



Bayesian Machine Learning

April 2022 - François HU
<https://curiousml.github.io/>

Outline

PREREQUISITE

1

Bayesian statistics

- Bayesian statistics and probabilistic model
- Analytical inference
- Conjugate priors

2

Latent Variable Models

3

Variational Inference

4

Markov Chain Monte Carlo

5

Extensions and oral presentations

THEORY

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

APPLICATION

Python (or at least R)

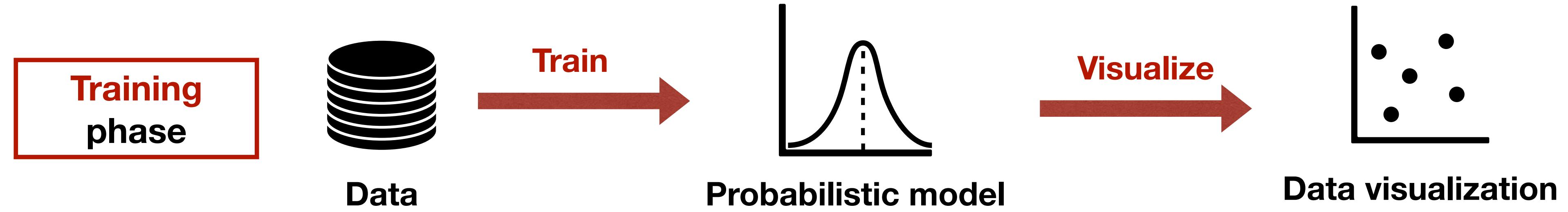
ALGORITHM

Some « classical » supervised & unsupervised models

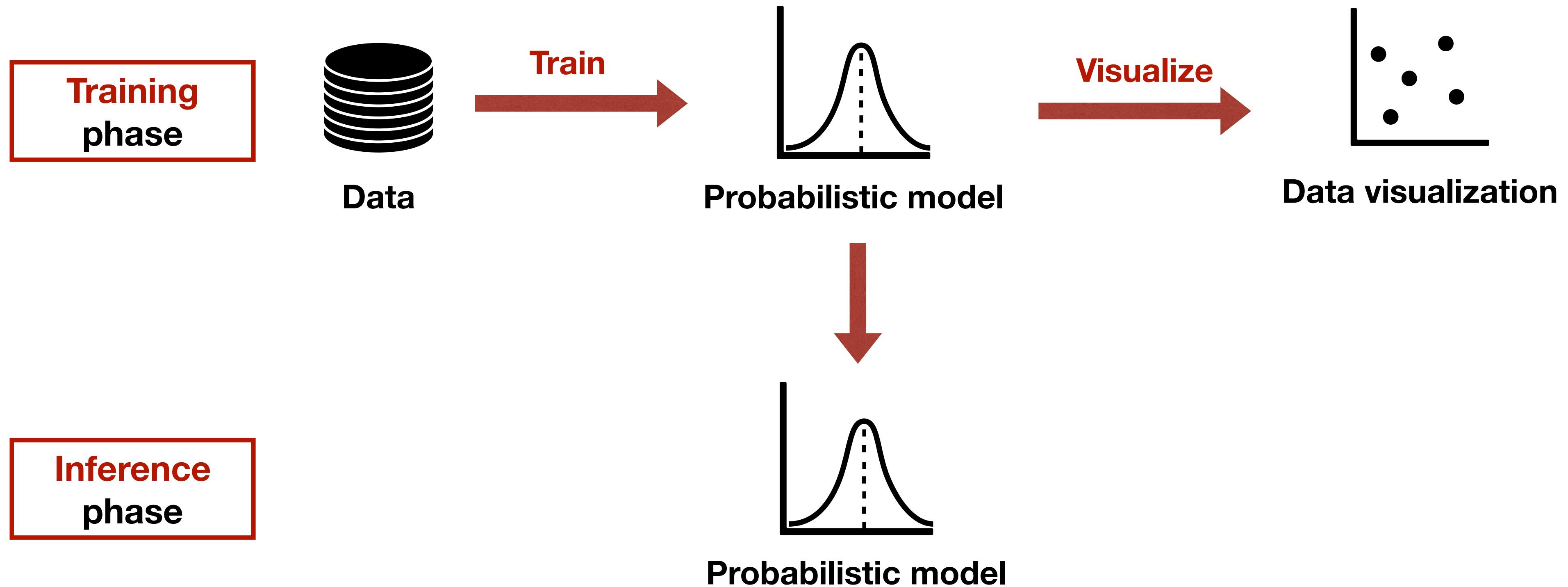
0

Gentle introduction to statistical learning

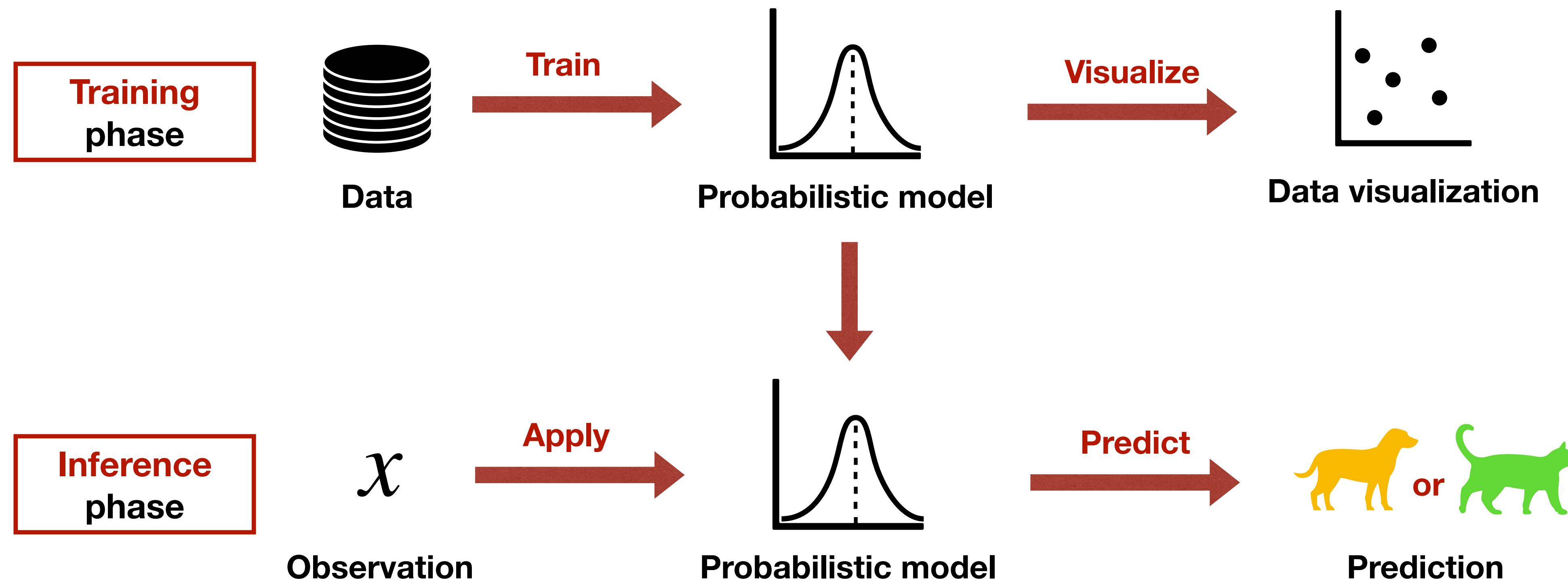
Simplified statistical learning process



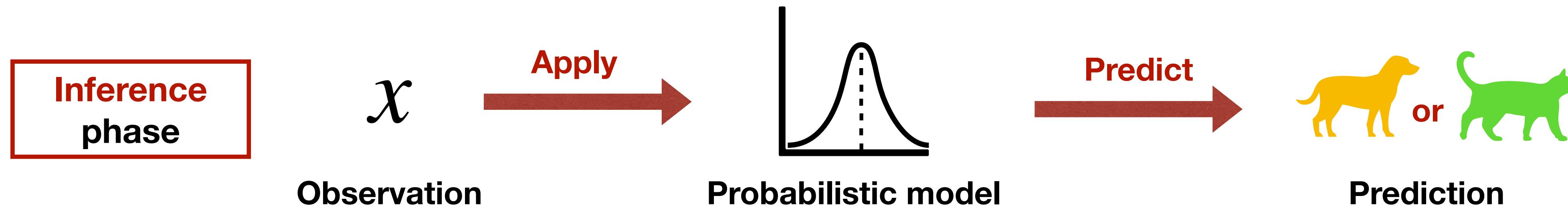
Simplified statistical learning process



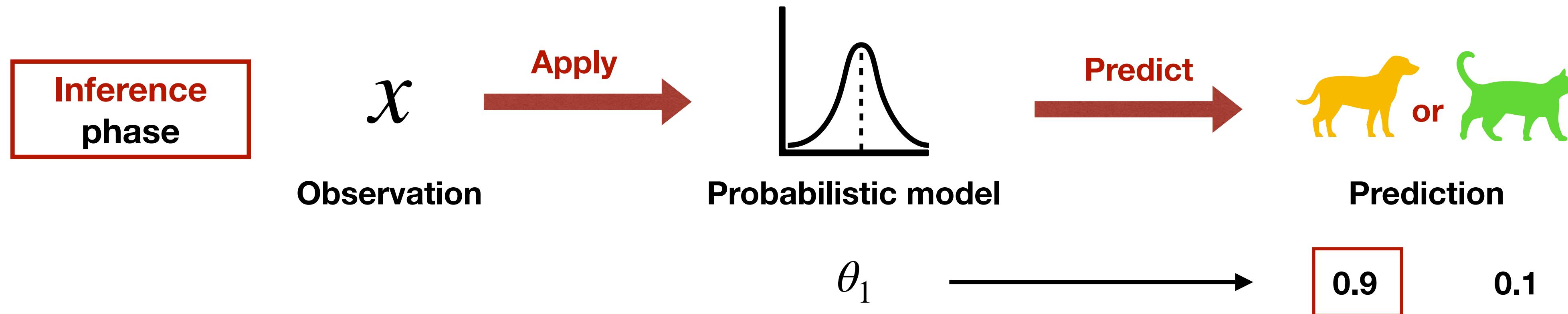
Simplified statistical learning process



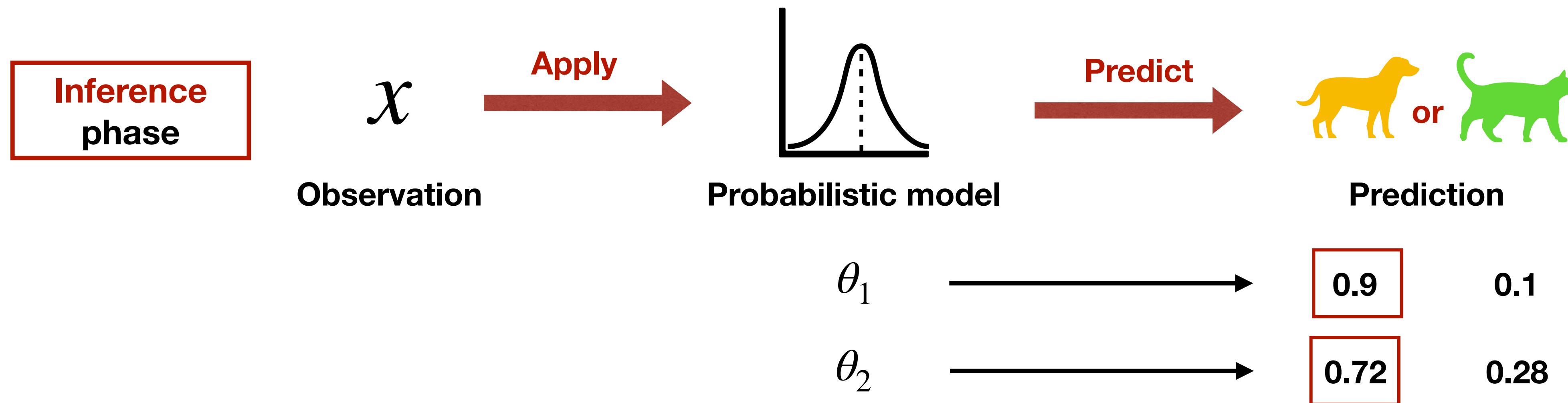
Inference phase



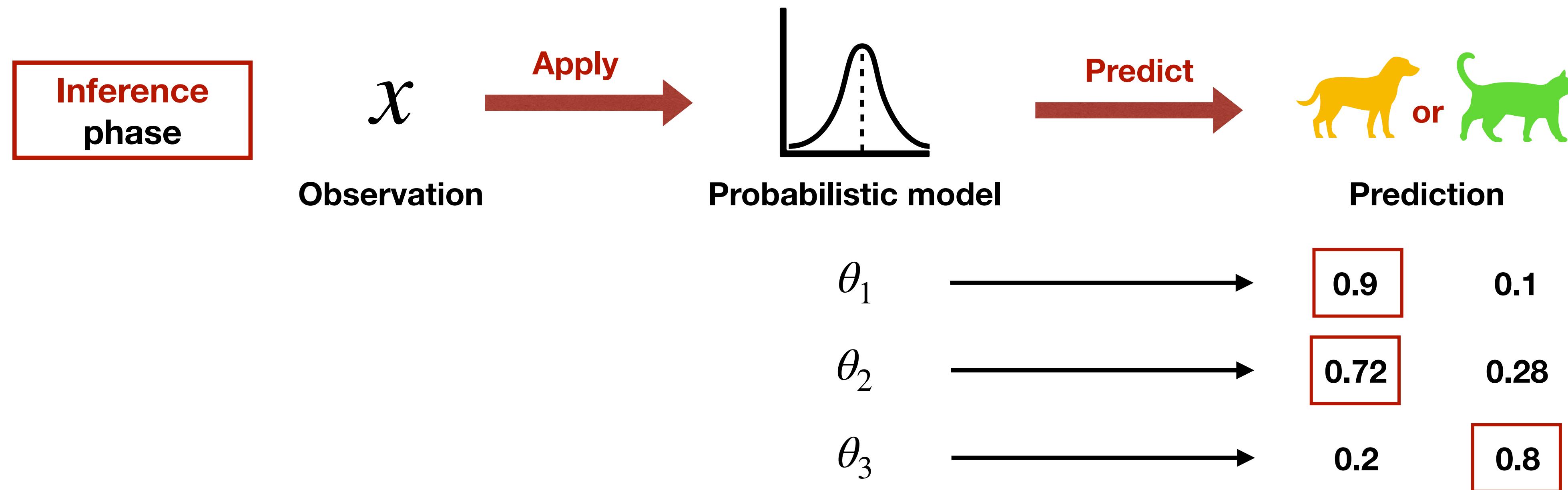
Inference phase



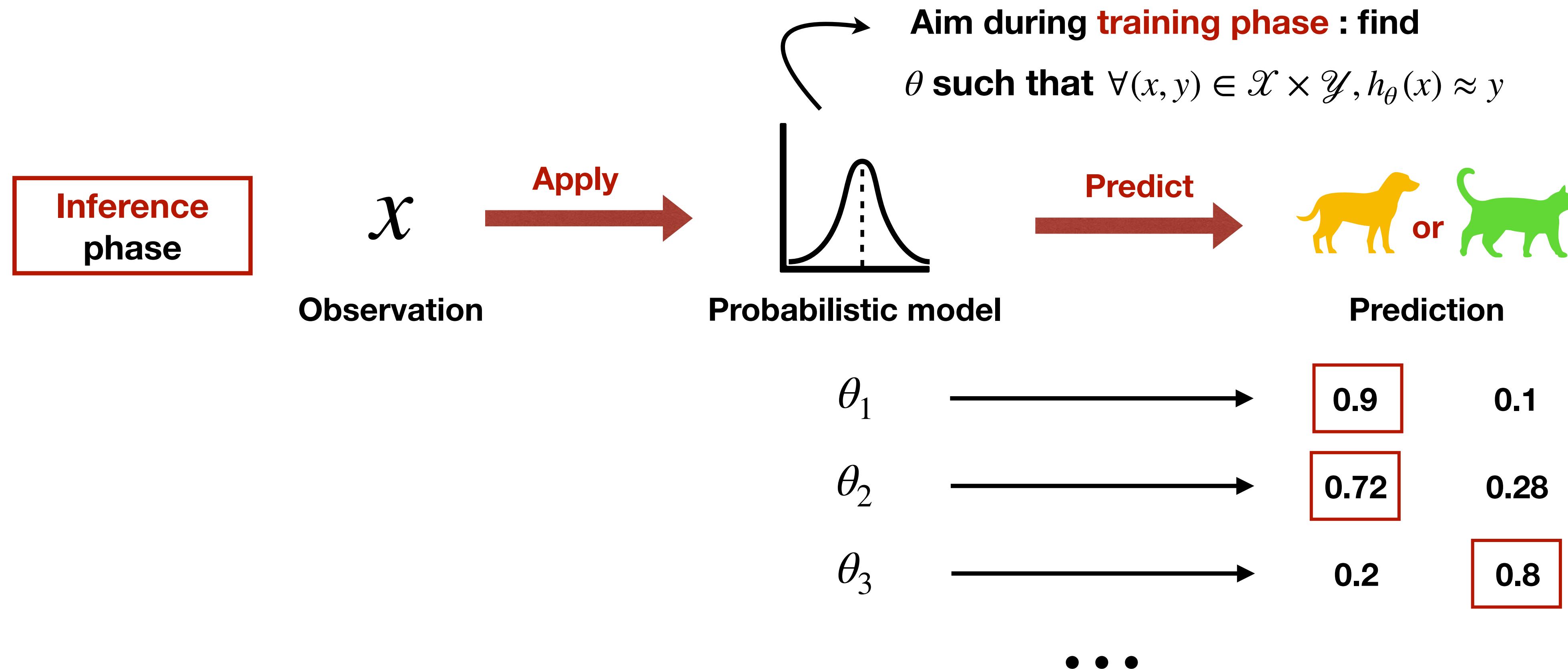
Inference phase



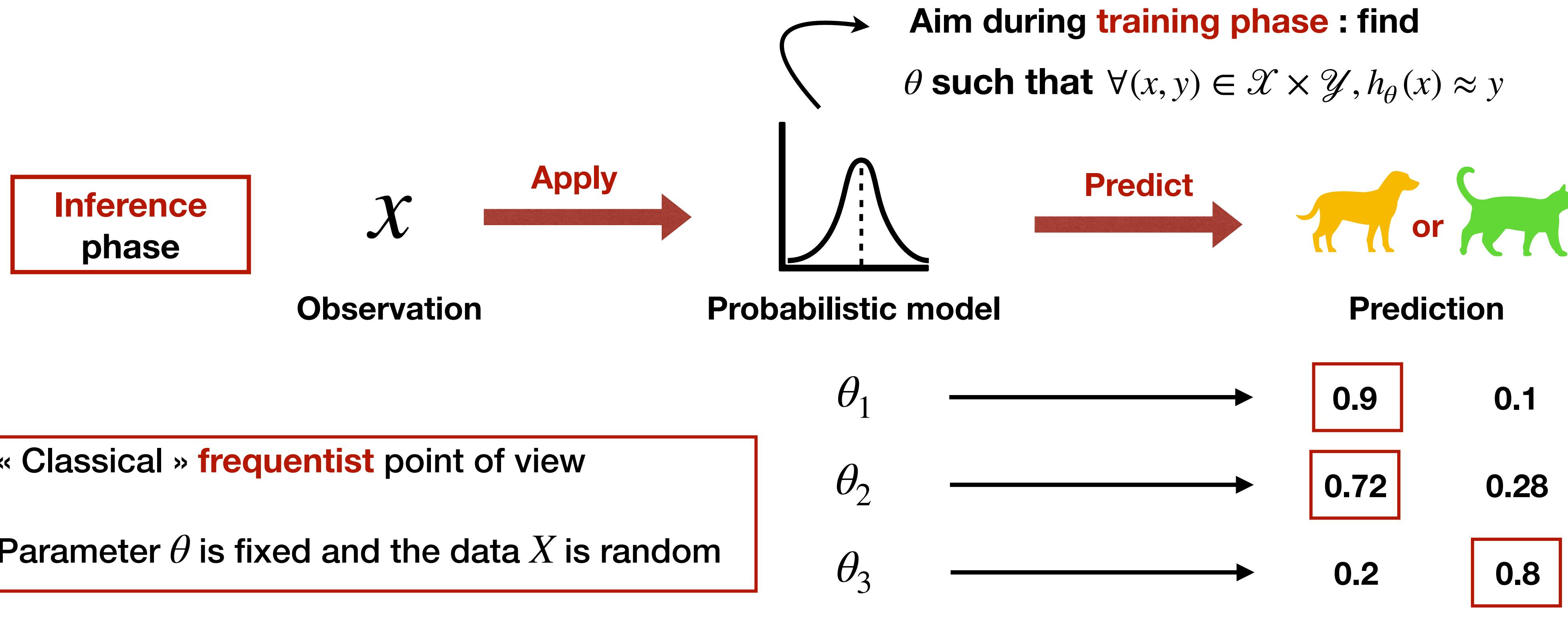
Inference phase



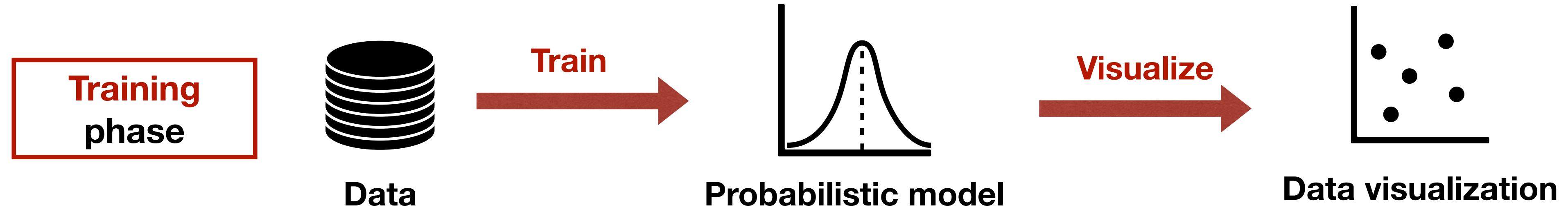
Inference phase



Inference phase



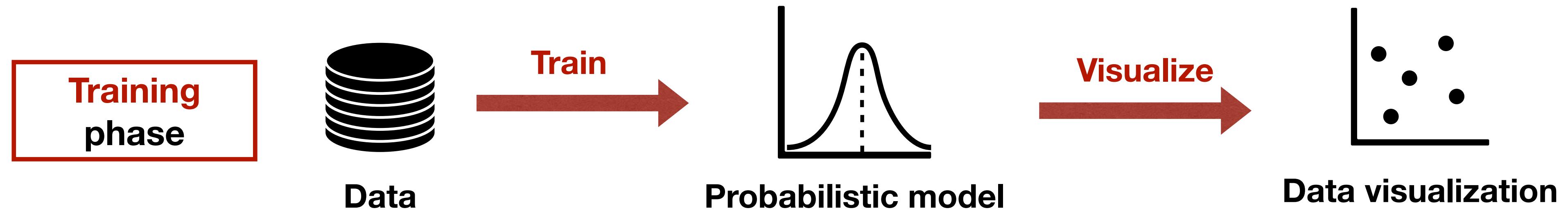
Training phase



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

Training phase

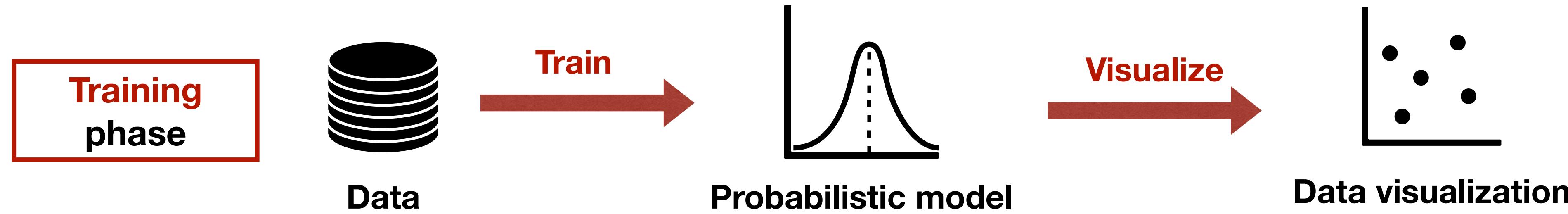


Aim during **training phase** : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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

Training phase



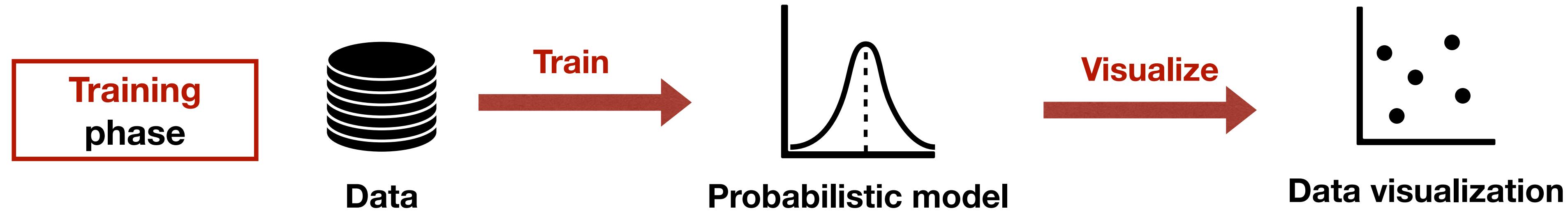
Aim during **training phase** : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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



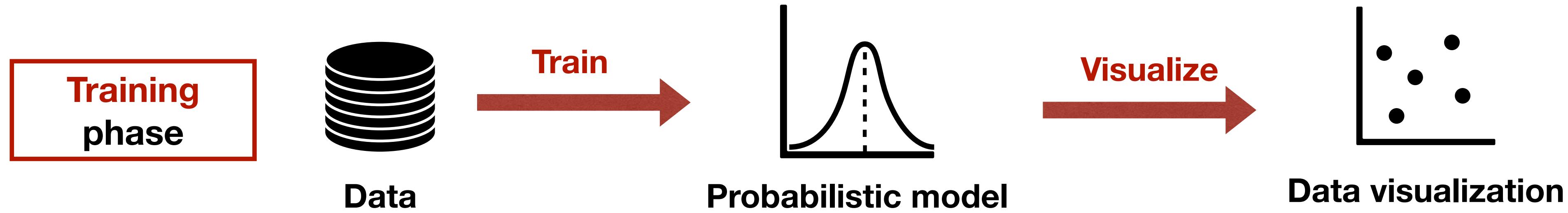
Aim during **training phase** : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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



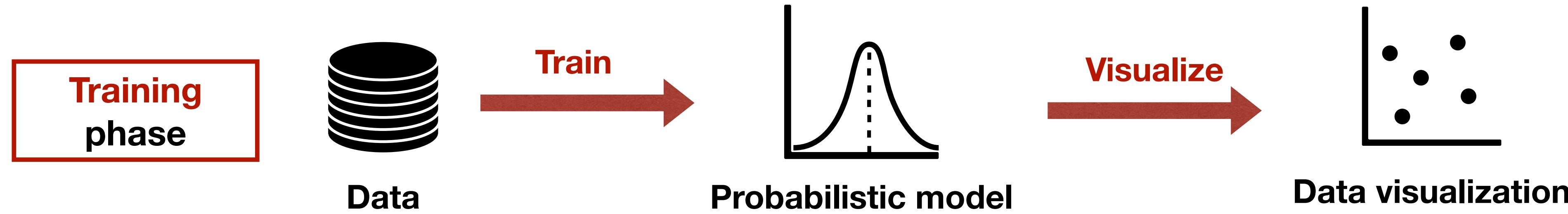
Aim during **training phase** : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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



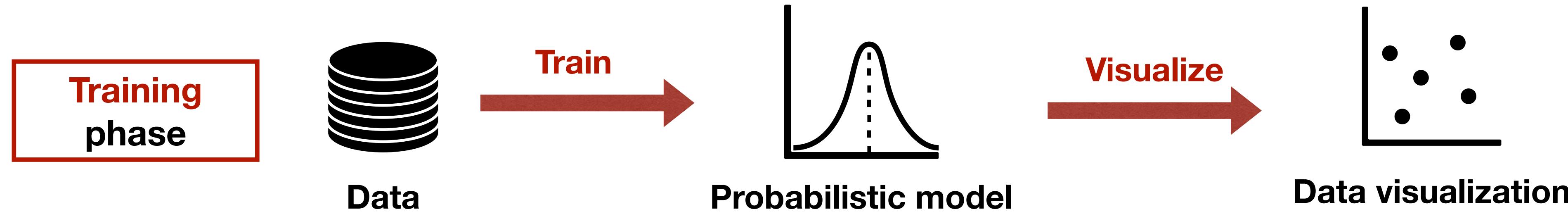
Aim during **training phase** : find θ such that $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, h_\theta(x) \approx y$

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

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

Problems :

- Only works well if we have big data : $|X| \gg |\theta|$
- Cannot start without 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

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)



Supervised machine learning

Unsupervised machine learning

Others

Why Bayesian methods ?

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)



Supervised machine learning

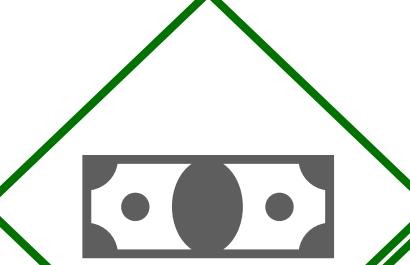
Unsupervised machine learning

Others

Why Bayesian methods ?

Examples of application of Bayesian Machine Learning

Credit card default detection
Bayes theorem



Medical diagnosis
Bayes theorem



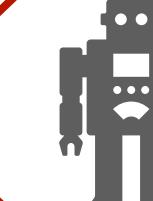
Spam filtering
Bayesian Neural Network



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)



Supervised machine learning

Unsupervised machine learning

Others using more advanced techniques

1

Introduction to bayesian statistics

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

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



$$P(\text{ } \begin{array}{c} \text{K} \\ \text{H} \end{array} \text{ }) = \frac{1}{52}$$
$$P(\text{ } \begin{array}{c} \spadesuit \\ \text{ } \end{array} \text{ }) = \frac{1}{4}$$

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

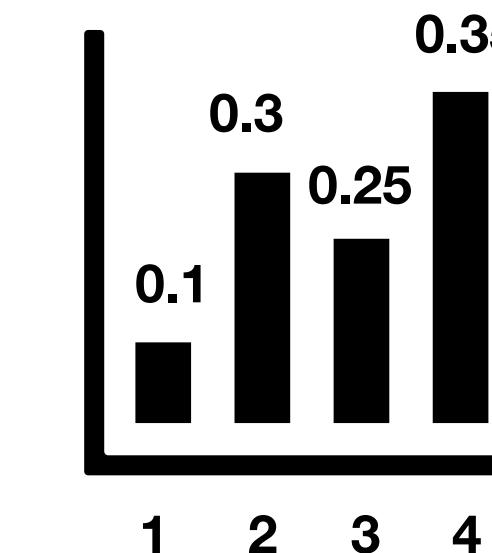
Relative **frequency** of an event in an infinite trials



$$P(\text{King of Hearts}) = \frac{1}{52}$$
$$P(\text{Spade}) = \frac{1}{4}$$

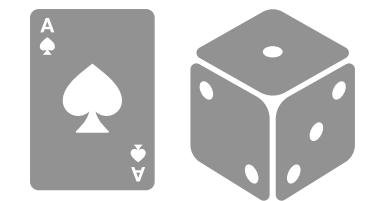
Random variable

- Discrete variable

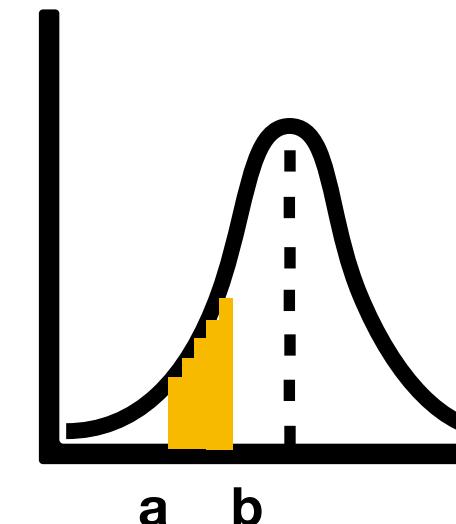


Probability Mass Function (PMF)

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

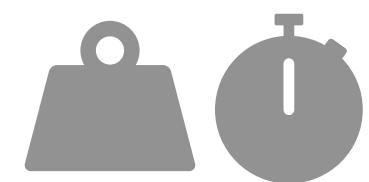


- Continuous variable



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

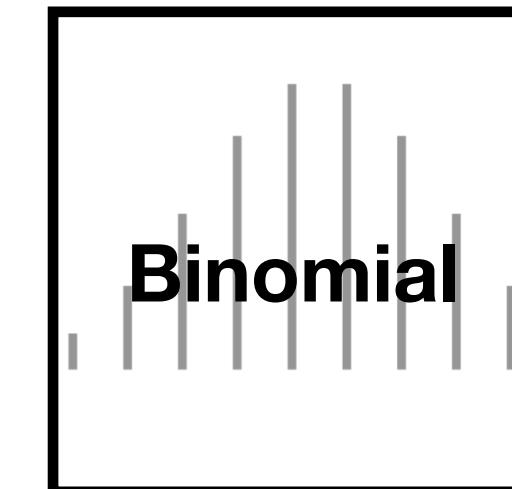
Relative **frequency** of an event in an infinite trials



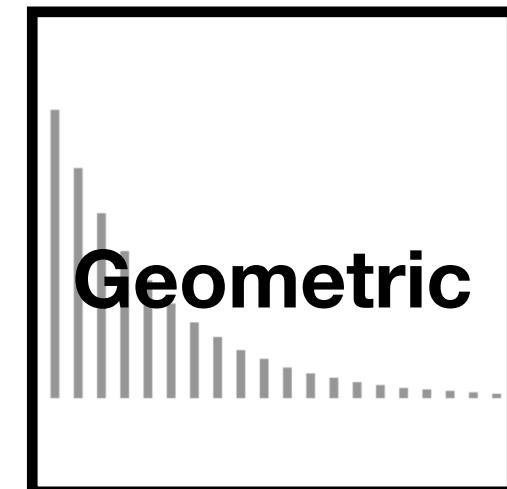
$$P(\text{King of Hearts}) = \frac{1}{52}$$
$$P(\text{Spade}) = \frac{1}{4}$$

Random variable

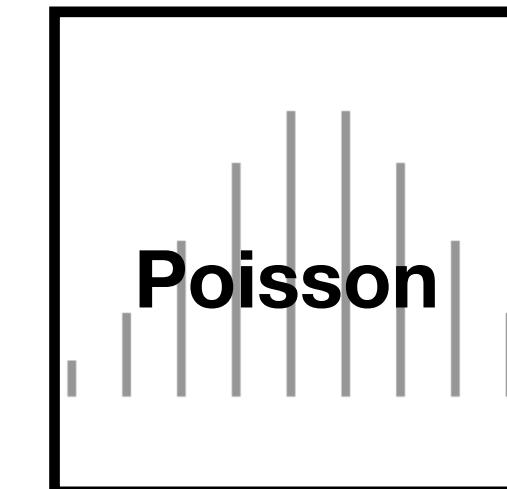
- **Discrete** variable



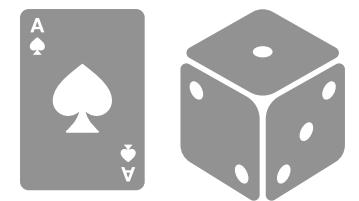
Binomial



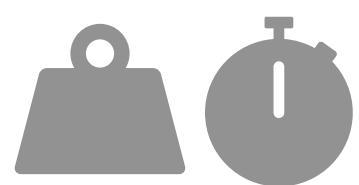
Geometric



Poisson

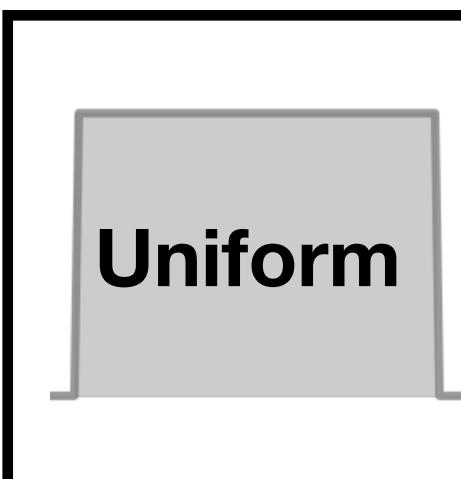


Usual distributions

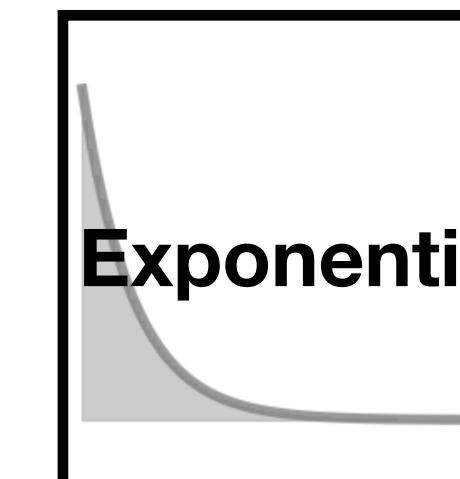


Usual distributions

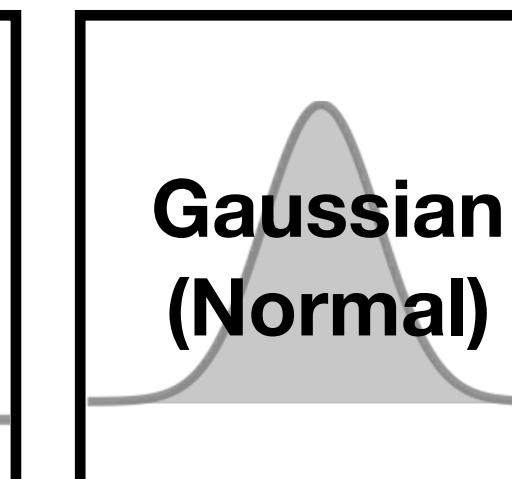
- **Continuous** variable



Uniform



Exponential



Gaussian
(Normal)

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

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



$$P(\text{King of hearts}) = \frac{1}{52}$$
$$P(\text{Spade}) = \frac{1}{4}$$

Independence

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

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

joint probability

marginals

dependency : one dice

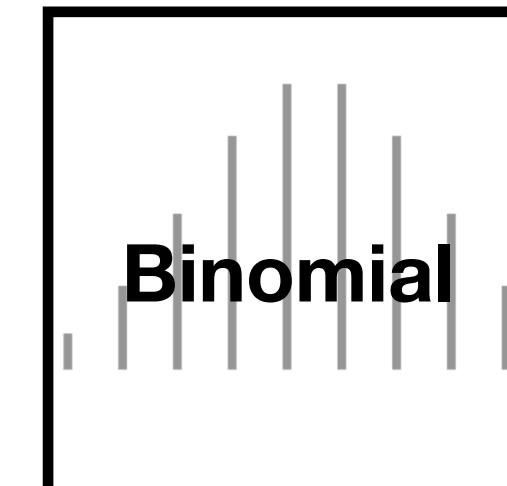
$$P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}, \begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = 0 \neq P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix})P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = 1/6^2$$

independency : two dices

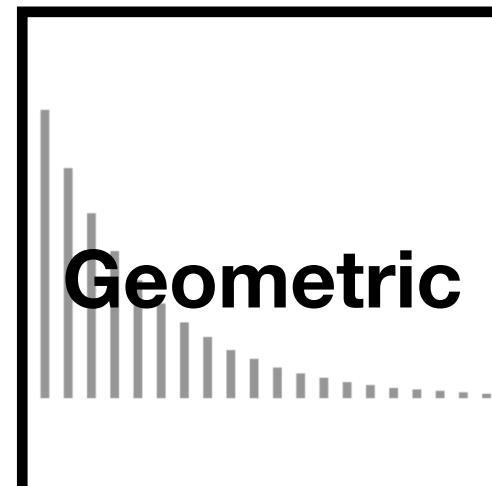
$$P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}, \begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix})P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = 1/6^2$$

Random variable

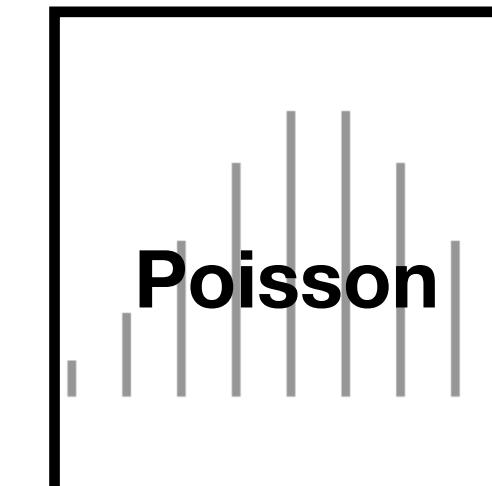
- Discrete variable



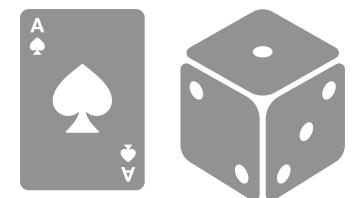
Binomial



Geometric



Poisson



Usual distributions



Usual distributions

1. Introduction to bayesian statistics

Probability & statistics : basic definitions

Probability

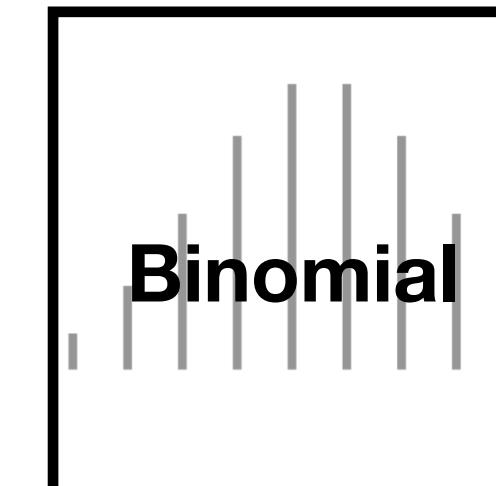
Relative **frequency** of an event in an infinite trials



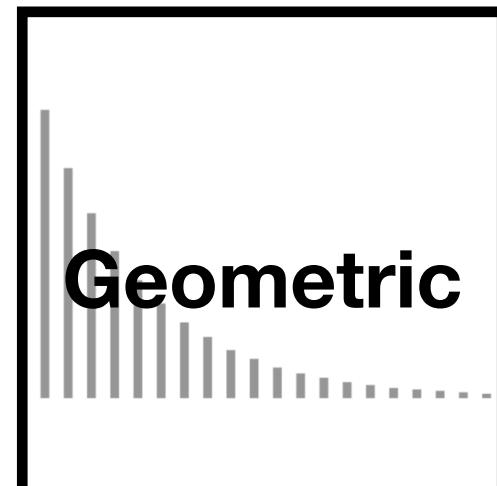
$$P(\text{King of hearts}) = \frac{1}{52}$$
$$P(\text{Spade}) = \frac{1}{4}$$

Random variable

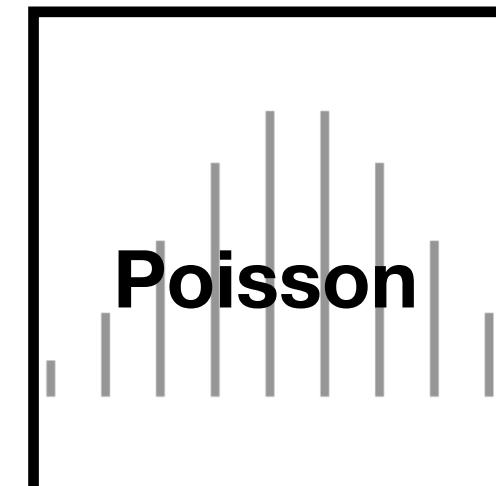
- Discrete variable



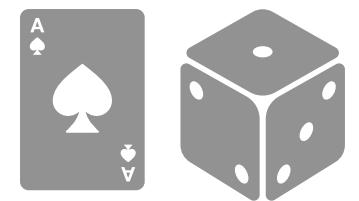
Binomial



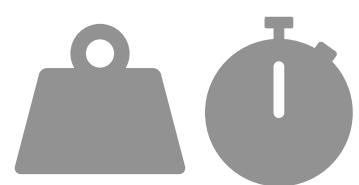
Geometric



Poisson



Usual distributions



Usual distributions

Independence

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

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

joint probability

marginals

dependency : one dice

$$P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}, \begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = 0 \neq P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix})P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = 1/6^2$$

independency : two dices

$$P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}, \begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix})P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = 1/6^2$$

Conditional probability

probability of X **given that** Y happened

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

joint probability
conditional

$$P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix} | \begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}) = \frac{P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix}, \begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix})}{P(\begin{smallmatrix} \bullet & \bullet \\ \bullet & \bullet \end{smallmatrix})}$$

marginal

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

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

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

$$P(X_1, \dots, X_n) = \dots$$

Sum rule

discrete

$$P(X) = \dots$$

continuous

$$P(X) = \dots$$

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

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

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

$$P(X_1, \dots, X_n) = \dots$$

Sum rule

discrete

$$P(X) = \dots$$

continuous

$$P(X) = \dots$$

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

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

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

$$P(X_1, \dots, X_n) = \dots$$

Sum rule

discrete

$$P(X) = \dots$$

continuous

$$P(X) = \dots$$

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 **Marginal**

Joint probability

Chain rule

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

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

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

Sum rule

discrete

$$P(X) = \dots$$

continuous

$$P(X) = \dots$$

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 **Marginal**

Joint probability

Chain rule

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

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

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

Sum rule

discrete

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

continuous

$$P(X) = \dots$$

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 **Marginal**

Joint probability

Chain rule

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

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

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

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

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

ConditionalJoint probabilityMarginal

Chain rule

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

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

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

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 Observations (data)

Likelihood Prior

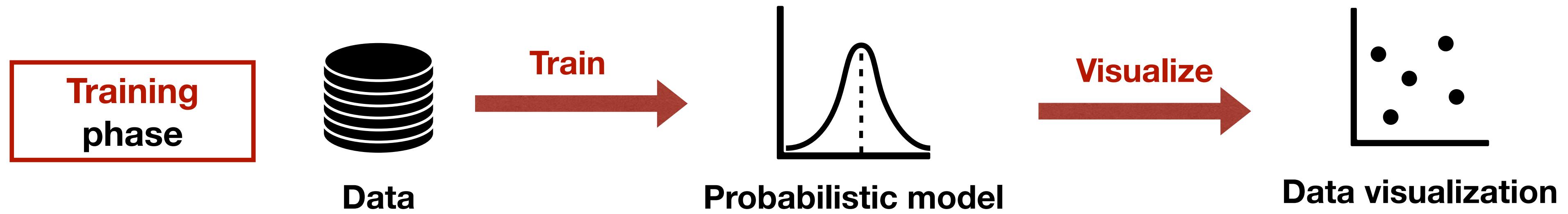
$$P(\theta | X) = \frac{P(X, \theta)}{P(X)} = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$

Posterior

Evidence

1. Introduction to bayesian statistics

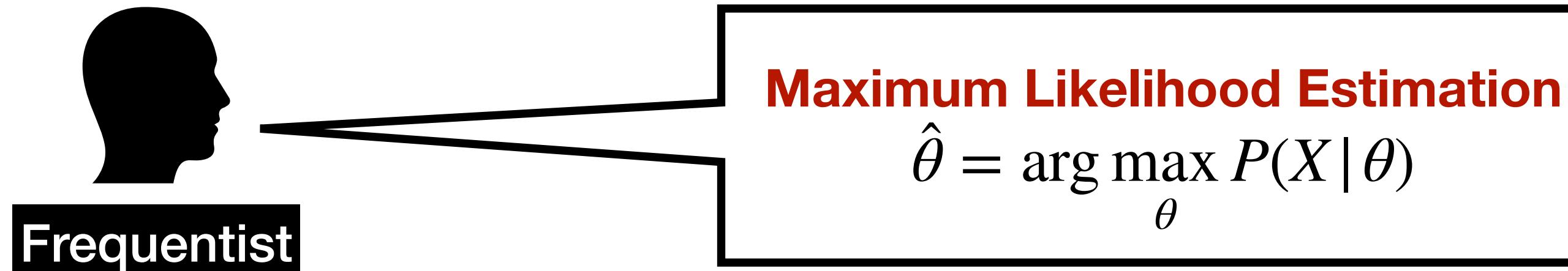
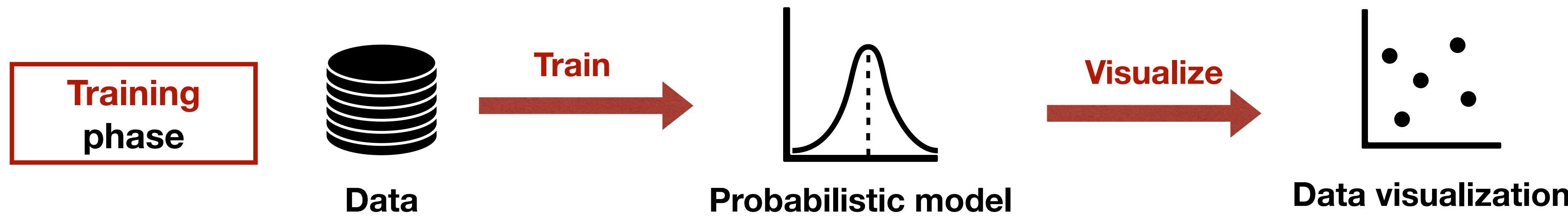
Frequentist VS Bayesian point of view



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

1. Introduction to bayesian statistics

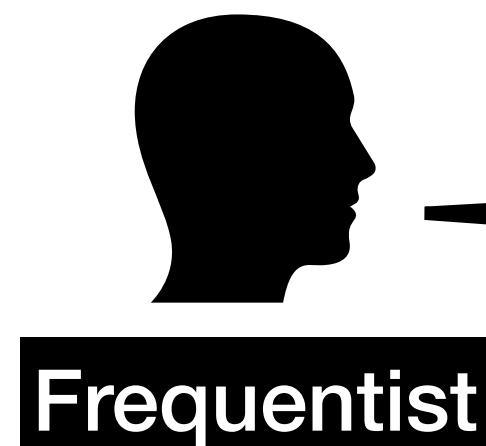
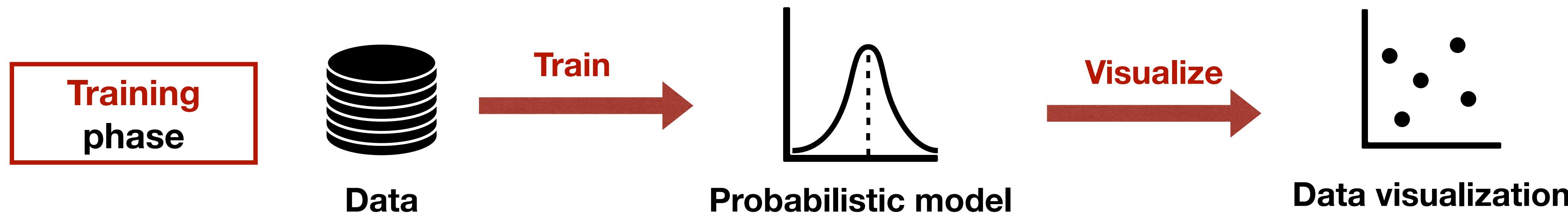
Frequentist VS Bayesian point of view



θ is fixed
 X is random

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

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

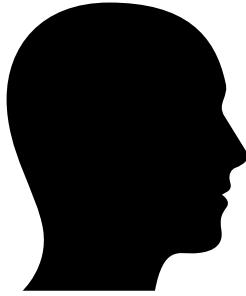


Bayesian

θ is random
 X is fixed

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view



Frequentist

Maximum Likelihood Estimation

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



Bayesian

Bayes theorem

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

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view



Frequentist

Maximum Likelihood Estimation

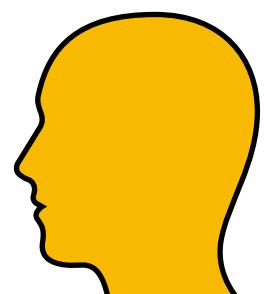
$$\hat{\theta} = \arg \max_{\theta} P(X | \theta)$$

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

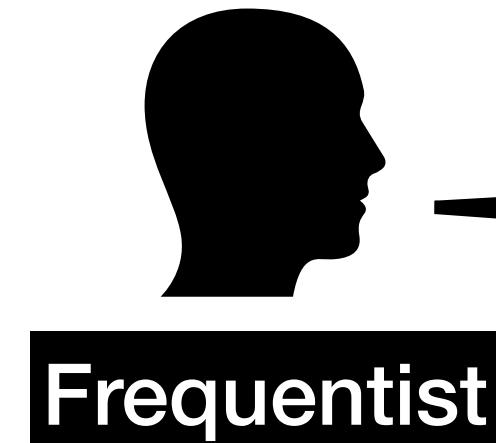
$$P(\theta | X) = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$



Bayesian

1. Introduction to bayesian statistics

Frequentist VS Bayesian point of view



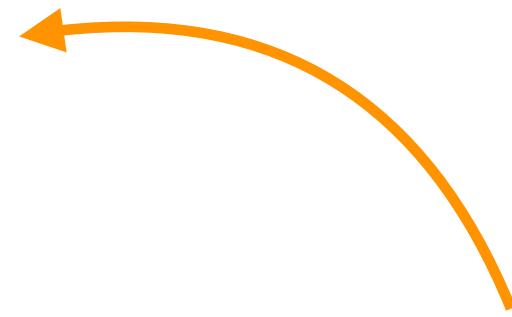
Frequentist

Maximum Likelihood Estimation

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

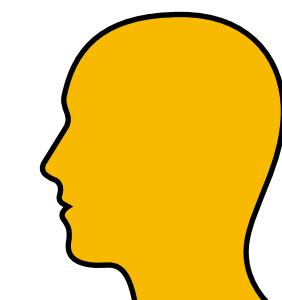
Problems in frequentist estimation :

- ~~Only works well if we have big data : $|X| \gg |\theta|$~~
- ~~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

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

Training phase

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

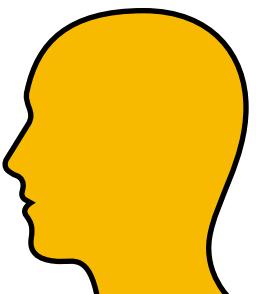


1. Introduction to bayesian statistics

Bayesian point of view : training

Bayes theorem

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



Bayesian

Training
phase

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

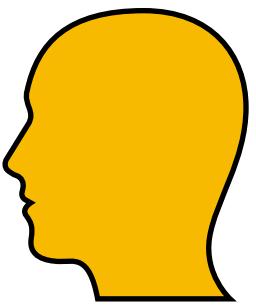
Can regularize your model when training on your data

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

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

Can regularize your model when training on your data

Inference phase

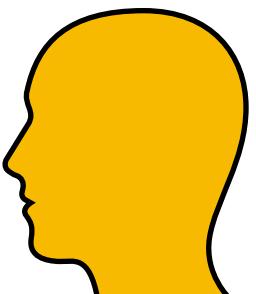
$$P(y_{new} | X_{new}, X_{train}, y_{train}) = \int P(y_{new} | X_{train}, \theta) \times P(\theta | X_{train}, y_{train}) d\theta$$

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

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

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

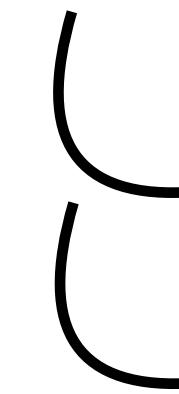
- **Nodes** : random variables
- **Links** : probabilistic relationships

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

Bayesian networks
(Directed graphical models)



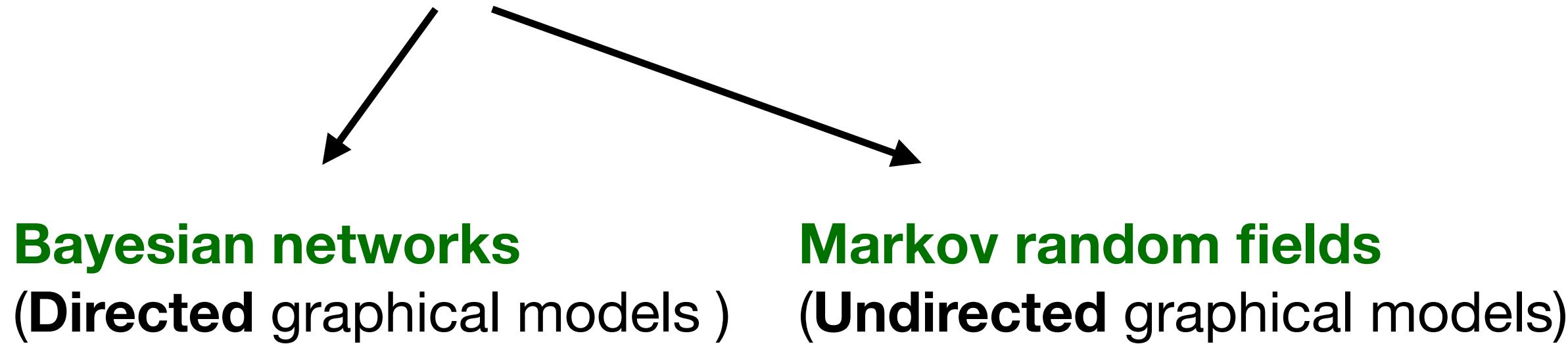
Nodes : random variables

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

2. Probabilistic model

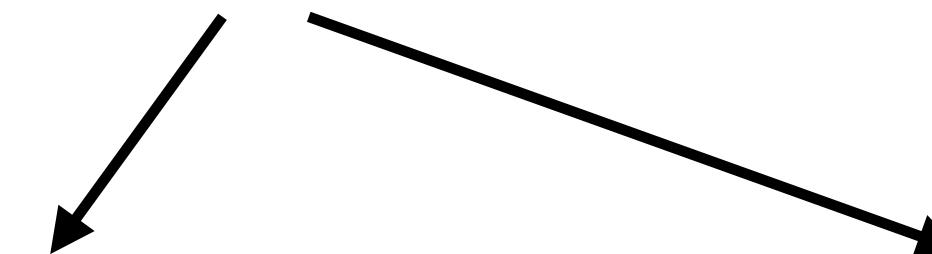
Probabilistic Graphical Model (PGM)

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

Bayesian networks
(Directed graphical models)

Markov random fields
(Undirected graphical models)

The focus of our course !



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

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

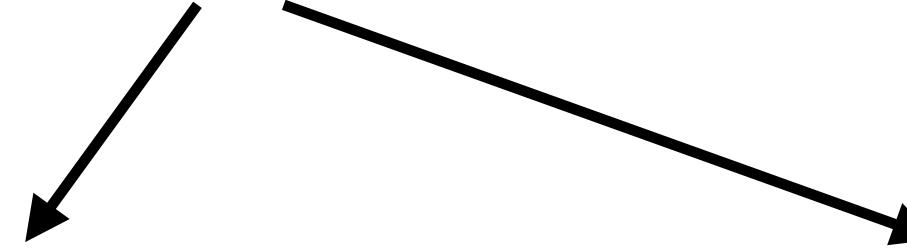
Bayesian networks
(Directed graphical models)

Markov random fields
(Undirected graphical models)

The focus of our course !

Model : joint probability over all variables

$$P(X_1, \dots, X_N) = \dots$$

- 
- **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

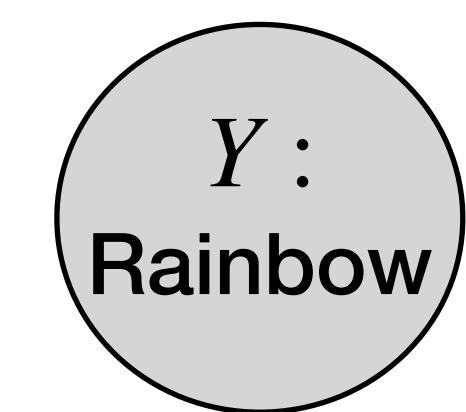
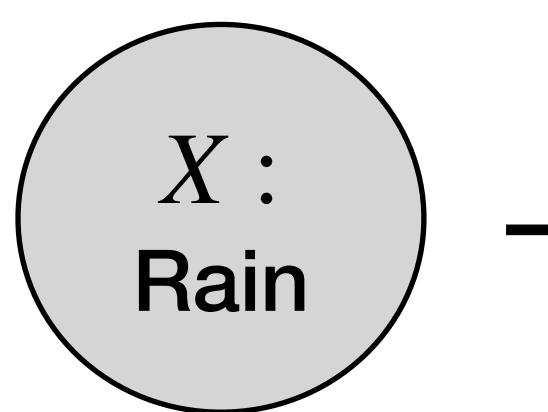
(Directed graphical models)

The focus of our course !

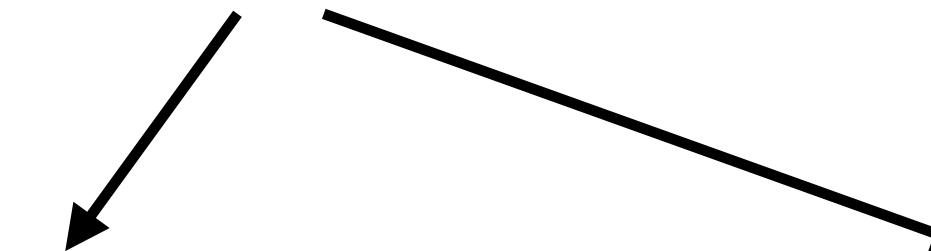
Model : joint probability over all variables

$$P(X_1, \dots, X_N) = \dots$$

Example :



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



- **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

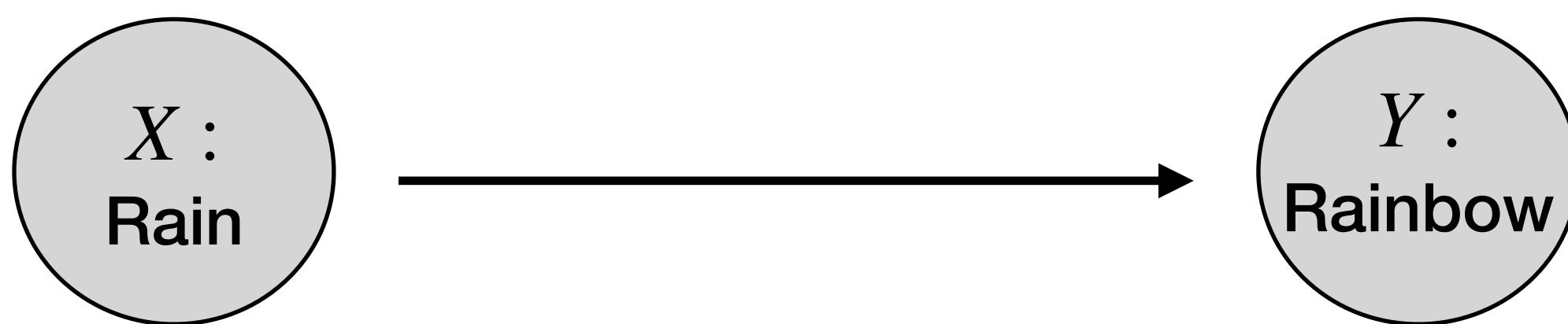
(Directed graphical models)

The focus of our course !

Model : joint probability over all variables

$$P(X_1, \dots, X_N) = \dots$$

Example :



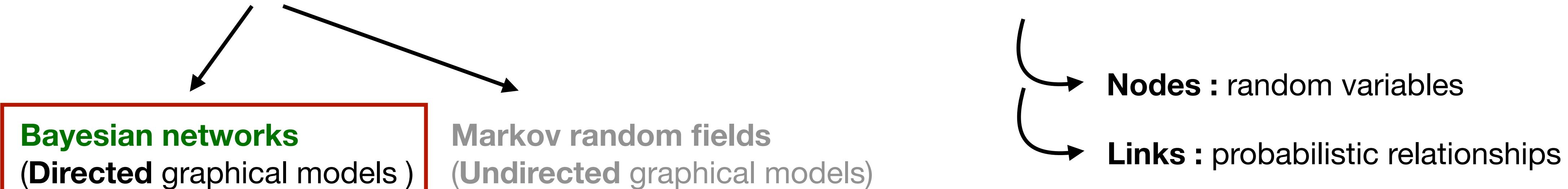
$$P(X, Y) = P(Y | X) \times P(X)$$

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

2. Probabilistic model

Probabilistic Graphical Model (PGM)

Probabilistic graphical models : analysis using diagrammatic representations of probability distributions

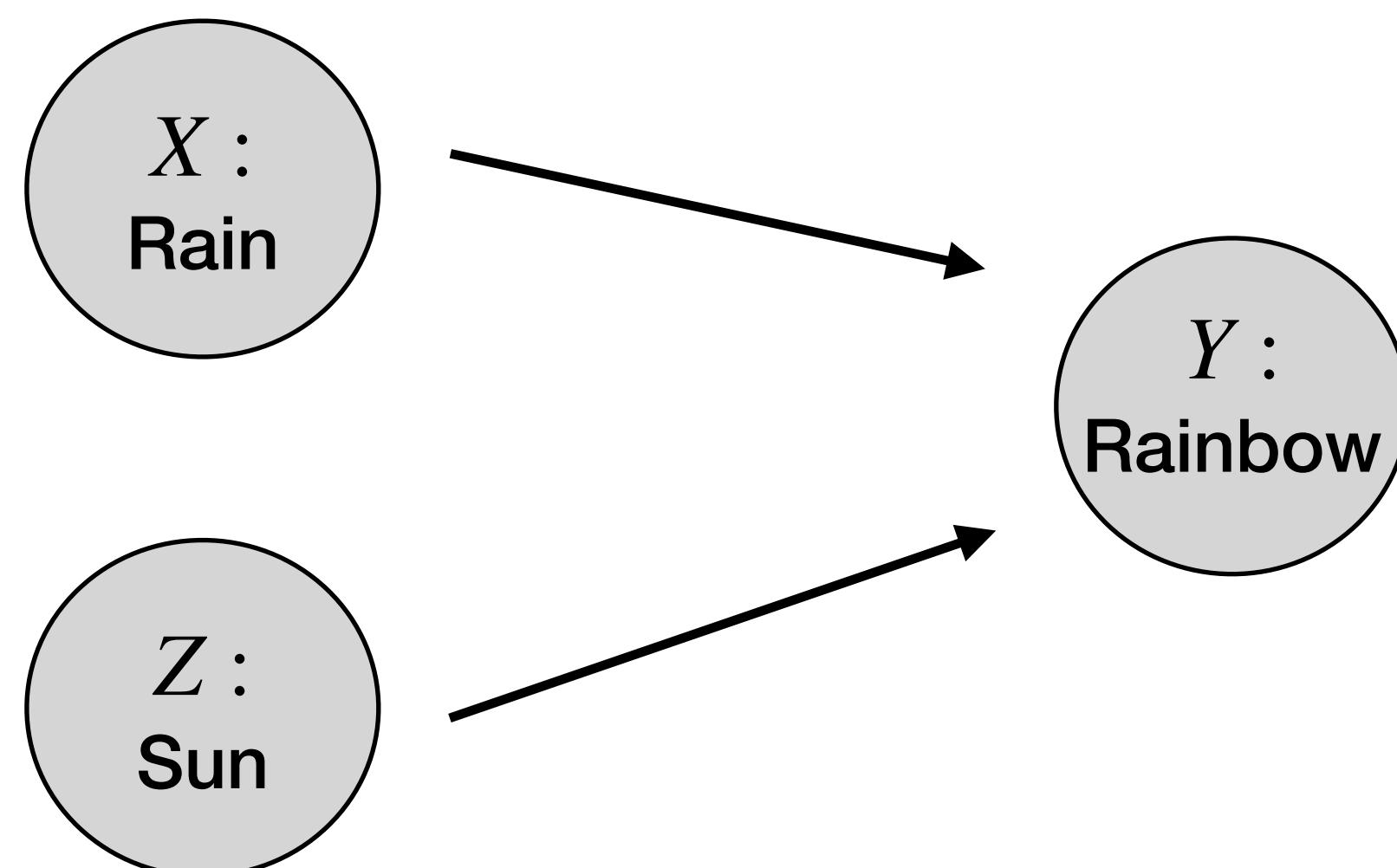


The focus of our course !

Model : joint probability over all variables

$$P(X_1, \dots, X_N) = \dots$$

Example :



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

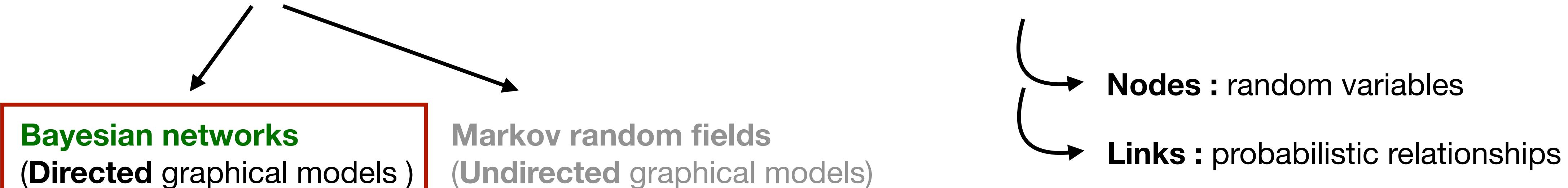
Features

X , Z conditionally
independent given Y

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

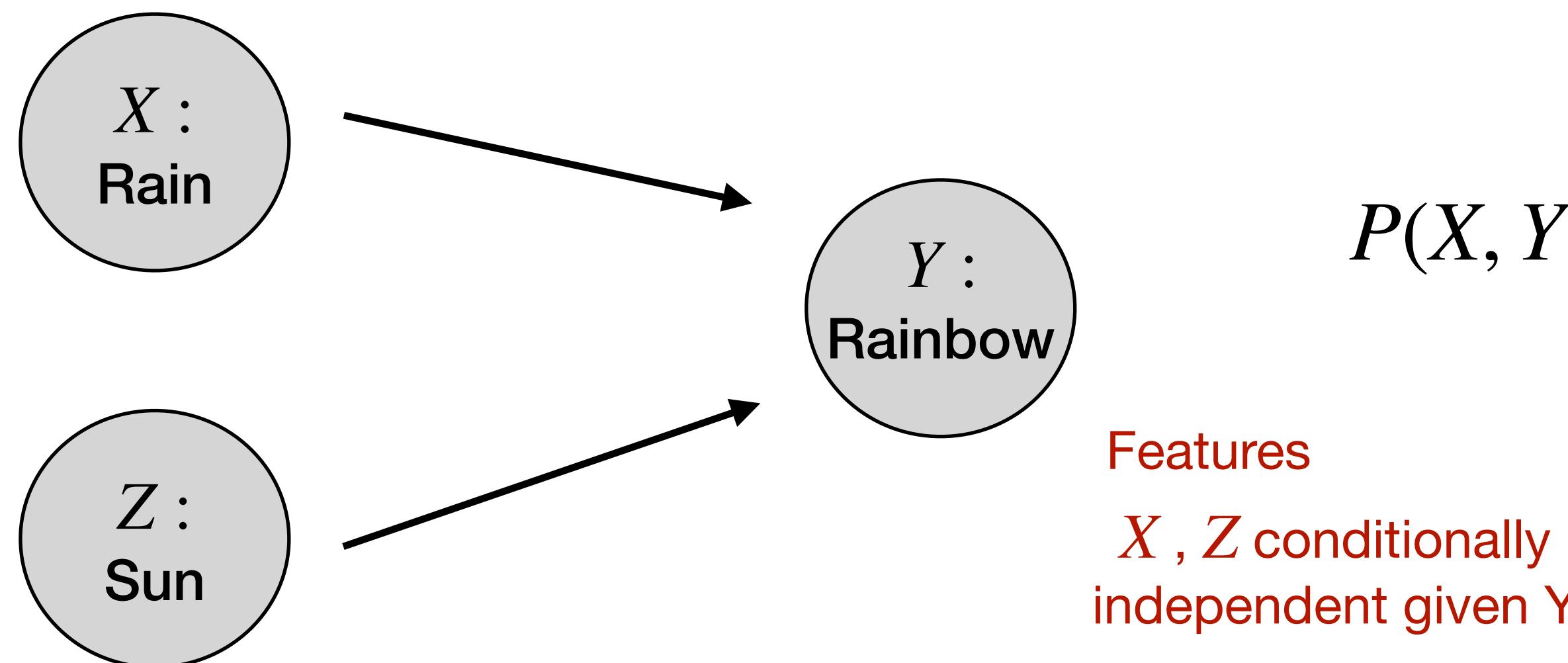


The focus of our course !

Model : joint probability over all variables

$$P(X_1, \dots, X_N) = \dots$$

Example :



Features

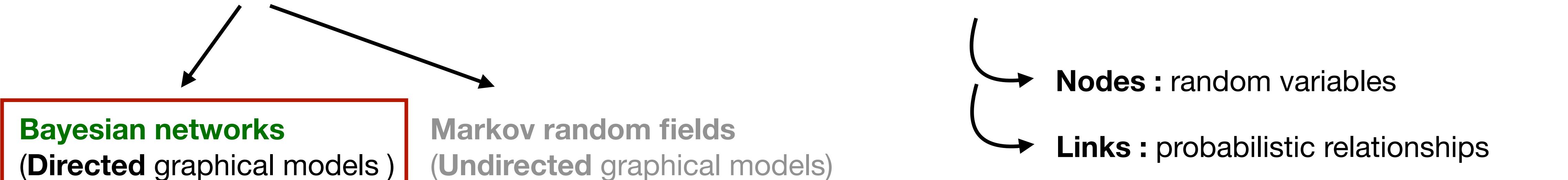
X , Z conditionally
independent given Y

$$\begin{aligned} P(X, Y, Z) &= P(Y|X, Z) \times P(X, Z) \\ &= P(Y|X, Z) \times P(X) \times P(Z) \end{aligned}$$

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

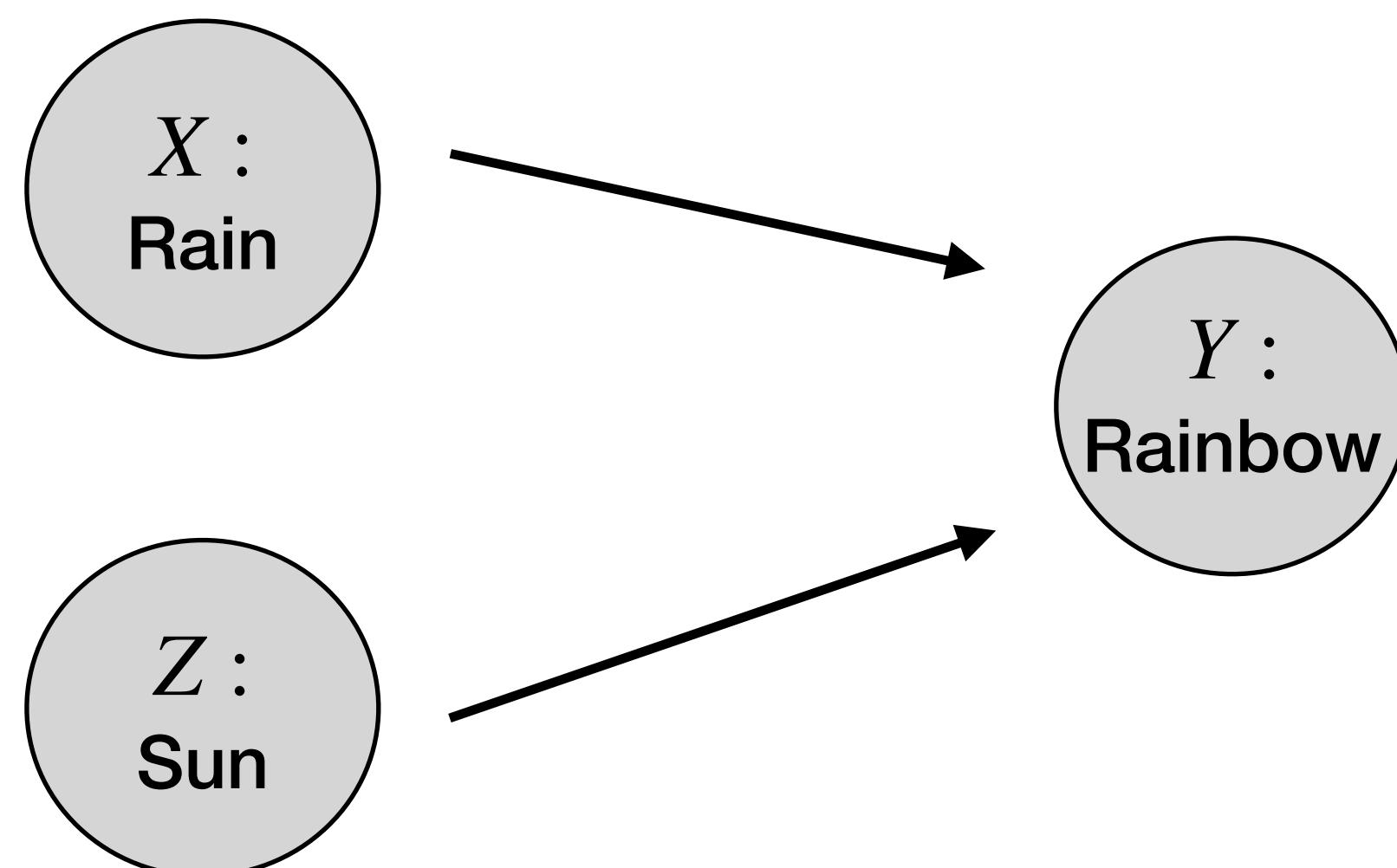


The focus of our course !

Model : joint probability over all variables

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

Example :



Features
 X, Z conditionally
independent given Y

$$\begin{aligned} P(X, Y, Z) &= P(Y|X, Z) \times P(X, Z) \\ &= P(Y|X, Z) \times P(X) \times P(Z) \end{aligned}$$

2. Probabilistic model

Probabilistic Graphical Model (PGM)

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

Bayesian networks
(Directed graphical models)

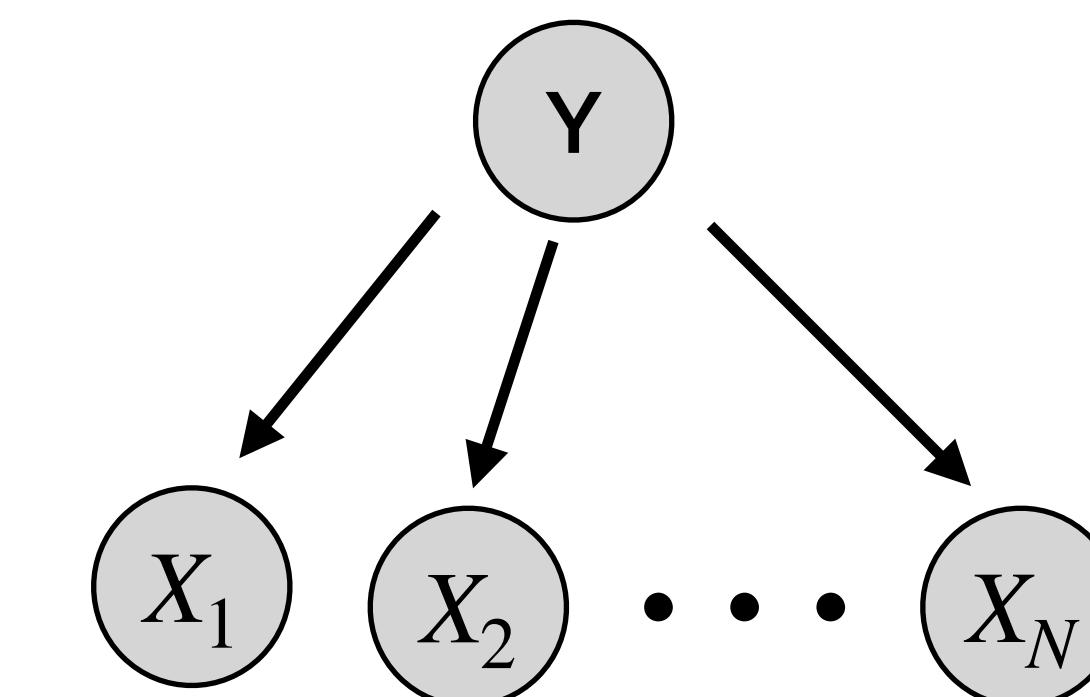
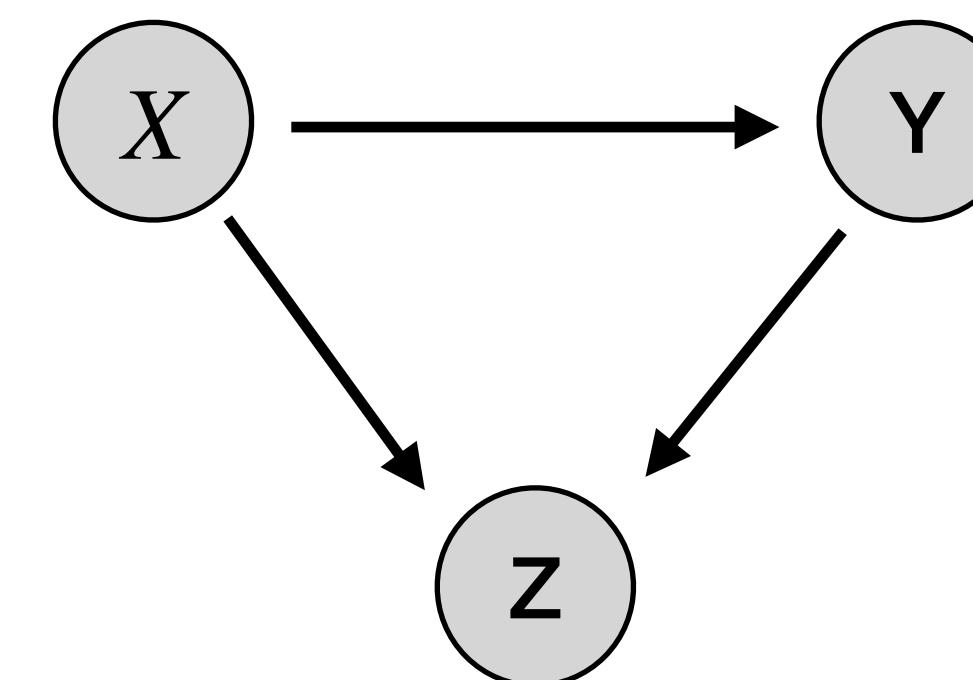
Markov random fields
(Undirected graphical models)

The focus of our course !

Model : joint probability over all variables

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

Example :



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

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

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

- **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)

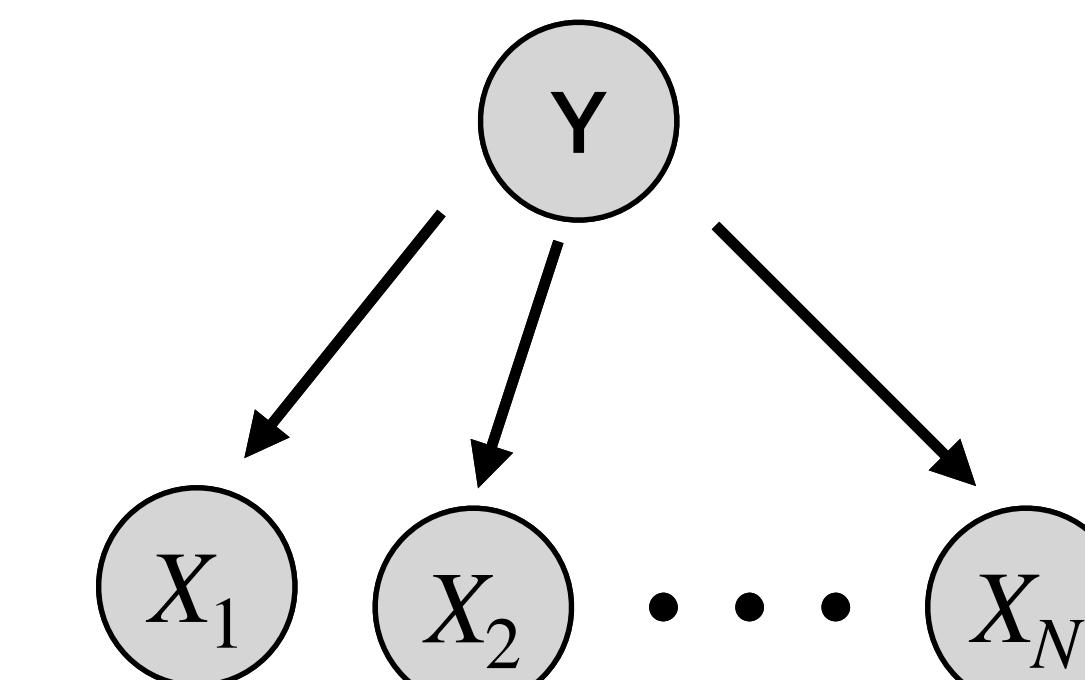
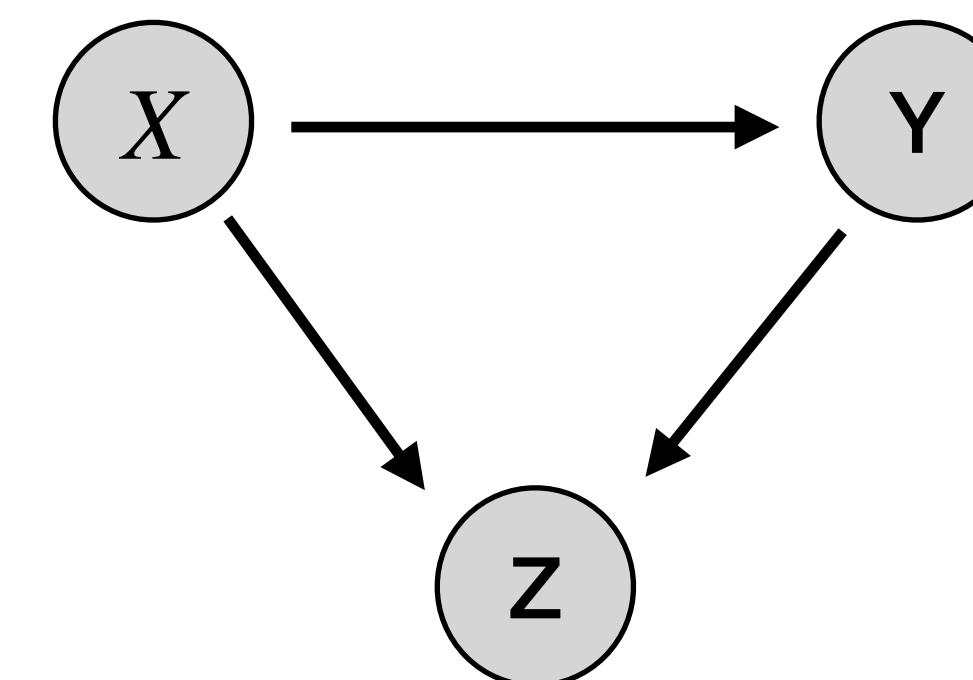
Markov random fields
(Undirected graphical models)

The focus of our course !

Model : joint probability over all variables

$$P(X_1, \dots, X_N) = \prod_{i=1, \dots, N} P(X_i | \text{parents}(X_i))$$

Example :



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

$$\begin{aligned} P(X, Y, Z) &= P(Z|X, Y) \times P(X, Y) \\ &= P(Z|X, Y) \times P(Y|X) \times P(X) \end{aligned}$$

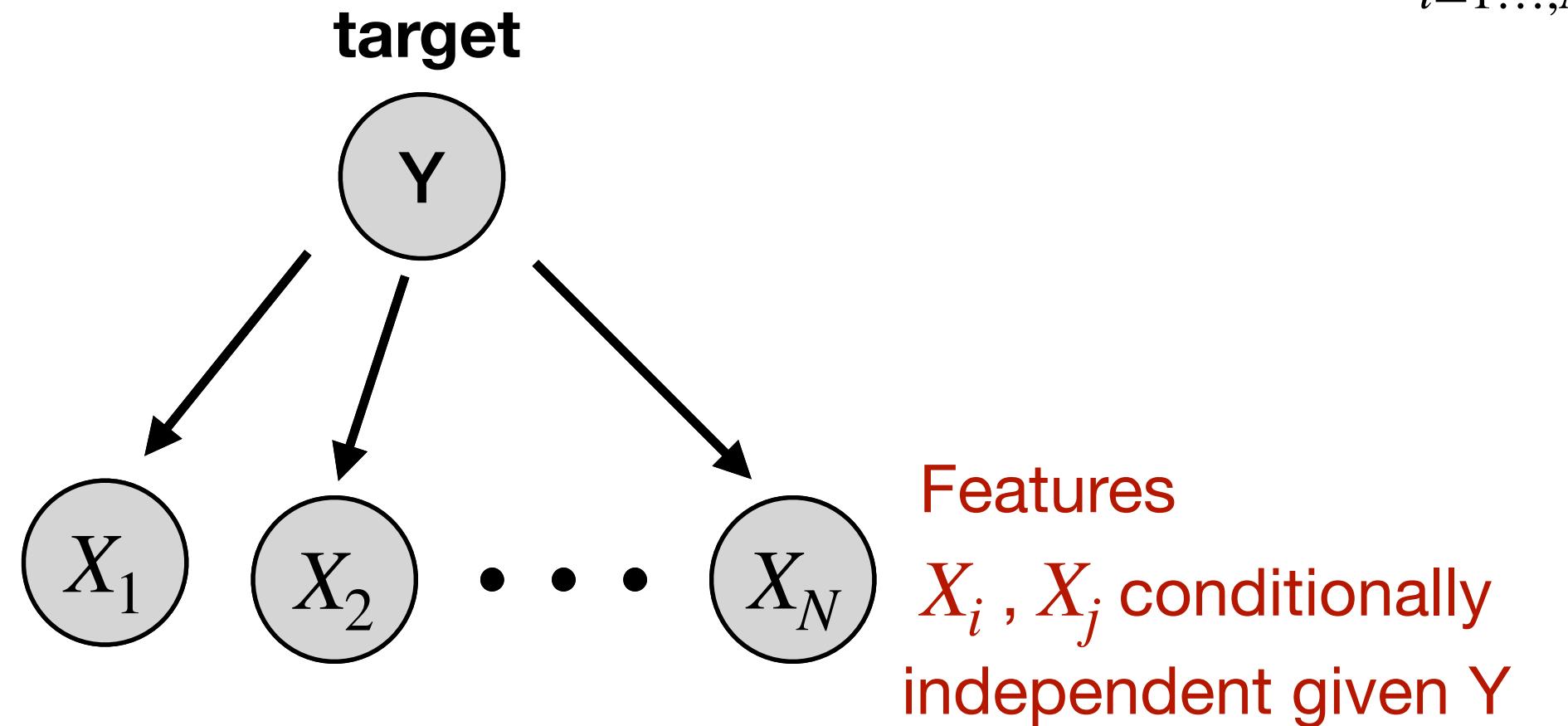
$$P(Y, X_1, \dots, X_N) = P(Y) \prod_{i=1, \dots, N} P(X_i | Y)$$

2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier

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

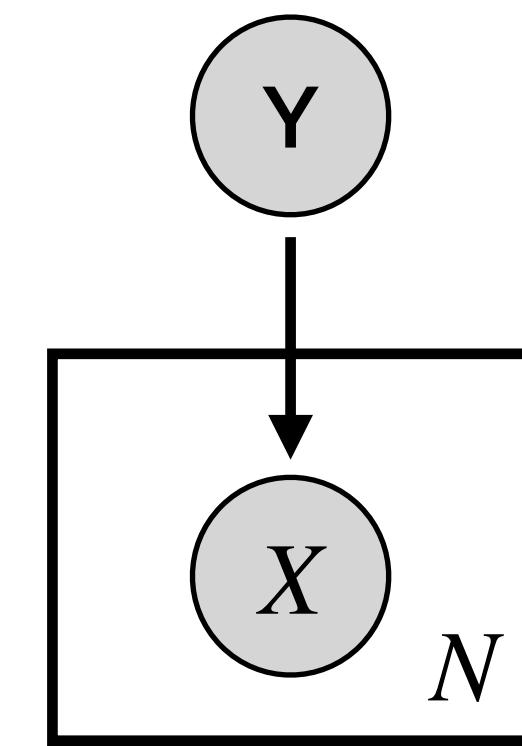
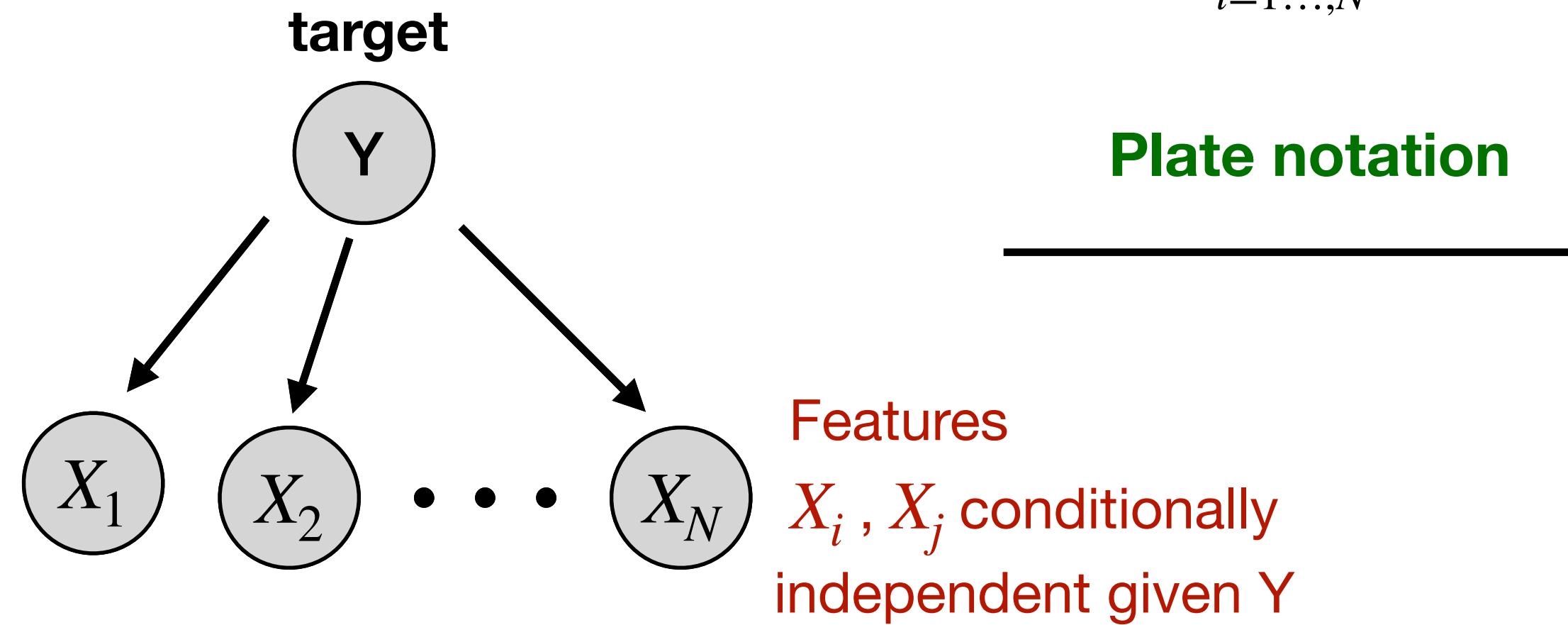


2. Probabilistic model

Plates and examples of probabilistic model

Naive Bayes Classifier

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



2. Probabilistic model

Frequentist linear regression

Reminder : Frequentist linear regression

$$x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$

Scalar notation :

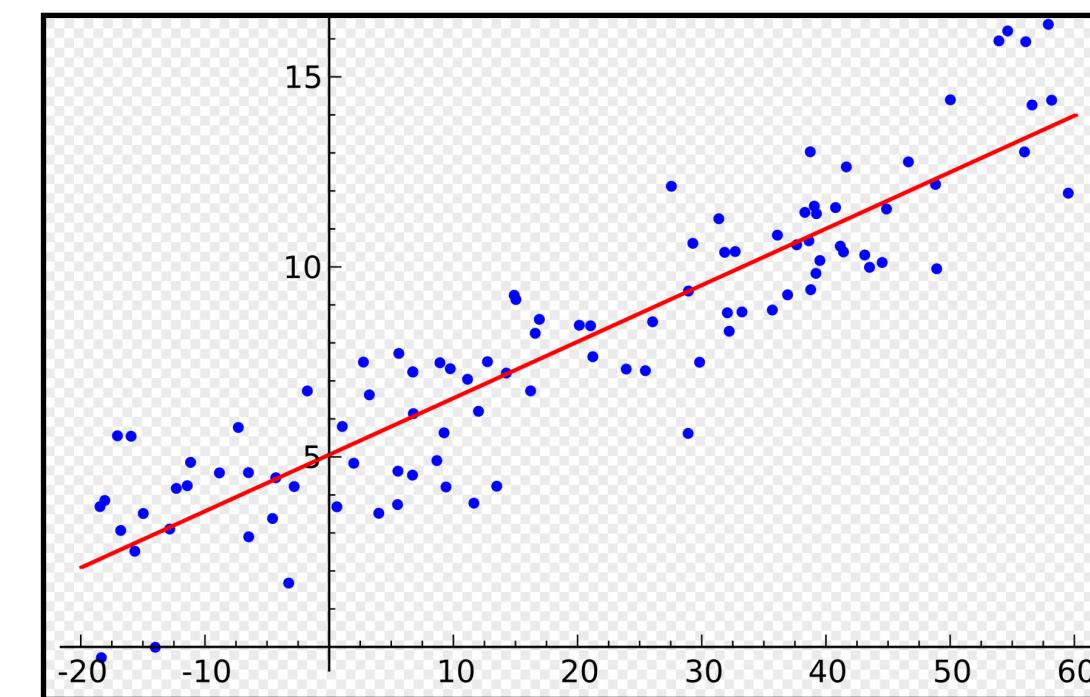
$$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

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

Matrix notation :

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

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



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

2. Probabilistic model

Frequentist linear regression

Reminder : Frequentist linear regression

$$x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$$

Scalar notation :

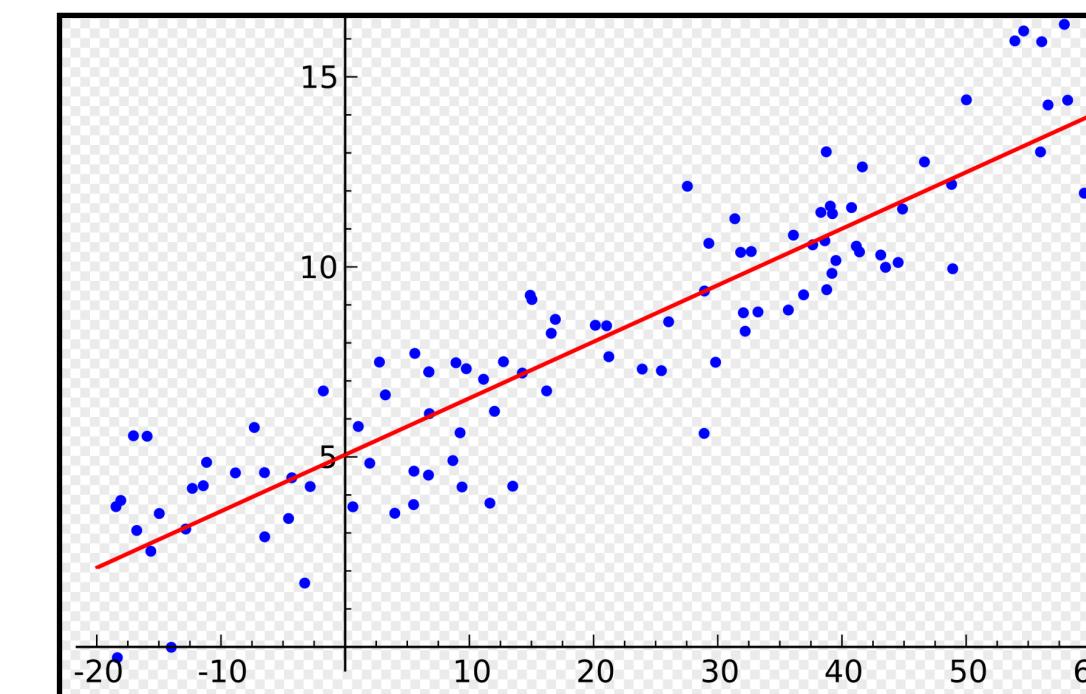
$$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

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

Matrix notation :

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

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



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

Proof : MLE for linear regression

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

$$\mathbf{y} \sim N(\mathbf{w}^T \mathbf{x}, \sigma^2)$$

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

$$y_1, \dots, y_n \text{ indep, } y_i \sim N(w^T x_i, \sigma^2) \text{ OR } y_i = w^T x_i + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2)$$

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

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

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

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

"vector
matrix
way"

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

$$l(\theta) = \log p(y | x, \theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x}^T \theta)^T (\mathbf{y} - \mathbf{x}^T \theta)$$

$$\frac{\partial l(\theta)}{\partial \theta} = 0 - \frac{1}{2\sigma^2} [\mathbf{0} - 2\mathbf{x}^T \mathbf{y} + 2\mathbf{x}^T \mathbf{x} \theta]$$

$$\frac{\partial l(\theta)}{\partial \theta} = 0 \Rightarrow \hat{\theta}_{MLE} = \theta = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}$$

2. Probabilistic model

Bayesian linear regression

Bayesian Linear regression

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

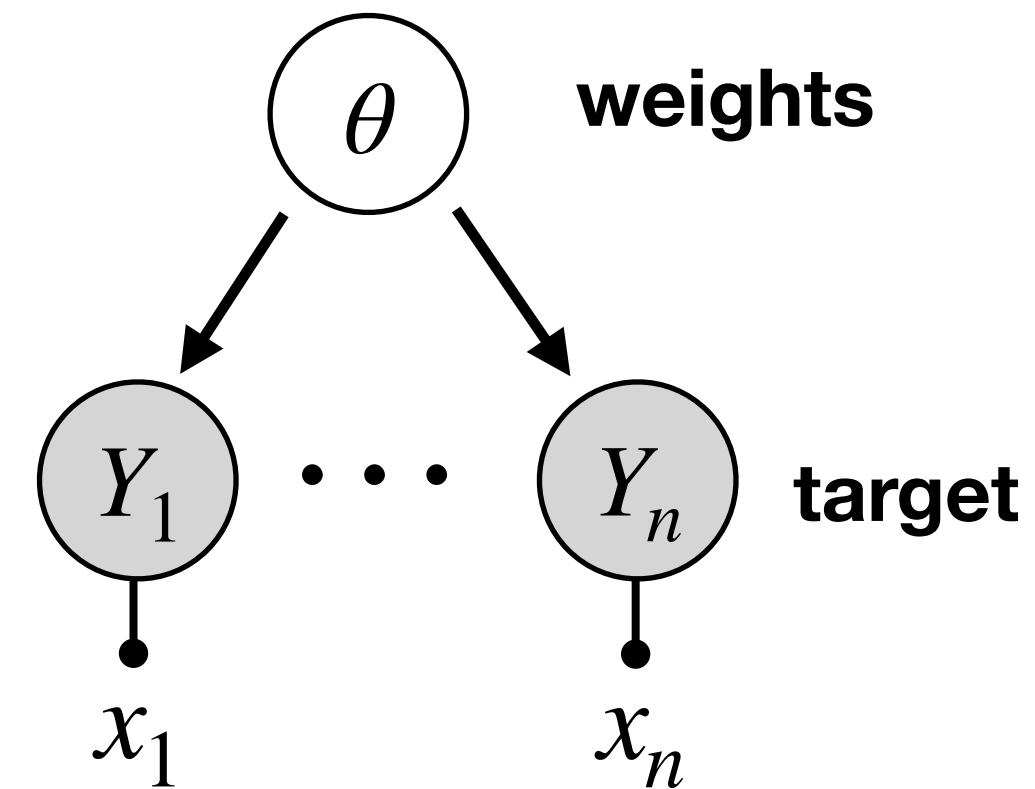
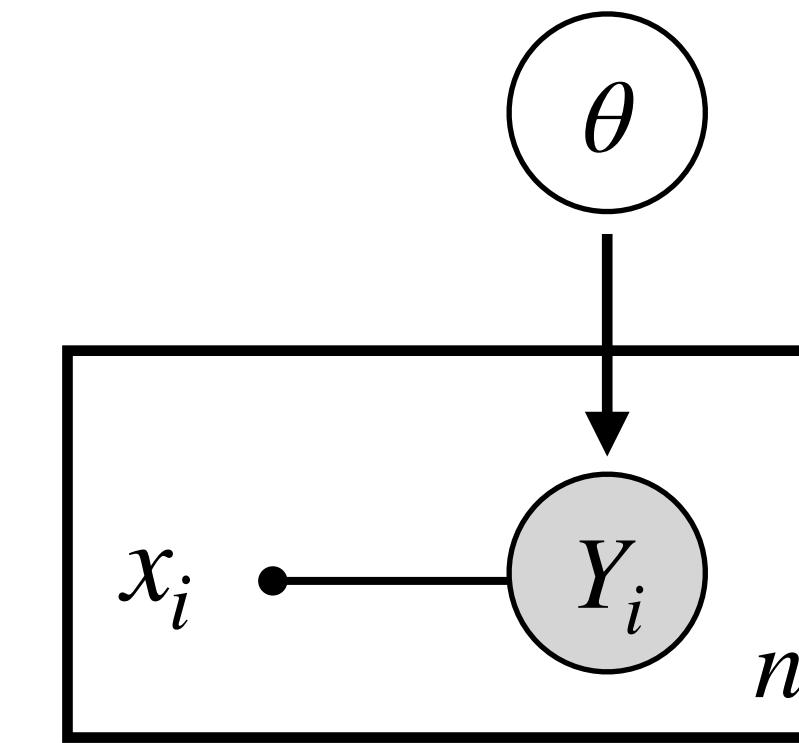


Plate notation

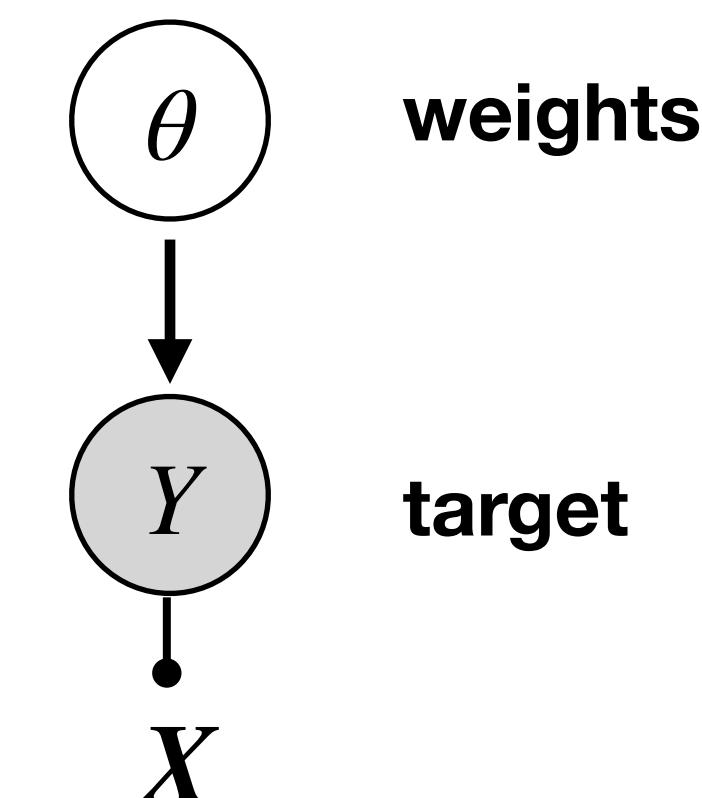


$$P(\theta, y_i | x_i) = P(y_i | \theta, x_i) \times P(\theta)$$

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

$$P(\theta) = \dots$$

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



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

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

$$P(\theta) = \dots$$

Legend :

— Fixed variable

○ Hidden (latent) r.v.

● Observed r.v.

2. Probabilistic model

Bayesian linear regression

Bayesian Linear regression

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

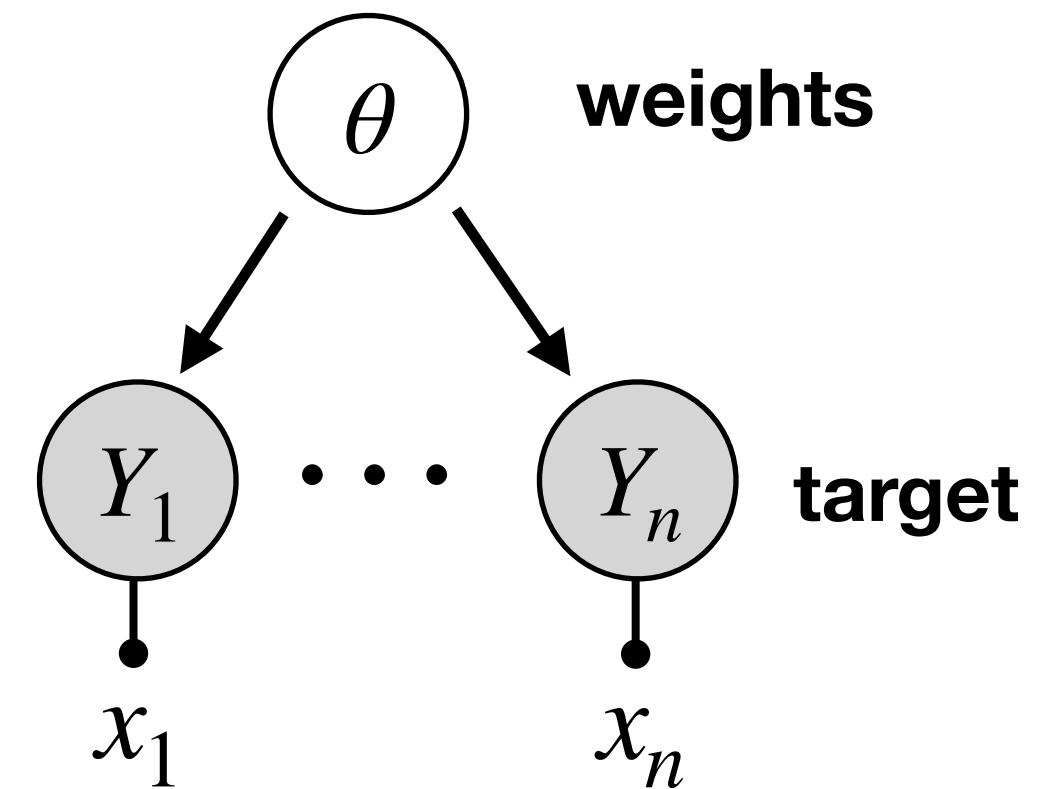
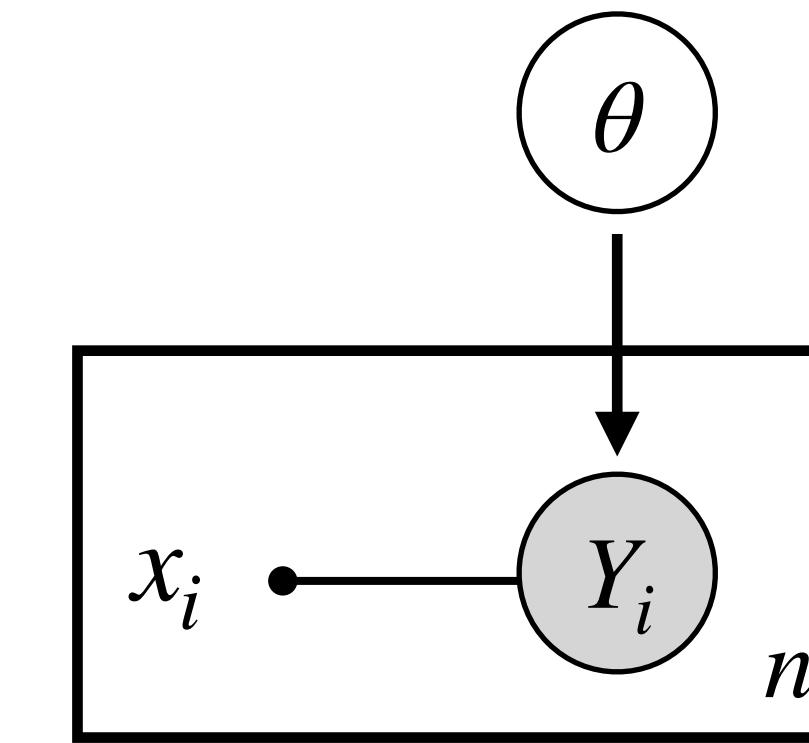


Plate notation

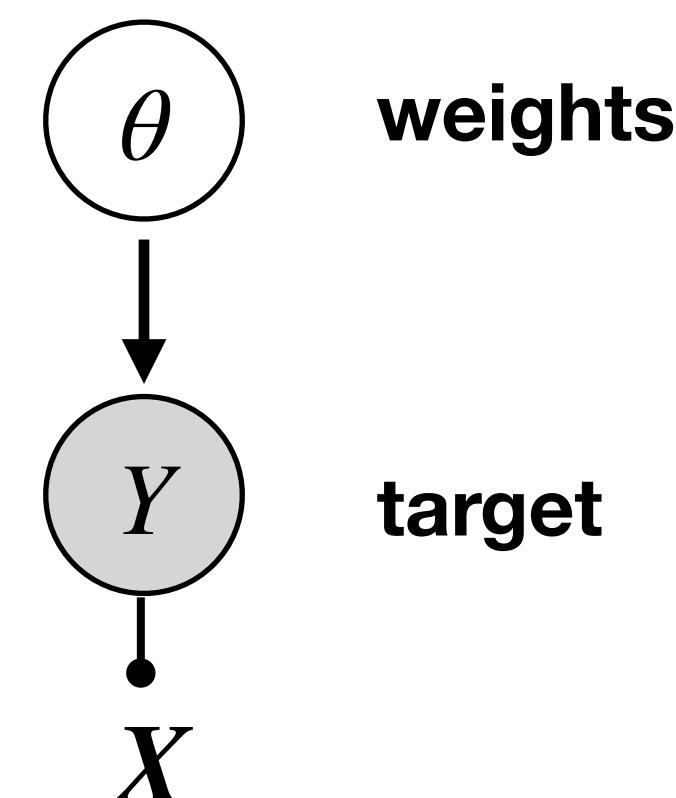


$$P(\theta, y_i | x_i) = P(y_i | \theta, x_i) \times P(\theta)$$

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

$$P(\theta) = \dots$$

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



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

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

$$P(\theta) = \dots$$

Legend :

— Fixed variable

○ Hidden (latent) r.v.

● Observed r.v.

2. Probabilistic model

Bayesian linear regression

Bayesian Linear regression

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

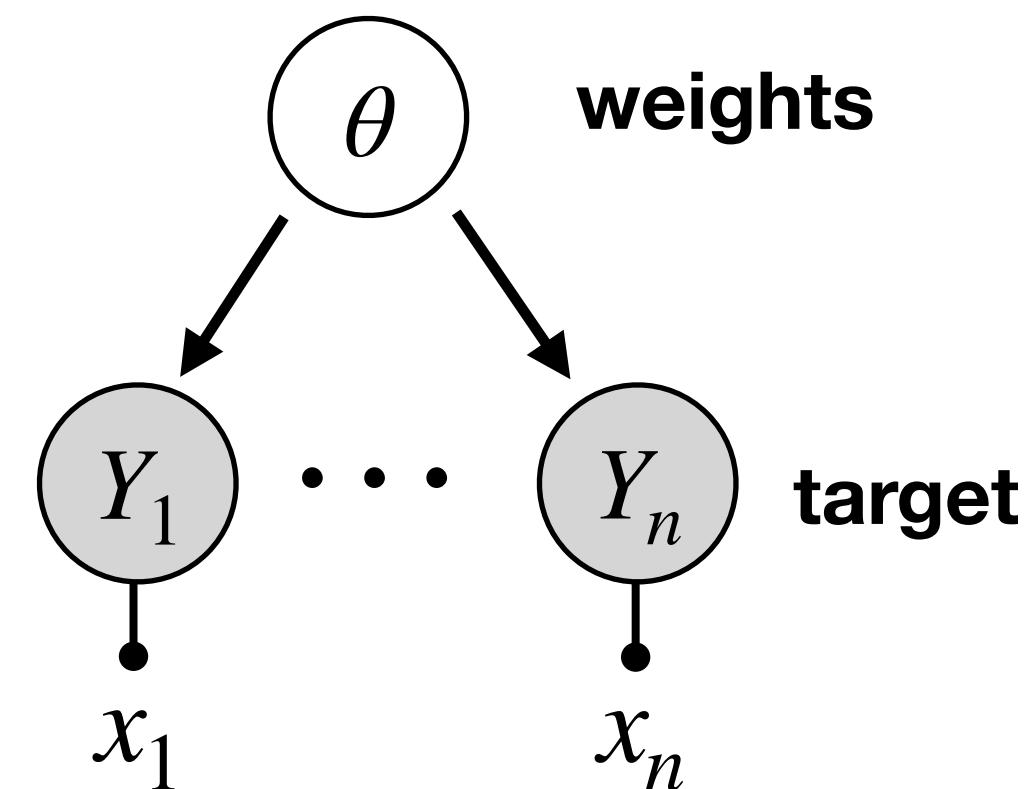
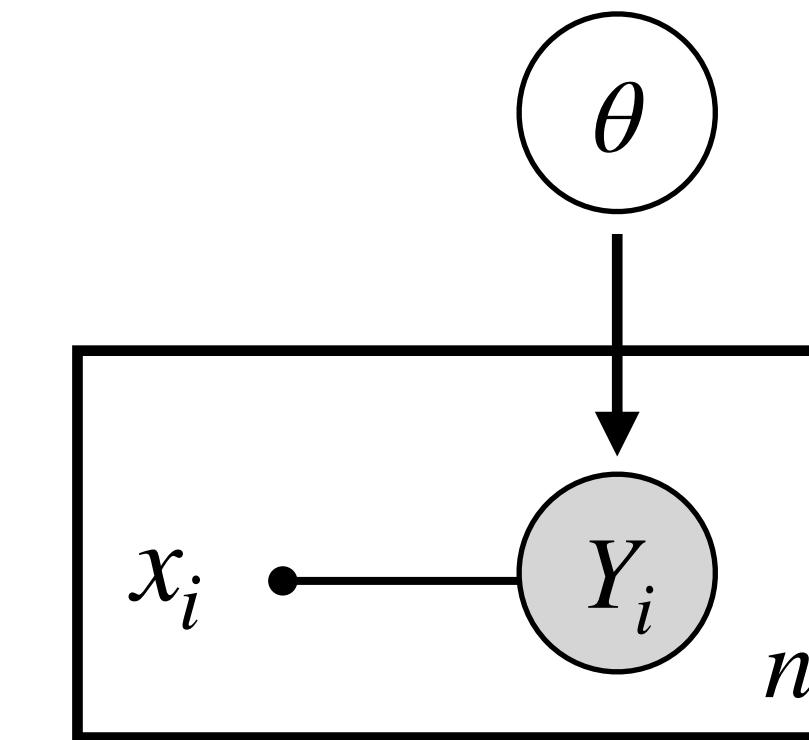


Plate notation

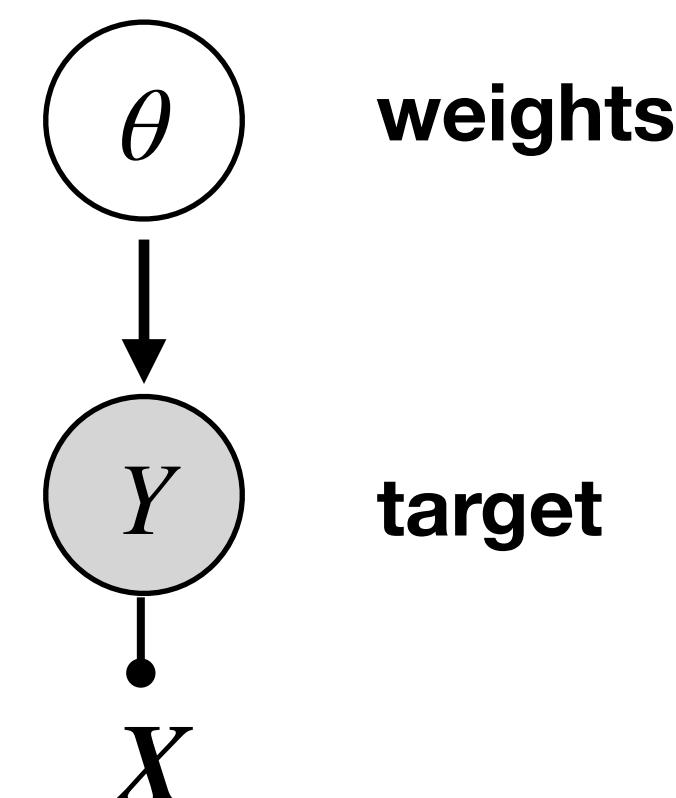


$$P(\theta, y_i | x_i) = P(y_i | \theta, x_i) \times P(\theta)$$

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

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

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



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

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

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

Legend :

| — Fixed variable

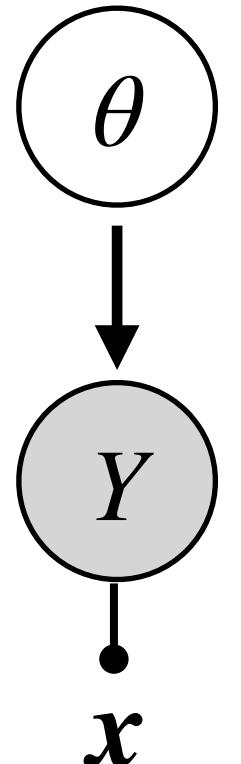
○ Hidden (latent) r.v.

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

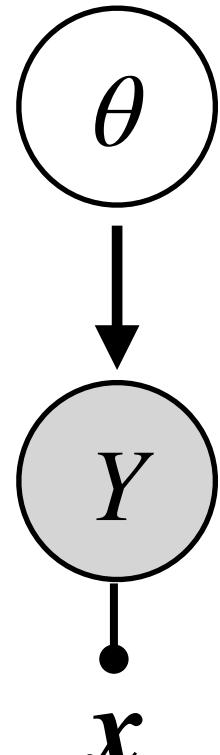
$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}(\theta)$$

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

2. Probabilistic model

Linear regression

Bayesian Linear regression



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

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

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

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

Frequentist linear regression

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

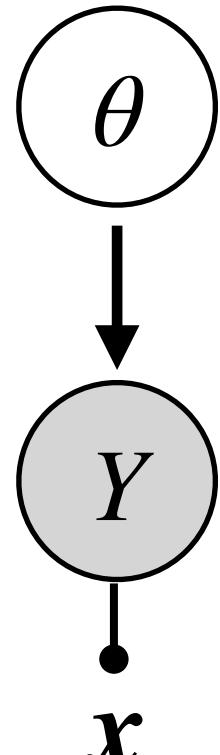
$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}(\theta)$$

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

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 :

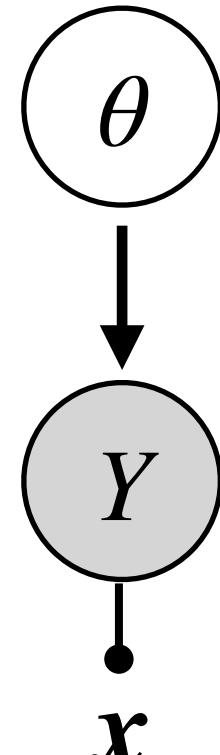
There exists $\lambda \in \mathbb{R}$ such that : $\arg \max_{\theta} P(\theta | X, y) = \arg \min_{\theta} \left\{ \|\theta^T X - y\|^2 + \lambda \|\theta\|^2 \right\}$

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

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 or left as an exercise



3

Analytical Inference

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

$$P(\theta | X) = \frac{P(X, \theta)}{P(X)} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

The diagram illustrates the formula for the posterior distribution:

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

The components are labeled as follows:

- Likelihood**: Fixed by model
- Prior**: Fixed by us
- Evidence**: Fixed by data
- Posterior**: The resulting distribution

3. Analytical Inference

Reminder of posterior distribution

Posterior distribution

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

Posterior

Fixed by model Likelihood Prior Fixed by us

Evidence
HARD TO COMPUTE

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

Fixed by data

3. Analytical Inference

Maximum a posteriori (MAP) : definition & remarks

Posterior distribution

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

Posterior

Fixed by model

Likelihood

Prior

Fixed by us

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

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

Posterior

Likelihood **Prior**

Evidence
HARD TO COMPUTE

Fixed by model

Fixed by us

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

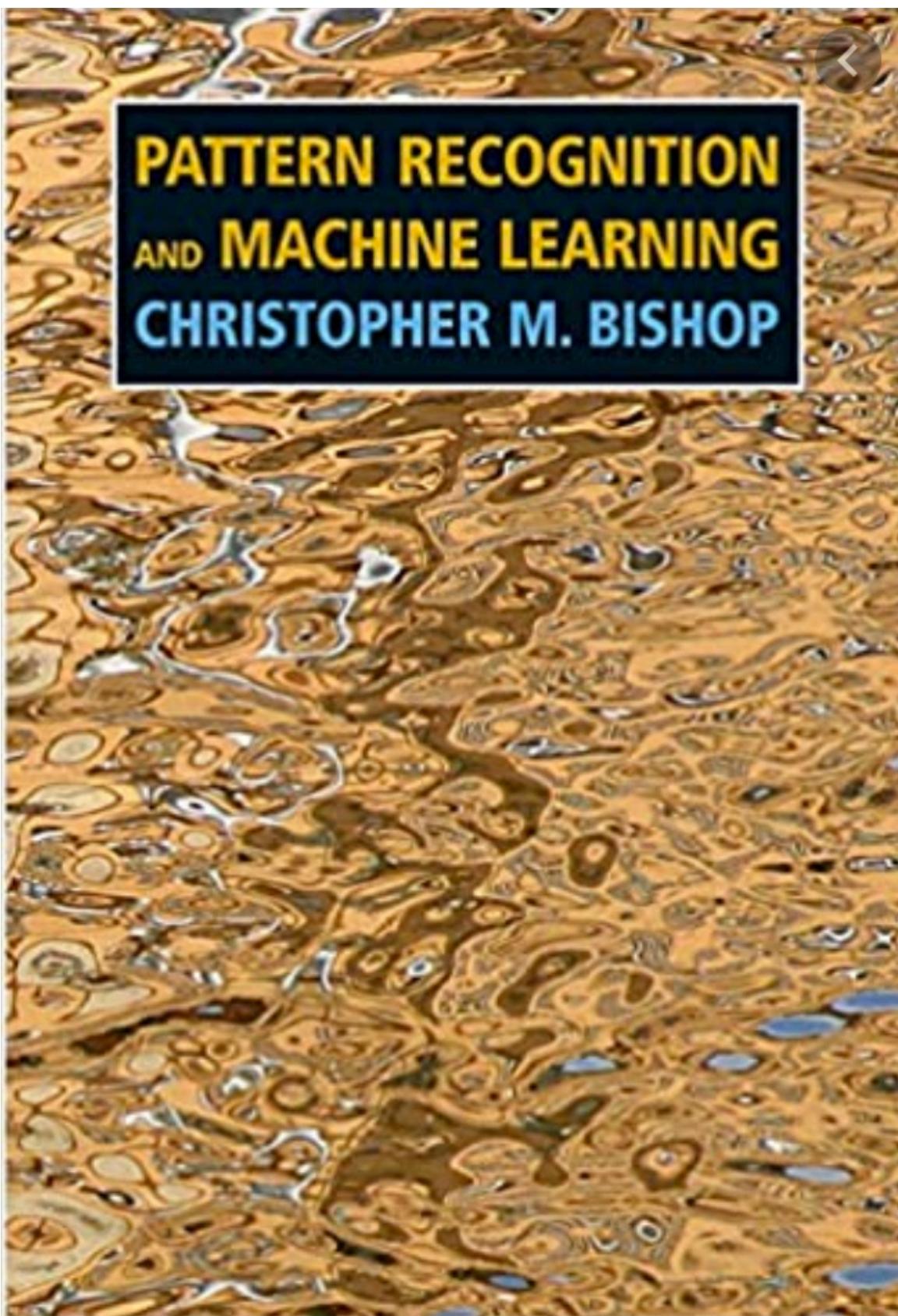
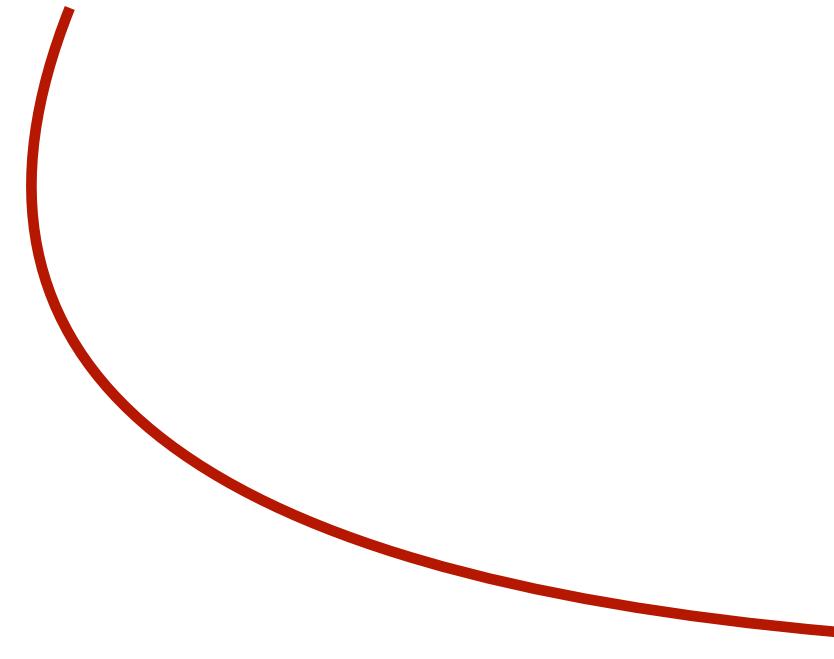
Limitations (among many others)

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

The diagram illustrates the formula for the posterior distribution:

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

The components are labeled as follows:

- Likelihood**: Fixed by model
- Prior**: Fixed by us
- Evidence**: Fixed by data

4. Conjugate distributions

Conjugate distributions : avoid computing evidence

Posterior distribution

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

Posterior

Fixed by model

Likelihood

Prior

Fixed by us

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

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

Posterior

Fixed by model

Likelihood

Prior

Fixed by us

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

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

Posterior

Likelihood Prior

Evidence

Fixed by model Fixed by us 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

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

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

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

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

4. Conjugate distributions

Limitations

Posterior distribution

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

Posterior

Fixed by model

Likelihood

Prior

Fixed by us

Evidence

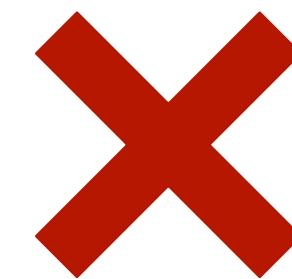
Fixed by data

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

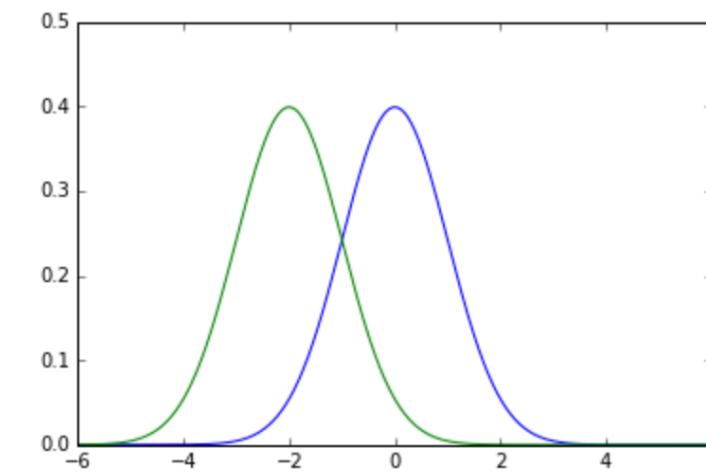


If you do these exercices before the next lecture,
you'll have **bonus points**:

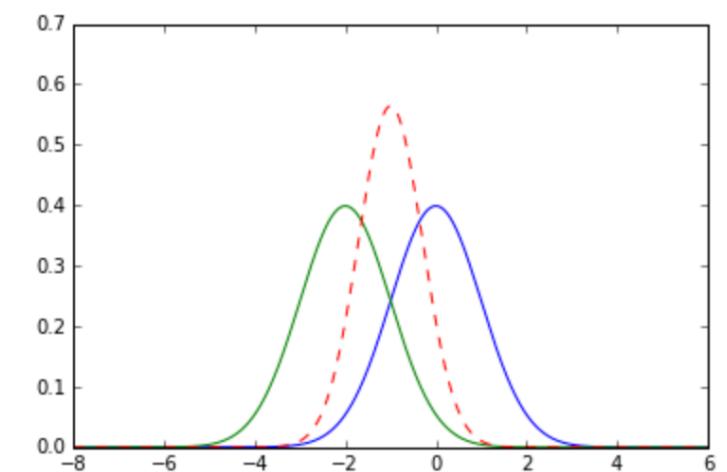
- 0.25 pts if only the reasoning is good
- 0.5 pts if only some of them are correct
- 0.75 pts if most of them are correct
- 1 pts if all of them are correct

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

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



→ **pointwise product**



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

show the following equation

$$\Gamma(\gamma | \alpha_{posterior}, \beta_{posterior}) = P(\gamma | x) = \frac{\mathcal{N}(x | \mu, \gamma^{-1}) \times P(\gamma)}{P(x)} \quad \Gamma(\gamma | \alpha_{prior}, \beta_{prior})$$
$$\Gamma(\gamma | \alpha_{prior} + 1/2, \beta_{prior} + (x - \mu)^2/2)$$

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

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

show the following equation

$$B(\theta | \alpha_{posterior}, \beta_{posterior}) = P(\theta | x) = \frac{Ber(x | \theta) \times P(\theta)}{P(x)} \quad B(\theta | \alpha_{prior}, \beta_{prior})$$
$$B(\theta | n_1 + \alpha_{prior}, n_0 + \beta_{prior})$$

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

!

Road map

Bayesian statistics



1

Bayesian perspective :

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

Posterior distribution

θ parameters

X observations

Exemple :
Naive Bayes classifier,
Linear regression,

Likelihood Prior distribution

Evidence

Hard to compute !

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

Conjugate distribution

Pros :
- exact posterior

Cons :
- conjugate prior
maybe inadequate

Latent variable models

2

Variational Inference

3

Markov Chain Monte Carlo

4

Oral presentations

5