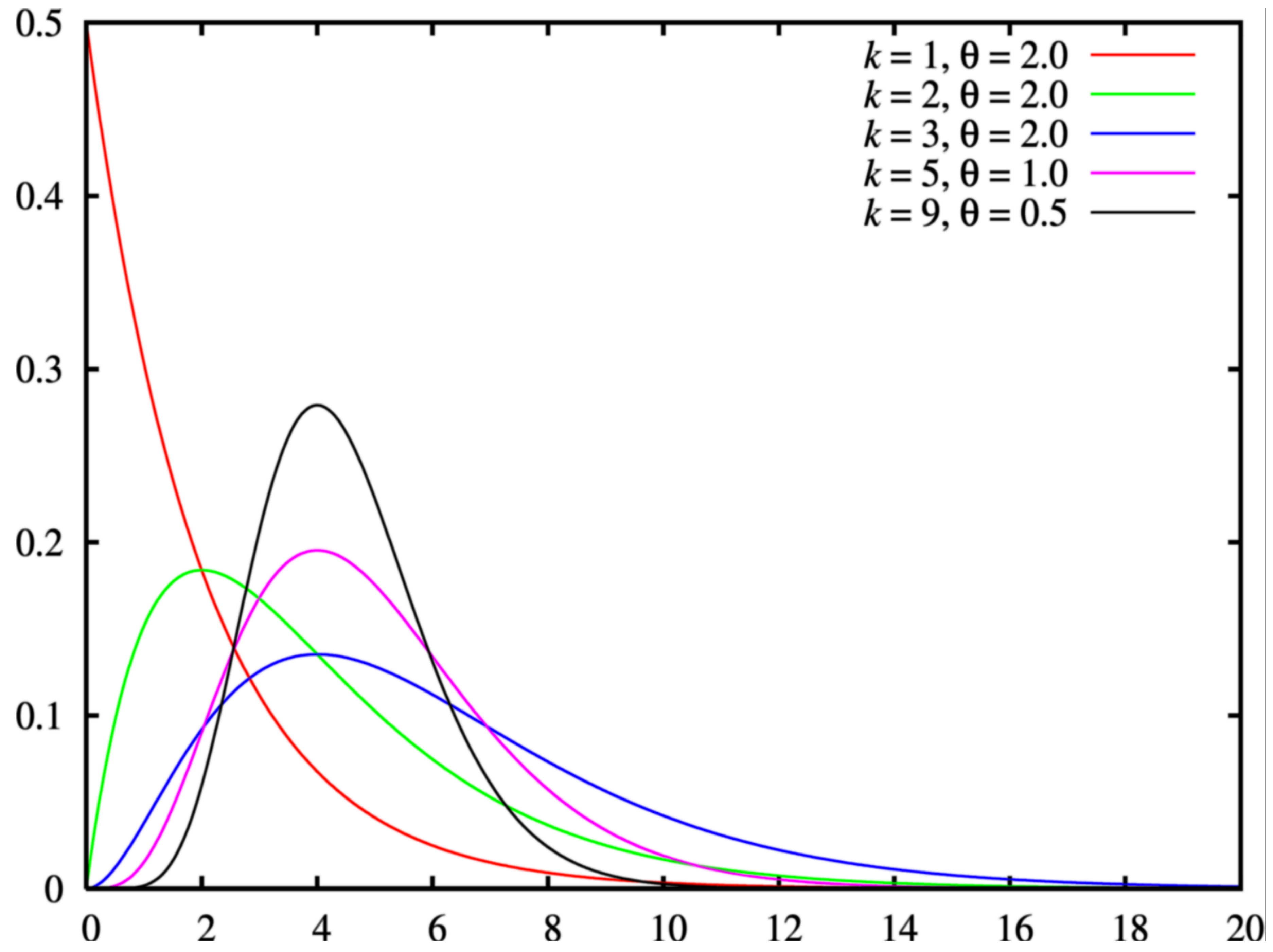
PDF : $\Gamma(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ with $x, \alpha, \beta > 0$

$\Gamma(\alpha) = (\alpha - 1)!$

mean : $\mathbb{E}[x] = \frac{\alpha}{\beta}$

variance : $V(x) = \frac{\alpha}{\beta^2}$

mode : $Mode\,[x] = \frac{\alpha - 1}{\beta}$



Exercice (left as an exercice, correction in the next lecture)

$$P(\gamma \mid x) = \frac{\mathcal{N}(x \mid \mu, \gamma^{-1}) \times P(\gamma)}{P(x)}$$

$\Gamma(\gamma \mid \alpha_{prior}, \beta_{prior})$

$\Gamma(\gamma \mid \alpha_{posterior}, \beta_{posterior})$

$\Gamma(\gamma \mid \alpha_{prior} + 1/2,\ \beta_{prior} + (x - \mu)^2/2)$

In the context of a gaussian, the prior for the **precision** is a gamma !

4. Conjugate distributions

Usual distributions : Beta distribution

Beta distribution

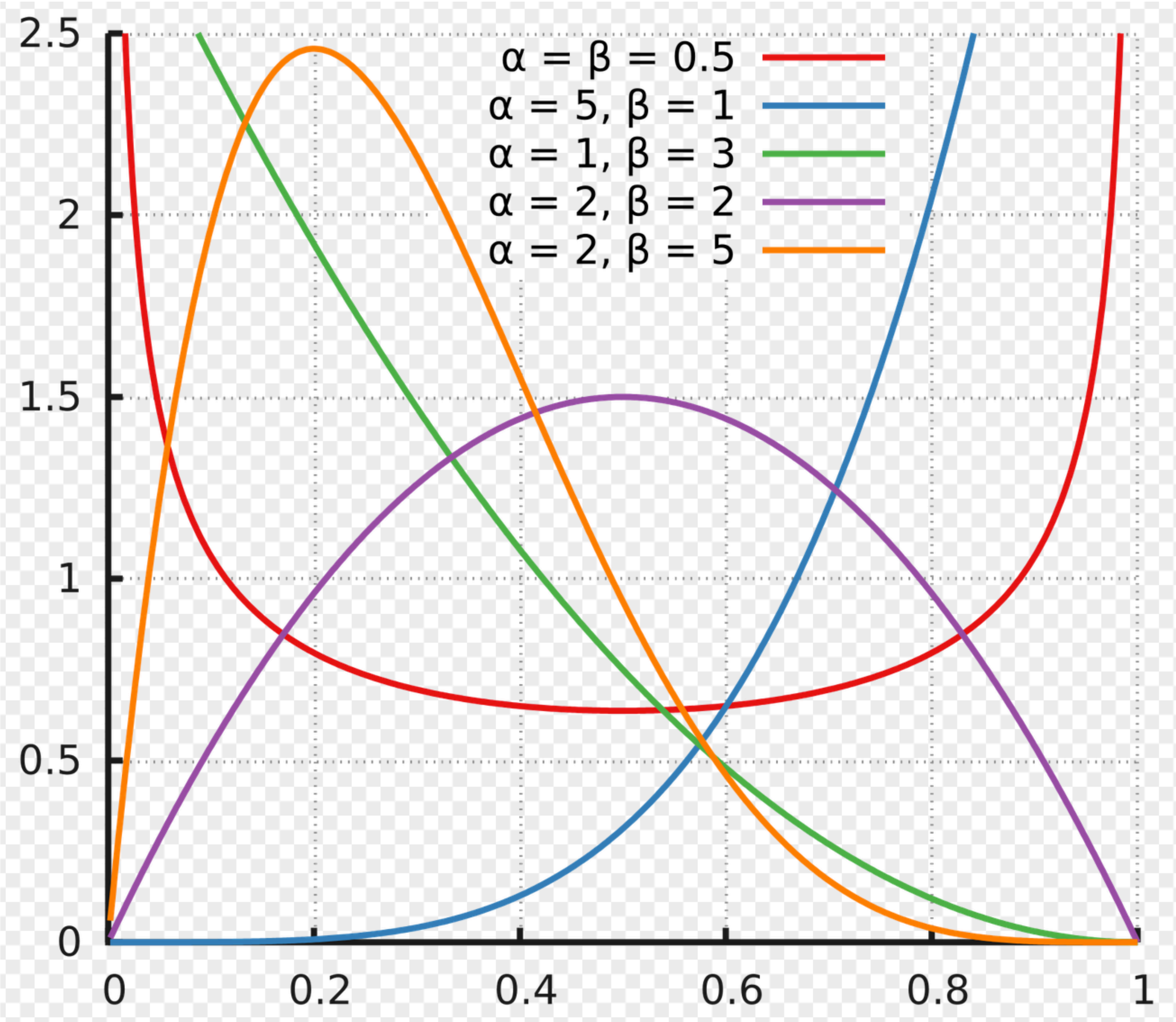
PDF : $$B(x\,|\,\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ with } \alpha,\beta > 0 \text{ and } x \in [0,1]$$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

mean : $$\mathbb{E}[x] = \frac{\alpha}{\alpha+\beta}$$

variance : $$V(x) = \frac{\alpha\beta}{(\alpha+\beta)^2\cdot(\alpha+\beta-1)}$$

mode : $$Mode\,[x] = \frac{\alpha-1}{\alpha+\beta-2}$$



Example

$$\underbrace{B(\theta\,|\,\alpha_{posterior},\beta_{posterior})}\ —\ P(\theta\,|\,x) = \frac{Ber(x\,|\,\theta)\times P(\theta)}{P(x)} \ —\ B(\theta\,|\,\alpha_{prior},\beta_{prior})$$

In the context of a Bernoulli distribution, the prior is a beta !

4. Conjugate distributions

Usual distributions : Beta distribution

Beta distribution

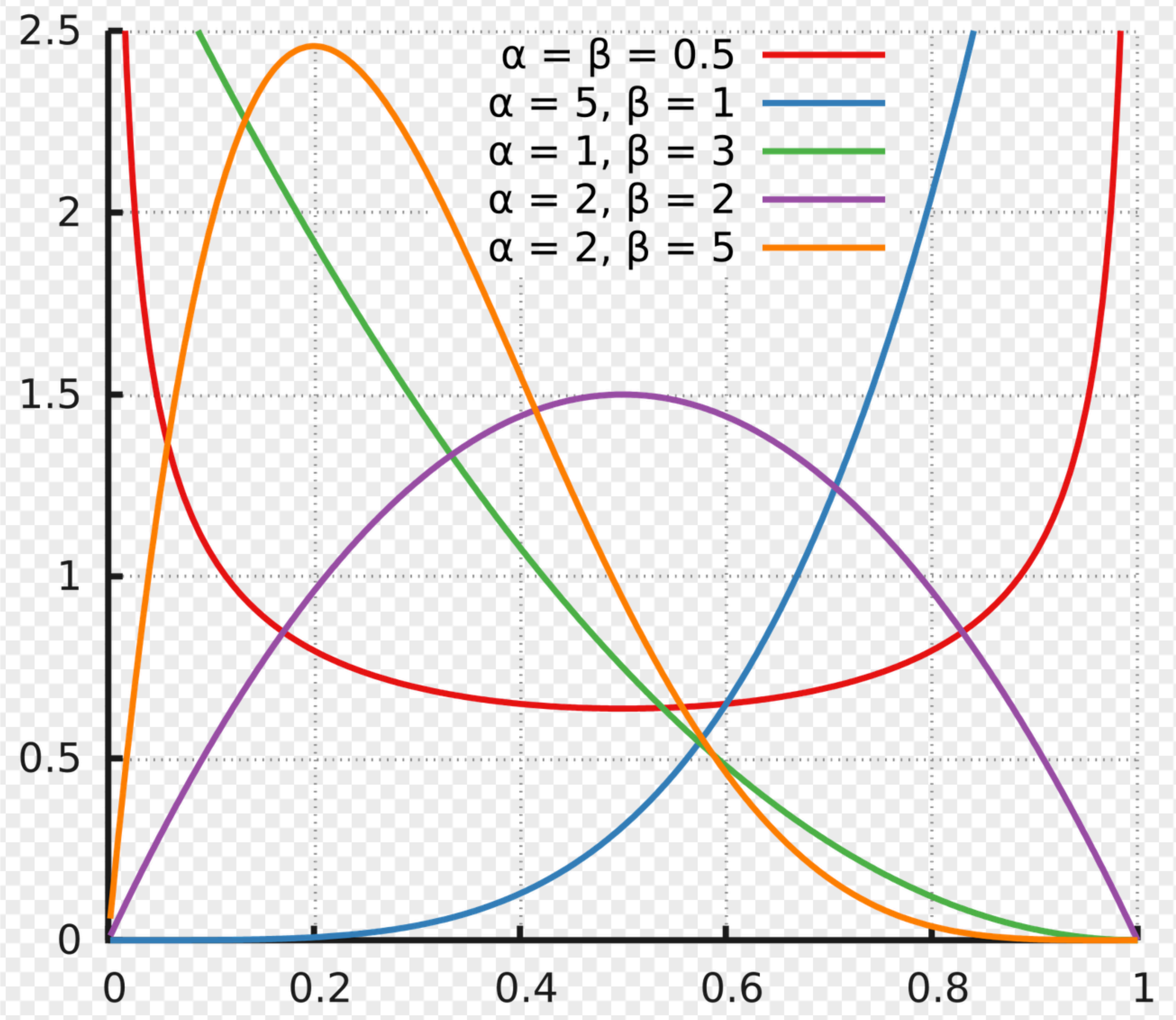
PDF : $B(x\,|\,\alpha,\beta) = \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ with $\alpha,\beta > 0$ and $x \in [0,1]$

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

mean : $\mathbb{E}[x] = \dfrac{\alpha}{\alpha+\beta}$

variance : $V(x) = \dfrac{\alpha\beta}{(\alpha+\beta)^2\cdot(\alpha+\beta-1)}$

mode : $Mode\,[x] = \dfrac{\alpha-1}{\alpha+\beta-2}$



Exercice (left as an exercice, correction in the next lecture)

$$\theta^{n_1}\cdot(1-\theta)^{n_0}$$

$$B(\theta\,|\,\alpha_{prior},\beta_{prior})$$

$$B(\theta\,|\,\alpha_{posterior},\beta_{posterior}) \qquad P(\theta\,|\,x) = \frac{Ber(x\,|\,\theta)\times P(\theta)}{P(x)}$$

$$B(\theta\,|\,n_1+\alpha_{prior}, n_0+\beta_{prior})$$

In the context of a Bernoulli distribution, the prior is a beta !

4. Conjugate distributions

Limitations

Posterior distribution

Fixed by **model**

Fixed by **us**

Likelihood Prior

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

Fixed by **data**

Remarks

- We have to avoid computing the evidence
- We can choose a convenient prior which enable us to compute the **posterior :**

  Conjugate prior

- It computes the exact posterior
- Easy for online learning

- For some (complex) models, the conjugate prior can be inadequate (improper prior)
- Can be unrealistic (non-informative prior)

! Road map

Bayesian statistics (03/05/21)

1

Bayesian perspective :

Likelihood

Prior distribution

$$P(\theta | X) = \frac{P(X, \theta)}{P(X)} = \frac{P(X | \theta) \cdot P(\theta)}{P(X)}$$

Posterior distribution

Evidence

Hard to compute !

θ parameters

X observations

Exemple :
Naive Bayes classifier,
Linear regression, ….

MAP : $\arg\max_{\theta} P(X | \theta) \cdot P(\theta)$

Conjugate distribution

Pros :	Cons :
- exact posterior	- conjugate prior maybe inadequate

Latent variable models (17/05/21)

2

Variational Inference (31/05/21)

3

Markov Chain Monte Carlo (07/06/21)

4

Extensions (14/06/21)

5