HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Introduction to

Machine Learning and Data Mining

IT3190

Lecture: Probabilistic models

ONE LOVE. ONE FUTURE.

Contents

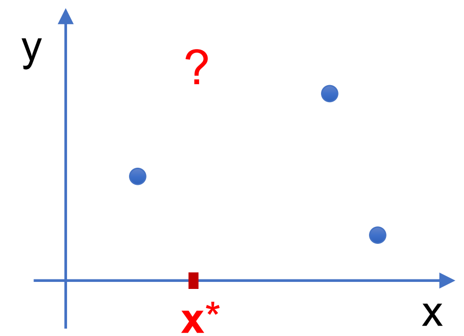
- Lecture 1: Introduction to Machine Learning & Data Mining
- Lecture 2: Data crawling and pre-processing
- Lecture 3: Linear regression
- Lecture 4+5: Clustering
- Lecture 6: Decision tree and Random forest
- Lecture 7: Neural networks
- Lecture 8: Support vector machines
- Lecture 9: Performance evaluation
- **Lecture 10: Probabilistic models**
- Lecture 11: Basics of data mining
- Lecture 12: Association rule mining
- Lecture 13: Regularization and advanced topics

Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
  (suy diễn từ dữ liệu thường không chắc chắn)
- Probability theory: *model uncertainty* instead of ignoring it!
- Inference or prediction can be done by using probabilities.
- Applications: Machine learning, Data Mining, Computer Vision, NLP, Bioinformatics, …
- The goal of this lecture
  - Overview about probabilistic modeling
  - Key concepts
  - Application to classification

Data

- Let **D** = $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_M, y_M)\}$ be a dataset with *M* instances.
  - Each $\mathbf{x}_i$ is a vector in an *n*-dimensional space, e.g., $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$. Each dimension represents an attribute.
  - y is the output (response), univariate
- **Prediction:** given data **D**, what can we say about y^* at an unseen input $\boldsymbol{x}^*$?



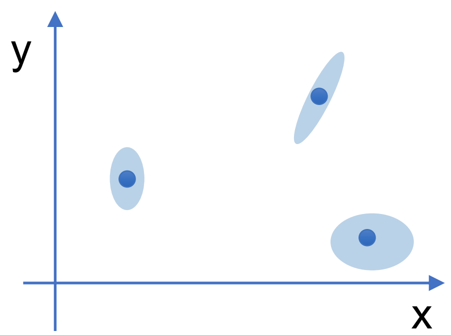
- To make predictions, we need to make **assumptions**
- A **model H** (mô hình) encodes these assumptions, and often depends on some parameters $\boldsymbol{\theta}$, e.g.,

$$y = f(\boldsymbol{x}|\boldsymbol{\theta})$$

- **Learning** (estimation) is to find an **H** from a given **D**.

Uncertainty

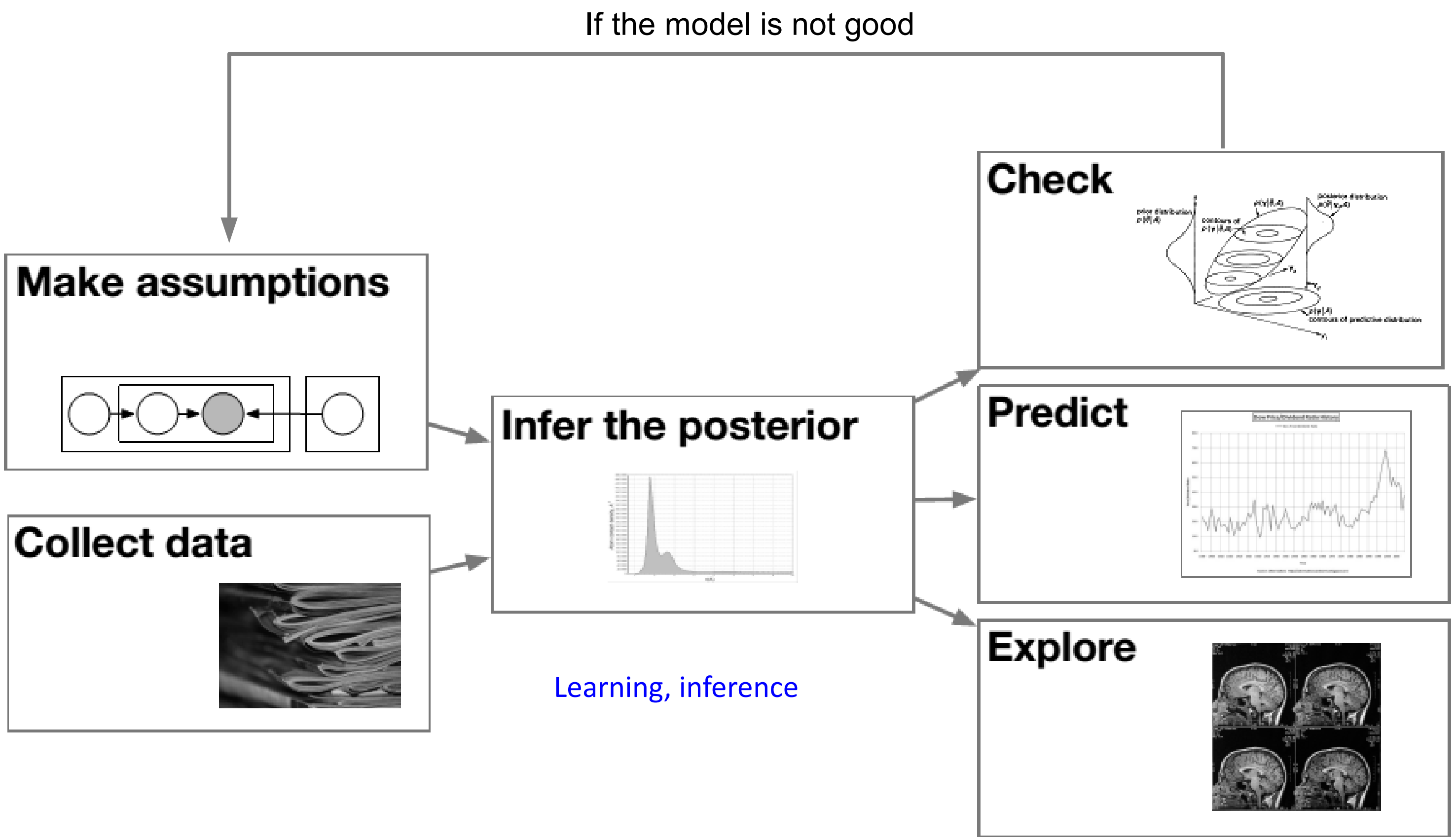
- Uncertainty apprears in any step
  - Measurement uncertainty (**D**)
  - Parameter uncertainty (**θ**)
  - Uncertainty regarding the correct model (**H**)
- Measurement uncertainty
  - Uncertainty can occur in both inputs and outputs.



- How to represent uncertainty?

→ **Probability theory**

The modeling process



Model making

Learning, inference

[Blei, 2012]

Basics of Probability Theory

HUST

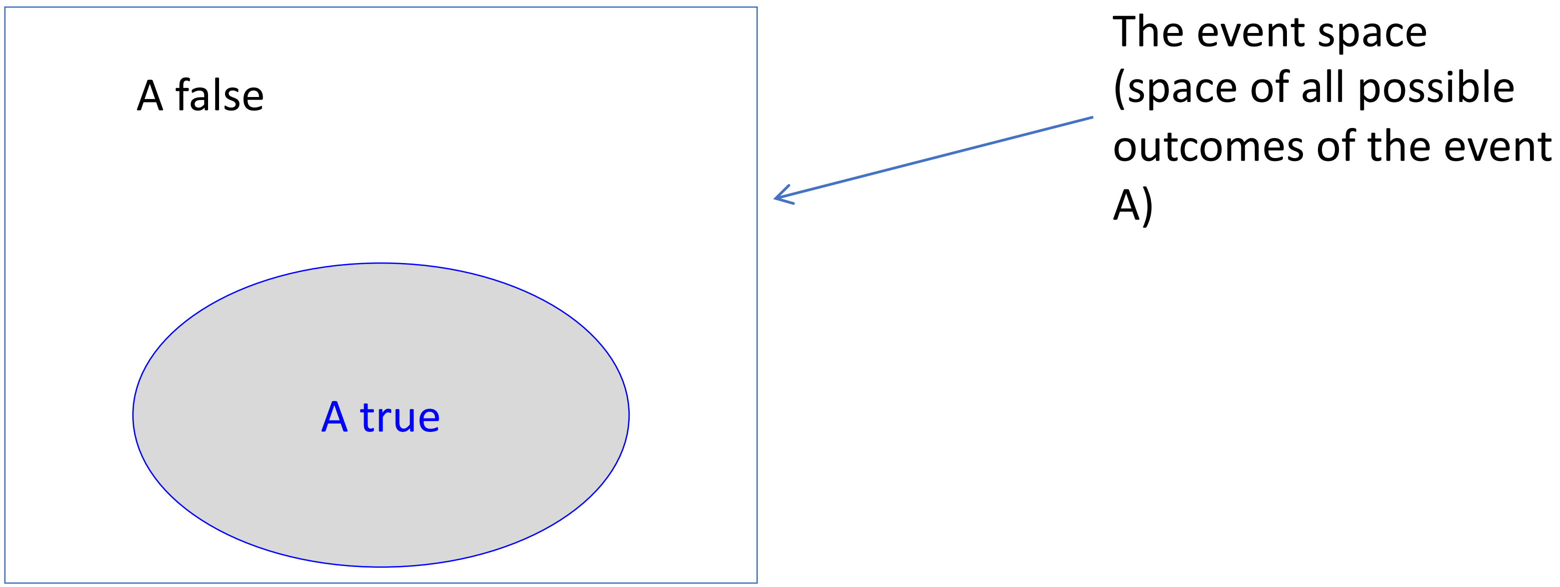
hust.edu.vn fb.com/dhbkhn

Basic concepts in Probability Theory

- Assume we do an experiment with random outcomes, e.g., tossing a die.
- *Space S of outcomes:* the set of all possible outcomes of an experiment
  - Ex: S = {1, 2, 3, 4, 5, 6} for tossing a die
- *Event E:* a subset of the outcome space S.
  - Ex: E = {1} the event that the die appears 1.
  - Ex: E = {1, 3, 5} the event that the die appears odd.
- *Space W of events:* the space of all possible events
  - Ex: W contains all possible tosses
- *Random variable:* represents a random event, and has an associated probability of occurrence of that event.

Probability visualization

- Probability represents the likelihood/possibility that an event A occurs.
  - Denoted by P(A).
- P(A) is the proportion of the subspace that A is true.



Binary random variables

- A binary (boolean) random variable can receive only value of either *True* or *False*.
- Some axioms:
  - $0 \leq P(A) \leq 1$
  - P(true)= 1
  - P(false)= 0
  - $P(A \text{ or } B) = P(A) + P(B) - P(A, B)$
- Some consequences:
  - P(not A) = P(~A)= 1 - P(A)
  - P(A)= P(A, B) + P(A, ~B)

Multinomial random variables

- A multinomial random variable can receive one from K possible values of $\{v_1, v_2, \dots, v_k\}$.

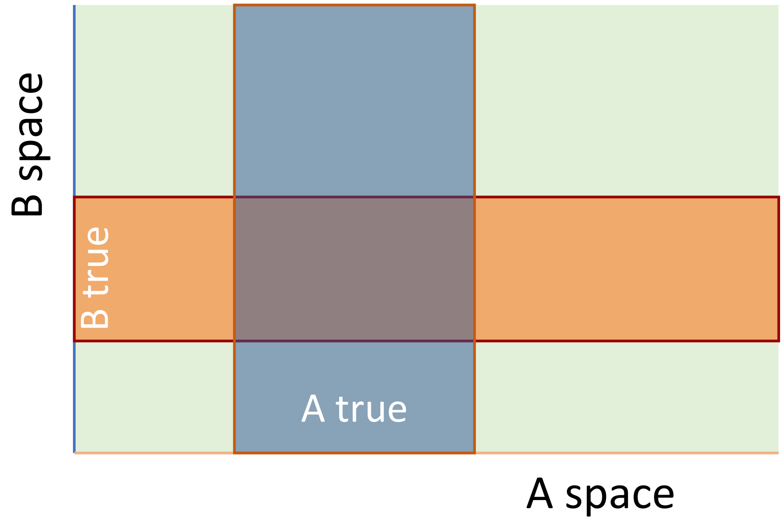
$$P\left(A = v_i, A = v_j\right) = 0 \text{ if } i \neq j$$

$$P\left(\bigcup_{n=1}^{m} (A = v_n)\right) = \sum_{n=1}^{m} P(A = v_n)$$

$$P\left(\bigcup_{n=1}^{k} (A = v_n)\right) = \sum_{n=1}^{k} P(A = v_n) = 1$$

Joint probability (1)

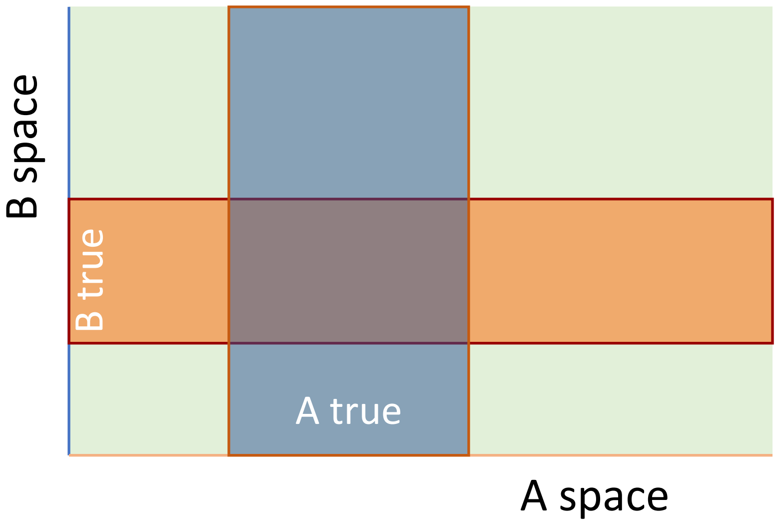
- Joint probability:
  - The possibility of A and B that occur simutaneously.
  - P(A,B) is the proportion of the space in which both A and B are true.
- Ex:
  - A: I will play football tomorrow.
  - B: John will not play football.
  - P(A,B): the probability that I will but John will not play football tomorrow.



Joint probability (2)

- Denote S_A the space of A.
- Denote S_B the space of B.
- Denote S_{AB} the space of (A, B).

$$S_{AB} = S_A \times S_B$$



- Then:

$$P(A,B) = |T_{AB}| / |S_{AB}|$$

  - T_{AB} is the space in which both A and B are true.
  - $|X|$ denotes the volumn of the set X.

Conditional probability (1)

- Conditional probability:
  - P(A|B): the possibility that A happens given that B has already occurred.
  - P(A|B) is the proportion of the space in which A occurs, knowing that B is true.
- Ex:
  - A: I will play football tomorrow.
  - B: it will not rain tomorrow.
  - P(A|B): the probability that I will play football, provided that it will not rain tomorrow.
- What is different between joint and conditional probabilities?

Conditional probability (2)

- We have:

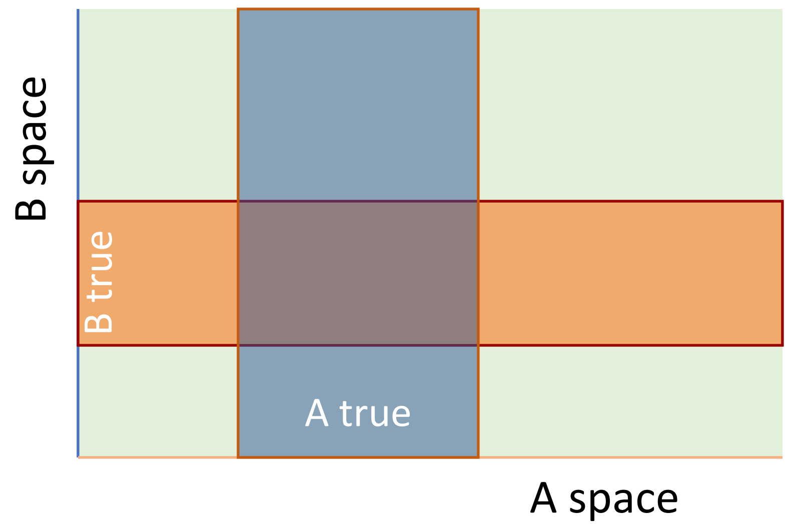
$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

- Some consequences:

$$P(A,B) = P(A|B) \,.\, P(B)$$

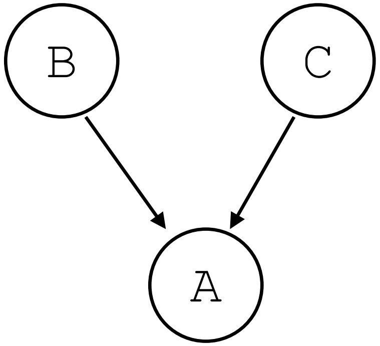
$$P(A|B) + P(\sim A|B) = 1$$

$$\sum_{i=1}^{k} P(A = v_i \mid B) = 1$$



Conditional probability (3)

- P(A|B, C) shows the probability of A given that B and C already has occurred.
- Ex:
  - A: I will wander over the near river tomorrow morning.
  - B: it will be very nice tomorrow morning.
  - C: I will wake up early tomorrow morning.
  - P(A|B, C): the probability that wander over the near river, provided that it will be very nice and I will wake up early tomorrow morning.



P(A|B,C)

Statistical independence (1)

- Two events A and B are called ***Statistically Independent*** if the the probability that A occurs does not change with respect to the occurrence of B.
  - $P(A|B) = P(A)$.
- Ex:
  - A: I will play football tomorrow.
  - B: the pacific ocean contains many fishes.
  - $P(A|B) = P(A)$: the fact that the pacific ocean contains many fishes does not affect my decision to play football tomorrow.

Statistical independence (2)

- Assume $P(A|B) = P(A)$, we have:
  - $P(\sim A|B) = P(\sim A)$
  - $P(B|A) = P(B)$
  - $P(A,B) = P(A).\ P(B)$
  - $P(\sim A,B) = P(\sim A).\ P(B)$
  - $P(A,\sim B) = P(A).\ P(\sim B)$
  - $P(\sim A,\sim B) = P(\sim A).\ P(\sim B)$.

Conditional independence

- Two events A and C are called ***Conditionally Independent*** given B if P(A|B, C) = P(A|B).
- Ex:
  - A: I will play football tomorrow.
  - B: the football match will happen in-house tomorrow.
  - C: it will not rain tomorrow.
  - P(A|B, C) = P(A|B).

Some rules in probability theory

- Chain rules:
  - $P(A,B) = P(A|B).P(B) = P(B|A).P(A)= P(B,A)$
  - $P(A|B) = P(A,B)/P(B) = P(B|A).P(A)/P(B)$
  - $P(A,B|C) = P(A,B,C)/P(C) = P(A|B,C).P(B,C)/P(C)$
    $= P(A|B,C).P(B|C).$
- Independence:
  - $P(A|B) = P(A)$
    if A and B are statistically independent.
  - $P(A,B|C) = P(A|C).P(B|C)$
    if A and B are statistically independent, conditioned on C.
  - $P(A_1,\ldots,A_n|C) = P(A_1|C)\ldots P(A_n|C)$
    if $A_1,\ldots,A_n$ are statistically independent, conditioned on C.

Product and sum rules

- Consider x and y are discrete random variables.
  Their domains are X and Y respectively
- **Product rule:**

$$P(x, y) = P(x|y)P(y)$$

- **Sum rule**

$$P(x) = \sum_{y \in Y} P(x, y)$$

- The summation (tổng) should be integration (tích phân) if y is continuous
  (tổng sẽ được thay bằng tích phân nếu biến y liên tục)

Bayes' rule

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{D})}$$

- P($\boldsymbol{\theta}$): *prior probability* (xác suất tiên nghiệm) of the variable $\boldsymbol{\theta}$.
  - Our uncertainty about $\boldsymbol{\theta}$ before observing data.
- P(**D**): prior probability that we can observe data **D**.
- P(**D**|$\boldsymbol{\theta}$): probability (*likelihood*) that we can observe data **D** provided that $\boldsymbol{\theta}$ is known.
- P($\boldsymbol{\theta}$|**D**): *posterior probability* (xác suất hậu nghiệm) of $\boldsymbol{\theta}$ if we already have observed data **D**.
  - Bayesian approach bases on this quatity.

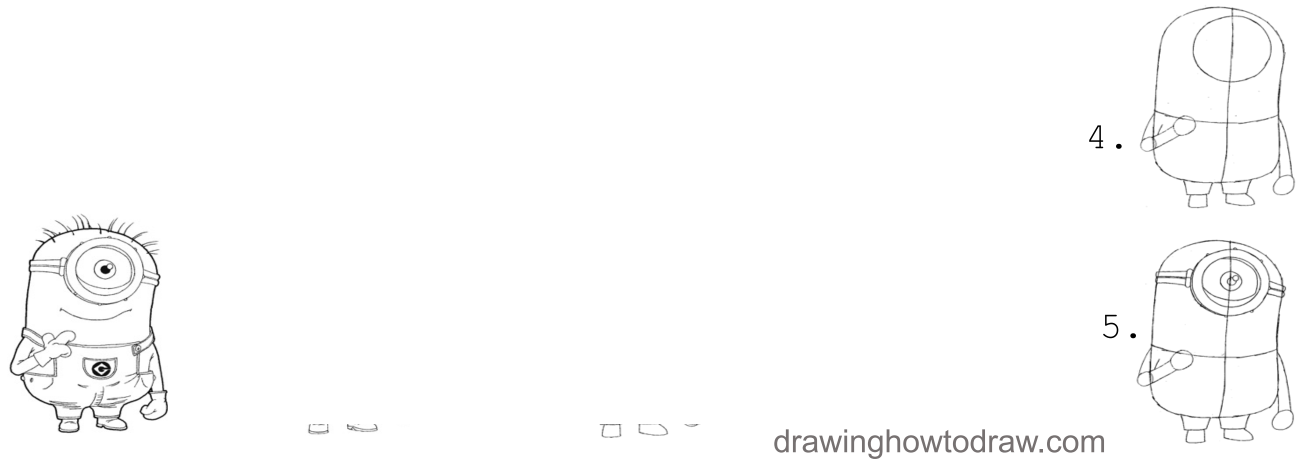
Probabilistic models

Model, inference, learning

hust.edu.vn fb.com/dhbkhn

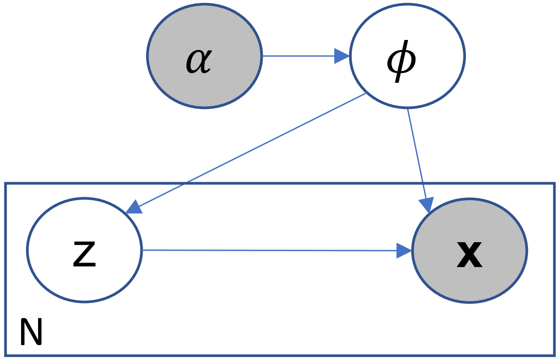
Probabilistic model

- Our assumption on how the data were generated
  (giả thuyết của chúng ta về quá trình dữ liệu đã được sinh ra như thế nào)
- Example: how a sentence is generated?
  - We assume our brain does as follow:
  - *First choose the topic of the sentence*
  - *Generate the words one-by-one to form the sentence*
- How will TIM be drawn?



Probabilistic model

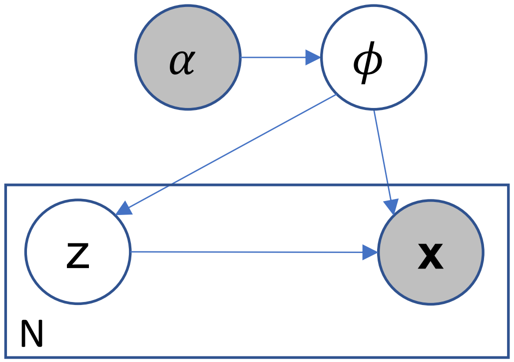
- A model sometimes consists of
  - **Observed variable** (e.g., $\boldsymbol{x}$) which models the observation (data instance) (biến quan sát được)
  - **Hidden variable** which describes the hidden things (e.g., z, ϕ) (biến ẩn)
  - **Local variable** (e.g., $z, \boldsymbol{x}$) which associates with one data instance
  - **Global variable** (e.g., ϕ) which is shared across the data instances, and is the representative of the model
  - **Relations** between the variables



- Each variable follows some probability distribution (mỗi biến tuân theo một phân bố xác suất nào đó)

Different types of models

- **Probabilistic graphical model (PGM):**
  Graph + Probability Theory
  (mô hình đồ thị xác suất)
  - Each vertex represents a random variable, grey circle means "observed", white circle means "latent"
  - Each edge represents the conditional dependence between two variables
  - *Directed graphical model:* each edge has a direction
  - *Undirected graphical model:* no direction in the edges

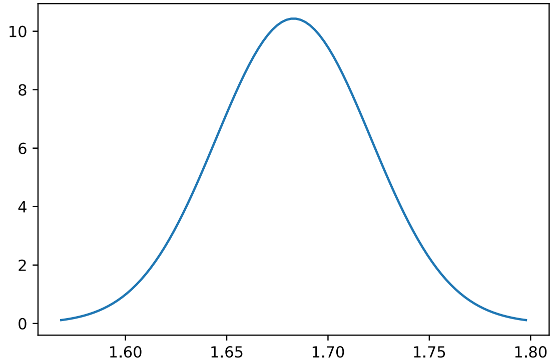


- Latent variable model: a PGM which has at least one latent variable
- Bayesian model: a PGM which has a prior distribution on its parameter

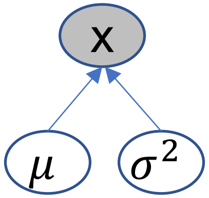
Univariate normal distribution

- We wish to model the height of a person
  - We had collected a dataset from 10 people in Hanoi: **D**={1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62}
- Let x denote the random variable that represents the height of a person
- **Assumption:** x follows a Normal distribution (Gaussian) with the following *probability density function* (PDF)

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



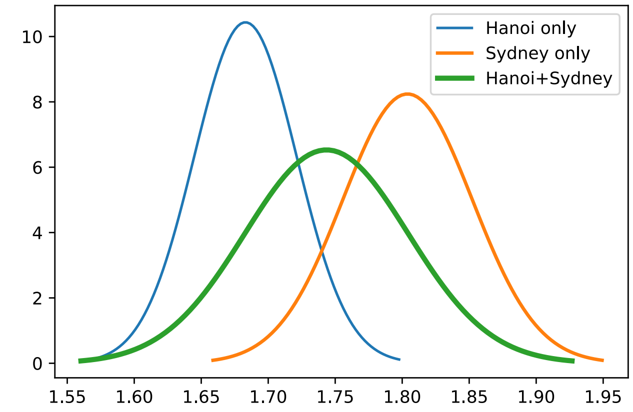
  - where $\{\mu,\sigma^2\}$ are the mean and variance
- Note:
  - $\mathcal{N}(x|\mu,\sigma^2)$ represents the class of normal distributions
  - This class is parameterized by $\boldsymbol{\theta} = (\mu,\sigma^2)$



- **Learning:** we need to know specific values of $\{\mu,\sigma^2\}$

Univariate Gaussian mixture model (1)

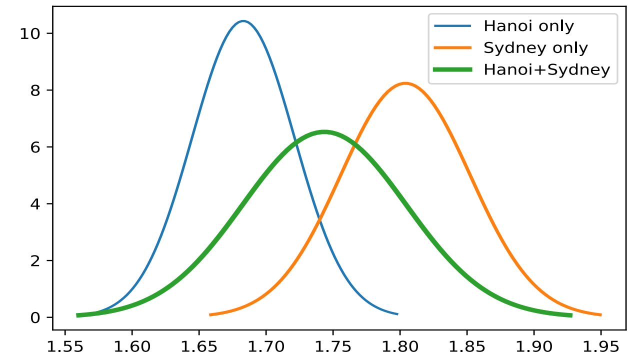
- We wish to model the height of a person
  - We had collected a dataset from 10 people in Hanoi + 10 people in Sydney
    D={1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62, 1.75, 1.80, 1.85, 1.65, 1.91, 1.78, 1.88, 1.79, 1.82, 1.81}
- Let x denote the random variable that represents the height
- If we use Normal distribution:
  - Blue curve models the height in Hanoi
  - Orange curve models the height in Sydney
  - Green curve models the whole **D**



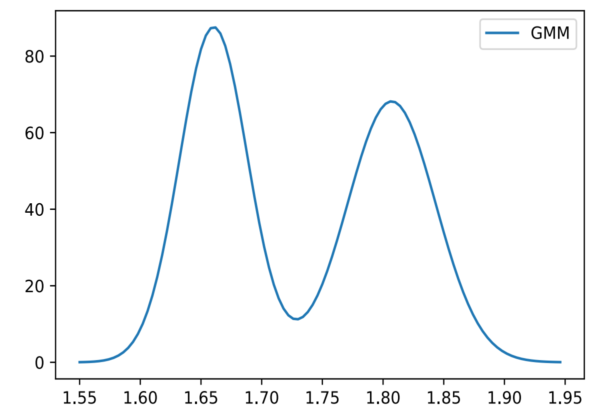
- Univariate Gaussian does not model well the underlying distribution
  - Mixture model?
    (mô hình hỗn hợp)

Univariate Gaussian mixture model (2)

- **Assumption**: the data are generated from two different Gaussians, and each instance is generated from one of the two Gaussians.
  Generative process:
  - *Pick the conponent index:* $z \sim Multinomial(z|\phi)$
  - *Generate sample* $x \sim Normal(x \mid \mu_z, \sigma_z^2)$



- This is **Gaussian mixture model** (GMM) (mô hình hỗn hợp Gauss)
  - (μ_1, σ_1^2) represents the first Gaussian
  - (μ_2, σ_2^2) represents the second Gaussian
  - $\phi \in [0,1]$ is the parameter of the Multinomial distribution, $P(z = 1 \mid \phi) = \phi = 1 - P(z = 2 \mid \phi)$



- Density of the GMM:

$$\phi \mathcal{N}(x|\mu_1, \sigma_1^2) + (1 - \phi)\mathcal{N}(x|\mu_2, \sigma_2^2)$$

Note: “~” means “follows” (tuân theo)

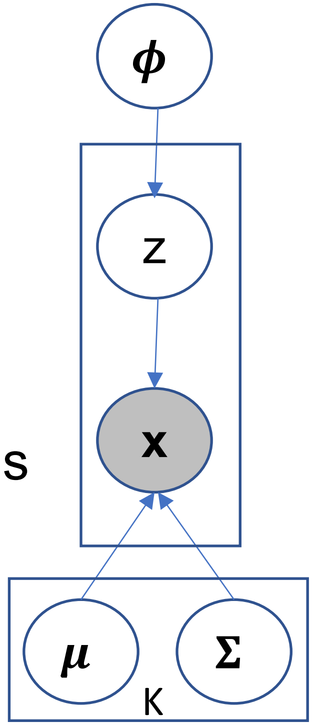
GMM: Multivariate case

- Consider the case each **x** belongs to the *n*-dimensional space.
- GMM: we assume that the data are samples from *K* different Gaussian distributions.
- Each instance **x** is generated from one of those K Gaussians by the following ***generative process***:
  - *Take the component index* $z \sim Multinomial(z|\boldsymbol{\phi})$
  - *Generate* $\boldsymbol{x} \sim Normal(\boldsymbol{x} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- The density function is

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \phi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  - $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$ represents the weights of the Gaussians
  - Each multivariate Gaussian has density

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

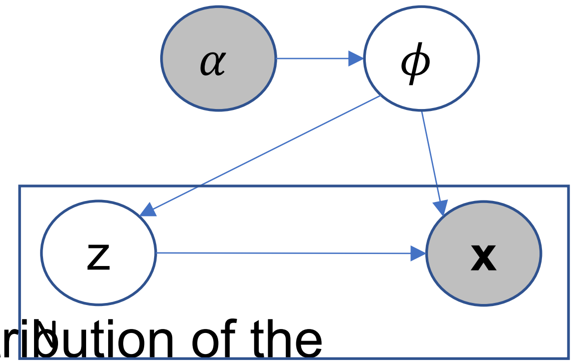


PGM: some well-known models

- Gaussian mixture model (GMM)
  - Modeling real-valued data
- Latent Dirichlet allocation (LDA)
  - Modeling the topics hidden in textual data
- Hidden Markov model (HMM)
  - Modeling time-series, i.e., data with time stamps or sequential nature
- Conditional Random Field (CRF)
  - for structured prediction
- Deep generative models
  - Modeling the hidden structures, generating artificial data

Probabilistic model: two problems

- **Inference** for a given instance $\boldsymbol{x}_n$
  - Recovery of the local variable (e.g., z_n), or
  - The distribution of the local variables (e.g., $P(z_n, \boldsymbol{x}_n | \phi)$)
  - Example: for GMM, we want to know z_n indicating which Gaussian did generate $\boldsymbol{x}_n$



- **Learning (estimation)**
  - Given a training dataset, estimate the joint distribution of the variables
    - E.g., estimate $P(\phi, z_1, \ldots, z_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \alpha)$
    - E.g., estimate $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \alpha)$
    - E.g., estimate α
    - Inference of local variables is often needed

Inference and Learning

MLE, MAP

hust.edu.vn fb.com/dhbkhn

Some inference approaches (1)

- Let D be the data, and *h* be a hypothesis
  - hypothesis: unknown parameter, hidden variables, …
- **Maximum Likelihood Estimation (MLE, cực đại hoá khả năng)**

$$h^* = \arg\max_{h \in \boldsymbol{H}} P(D|h)$$

  - Finds h* (in the hypothesis space **H**) that maximizes the likelihood of the data.
  - *Other words: MLE makes inference about the model that is most likely to have generated the data.*
- **Bayesian inference** (suy diễn Bayes) considers the transformation of our prior knowledge $P(h)$, through the data D, into the posterior knowledge $P(h|D)$.
  - Remember the Bayes'rule: $P(h|D) = P(D|h)P(h)/P(D)$. So

$$P(h|D) \propto P(D|h) * P(h)$$

(Posterior ∝ Likelihood * Prior)

Some inference approaches (2)

- In some cases, we may know the prior distribution of *h*.
- **Maximum a Posterior Estimation (MAP, cực đại hoá hậu nghiệm)**

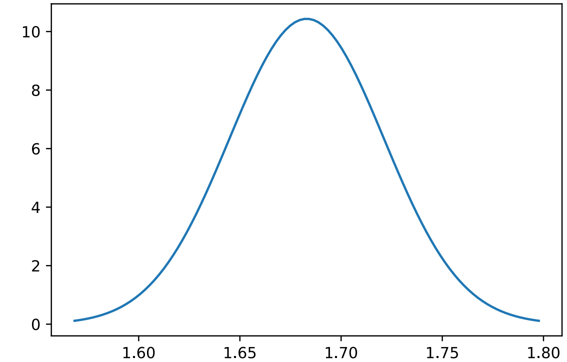
$$h^* = \arg\max_{h \in \boldsymbol{H}} P(h|\boldsymbol{D}) = \arg\max_{h \in \boldsymbol{H}} P(\boldsymbol{D}|h)\, P(h)/P(\boldsymbol{D}) \quad = \arg\max_{h \in \boldsymbol{H}} P(\boldsymbol{D}|h)\, P(h)$$

  - Finds h* that maximizes the posterior probability of h.
  - MAP finds a point (posterior mode), not a distribution → point estimation
- MLE is a special case of MAP, when using uniform prior over *h*.
- *Full Bayesian inference* tries to estimate the full posterior distribution $P(h|\boldsymbol{D})$, not just a point *h**.
- Note:
  - MLE, MAP, or full Bayesian approaches can be applied to both learning and inference.

MLE: Gaussian example (1)

- We wish to model the height of a person, using the dataset **D** = {1.6, 1.7, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62}
  - Let x be the random variable representing the height of a person.
  - **Model**: assume that x follows a Gaussian distribution with ***unknown*** mean μ and variance σ^2
  - **Learning:** estimate (μ, σ) from the given data $\boldsymbol{D} = \{x_1, \dots, x_{10}\}$.
- Let $f(x|\mu, \sigma)$ be the density function of the Gaussian family, parameterized by (μ, σ).
  - $f(x_n|\mu, \sigma)$ is the likelihood of instance x_n.
  - $f(\boldsymbol{D}|\mu, \sigma)$ is the likelihood function of **D**.
- Using MLE, we will find

$$(\mu_*, \sigma_*) = \arg\max_{\mu,\sigma} f(\boldsymbol{D}|\mu, \sigma)$$



MLE: Gaussian example (2)

- **i.i.d assumption:** we assume that the data are independent and identically distributed (dữ liệu được sinh ra một cách độc lập)
  - As a result, we have $P(\boldsymbol{D}|\mu,\sigma) = P(x_1,\dots,x_{10}|\mu,\sigma) = \prod_{i=1}^{10} P(x_i|\mu,\sigma)$
- Using this assumption, MLE will be

$$(\mu_*,\sigma_*) = \arg\max_{\mu,\sigma} \prod_{i=1}^{10} f(x_i|\mu,\sigma) = \arg\max_{\mu,\sigma} \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \arg\max_{\mu,\sigma} \log \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$= \arg\max_{\mu,\sigma} \sum_{i=1}^{10} \left(-\frac{1}{2\sigma^2}(x_i-\mu)^2 - \log\sqrt{2\pi\sigma^2}\right)$$

Log trick, $\log \stackrel{\text{def}}{=} \ln$

- Using gradients (w.r.t μ, σ), we can find

$$\mu_* = \frac{1}{10}\sum_{i=1}^{10} x_i = 1.683, \qquad \sigma_*^2 = \frac{1}{10}\sum_{i=1}^{10} (x_i - \mu_*)^2 \approx 0.0015$$

MAP: Gaussian Naïve Bayes (1)

- Consider the **classification problem**
  - Training data **D** = {($\mathbf{x}_1$, y_1), ($\mathbf{x}_2$, y_2), …, ($\mathbf{x}_M$, y_M)} with *M* instances, *C* classes.
  - Each $\mathbf{x}_i$ is a vector in the *n*-dimensional space $\mathbb{R}^n$, e.g., $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$.
- **Gaussian Naive Bayes (GNB):** *we assume there are C different Gaussian distributions that generate the data in **D**, and each instance is generated by the following generative process:*
  - *Pick a class index* $c \sim Cat(\boldsymbol{\phi})$
  - *Generate* $\boldsymbol{x} \sim Normal(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
    - Where $\boldsymbol{\mu}_c$ is the mean vector, $\boldsymbol{\Sigma}_c$ is the covariance matrix of size $n \times n$, $Cat(\boldsymbol{\phi})$ is the Categorical distribution with parameter $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_C) \geq \mathbf{0}$ so that $||\boldsymbol{\phi}||_1 = 1$.

A class is dominated by a Normal distribution

MAP: Gaussian Naïve Bayes (2)

- *Learning:* estimate the model with parameter $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C)$
- Let c be the random variable to represent the class label for each **x**. $P(y_1, \ldots, y_M | \boldsymbol{D}, \boldsymbol{\theta})$ denotes the posterior $P(c = y_1, \ldots, c = y_M | \boldsymbol{D}, \boldsymbol{\theta})$
- Following MAP, we find

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} P(y_1, \ldots, y_M | \boldsymbol{D}, \boldsymbol{\theta})$$

- Using Bayes rule, i.i.d, log trick, and some reformulations:

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{C} \sum_{\boldsymbol{x} \in \boldsymbol{D}_k} \log P(\boldsymbol{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{k=1}^{C} |\boldsymbol{D}_k| \log \phi_k$$

  - Where $\boldsymbol{D}_k$ contains all the training examples in class k and has size $|\boldsymbol{D}_k|$.

MAP: Gaussian Naïve Bayes (3)

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{C} \sum_{\boldsymbol{x} \in \boldsymbol{D}_k} \log P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{k=1}^{C} |\boldsymbol{D}_k| \log \phi_k$$

- To find $\boldsymbol{\phi}$, we need to solve

$$\max_{\boldsymbol{\phi}} \sum_{k=1}^{C} |\boldsymbol{D}_k| \log \phi_k \ \text{ such that } \sum_{k=1}^{C} \phi_k = 1 \text{ and } \phi_k \geq 0, \forall k$$

  - By using Lagrange multiplier method, we can obtain

$$\phi_k^* = \frac{|\boldsymbol{D}_k|}{\mathrm{M}}$$

- To find $(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, we can solve for:

$$(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) = \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \log P(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

MAP: Gaussian Naïve Bayes (4)

- Note

$$(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) = \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \log \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

$$= \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \log \left[\frac{1}{\sqrt{\det(2\pi \boldsymbol{\Sigma}_c)}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \right) \right]$$

$$= \arg\max_{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) - \log \sqrt{\det(2\pi \boldsymbol{\Sigma}_c)} \right]$$

- Using gradients (w.r.t $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$), we can arrive at

$$\boldsymbol{\mu}_{c*} = \frac{1}{|\boldsymbol{D}_c|} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} \boldsymbol{x}, \qquad \boldsymbol{\Sigma}_{c*} = \frac{1}{|\boldsymbol{D}_c|} \sum_{\boldsymbol{x} \in \boldsymbol{D}_c} (\boldsymbol{x} - \boldsymbol{\mu}_{c*})(\boldsymbol{x} - \boldsymbol{\mu}_{c*})^T$$

- **So, after training we obtain the $(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, \phi_c^*)$ for each class *c*.**

MAP: Gaussian Naïve Bayes (5)

- Trained model: $(\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, \phi_c^*)$ for each class *c*
- **Prediction** for a new instance **z** by finding the class label that has the highest posterior probability:

$$c_z = \arg\max_{c\in\{1,\ldots,C\}} P(c|\mathbf{z}, \boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}, \phi_c^*) = \arg\max_{c\in\{1,\ldots,C\}} P(\mathbf{z}|\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) P(c|\phi_c^*)$$

$$= \arg\max_{c\in\{1,\ldots,C\}} [\log P(\mathbf{z}|\boldsymbol{\mu}_{c*}, \boldsymbol{\Sigma}_{c*}) + \log P(c|\phi_c^*)]$$

Bayes' rule

$$= \arg\max_{c\in\{1,\ldots,C\}} \left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{c*})^T \boldsymbol{\Sigma}_{c*}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{c*}) - \log\sqrt{\det(2\pi\boldsymbol{\Sigma}_{c*})} + \log \phi_c^*\right]$$

- If using MLE, we do not need to use/estimate the prior P(c)

MAP: Multinomial Naïve Bayes (1)

- Consider the **text classification** problem (dữ liệu có thuộc tính rời rạc)
  - Training data **D** = {($\mathbf{x}_1$, y_1), ($\mathbf{x}_2$, y_2), …, ($\mathbf{x}_M$, y_M)} with *M* documents, *C* classes.
  - TF: each document $\mathbf{x}_i$ is represented by a vector of *V* dimensions, e.g., $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iV})^T$, each x_{ij} is the *frequency* of term *j* in document $\mathbf{x}_i$
- **Multinomial Naive Bayes (MNB):** *we assume there are C different Multinomial distributions that generate the data in* ***D****, and each instance is generated by the following generative process:*
  - *Pick a class index* $c \sim Cat(\boldsymbol{\phi})$
  - *Generate* $\boldsymbol{x} \sim Multinomial(\boldsymbol{\theta}_c)$
    - $Cat(\boldsymbol{\phi})$ is the Categorical distribution with parameter $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_C) \geq \mathbf{0}$ s.t. $\|\boldsymbol{\phi}\|_1 = 1$.

A class is dominated by a Multinomial distribution

MAP: Multinomial Naïve Bayes (2)

- A multinomial distribution, which is parameterized by $\boldsymbol{\theta}_c$, has probability mass function

$$f(x_1, \dots, x_V | \theta_{c1}, \dots, \theta_{cV}) = \frac{\Gamma(\sum_{j=1}^{V} x_j + 1)}{\prod_{j=1}^{V} \Gamma(x_j + 1)} \prod_{k=1}^{V} \theta_{ck}^{x_k}$$

  - $\theta_{cj} = P(x = j|\theta_{cj})$ is the probability that term $j \in \{1, \dots, V\}$ appears, satisfying $\sum_{k=1}^{V} \theta_{ck} = 1$. Γ is the gamma function.

- *Learning MNB:* we can do similarly with Gaussian Naïve Bayes to estimate $\boldsymbol{\theta}_c = (\theta_{c1}, \dots, \theta_{cV})$ and ϕ_c for each class c.

Homework ?

MAP: Multinomial Naïve Bayes (3)

- Trained model: $(\boldsymbol{\theta}_{c*}, \phi_c^*)$ for each class *c*
- Prediction for a new instance $\boldsymbol{z} = (z_1, \dots, z_V)^T$ by

$$c_z = \arg \max_{c \in \{1,\dots,C\}} P(c|\boldsymbol{z}, \boldsymbol{\theta}_{c*}) = \arg \max_{c \in \{1,\dots,C\}} P(\boldsymbol{z}|\boldsymbol{\theta}_{c*}, c) P(c)$$

$$= \arg \max_{c \in \{1,\dots,C\}} \log P(\boldsymbol{z}|\boldsymbol{\theta}_{c*}) + \log P(c) \qquad \text{(MNB.1)}$$

$$= \arg \max_{c \in \{1,\dots,C\}} \log \frac{\Gamma(\sum_{j=1}^{V} z_j + 1)}{\prod_{j=1}^{V} \Gamma(z_j + 1)} \prod_{k=1}^{V} \theta_{ck*}^{z_k} + \log \phi_c^*$$

$$= \arg \max_{c \in \{1,\dots,C\}} \log \prod_{k=1}^{V} \theta_{ck*}^{z_k} + \log \phi_c^*$$

$$= \arg \max_{c \in \{1,\dots,C\}} \log \prod_{k=1}^{V} P(z_k|\theta_{ck*}) + \log \phi_c^* \qquad \text{(MNB.2)}$$

Note: we implicitly assume that *the attributes are conditionally independent*, as shown in equations (MNB.1) and (MNB.2).
(ta ngầm giả thuyết rằng các thuộc tính độc lập với nhau)

Difficult situations

- No closed-form solution for the learning/inference problem? (không tìm được ngay công thức nghiệm)
  - The examples before are easy cases, as we can find solutions in a closed form by using gradient.
  - Many models (e.g., GMM) do not admit a closed-form solution.
- No explicit expression of the density/mass function? (không có công thức tường minh để tính toán)
- Intractable inference (bài toán suy diễn không khả thi)
  - Inference in many probabilistic models is NP-hard. [Sontag & Roy, 2011; Tosh & Dasgupta, 2019]

Reference

- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." *Journal of the American Statistical Association* 112, no. 518 (2017): 859-877.
- Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight Uncertainty in Neural Network." In *International Conference on Machine Learning (ICML)*, pp. 1613-1622. 2015.
- Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In *International Conference on Machine Learning*, pp. 1050-1059. 2016.
- Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." *Nature* 521, no. 7553 (2015): 452-459.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." In *International Conference on Learning Representations* (ICLR), 2014.
- Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349, no. 6245 (2015): 255-260.
- Tosh, Christopher, and Sanjoy Dasgupta. “The Relative Complexity of Maximum Likelihood Estimation, MAP Estimation, and Sampling.” In *Proceedings of the 32nd Conference on Learning Theory, in PMLR* 99:2993-3035, 2019.
- Sontag, David, and Daniel Roy, “Complexity of inference in latent dirichlet allocation” in: Proceedings of Advances in Neural Information Processing System, 2011.

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !