ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Bayesian Machine Learning

May 2024 - François HU
https://curiousml.github.io/

Outline

1. **Bayesian statistics**
   - Bayesian statistics and probabilistic model
   - Analytical inference
   - Conjugate priors
2. **Latent Variable Models**
3. **Variational Inference**
4. **Causal Inference**
5. **Extensions and oral presentations**

PREREQUISITE

THEORY

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

APPLICATION

Python

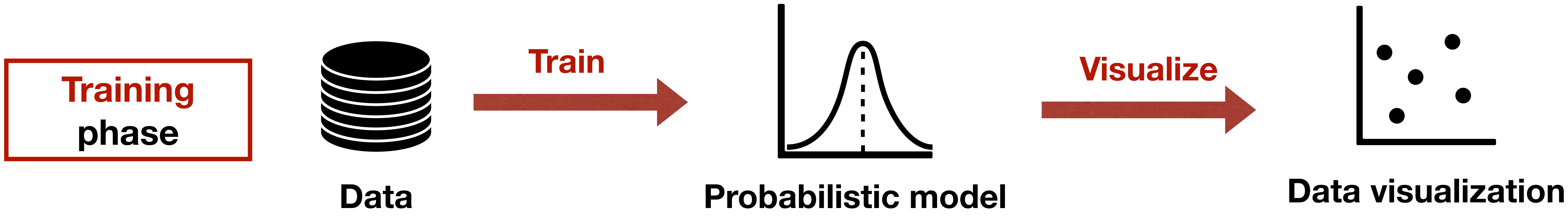
ALGORITHM

Some « classical » supervised & unsupervised models

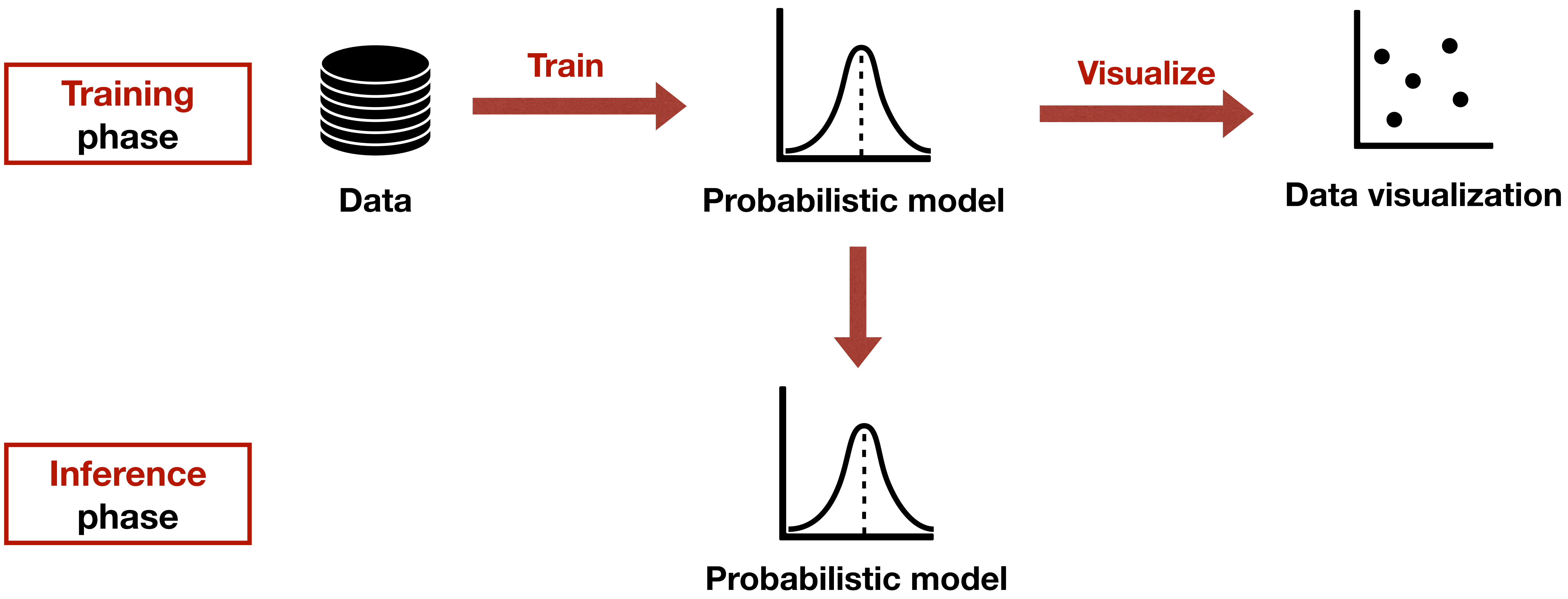
0

Gentle introduction to statistical learning

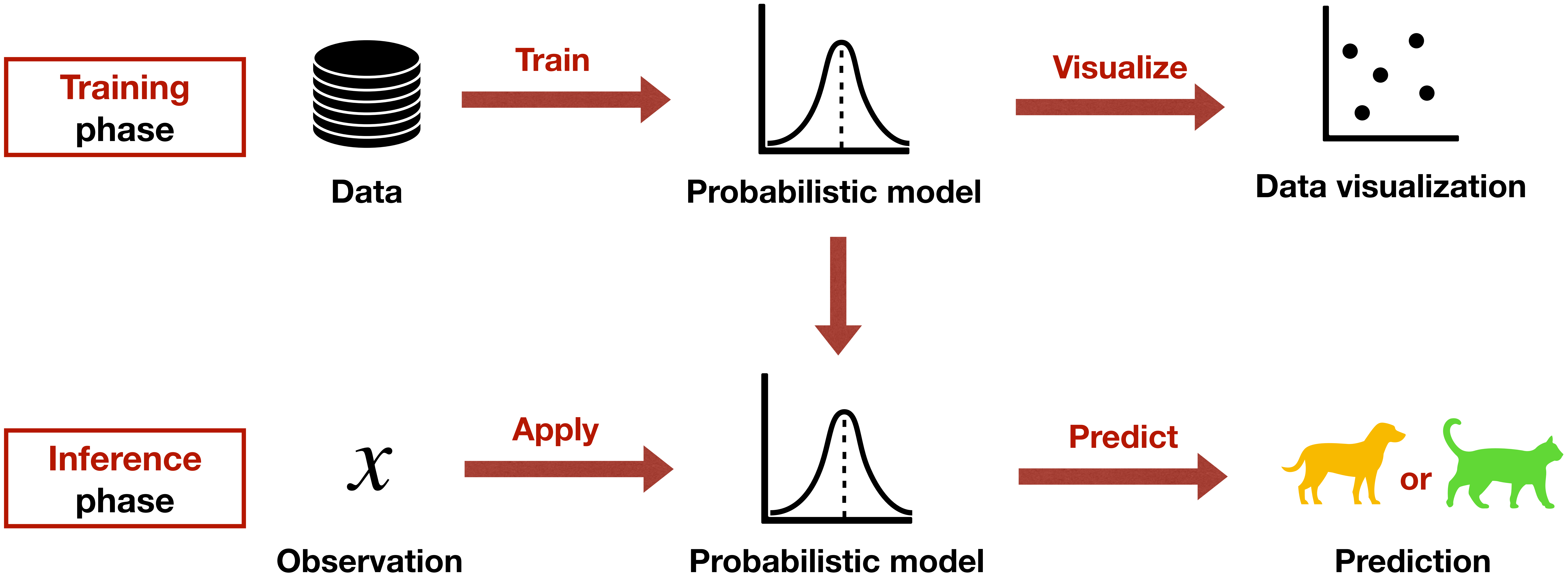
Simplified statistical learning process



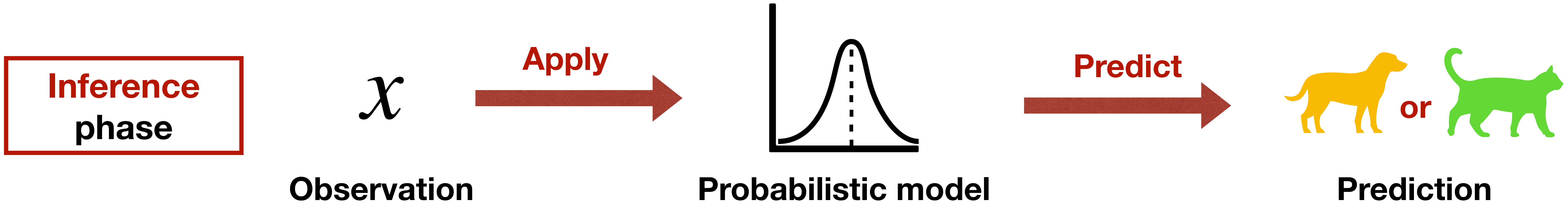
Simplified statistical learning process



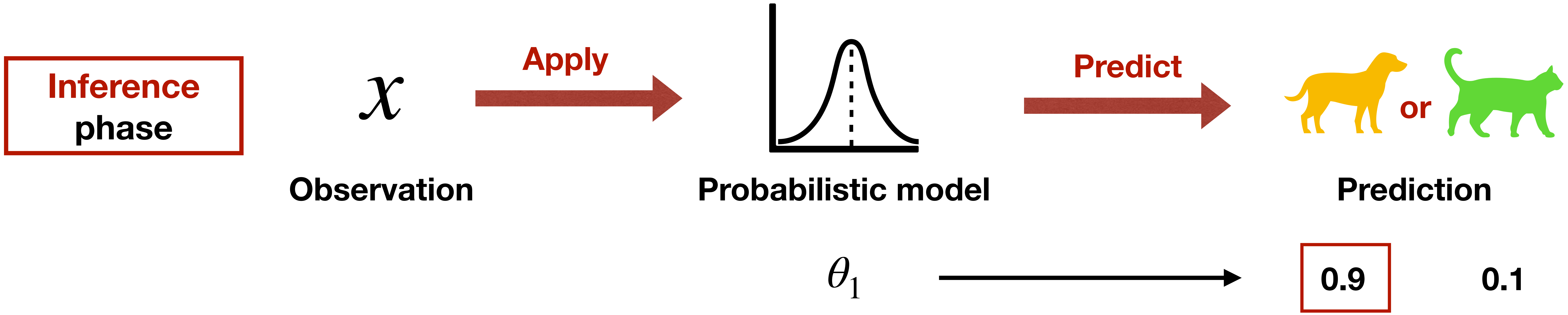
Simplified statistical learning process



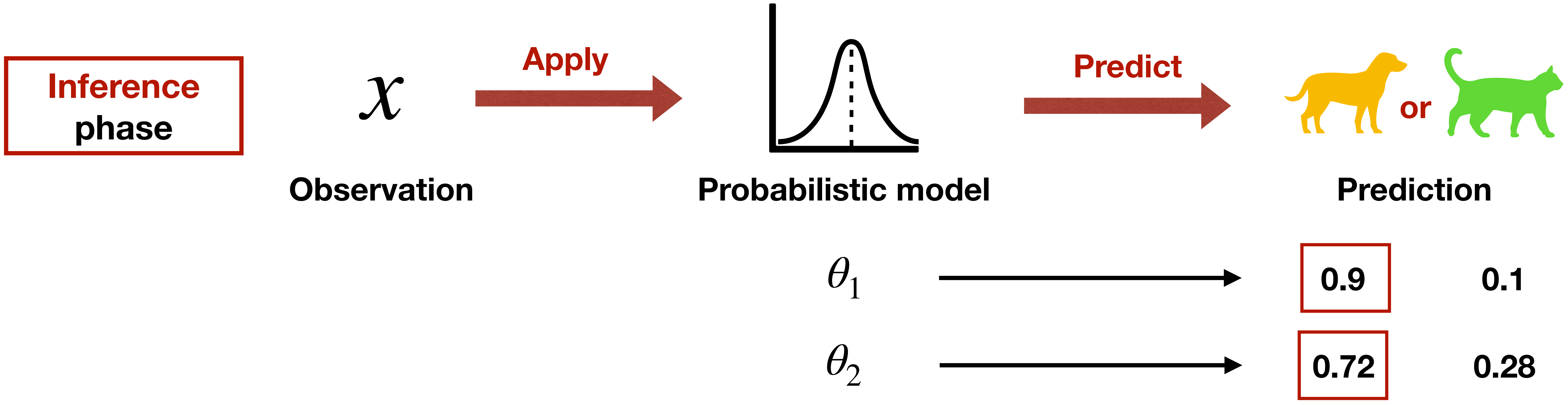
Inference phase



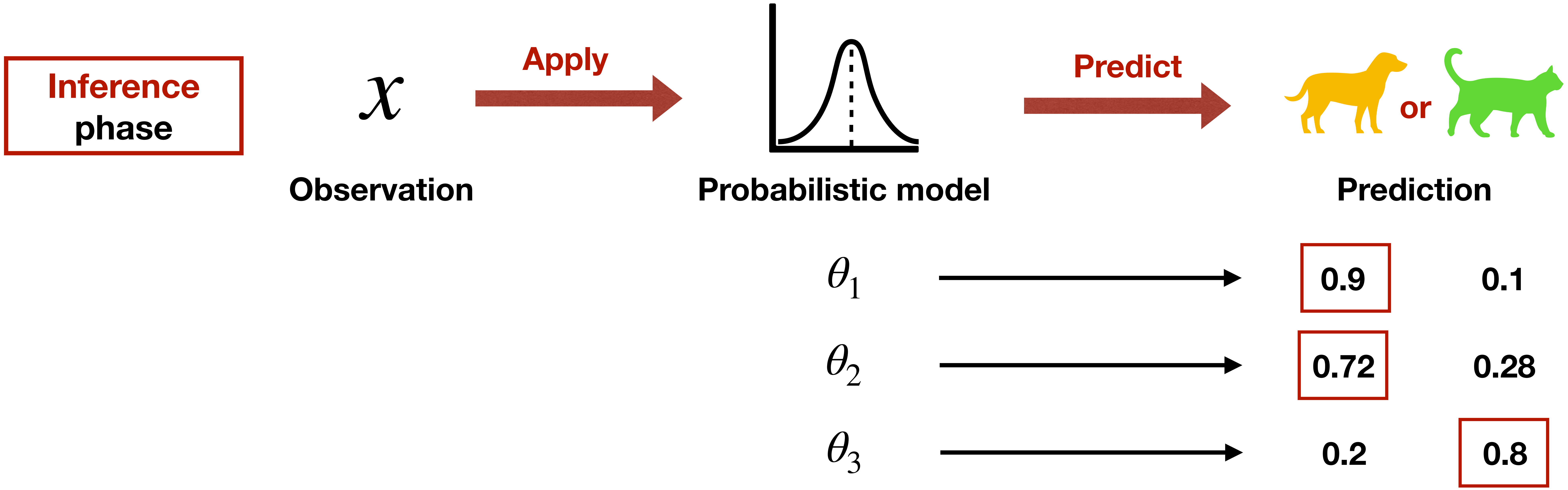
Inference phase



Inference phase



Inference phase



Inference phase

Aim during training phase : find

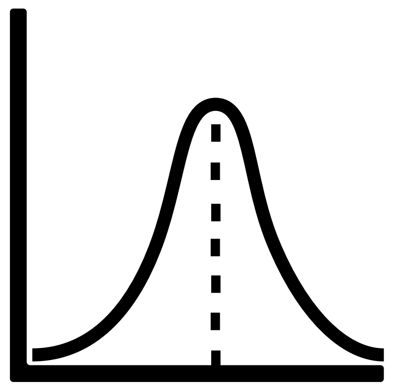
θ **such that** $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

Inference phase

x

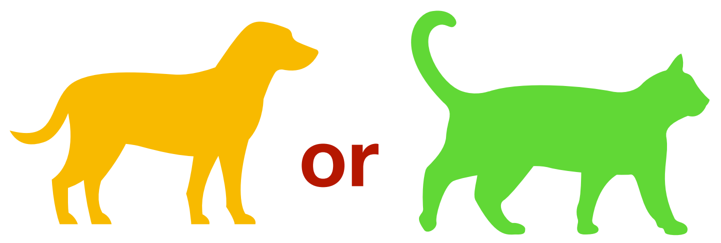
Observation

Apply $\longrightarrow$



Probabilistic model

Predict $\longrightarrow$



Prediction

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

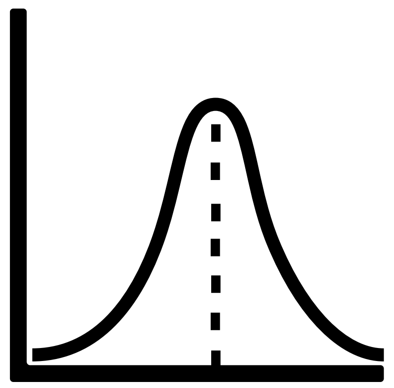
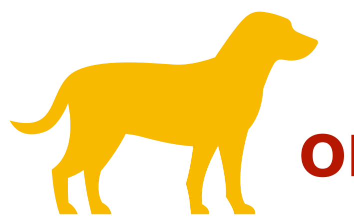
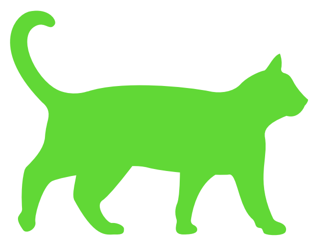
…

Inference phase

Aim during **training phase** : find

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

Inference phase

x → **Apply** →  → **Predict** →  **or** 

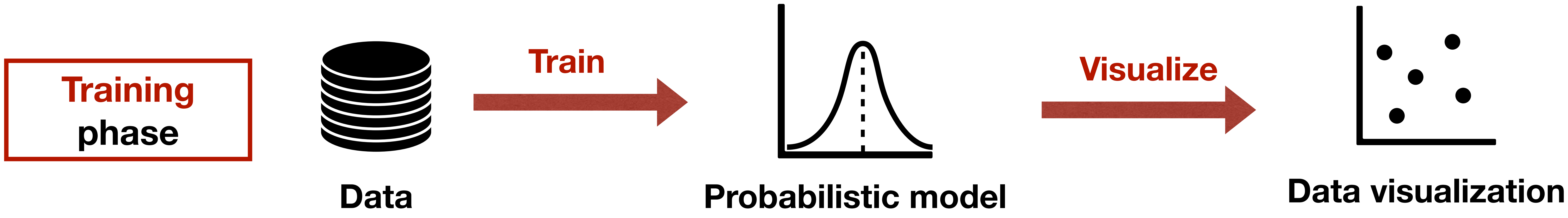
Observation | **Probabilistic model** | **Prediction**

θ_1 →	$\boxed{0.9}$	0.1
θ_2 →	$\boxed{0.72}$	0.28
θ_3 →	0.2	$\boxed{0.8}$

...

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

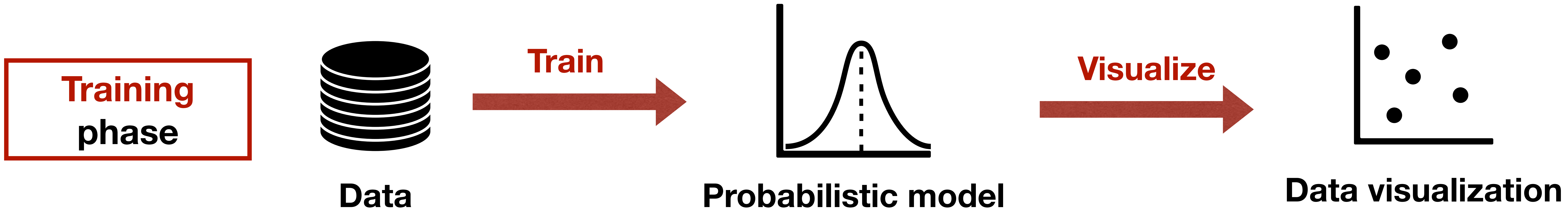
Training phase



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

Training phase

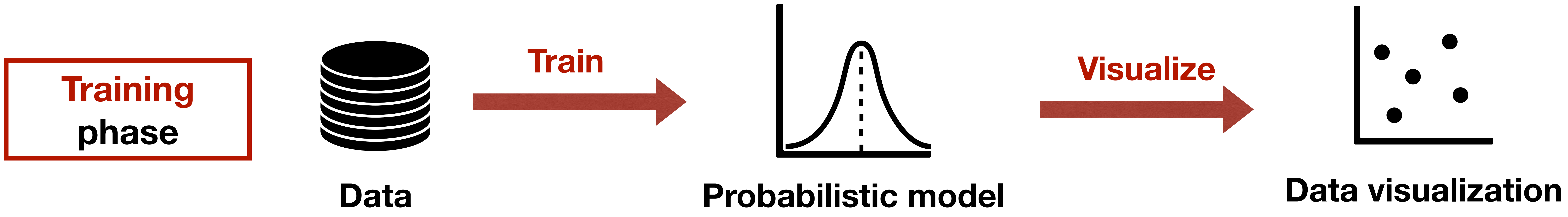


Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

Example : Linear regression example (proof later in the course if needed)

Training phase



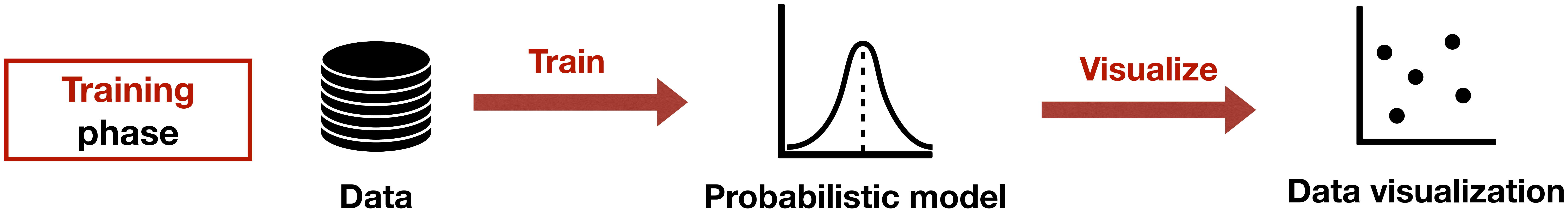
Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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

Training phase



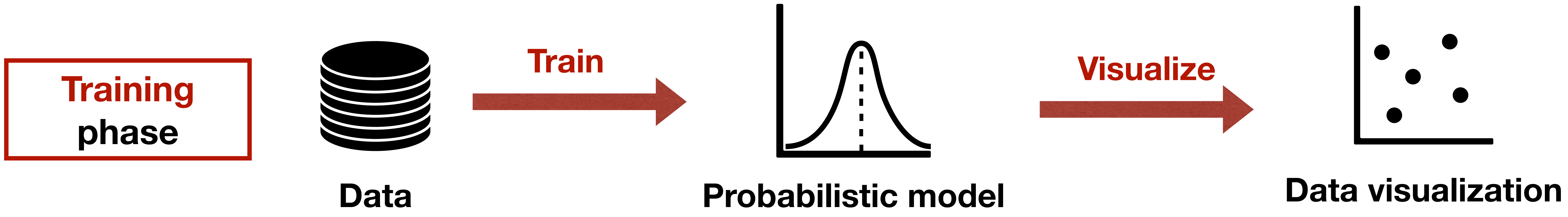
Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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

Training phase



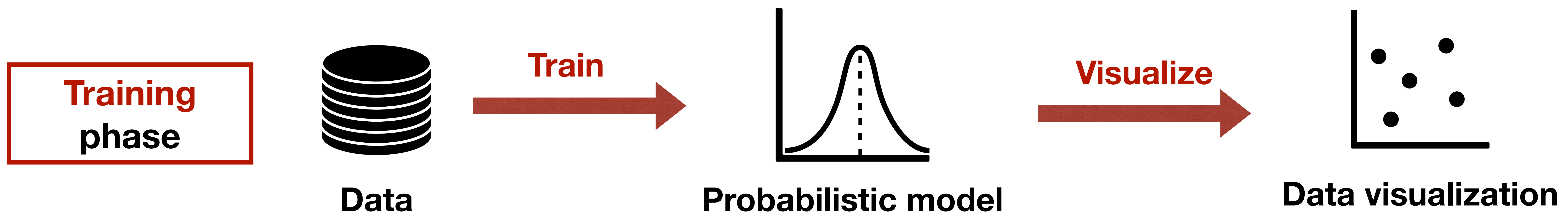
Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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

Training phase



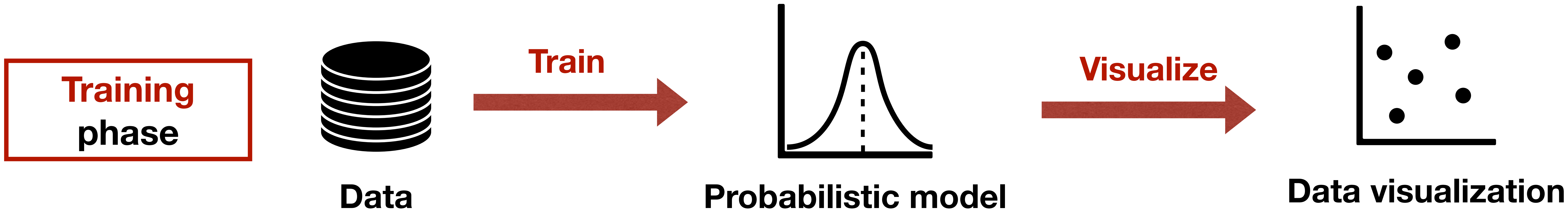
Aim during training phase : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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

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

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

SOLUTION : Bayesian statistics

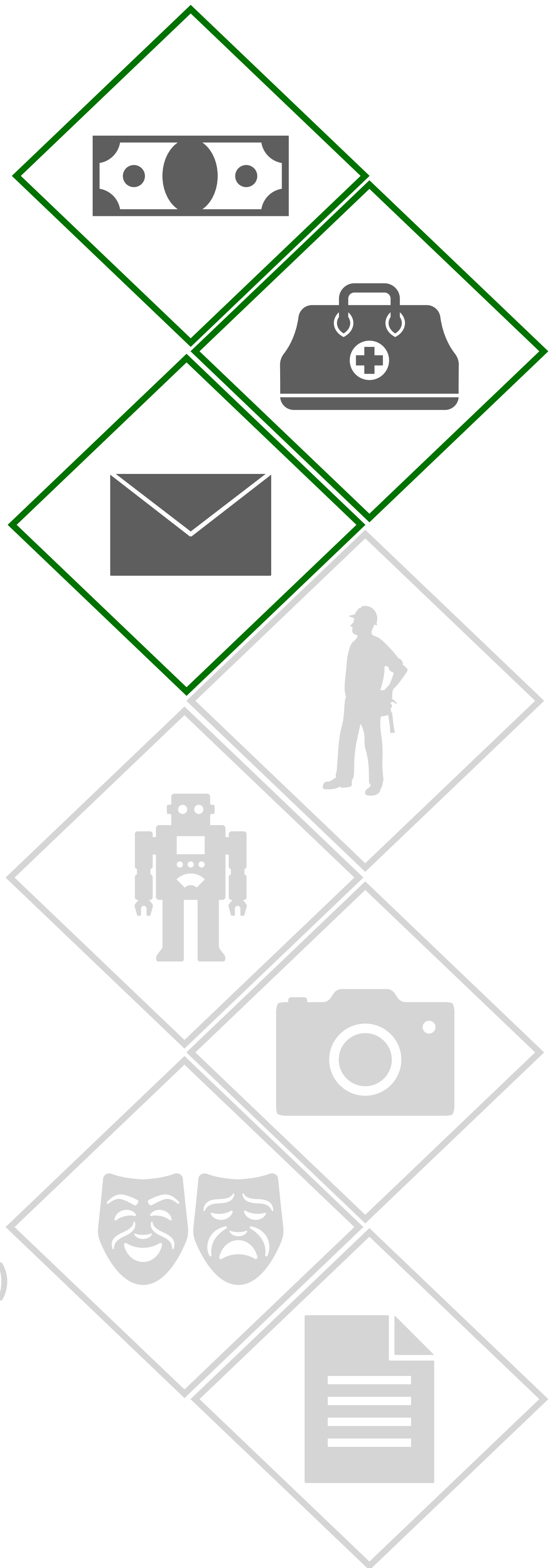
Why Bayesian methods ?

Examples of application of Bayesian Machine Learning

Supervised machine learning

Unsupervised machine learning

Others



Credit card default detection
Bayes theorem

Medical diagnosis
Bayes theorem

Spam filtering
Bayesian Neural Network

Patterns in customer dataset
Bayesian Non parametric Clustering (BNC)

Help robots make decision
Bayesian Reinforcement Learning (BRL)

Reconstructing images from noisy images
Bayes theorem + MCMC

Speech emotion recognition
Nonparametric hierarchical neural network (NHNN)

Optimal character recognition (OCR)

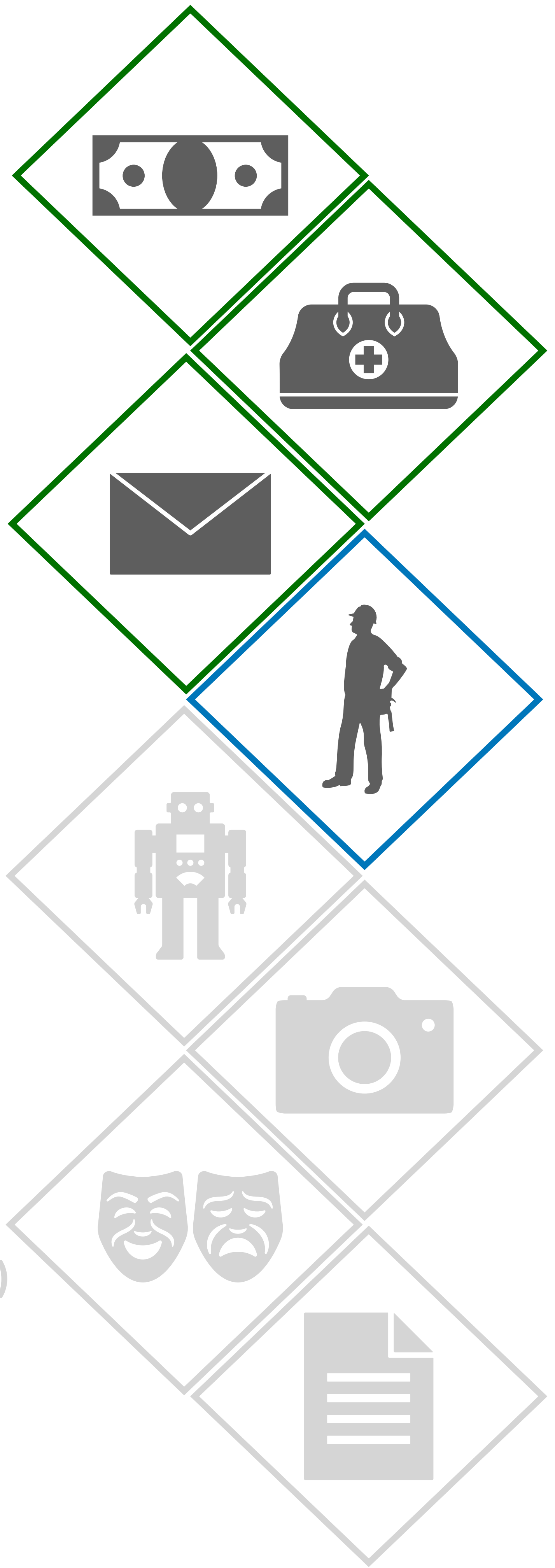
Why Bayesian methods ?

Supervised machine learning

Unsupervised machine learning

Others

Examples of application of Bayesian Machine Learning



Credit card default detection
Bayes theorem

Medical diagnosis
Bayes theorem

Spam filtering
Bayesian Neural Network

Patterns in customer dataset
Bayesian Non parametric Clustering (BNC)

Help robots make decision
Bayesian Reinforcement Learning (BRL)

Reconstructing images from noisy images
Bayes theorem + MCMC

Speech emotion recognition
Nonparametric hierarchical neural network (NHNN)

Optimal character recognition (OCR)

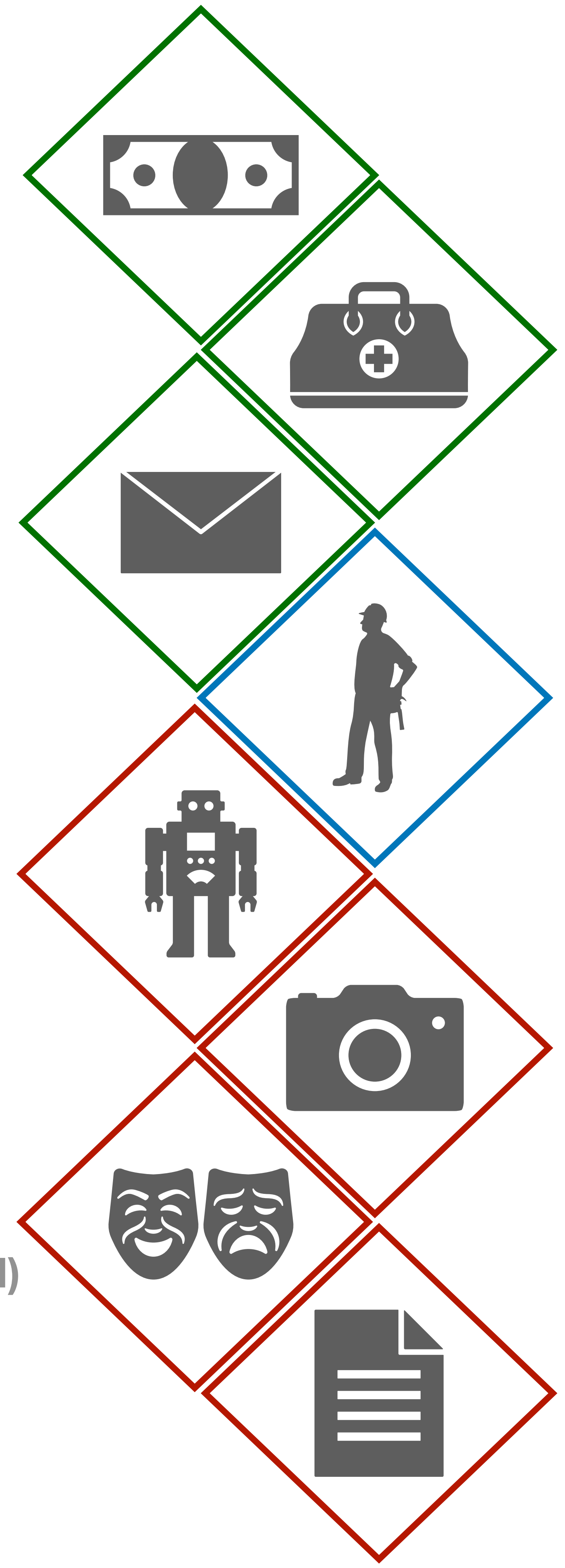
Why Bayesian methods ?

Supervised machine learning

Unsupervised machine learning

Others using more advanced techniques

Examples of application of Bayesian Machine Learning



Credit card default detection
Bayes theorem

Medical diagnosis
Bayes theorem

Spam filtering
Bayesian Neural Network

Patterns in customer dataset
Bayesian Non parametric Clustering (BNC)

Help robots make decision
Bayesian Reinforcement Learning (BRL)

Reconstructing images from noisy images
Bayes theorem + MCMC

Speech emotion recognition
Nonparametric hierarchical neural network (NHNN)

Optimal character recognition (OCR)

1 Introduction to bayesian statistics

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

Relative **frequency** of an event in an infinite trials

$$P(K\heartsuit) = \frac{1}{52}$$

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

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

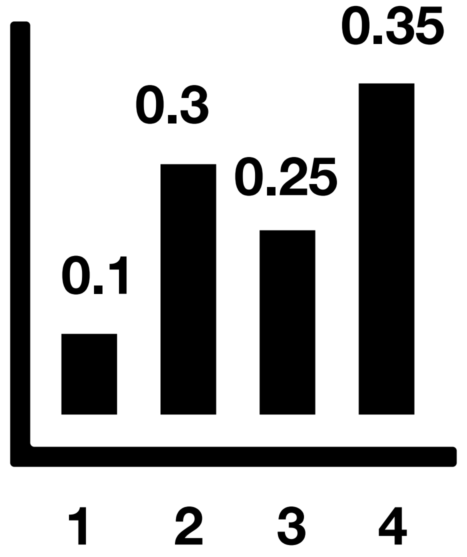
Relative **frequency** of an event in an infinite trials

$$P(K\heartsuit) = \frac{1}{52}$$

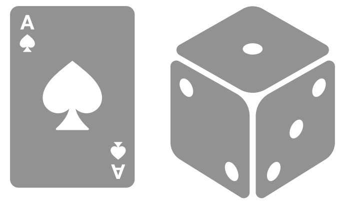
$$P(\spadesuit) = \frac{1}{4}$$

Random variable

Discrete variable

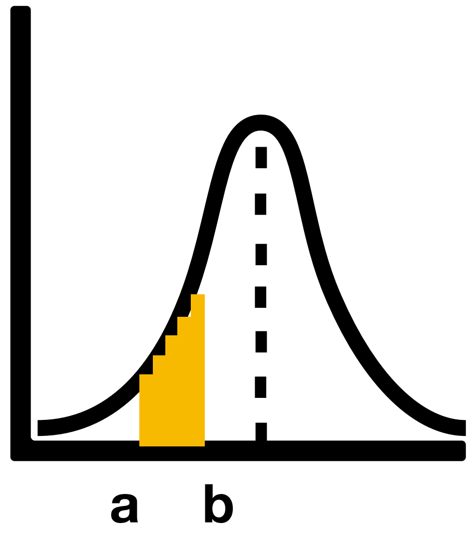


Probability Mass Function (PMF)

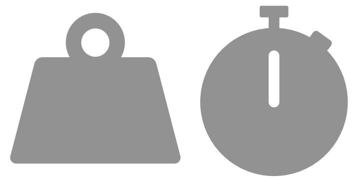


$$P(X) = \begin{cases} 0.1 & \text{if } X = 1 \\ 0.3 & \text{if } X = 2 \\ 0.25 & \text{if } X = 3 \\ 0.35 & \text{if } X = 4 \end{cases}$$

Continuous variable



Probability Density function (PDF)



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

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

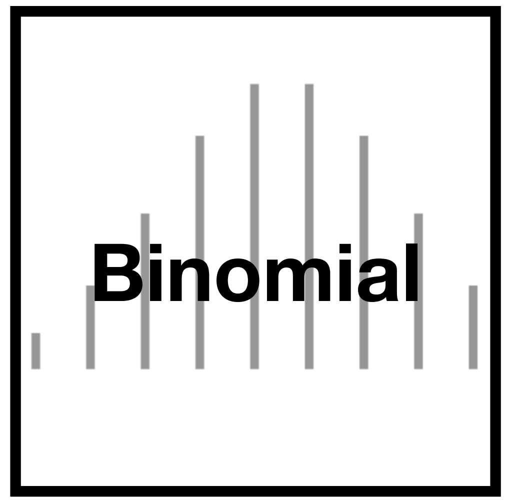
Relative **frequency** of an event in an infinite trials

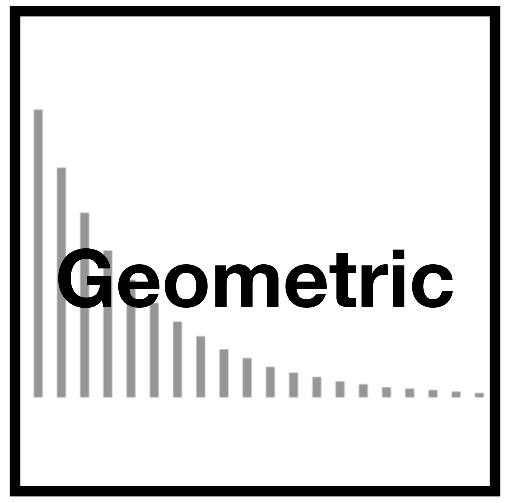
$P(K\heartsuit) = \frac{1}{52}$

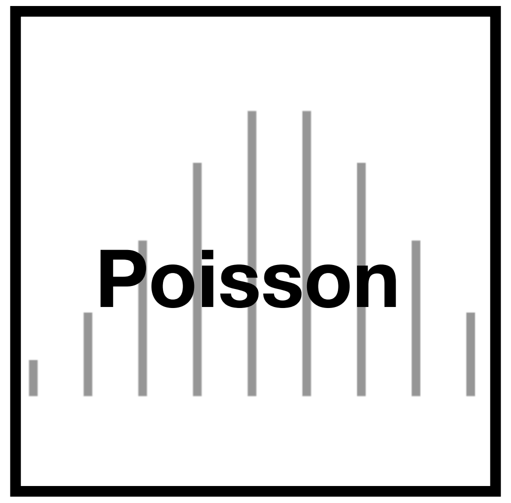
$P(\spadesuit) = \frac{1}{4}$

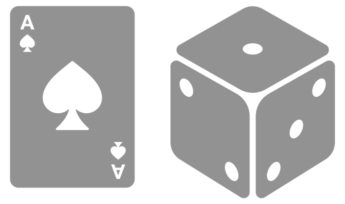
Random variable

Discrete variable



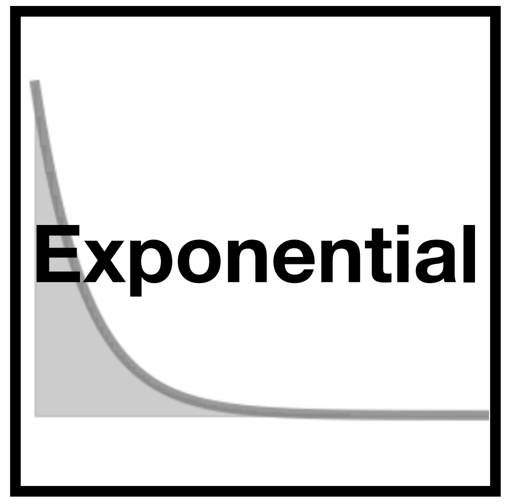


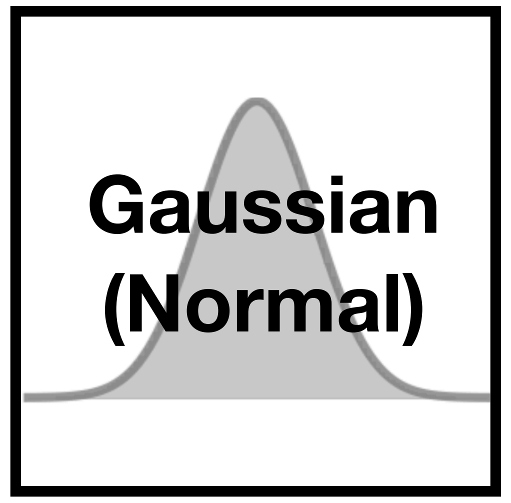


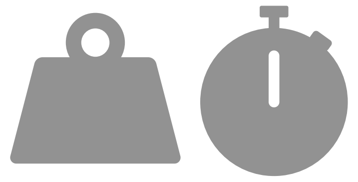


Usual distributions

Continuous variable







Usual distributions

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

Relative **frequency** of an event in an infinite trials

$$P(K\heartsuit) = \frac{1}{52}$$

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

Independence

Two random variables X and Y are **independent** if

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

joint probability **marginals**

dependency : **one** dice

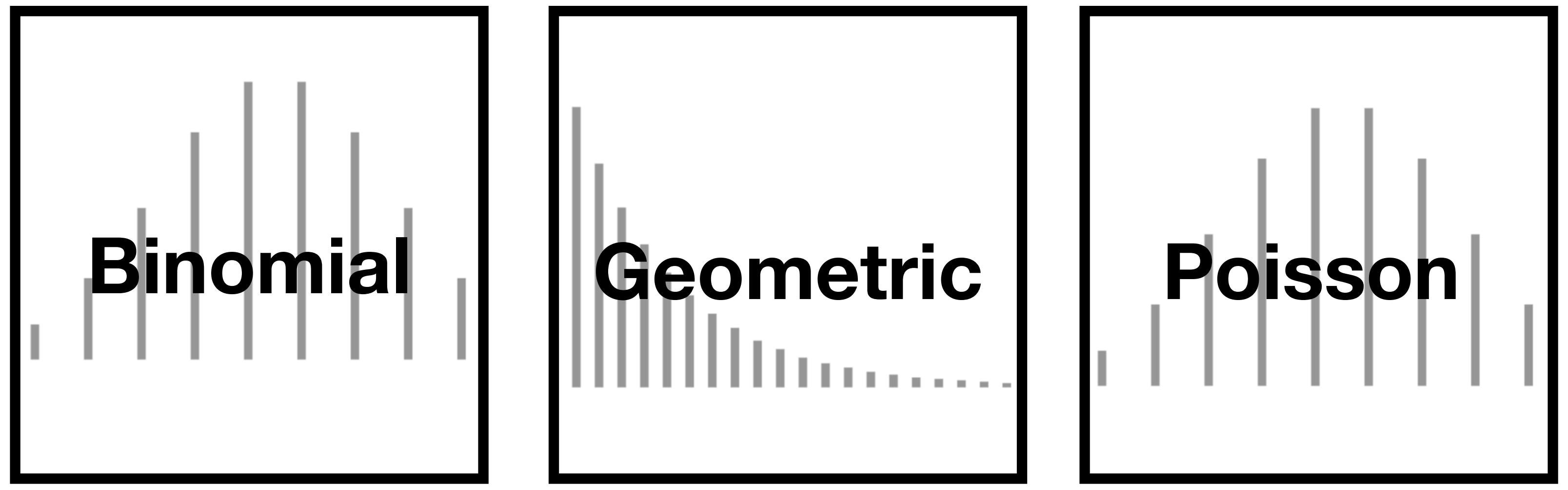
$$P(⚃,⚃) = 0 \neq P(⚃)P(⚃) = 1/6^2$$

independency : **two** dices

$$P(⚃,⚃) = P(⚃)P(⚃) = 1/6^2$$

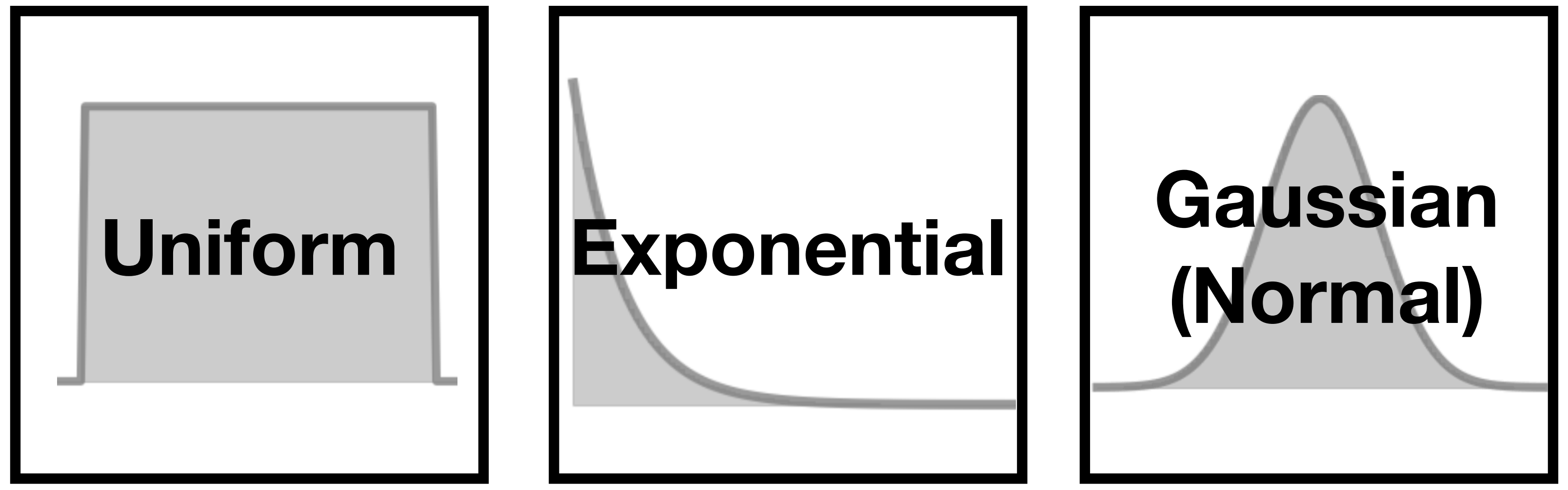
Random variable

Discrete variable



Usual distributions

Continuous variable



Usual distributions

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

Relative **frequency** of an event in an infinite trials

$P(\text{K}\heartsuit) = \frac{1}{52}$

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

Independence

Two random variables X and Y are **independent** if

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

joint probability — **marginals**

dependency : **one** dice

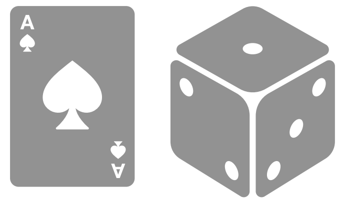
$$P(⚃,⚃) = 0 \neq P(⚃)P(⚃) = 1/6^2$$

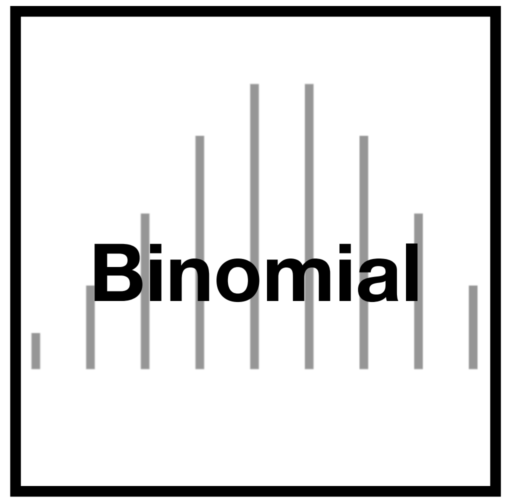
independency : **two** dices

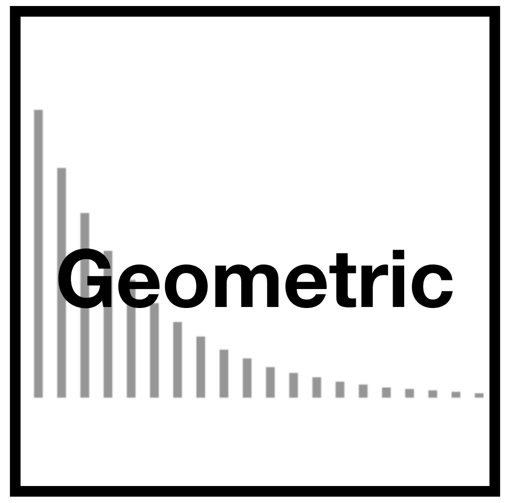
$$P(⚃,⚃) = P(⚃)P(⚃) = 1/6^2$$

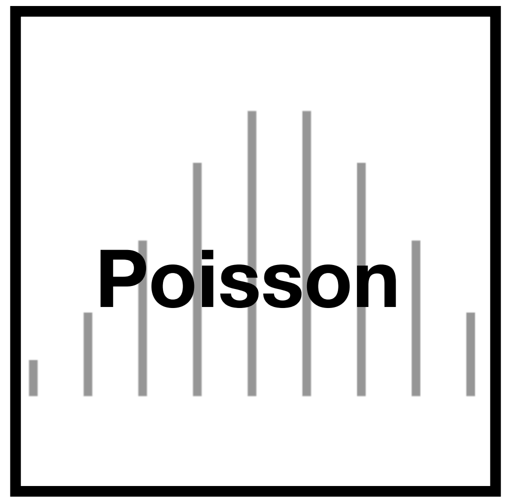
Random variable

Discrete variable



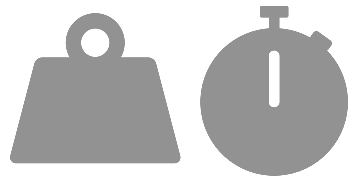


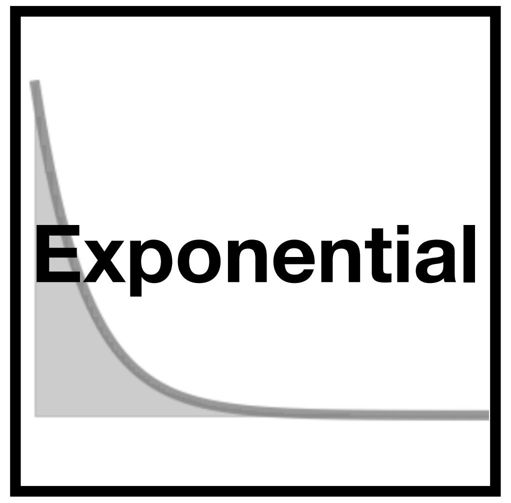


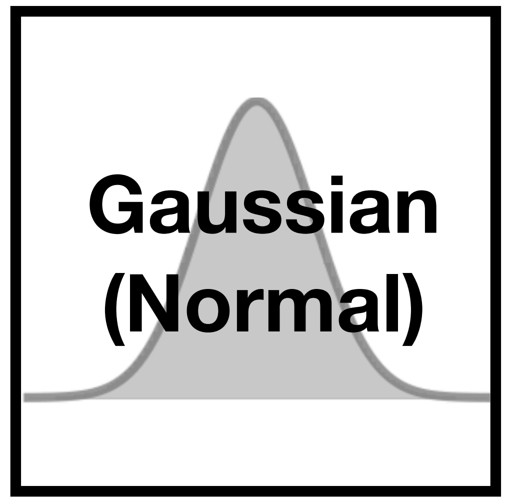


Usual distributions

Continuous variable







Usual distributions

Conditional probability

probability of X **given that Y happened**

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

joint probability — **marginal**

$$P(\text{A}\heartsuit \,|\, \text{A}\clubsuit) = \frac{P(\text{A}\heartsuit, \text{A}\clubsuit)}{P(\text{A}\clubsuit)}$$

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

$$\underbrace{P(X|Y)}_{\text{Conditional}} = \frac{\overbrace{P(X,Y)}^{\text{Joint probability}}}{\underbrace{P(Y)}_{\text{Marginal}}}$$

Chain rule

$$P(X_1, X_2) = \cdots$$

$$P(X_1, X_2, X_3) = \cdots$$

$$P(X_1, \ldots, X_n) = \cdots$$

Sum rule

discrete

$$P(X) = \cdots$$

continuous

$$P(X) = \cdots$$

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

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

Conditional — Joint probability — Marginal

Chain rule

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

$$P(X_1, X_2, X_3) = \cdots$$

$$P(X_1, \ldots, X_n) = \cdots$$

Sum rule

discrete

$$P(X) = \cdots$$

continuous

$$P(X) = \cdots$$

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

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

Joint probability

Conditional

Marginal

Chain rule

$$P(X_1, X_2) = P(X_1 \mid X_2) \times P(X_2)$$

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

$$P(X_1, \ldots, X_n) = \cdots$$

Sum rule

discrete

$$P(X) = \cdots$$

continuous

$$P(X) = \cdots$$

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

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

Conditional — Joint probability — Marginal

Chain rule

$$P(X_1, X_2) = P(X_1 \mid X_2) \times P(X_2)$$

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

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

Sum rule

discrete

$$P(X) = \cdots$$

continuous

$$P(X) = \cdots$$

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

Joint probability

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

Conditional **Marginal**

Chain rule

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

$$P(X_1, X_2, X_3) = P(X_1|X_2, X_3) \times P(X_2|X_3) \times P(X_3)$$

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

Sum rule

discrete **continuous**

$$P(X) = \sum_{Y \in \mathcal{Y}} P(X, Y) \qquad P(X) = \cdots$$

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

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

Joint probability

Conditional

Marginal

Chain rule

$$P(X_1, X_2) = P(X_1 \mid X_2) \times P(X_2)$$

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

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

Sum rule

discrete

$$P(X) = \sum_{Y \in \mathcal{Y}} P(X, Y)$$

continuous

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

1. Introduction to bayesian statistics

Probability & statistics : Bayes theorem

Conditional probability

probability of X **given that Y happened**

$$\underbrace{P(X|Y)}_{\text{Conditional}} = \frac{\overbrace{P(X,Y)}^{\text{Joint probability}}}{\underbrace{P(Y)}_{\text{Marginal}}}$$

Chain rule

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

$$P(X_1, X_2, X_3) = P(X_1 | X_2, X_3) \times P(X_2 | X_3) \times P(X_3)$$

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

Sum rule

discrete

$$P(X) = \sum_{Y \in \mathcal{Y}} P(X, Y)$$

continuous

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

Bayes theorem

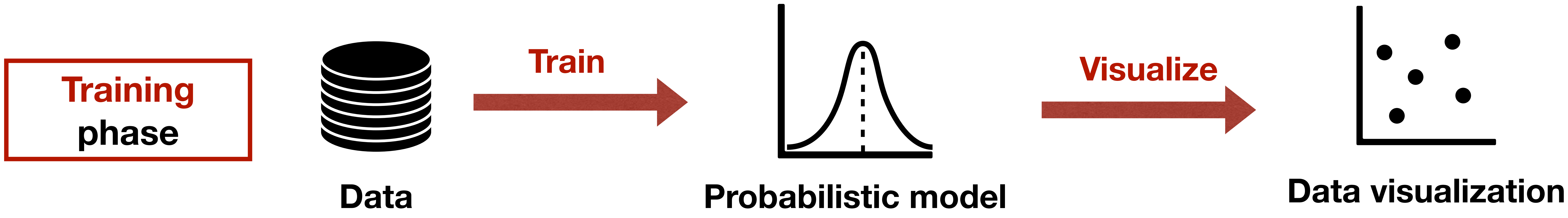
θ Parameters

X Data

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

1. Introduction to bayesian statistics

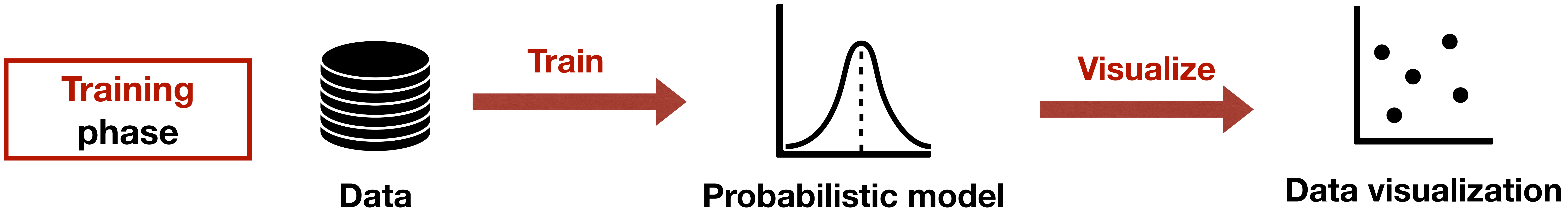
Frequentist VS Bayesian point of view



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



Frequentist

θ is fixed
X is random

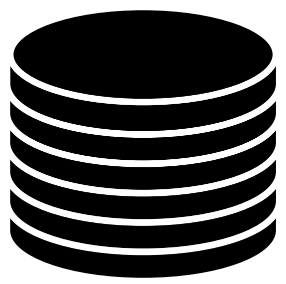
Maximum Likelihood Estimation

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

1. Introduction to bayesian statistics

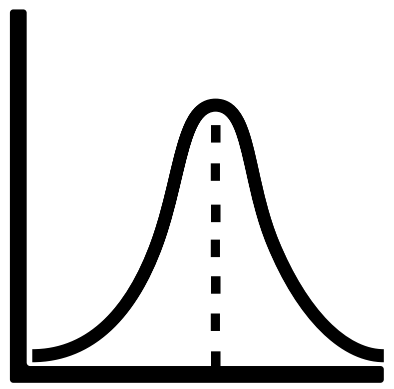
Frequentist VS Bayesian point of view

Training phase



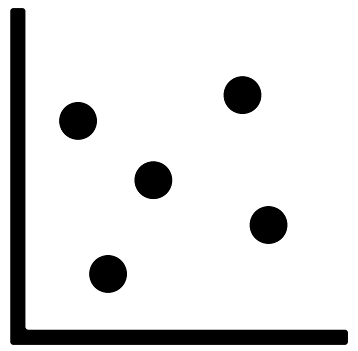
Data

Train



Probabilistic model

Visualize



Data visualization

Frequentist

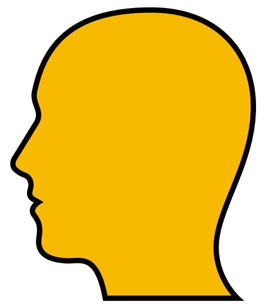
θ is fixed
X is random

Maximum Likelihood Estimation

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

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

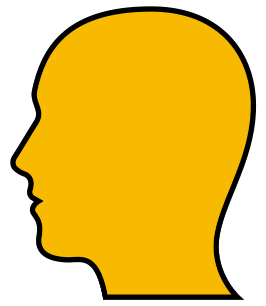
Frequentist

Maximum Likelihood Estimation

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

Bayes theorem

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



Bayesian

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view

Frequentist

Maximum Likelihood Estimation

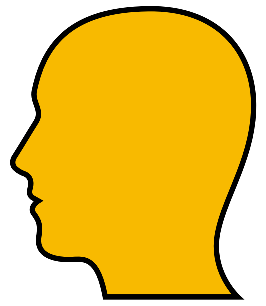
$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)$$

Problems in frequentist estimation :

- 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

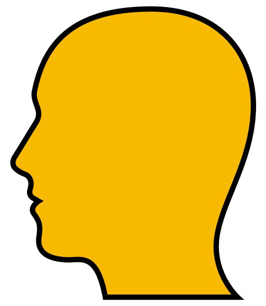
$$\hat{\theta} = \arg\max_{\theta} P(X\,|\,\theta)$$

Problems in frequentist estimation :

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

Bayes theorem

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



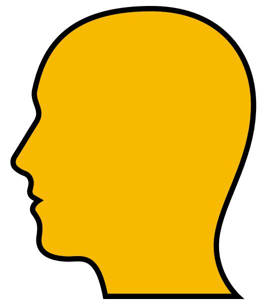
Bayesian

1. Introduction to bayesian statistics

Bayesian point of view : classification

Bayes theorem

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



Bayesian

Training phase

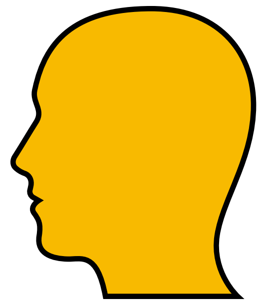
$$P(\theta | X_{train}, y_{train}) = \frac{P(y_{train} | X_{train}, \theta) \times P(\theta)}{P(y_{train} | X_{train})}$$

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

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

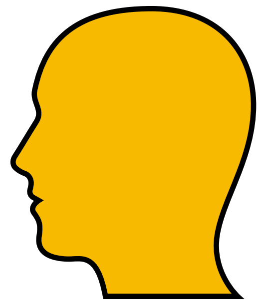
Can regularize your model when training on your data

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

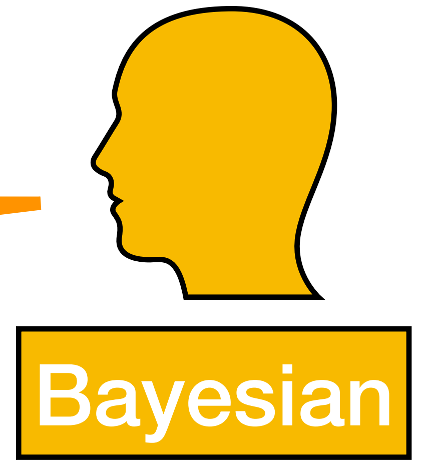
$$P(y_{new}|X_{new}, X_{train}, y_{train}) = \int P(y_{new}|X_{train}, \theta) \times P(\theta|X_{train}, y_{train})d\theta$$

1. Introduction to bayesian statistics

Bayesian point of view : online learning

Bayes theorem

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



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

Nodes : random variables

Links : probabilistic relationships

Bayesian networks
(**Directed** graphical models)

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

Bayesian networks
(**Directed** graphical models)

Markov random fields
(**Undirected** 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)

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)

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

The focus of our course !

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

Bayesian networks
(**Directed** graphical models)

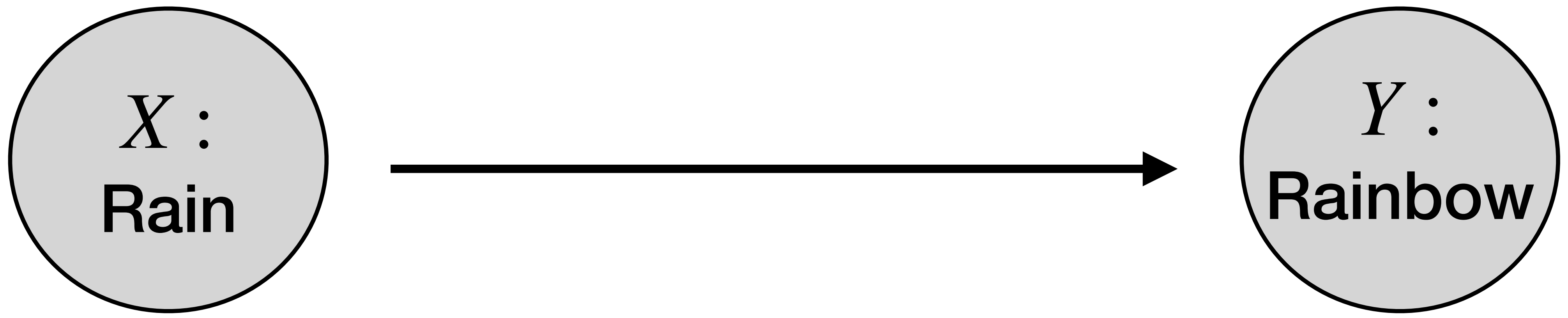
The focus of our course !

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

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

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

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

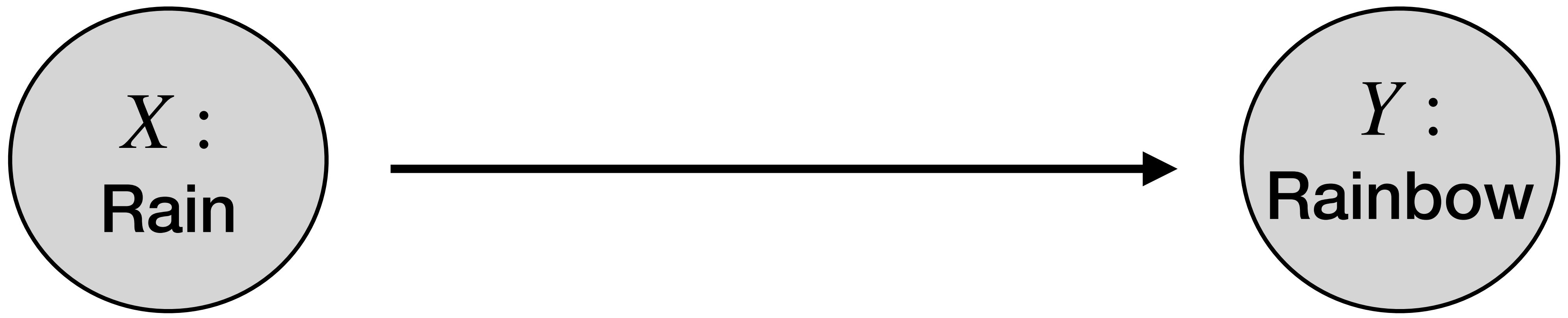
Bayesian networks
(**Directed** graphical models)

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 :



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

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

Bayesian networks
(**Directed** graphical models)

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

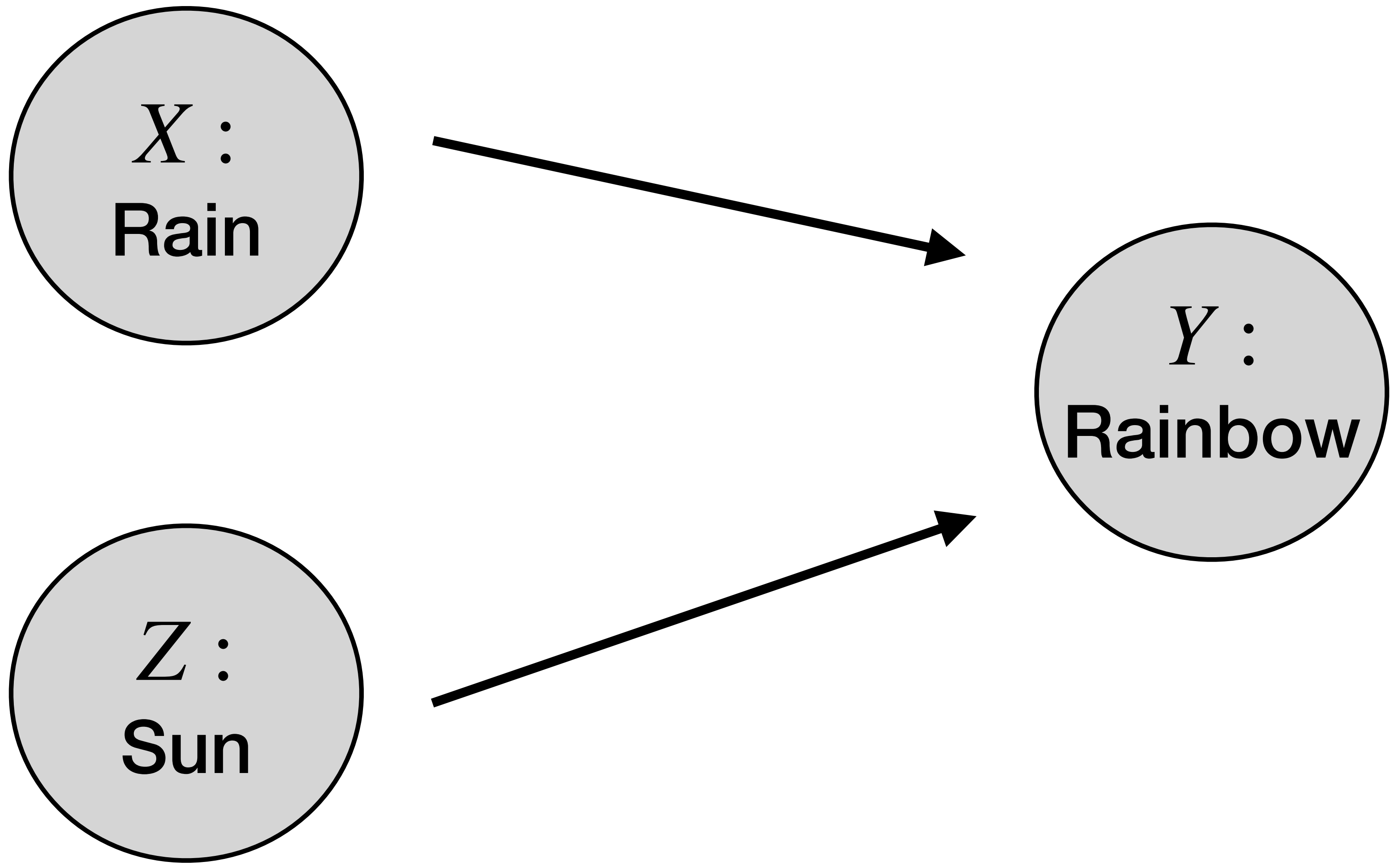
Nodes : random variables

Links : probabilistic relationships

The focus of our course !

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 !

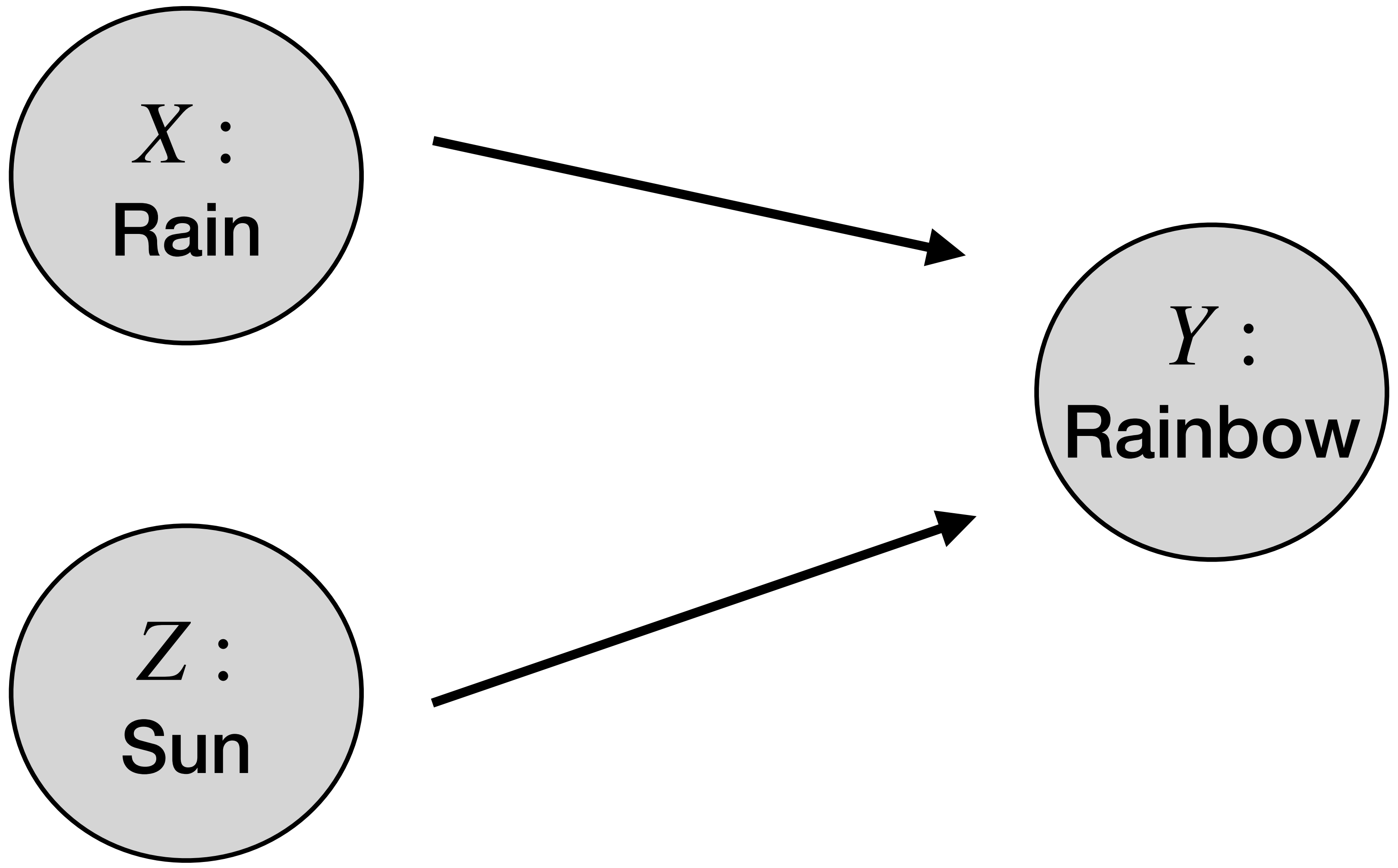
Markov random fields
(**Undirected** graphical models)

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)

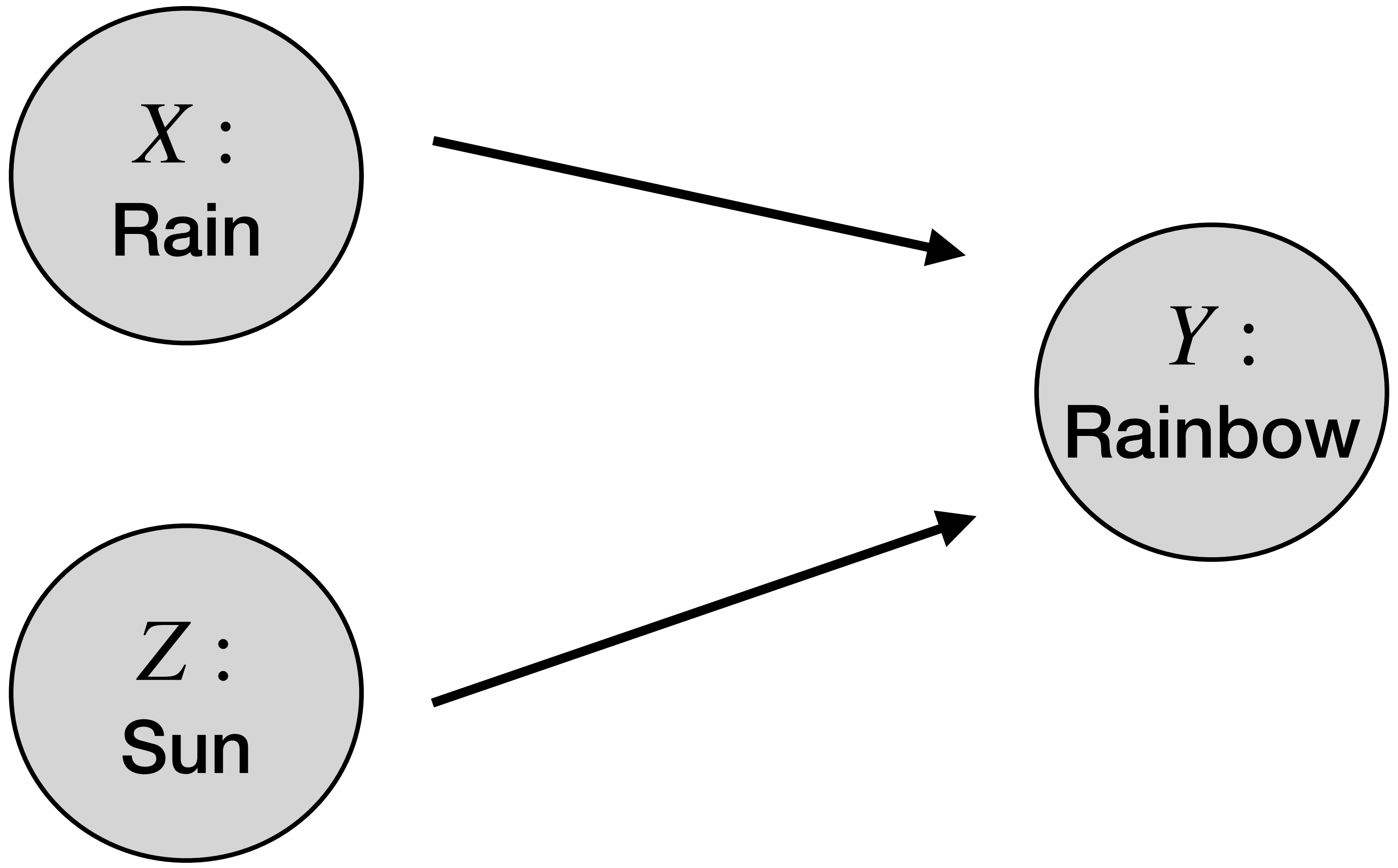
The focus of our course !

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)

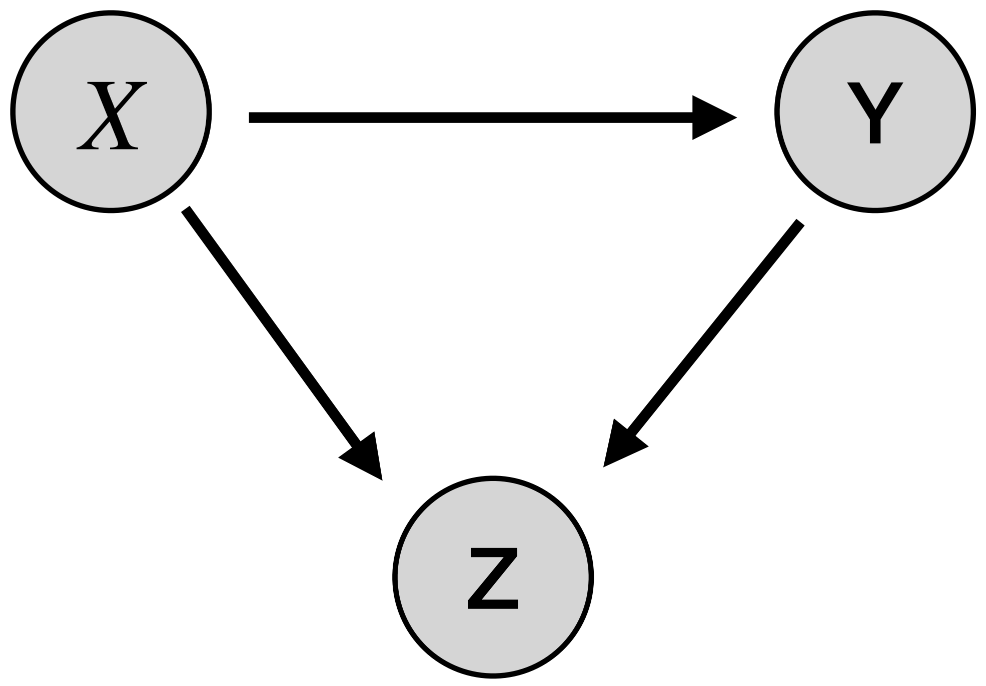
The focus of our course !

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

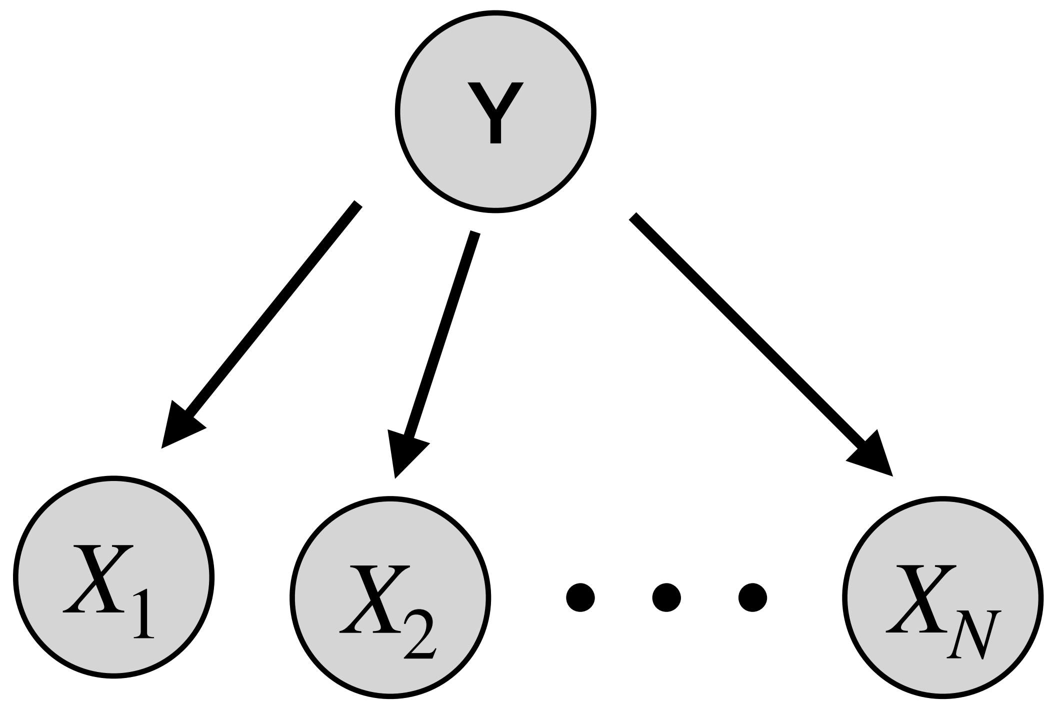
Model : joint probability over all variables $P(X_1, \cdots, X_N) = \prod_{i=1,\cdots,N} P(\, X_i \mid parents(X_i) \,)$

Example :

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



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



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

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)

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

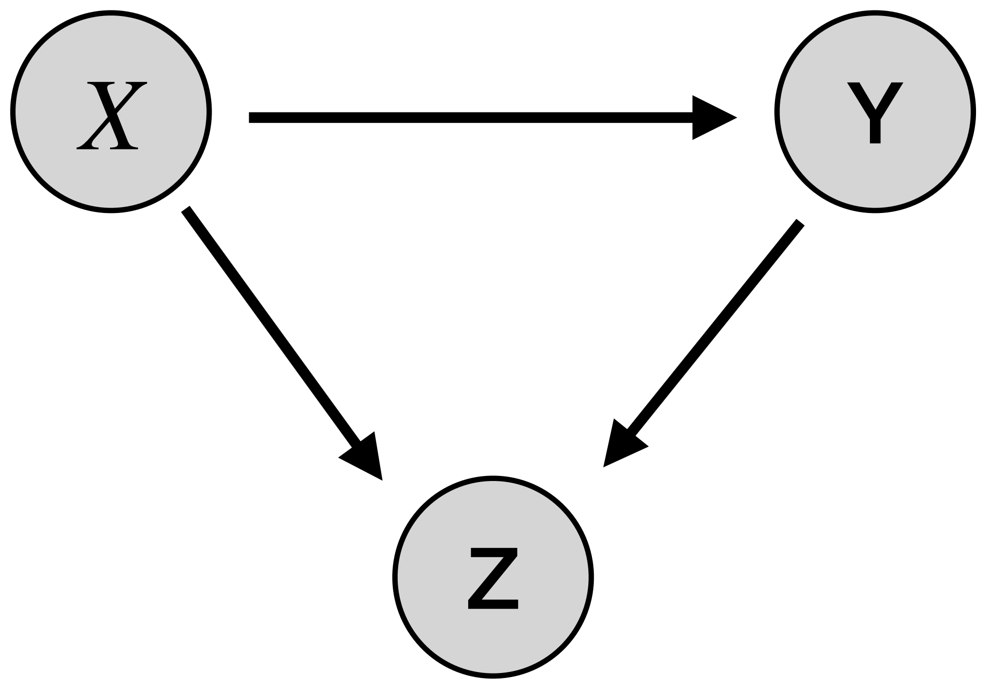
The focus of our course !

Model : joint probability over all variables

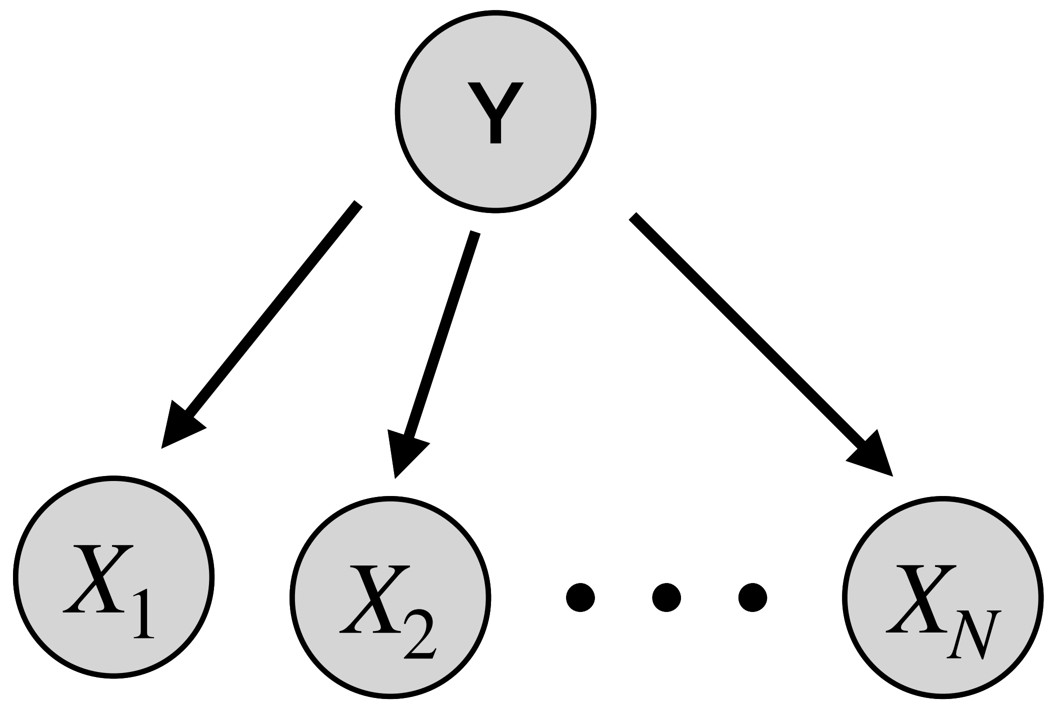
$$P(X_1, \cdots, X_N) = \prod_{i=1,\cdots,N} P(\ X_i \mid parents(X_i)\ )$$

Example :

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



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



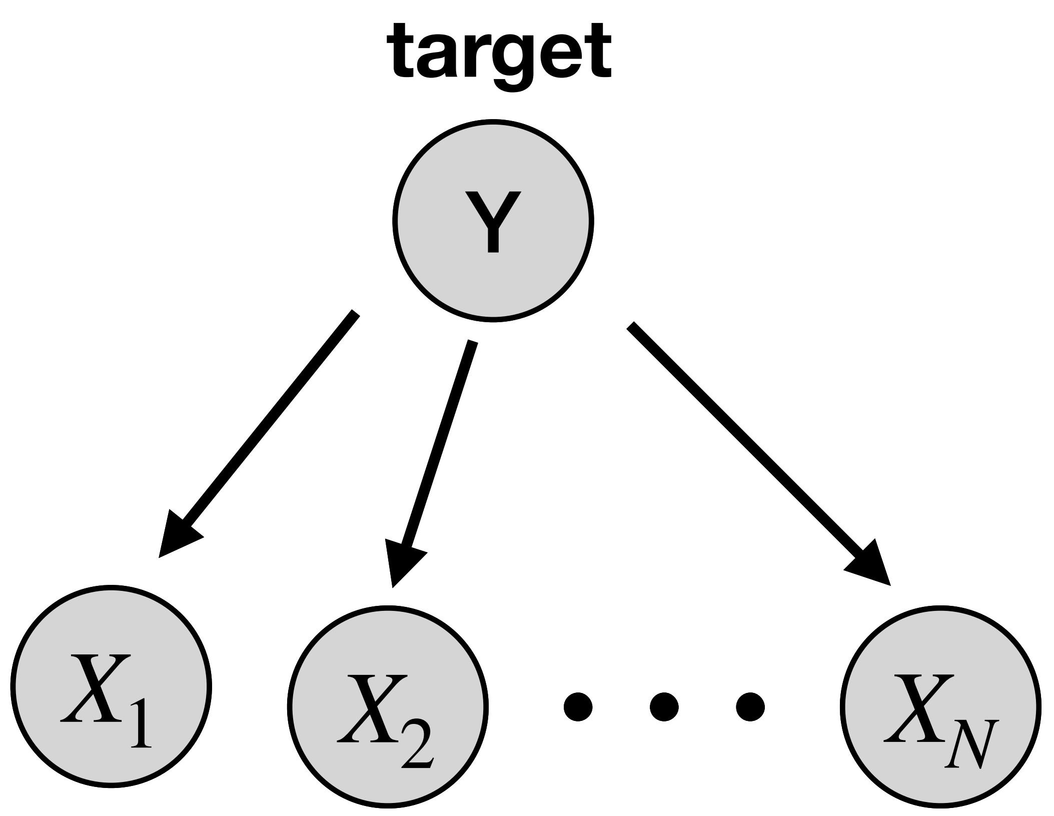
$$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i|Y)$$

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

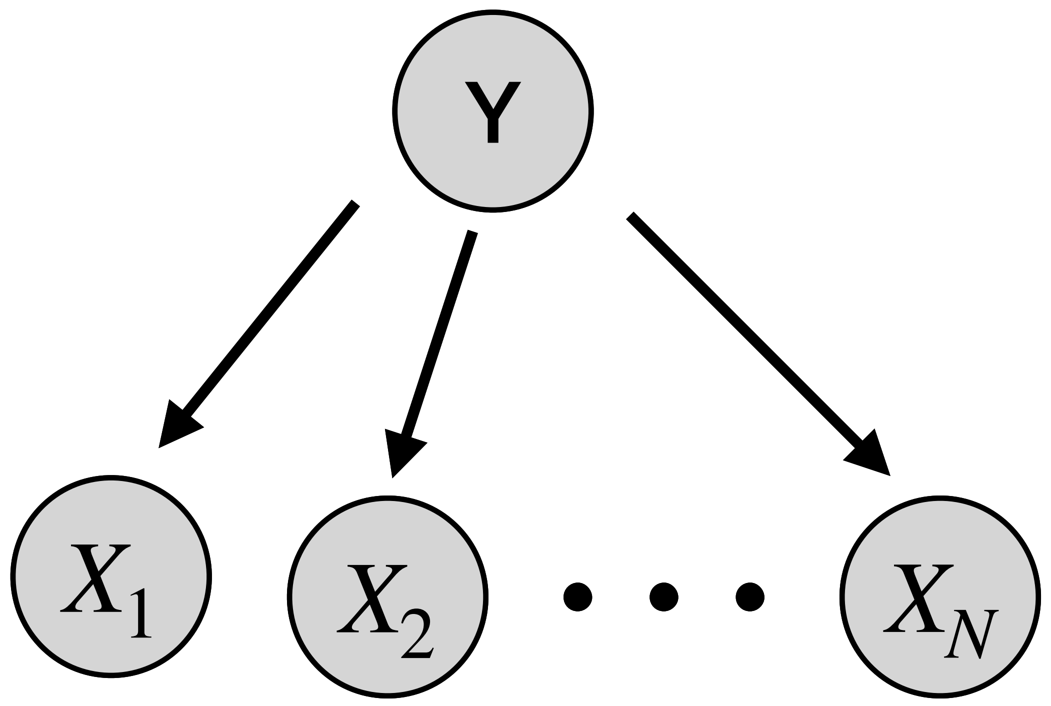
X_i , X_j conditionally independent given Y

2. Probabilistic model

Plates and examples of probabilistic model

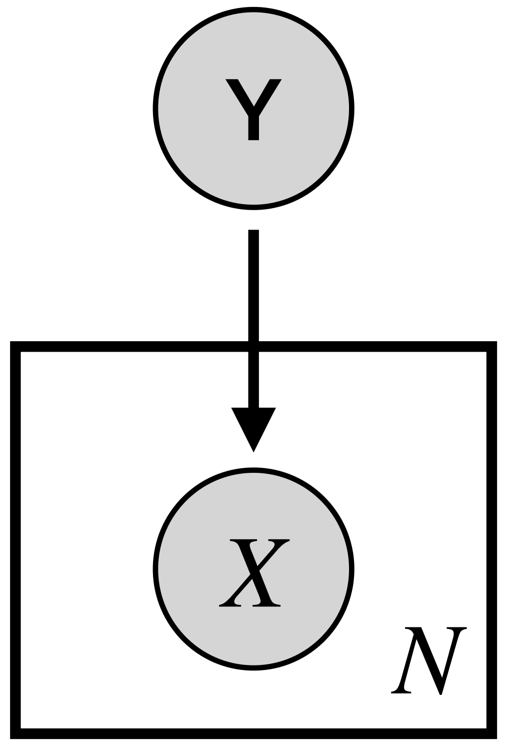
Naive Bayes Classifier $$P(Y, X_1, \cdots, X_N) = P(Y) \prod_{i=1\ldots,N} P(X_i \mid Y)$$

target



Features
X_i , X_j conditionally independent given Y

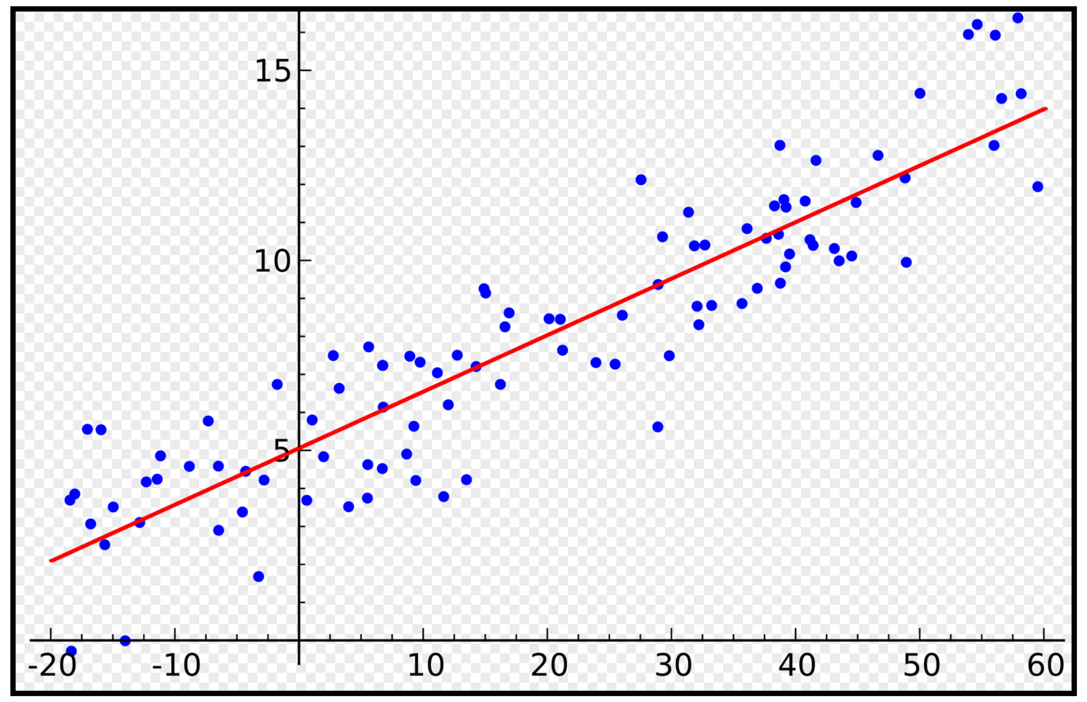
Plate notation



2. Probabilistic model

Frequentist linear regression

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



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

Scalar notation :

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

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

Matrix notation :

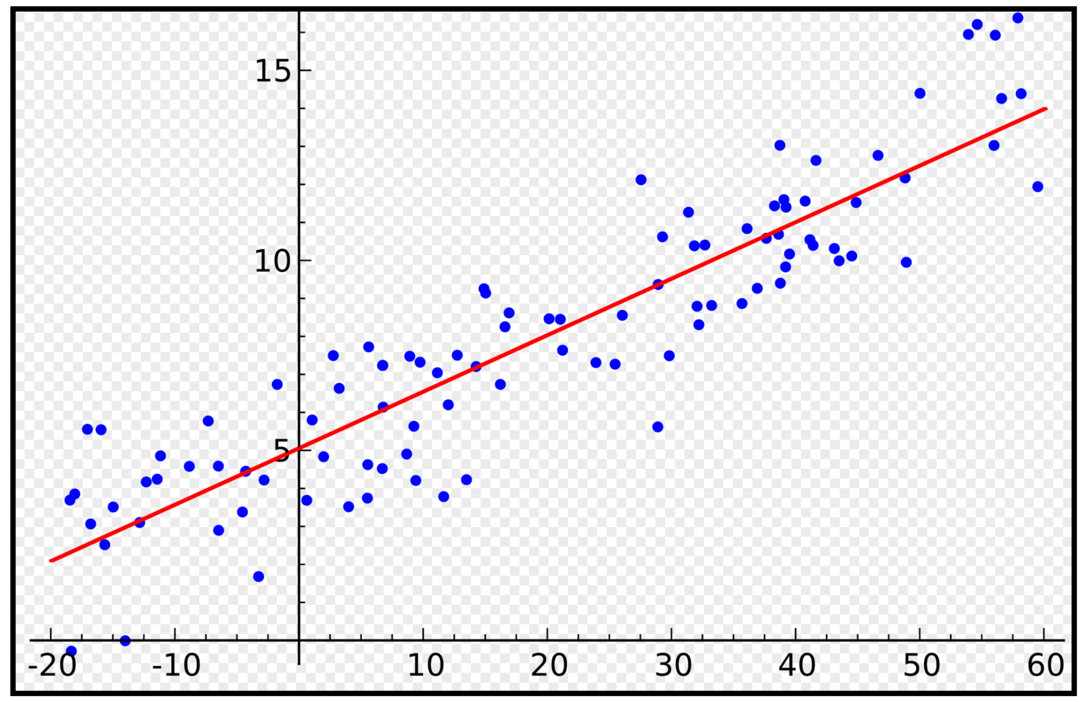
$\mathbf{X} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$

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

2. Probabilistic model

Frequentist linear regression

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



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

Scalar notation :

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

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

Matrix notation :

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

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

Proof : MLE for linear regression

$X = (x_1, \dots, x_n)$, $y = (y_1, \dots y_n)$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$Y \sim \mathcal{N}(w^T x, \sigma^2)$

$\theta = w$ we suppose that σ^2 is known

$Y_1, \dots, Y_n$ indep, $Y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$ OR $Y_i = w^T x_i + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$\hat{\theta}_{MLE} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}} \ \underbrace{p(y \mid x, \theta)}_{\text{likelihood}}$$

$$p(y \mid x, \theta) = p(y_1, \dots, y_n \mid x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} p(y_i \mid x_i, \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - x_i^T w)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T w)^2\right)$$

" vector matrix way "

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (y - x^T w)^T (y - x^T w)\right)$$

$$\ell(\theta) = \log p(y \mid x, \theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - Xw)^T (y - Xw)$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 - \frac{1}{2\sigma^2} \left[0 - 2X^T y + 2X^T X \theta\right]$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 \iff \hat{\theta}_{MLE} = \theta = (X^T X)^{-1} X^T y$$

2. Probabilistic model

Bayesian linear regression

Bayesian Linear regression

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

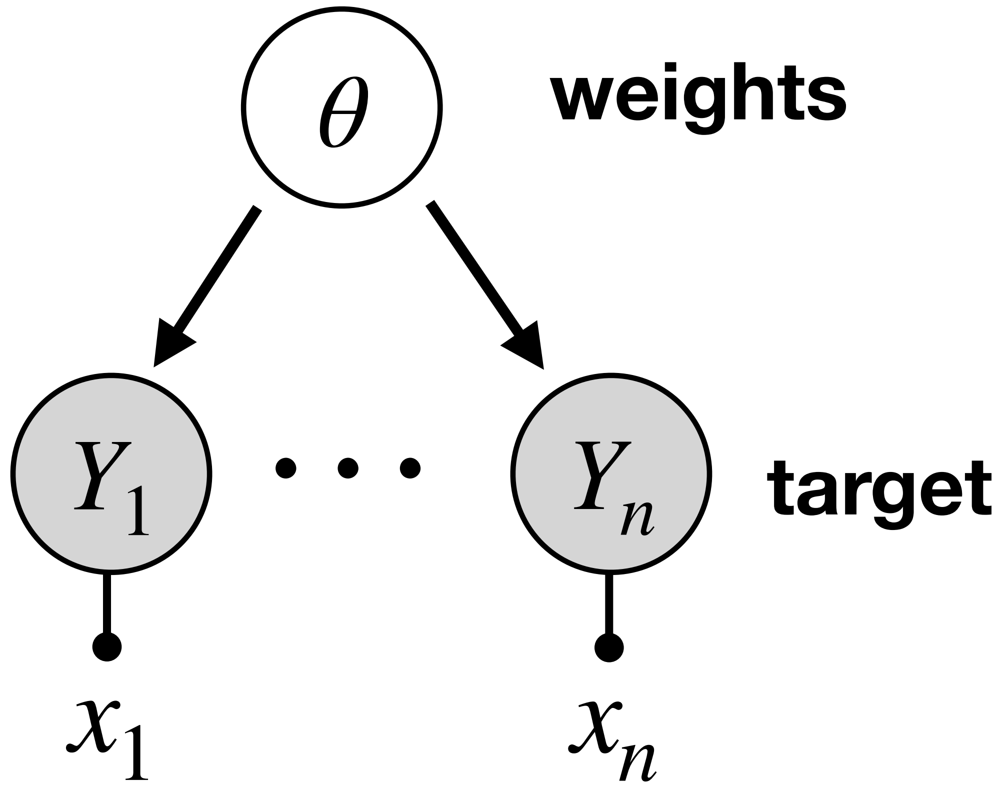
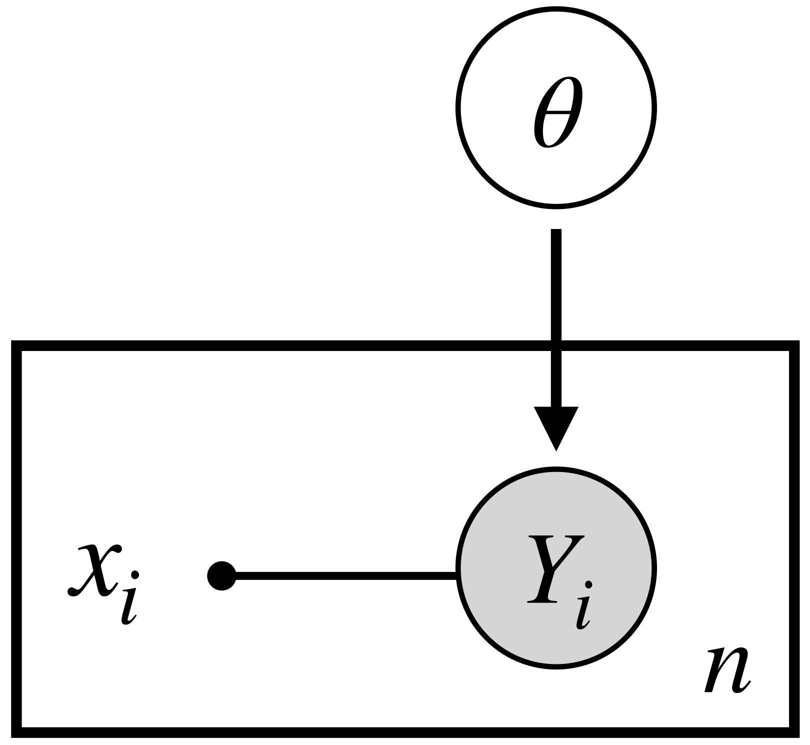


Plate notation

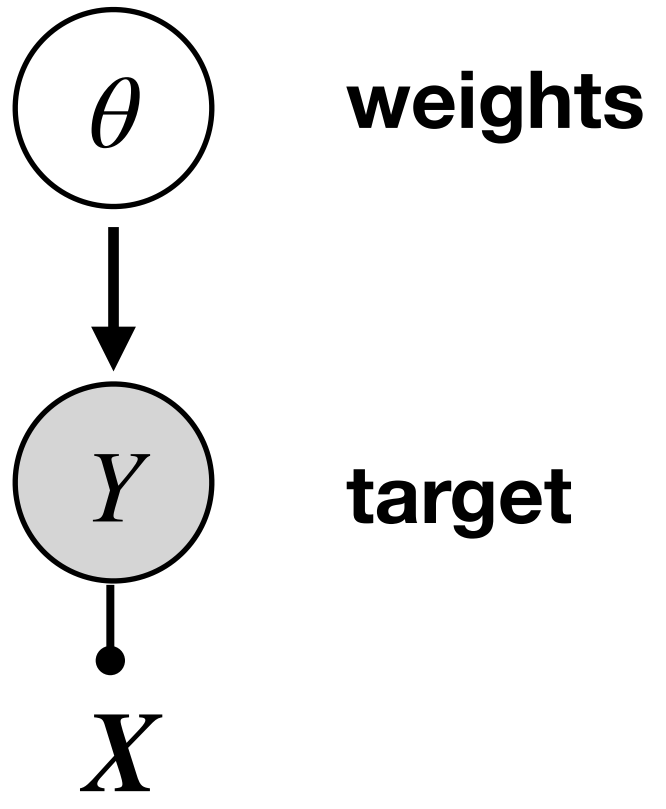


$$P(\theta, y_i \,|\, x_i) = P(y_i \,|\, \theta, x_i) \times P(\theta)$$

$$P(y_i \,|\, \theta, x_i) = \ldots$$

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

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



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

$$P(y \,|\, \theta, x) = \ldots$$

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

Legend :

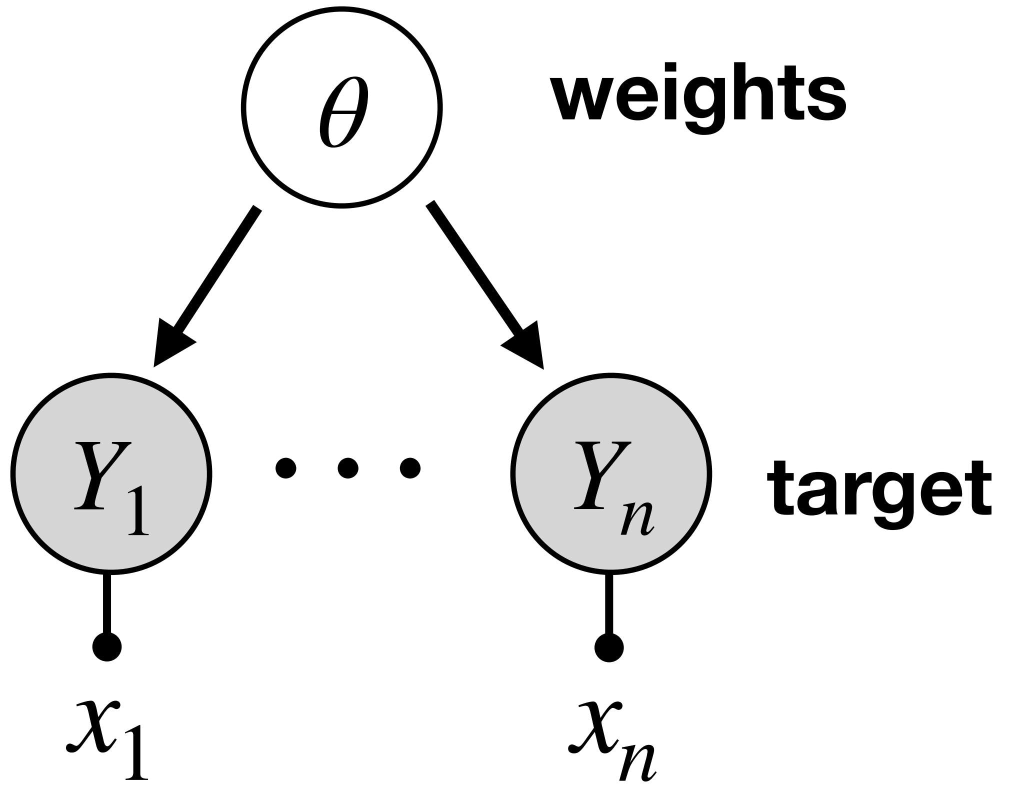
Fixed variable — Hidden (latent) r.v. — Observed r.v.

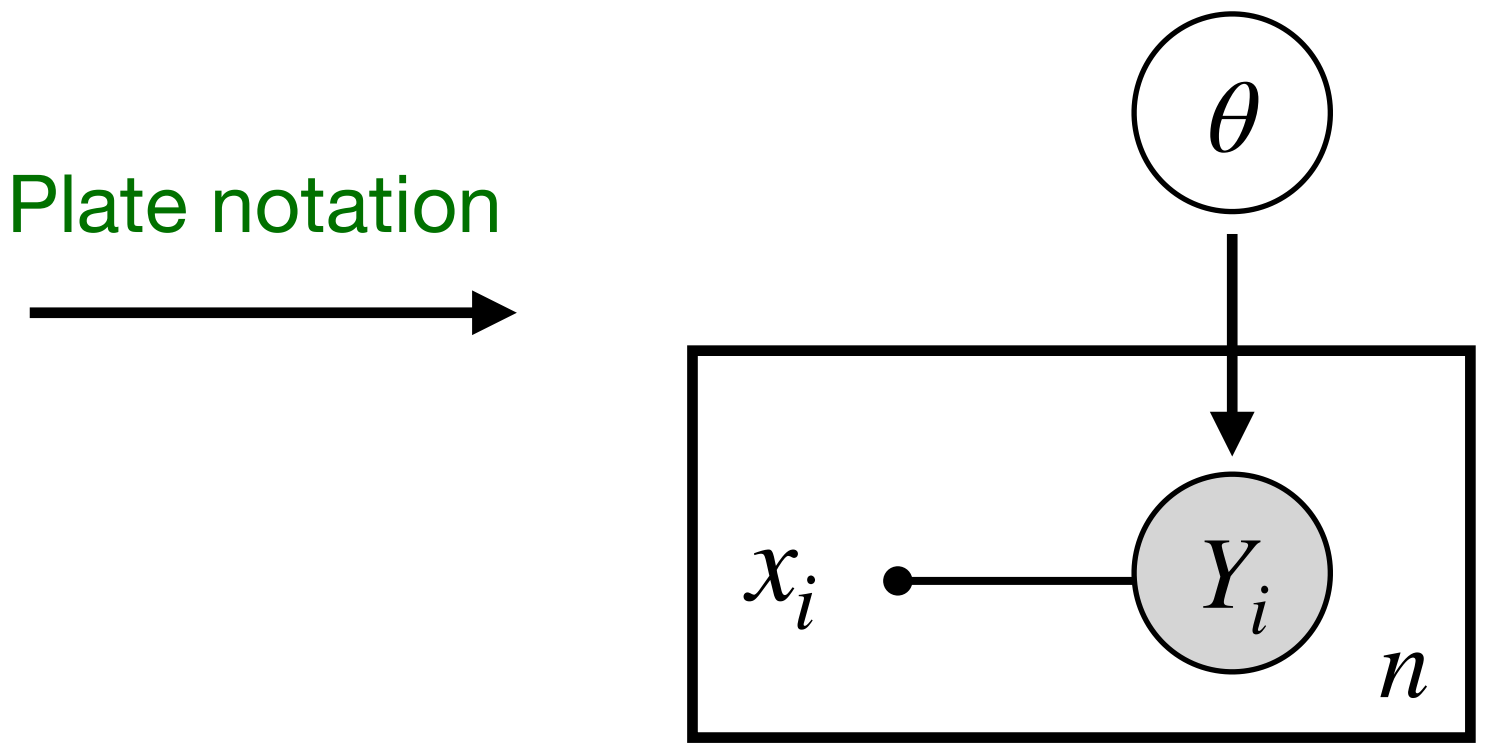
2. Probabilistic model

Bayesian linear regression

Bayesian Linear regression

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



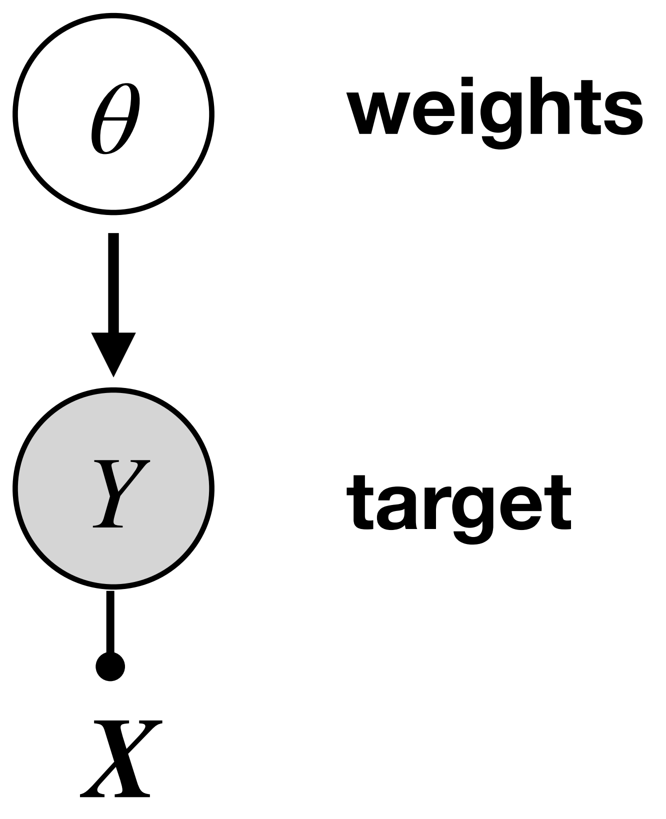


$$P(\theta, y_i \,|\, x_i) = P(y_i \,|\, \theta, x_i) \times P(\theta)$$

$$P(y_i \,|\, \theta, x_i) = \ldots$$

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

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



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

$$P(y \,|\, \theta, x) = \ldots$$

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

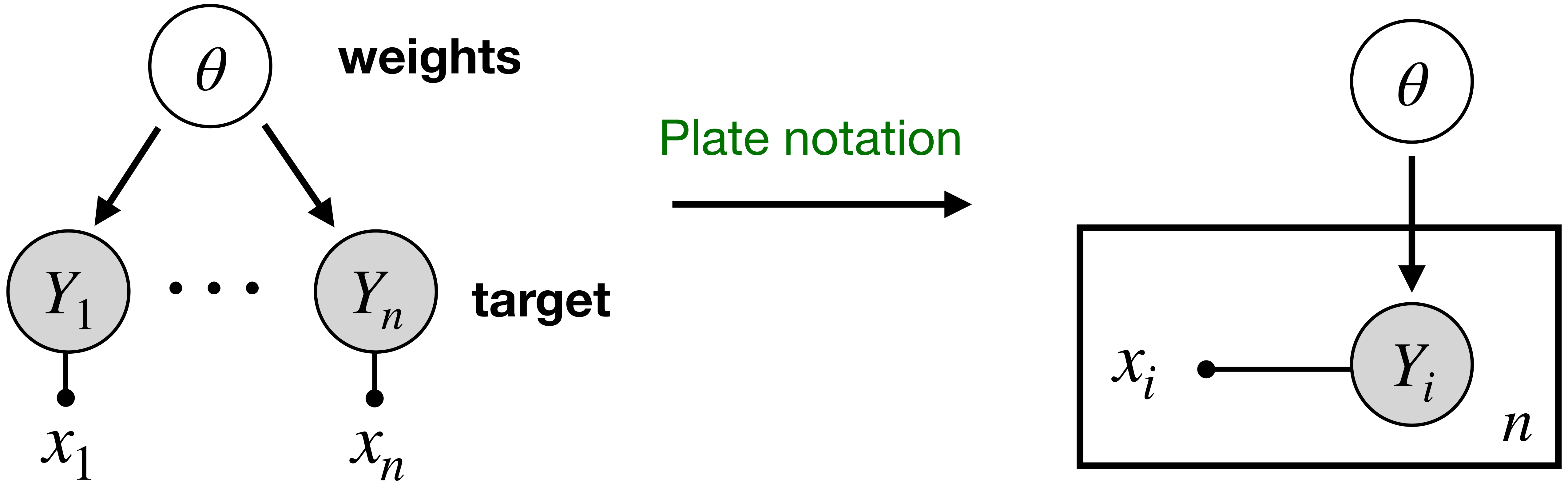
Legend :

2. Probabilistic model

Bayesian linear regression

Bayesian Linear regression

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

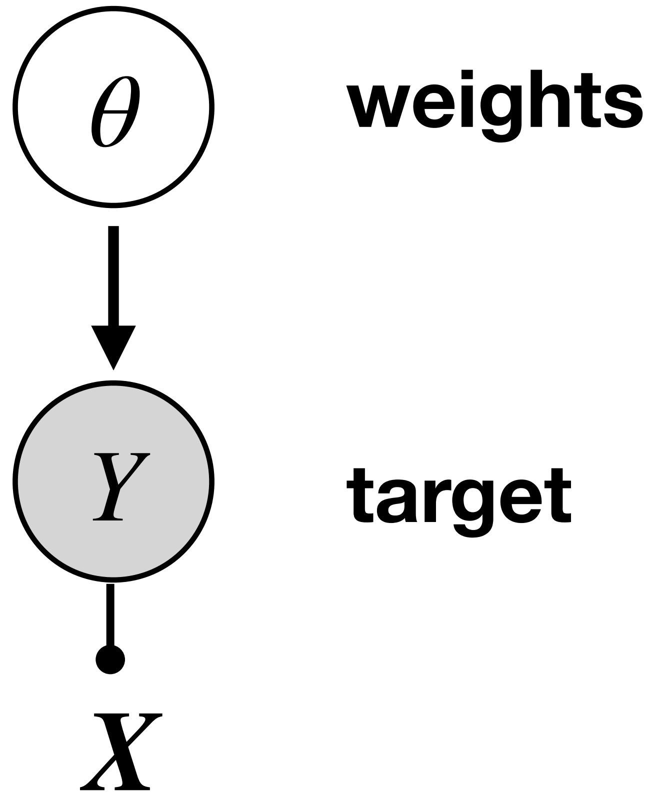


$$P(\theta, y_i \mid x_i) = P(y_i \mid \theta, x_i) \times P(\theta)$$

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

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

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



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

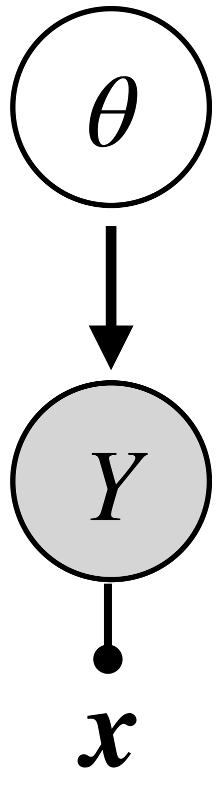
$$P(y \mid \theta, x) = \mathcal{N}(y \mid \theta^T x,\ \sigma^2 I_n)$$

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

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 :

Frequentist linear regression

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

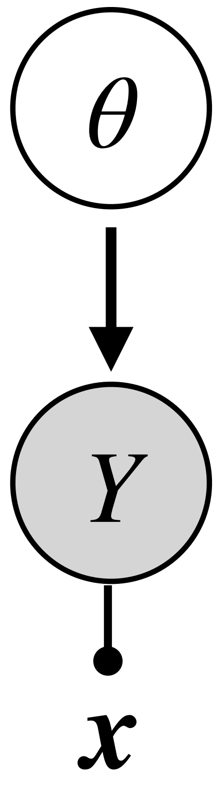
$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta)$$

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

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 : $\underset{\theta}{\arg\max}\, P(\theta \mid X, y) = \underset{\theta}{\arg\max}\, P(\theta, y \mid X)$

Frequentist linear regression

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

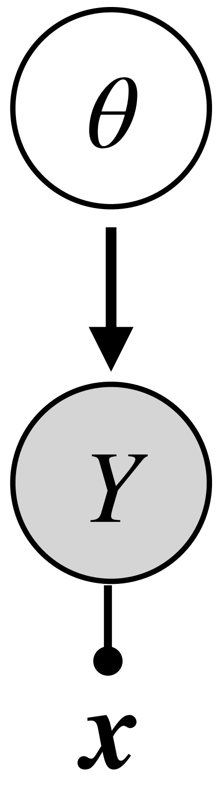
$$\hat{\theta} = \underset{\theta}{\arg\min}\, \mathcal{L}(\theta)$$

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

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

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

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

Theorem :

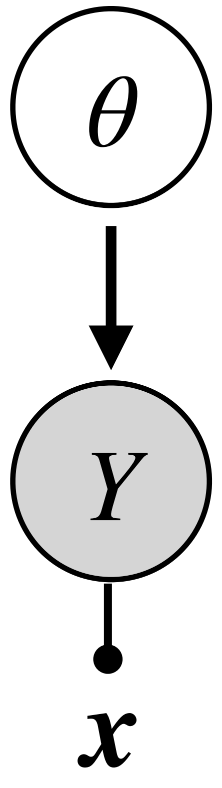
There exists $\lambda \in \mathbb{R}$ such that : $\arg\max_{\theta} P(\theta \,|\, X, y) = \arg\min_{\theta} \left\{ \|\theta^t X - y\|^2 + \lambda \|\theta\|^2 \right\}$

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



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

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

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

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

Frequentist linear regression

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

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

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

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**

Proof : left as an exercice

3 Analytical Inference

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

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

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

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

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

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

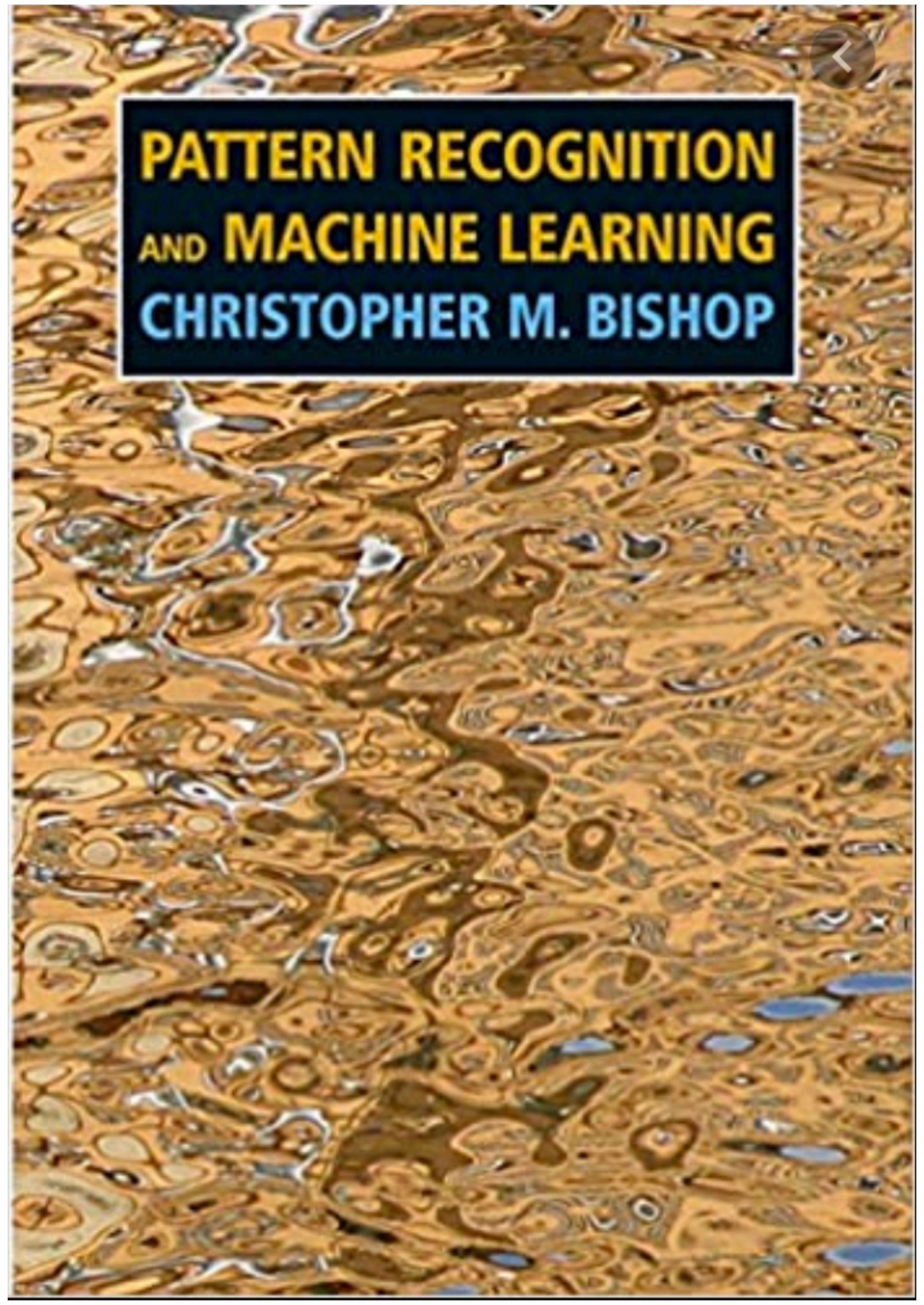
Limitations (among many others)

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 $P(\theta|X)$ lie in the same family of distributions (gaussian for example)

4. Conjugate distributions

Conjugate distributions : avoid computing evidence

Posterior distribution

Fixed by model — **Likelihood**

Fixed by us — **Prior**

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

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 $P(\theta|X)$ lie in the same family of distributions (gaussian for example)

Example

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

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

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

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

4. Conjugate distributions

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)**

5 Conjugate distributions : Exercices

5. Conjugate distributions

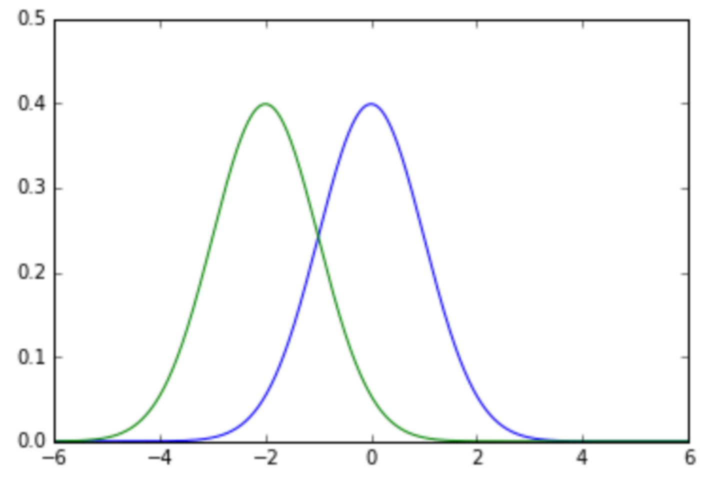
Exercices

If you do these exercices before the next lecture, you'll have **bonus points**:

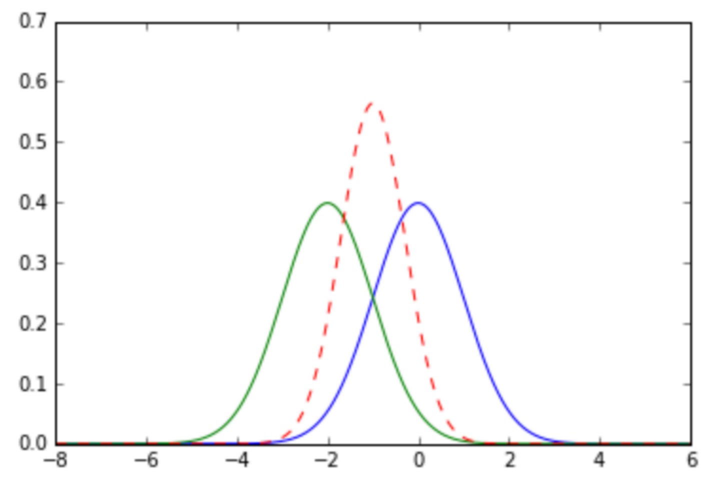
- **0.25 pts** if only the reasoning is good
- **0.5 pts** if only some of them are correct
- **0.75 pts** if most of them are correct
- **1 pts** if all of them are correct

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

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



pointwise product



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

show the following equation

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

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

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

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

In the context of a gaussian, the prior for the **<u>precision</u>** is a gamma !

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

show the following equation

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

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

$$B(\theta \mid \alpha_{posterior}, \beta_{posterior})$$

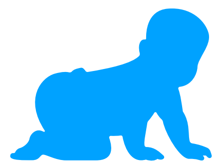
$$P(\theta \mid x) = \frac{Ber(x \mid \theta) \times P(\theta)}{P(x)}$$

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

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

! Road map

Bayesian statistics



1

Latent variable models

2

Variational Inference

3

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 :
- exact posterior

Cons :
- conjugate prior maybe inadequate

Causal Inference

4

Oral presentations

5