Computer Science

CSC580: Probabilistic Graphical Models

Probabilistic Graphical Models

Jason Pacheco

Administrivia

- Homework submission
  - Make sure questions are answered in PDF
  - Match pages to questions
  - Put code in PDF (relevant parts of code at least)
  - Doublecheck your submission

- Midterm Exam
  - Thursday 10/12
  - No coding
  - Probably closed-book

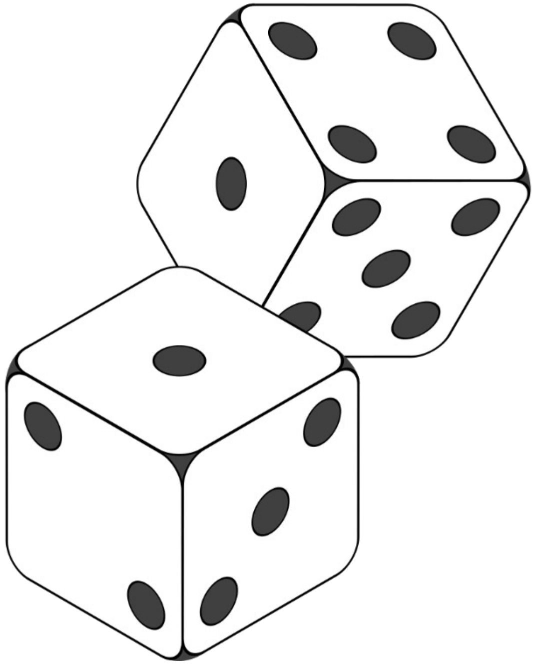
- Probability Refresher
- Probabilistic Graphical Models
- Naïve Bayes

- Probability Refresher
- Probabilistic Graphical Models
- Naïve Bayes

Before we learn about probabilistic graphical models, we need to review probability…

Random Events and Probability

Suppose we roll two fair dice…



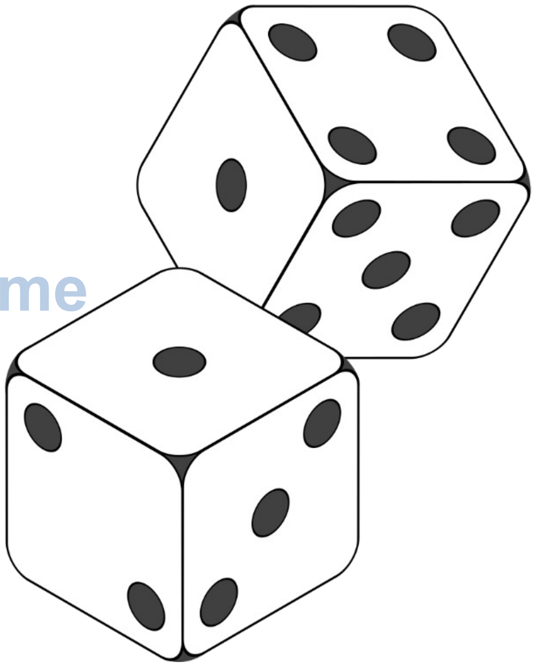
- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?
- What is the *probability* of rolling **odd** numbers?

…probability theory *gives a mathematical formalism to addressing such questions…*

Definition An **experiment** or **trial** is any process that can be repeated with well-defined outcomes. It is *random* if more than one outcome is possible.

Random Events and Probability

Definition An **outcome** is a possible result of an experiment or trial, and the collection of all possible outcomes is the **sample space** of the experiment,



Outcome

Example $\underbrace{(1,1),\ (1,2),\ \ldots,\ (6,1),\ (6,2),\ \ldots,\ (6,6)}_{\text{Sample Space}}$

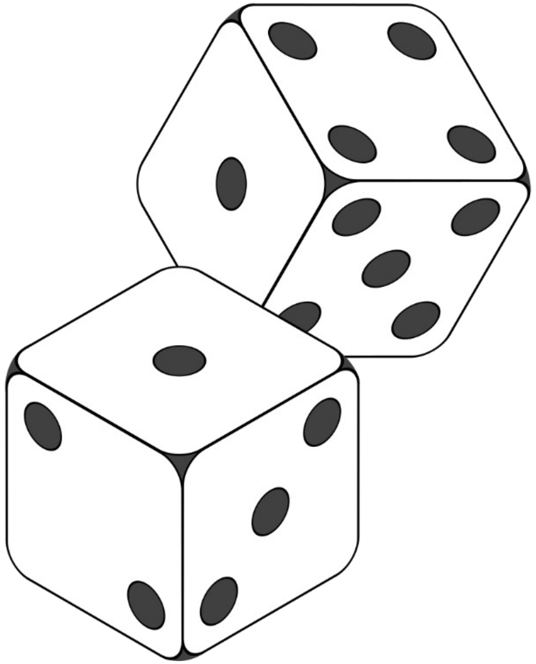
Definition An **event** is a *set* of outcomes (a subset of the sample space),

Example Event Roll at least a single 1

$\{(1,1),\ (1,2),\ (1,3),\ \ldots,\ (1,6),\ \ldots,\ (6,1)\}$

Random Variables

(Informally) A random variable is an unknown quantity that maps events to numeric values.



Example X is the *sum of two dice* with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

Example Flip a coin and let random variable Y represent the outcome,

$$Y \in \{\text{Heads}, \text{Tails}\}$$

Random Variables and Probability

Capitol letters represent random variables

Lowercase letters are realized *values*

$$X = x$$

$X = x$ is the **event** that X takes the value x

Example Let X be the random variable (RV) representing the sum of two dice with values,

$$X \in \{2, 3, 4, \ldots, 12\}$$

X=5 is the *event* that the dice sum to 5.

Probability Mass Function

A function $p(X)$ is a **probability mass function (PMF)** of a discrete random variable if the following conditions hold:

(a) It is nonnegative for all values in the support,

$$p(X = x) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x p(X = x) = 1$$

Intuition Probability mass is conserved, just as in physical mass. Reducing probability mass of one event must increase probability mass of other events so that the definition holds...

Probability Mass Function

Example Let X be the outcome of a single fair die. It has the PMF,

$$p(X = x) = \frac{1}{6} \qquad \text{for } x = 1, \ldots, 6$$

Uniform Distribution

Example We can often represent the PMF as a vector. Let S be an RV that is the *sum of two fair dice*. The PMF is then,

Observe that S does <u>not</u> follow a uniform distribution

$$p(S) = \begin{pmatrix} p(S=2) \\ p(S=3) \\ p(S=4) \\ \vdots \\ p(S=12) \end{pmatrix} = \begin{pmatrix} 1/36 \\ 1/18 \\ 1/2 \\ \vdots \\ 1/36 \end{pmatrix}$$

PMF Notation

- We use $p(X)$ to refer to the probability mass *function* (i.e. a function of the RV X)

- We use $p(X=x)$ to refer to the probability of the *outcome* $X=x$ (also called an "event")

- We will often use $p(x)$ as shorthand for $p(X=x)$

Joint Probability

Definition Two (discrete) RVs X and Y have a *joint PMF* denoted by $p(X,Y)$ and the probability of the event X=x and Y=y denoted by $p(X=x,Y=y)$ where,

(a) It is nonnegative for all values in the support,

$$p(X=x,Y=y) \geq 0$$

(b) The sum over all values in the support is 1,

$$\sum_x \sum_y p(X=x,Y=y) = 1$$

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

All values are nonnegative

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

X \ Y	0	1
0	0.04	0.36
1	0.30	0.30

The sum over all values is 1:
0.04 + 0.36 + 0.30 + 0.30 = 1

Joint Probability

Let X and Y be *binary RVs.* We can represent the joint PMF p(X,Y) as a 2x2 array (table):

		Y = 0	Y = 1
X	0	0.04	0.36
	1	0.30	0.30

P(X=1, Y=0) = 0.30

Fundamental Rules of Probability

Given two RVs X and Y the **conditional distribution** is:

$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X,Y)}{\sum_x p(X=x,Y)}$$

Multiply both sides by $p(Y)$ to obtain the **probability chain rule**:

$$p(X,Y) = p(Y)p(X \mid Y)$$

The probability chain rule extends to N RVs $X_1, X_2, \ldots, X_N$:

$$p(X_1, X_2, \ldots, X_N) = p(X_1)p(X_2 \mid X_1) \ldots p(X_N \mid X_{N-1}, \ldots, X_1)$$

$$= p(X_1) \prod_{i=2}^{N} p(X_i \mid X_{i-1}, \ldots, X_1)$$

Chain rule valid for any ordering

Fundamental Rules of Probability

Law of total probability

$$p(Y) = \sum_x p(Y, X = x)$$

- P(y) is a **marginal** distribution
- This is called **marginalization**

Proof

$$\sum_x p(Y, X = x) = \sum_x p(Y) p(X = x \mid Y) \qquad \text{(chain rule)}$$

$$= p(Y) \sum_x p(X = x \mid Y) \qquad \text{(distributive property)}$$

$$= p(Y) \qquad \text{(PMF sums to 1)}$$

Generalization for conditionals:

$$p(Y \mid Z) = \sum_x p(Y, X = x \mid Z)$$

Tabular Method

Let X, Y be binary RVs with the joint probability table

For Binomial use K-by-K probability table.

X \ Y	y_1	y_2	$P(x)$
x_1	0.04	0.36	0.4 $P(x_1)$
x_2	0.30	0.30	0.6 $P(x_2)$
$P(y)$	0.34 $P(y_1)$	0.66 $P(y_2)$	

$P(y_1)=P(x_1,y_1)+P(x_2,y_1)$
$P(y_2)=P(x_1,y_2)+P(x_2,y_2)$
[i.e., sum down columns]

$P(x_1)=P(x_1,y_1)+P(x_1,y_2)$
$P(x_2)=P(x_2,y_1)+P(x_2,y_2)$
[i.e., sum across rows]

Tabular Method

We don't care about event $Y=y_2$

		Y	
		y_1	y_2
X	x_1	0.04	Censored!
	x_2	0.30	
		0.34	

$P(y_1)$

$P(x|y_1)=?$

Tabular Method

		$Y=y_1$
X	x_1	0.04
	x_2	0.30
		0.34

$P(y_1)$

$P(x \mid y_1)$
0.04 / 0.34
0.30 / 0.34

These sum to one:
A conditional probability distribution is still a probability distribution

Intuition Check

<u>Question:</u> Roll two dice and let their outcomes be $X_1, X_2 \in \{1, \ldots, 6\}$ *for die 1 and die 2, respectively. Recall the definition of conditional probability,*

$$p(X_1 \mid X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Which of the following are true?

a) $p(X_1 = 1 | X_2 = 1) > p(X_1 = 1)$

b) $\boxed{p(X_1 = 1 | X_2 = 1) = p(X_1 = 1)}$ Outcome of die 2 doesn't *affect* die 1

c) $p(X_1 = 1 | X_2 = 1) < p(X_1 = 1)$

Intuition Check

<u>*Question:*</u> *Let* $X_1 \in \{1, \ldots, 6\}$ *be outcome of die 1, as before. Now let* $X_3 \in \{2, 3, \ldots, 12\}$ *be the* sum of both dice. Which of the following are true?

a) $\boxed{p(X_1 = 1 | X_3 = 3) > p(X_1 = 1)}$

b) $p(X_1 = 1 | X_3 = 3) = p(X_1 = 1)$

c) $p(X_1 = 1 | X_3 = 3) < p(X_1 = 1)$

Only 2 ways to get $X_3 = 3$ *, each with equal probability:*

$$(X_1 = 1, X_2 = 2) \quad \textit{or} \quad (X_1 = 2, X_2 = 1)$$

so

$$p(X_1 = 1 \mid X_3 = 3) = \frac{1}{2} > \frac{1}{6} = p(X_1 = 1)$$

Dependence of RVs

Intuition…

Consider $P(B|A)$ where you want to bet on B

Should you pay to know A?

In general you would pay something for A if it changed your belief about B. In other words if,

$$P(B|A) \neq P(B)$$

Independence of RVs

Definition *Two random variables* X *and* Y *are* *<u>independent</u>* *if and only if,*

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

for all values x and y, and we say $X \perp Y$.

Definition RVs $X_1, X_2, \ldots, X_N$ are <u>mutually independent</u> if and only if,

$$p(X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} p(X_i = x_i)$$

- Independence is *symmetric:* $X \perp Y \Leftrightarrow Y \perp X$
- Equivalent definition of independence: $p(X \mid Y) = p(X)$

Independence of RVs

Definition *Two random variables* X *and* Y *are* *<u>conditionally independent</u>* *given* Z *if and only if,*

$$p(X=x, Y=y \mid Z=z) = p(X=x \mid Z=z)p(Y=y \mid Z=z)$$

for all values x, y, and z, and we say that $X \perp Y \mid Z$.

- N RVs conditionally independent, given Z, if and only if:

$$p(X_1, \ldots, X_N \mid Z) = \prod_{i=1}^{N} p(X_i \mid Z)$$

Shorthand notation
Implies for all *x, y, z*

- Equivalent def'n of conditional independence: $p(X \mid Y, Z) = p(X \mid Z)$
- Symmetric: $X \perp Y \mid Z \Leftrightarrow Y \perp X \mid Z$

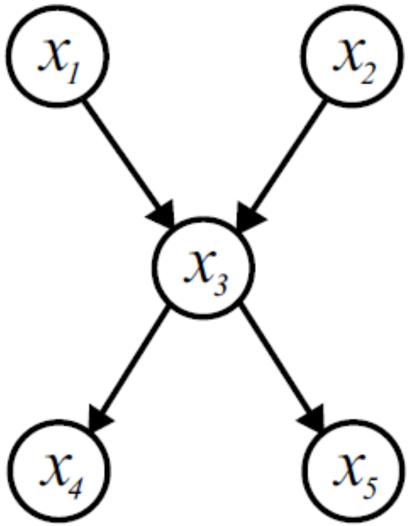
- Probability Refresher

- Probabilistic Graphical Models

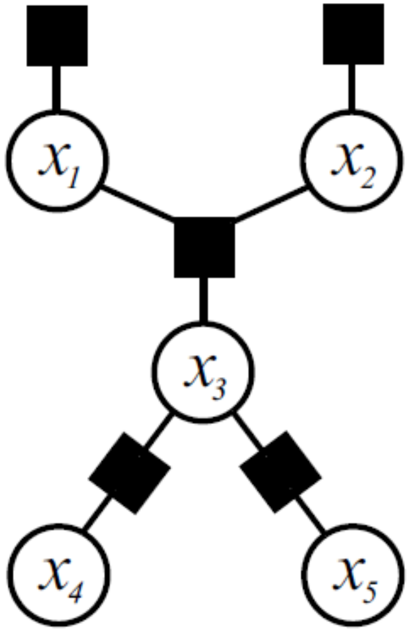
- Naïve Bayes

Graphical Models

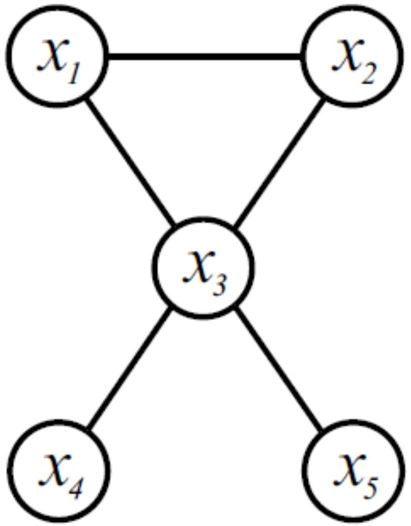
A variety of graphical models can represent the same probability distribution



Bayes Network



Factor Graph



Markov Random Field

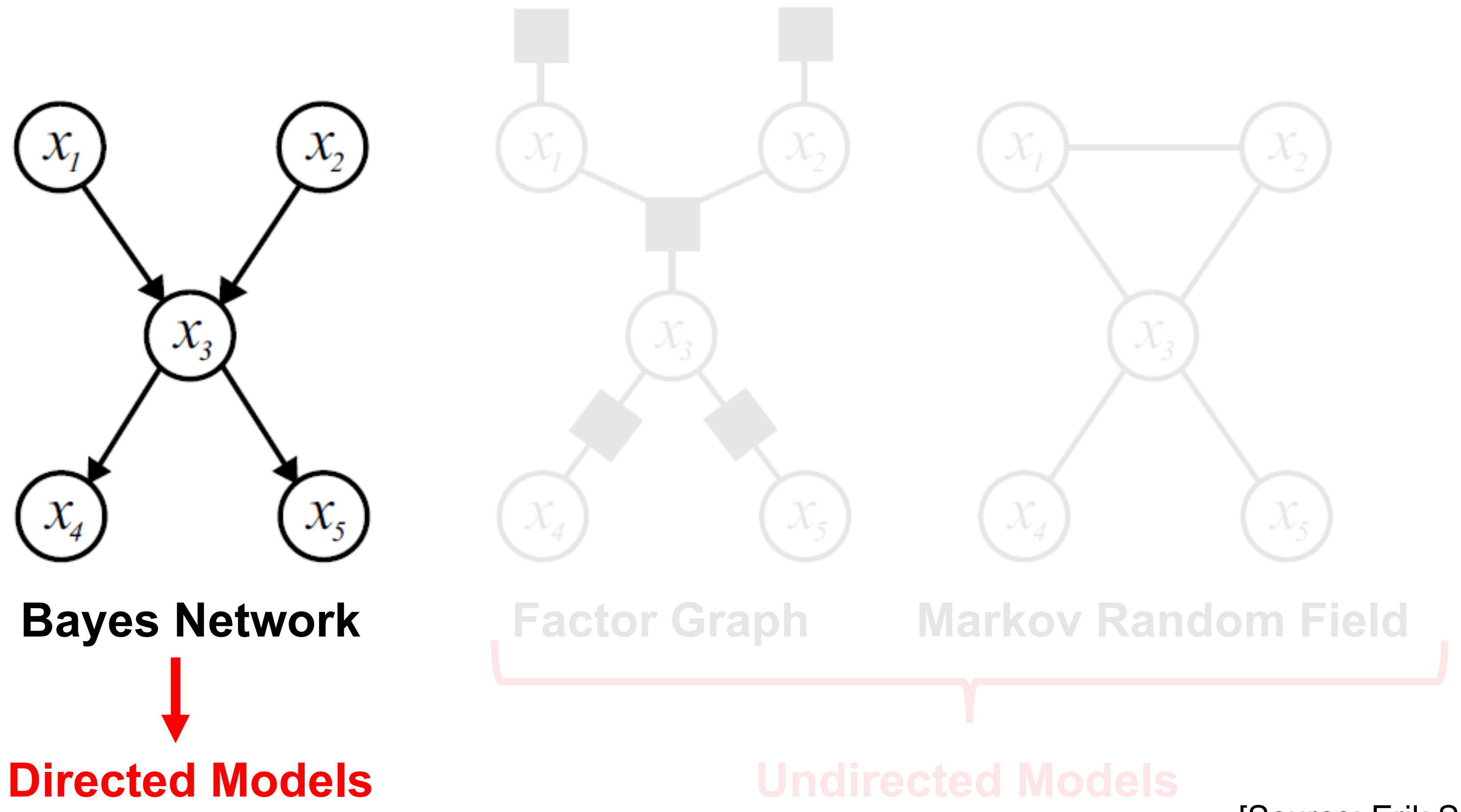
Directed Models

Undirected Models

[Source: Erik Sudderth, PhD Thesis]

Graphical Models

A variety of graphical models can represent the same probability distribution



[Source: Erik Sudderth, PhD Thesis]

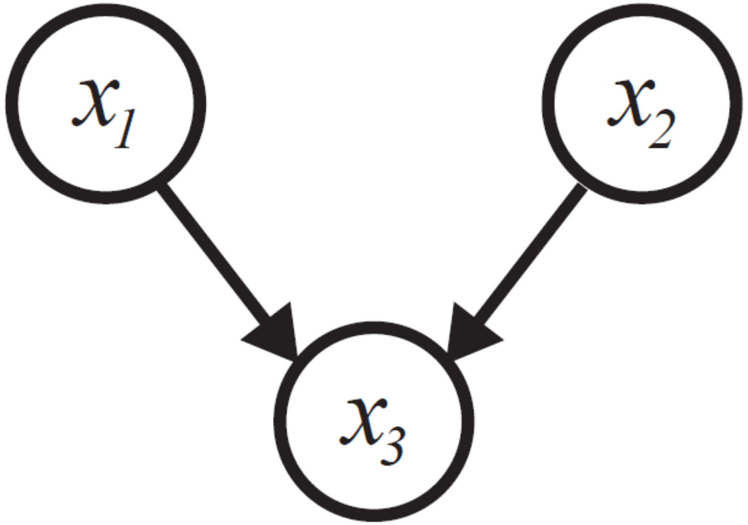
From Probabilities to Pictures

A probabilistic graphical model allows us to pictorially represent a probability distribution

Probability Model:

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$$

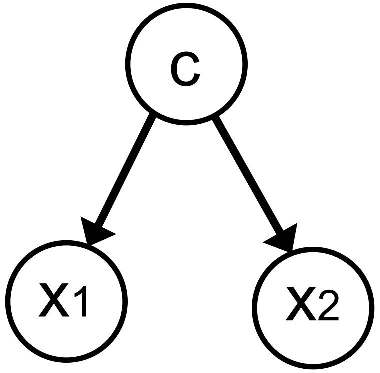
Graphical Model:



Conditional distribution on each RV is dependent on its parent nodes in the graph

Directed Graphical Models

Directed models are <u>generative models</u>…



$$p(C, X_1, X_2) = p(C)p(X_1 \mid C)p(X_2 \mid C)$$

The graph and the formula say exactly the same thing.
(The graph has very specific semantics.)

…tells how data are generated (called ***ancestral sampling***)

Step 1 Sample root node (prior): $c \sim p(C)$

Step 2 Sample children, given sample of parent (likelihood):

$$x_1 \sim p(X_1 \mid C = c) \qquad x_2 \sim p(X_2 \mid C = c)$$

Probability Chain Rule

Recall the **probability chain rule** says that we can decompose any joint distribution as a product of conditionals….

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3)$$

Valid for *any ordering* of the random variables…

$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_1, x_3)p(x_2 \mid x_1, x_3, x_4)$$

For a collection of N RVs and any permutation ρ :

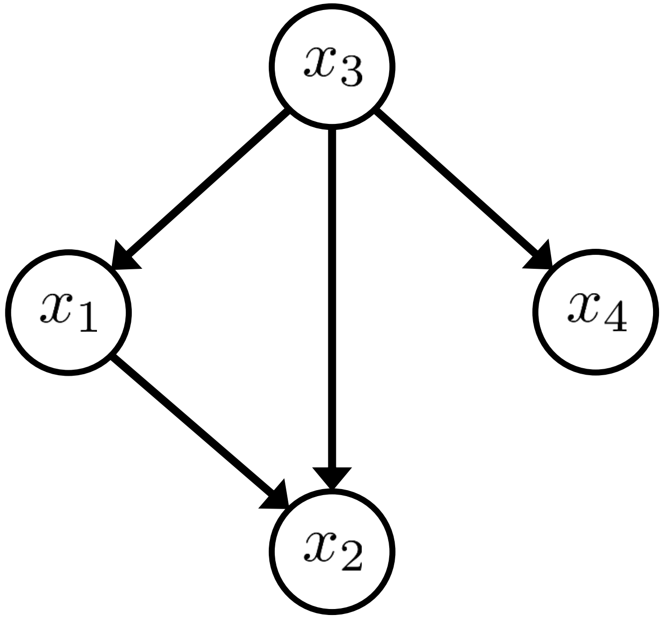
$$p(x_1, \ldots, x_N) = p(x_{\rho(1)}) \prod_{i=2}^{N} p(x_{\rho(i)} \mid x_{\rho(i-1)}, \ldots, x_{\rho(1)})$$

Conditional Independence

Recall two RVs X and Y are **conditionally independent** given Z (or $X \perp Y \mid Z$) iff:

$$p(X \mid Y, Z) = p(X \mid Z)$$

Idea Apply *chain rule* with ordering that exploits conditional independencies to simplify the terms



Ex. Suppose $x_4 \perp x_1 \mid x_3$ and $x_2 \perp x_4 \mid x_1$ then:

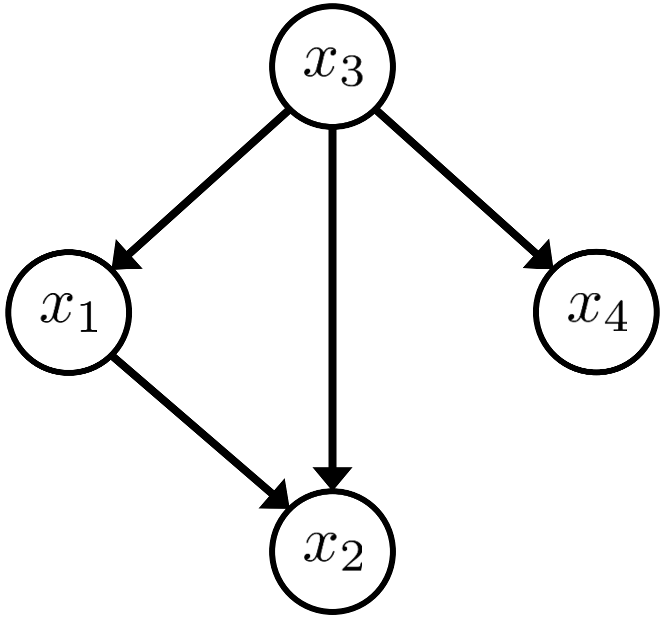
$$p(x) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_1, x_3)p(x_2 \mid x_1, x_3, x_4)$$

$$= p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3)$$

Can visualize conditional dependencies using ***directed acyclic graph*** *(DAG)*

General Directed Graphs

Def. A directed graph is a graph with edges $(s,t) \in \mathcal{E}$ (arcs) connecting parent vertex $s \in \mathcal{V}$ to a child vertex $t \in \mathcal{V}$



Def. Parents of vertex $t \in \mathcal{V}$ are given by the set of nodes with arcs pointing to t ,

$$\mathrm{Pa}(t) = \{s : (s,t) \in \mathcal{E}\}$$

Children of $t \in \mathcal{V}$ are given by the set,

$$\mathrm{Ch}(t) = \{t : (t,k) \in \mathcal{E}\}$$

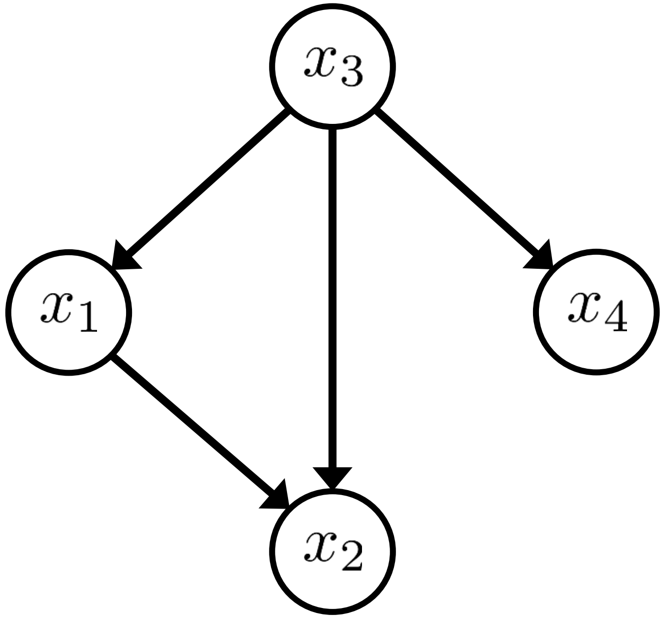
Ancestors are parents-of-parents.
Descendants are children-of-children.

Directed PGM = Bayes Network

Model factors are normalized conditional distributions:

$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid x_{\mathrm{Pa}(s)})$$

Parents of node *s*

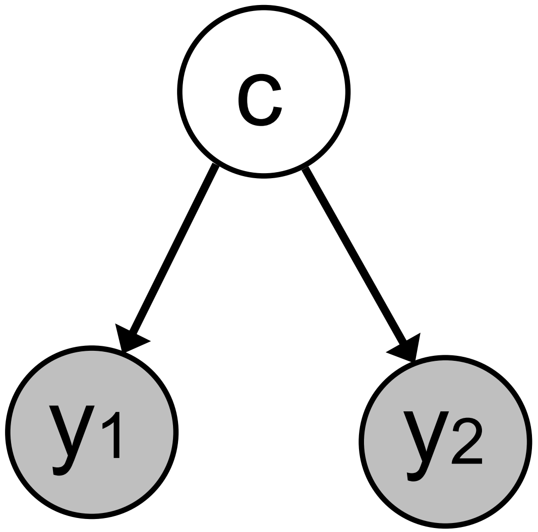


Directed acyclic graph (DAG) specifies factorized form of joint probability:

$$p(x) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3)$$

Locally normalized factors yield *globally normalized* joint probability

Inference



Denote observed data with shaded nodes,

$$Y_1 = y_1 \qquad Y_2 = y_2$$

Infer *latent* variable C via Bayes' rule:

$$p(c \mid y_1, y_2) = \frac{p(c)p(y_1 \mid c)p(y_2 \mid c)}{p(y_1, y_2)}$$

- This is (obviously) a simple example
- Models and inference task can get really complicated
- But the fundamental concepts and approach are the same

Bayes' Rule

Posterior represents all uncertainty after observing data…

prior probability

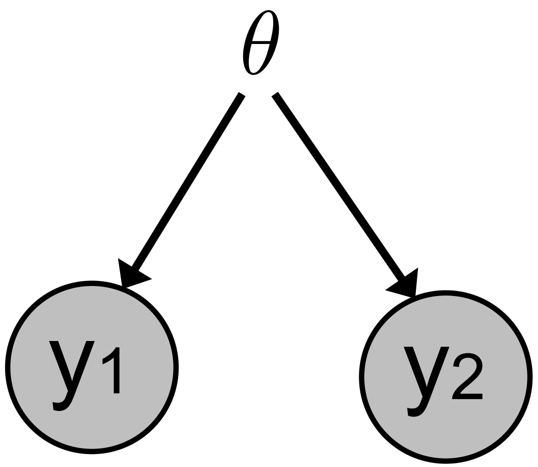
likelihood function for the parameters

$$p(c \mid y) = \frac{p(c)p(y \mid c)}{p(y)}$$

posterior probability

marginal likelihood
or: evidence
or: partition function
or: normalizer

Learning / Training



Model random data with hyperparameters θ :

$$y \sim p(y \mid \theta)$$

Sometimes we use:
$p(y; \theta)$

Given training data:

$$\{y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(y \mid \theta)$$

Learn parameters, e.g. via *maximum likelihood estimation*:

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log p(y_1, \ldots, y_n \mid \theta)$$

We will talk more about MLE in coming weeks

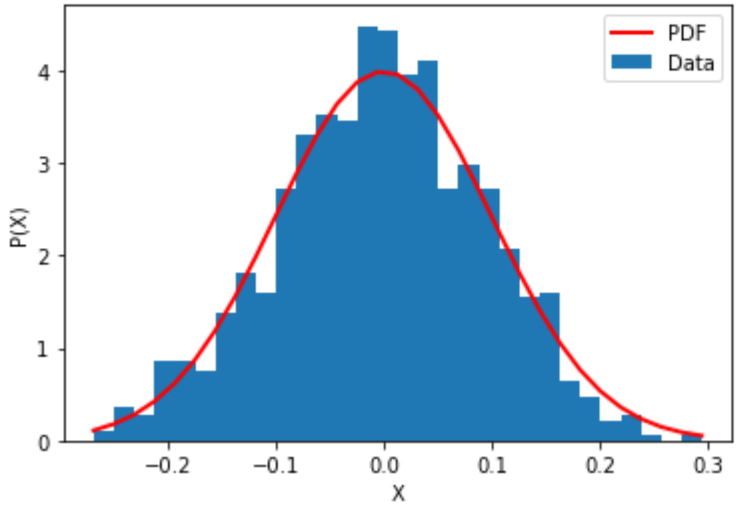
Other estimators are possible:

- *Maximum a posteriori* (MAP)
- *Minimum mean squared error (MMSE)*
- *Etc.*

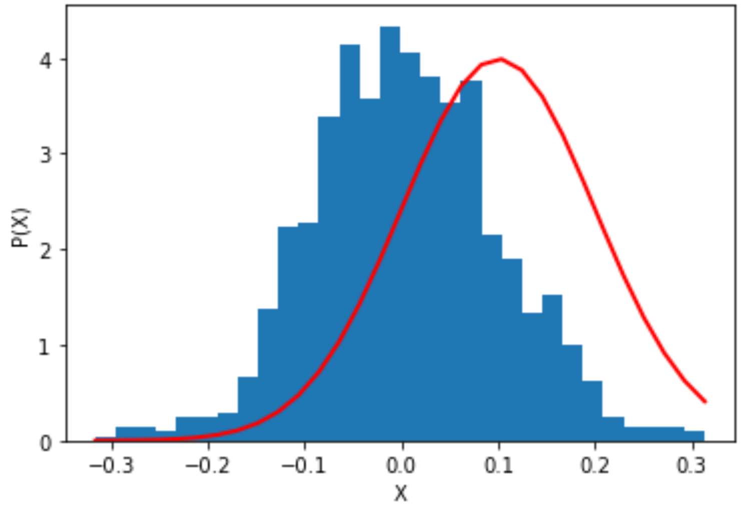
Likelihood (Intuitively)

Suppose we observe N data points from a Gaussian model and wish to estimate model parameters…

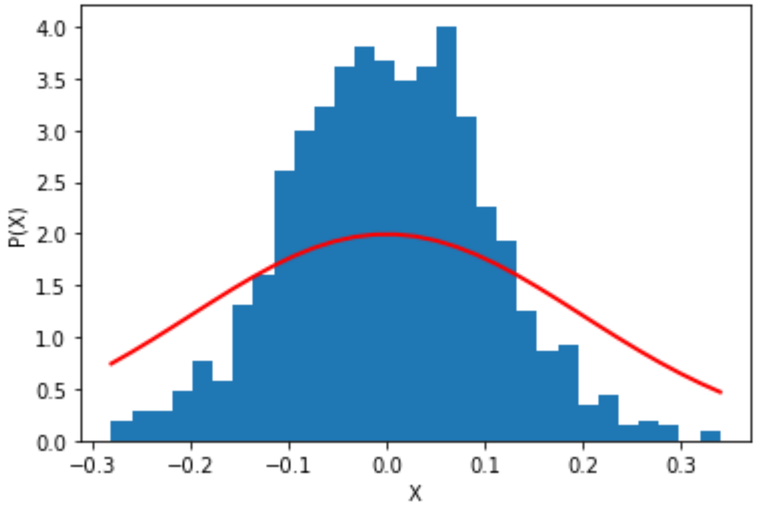
High Likelihood



Low Likelihood (mean)



Low Likelihood (variance)



Likelihood Principle *Given a statistical model, the likelihood function describes all evidence of a parameter that is contained in the data.*

Likelihood Function

Suppose $x_i \sim p(x;\theta)$, then what is the **joint probability** over N *independent identically distributed* (iid) observations $x_1, \ldots, x_N$?

$$p(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

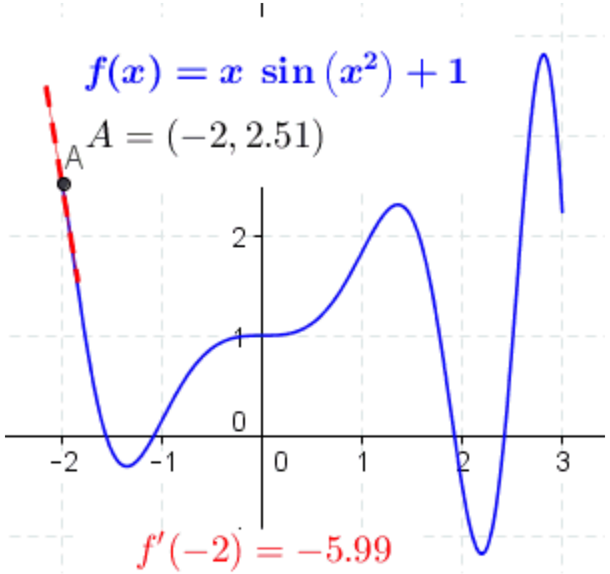
- We call this the **likelihood function**, often denoted $\mathcal{L}_N(\theta)$
- It is a function of the parameter θ, the data are fixed
- Measures how well parameter θ describes data (*goodness of fit*)

How could we use this to estimate a parameter θ?

Maximum Likelihood

Maximum Likelihood Estimator (MLE) as the name suggests, maximizes the likelihood function.

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \mathcal{L}_N(\theta) = \prod_{i=1}^{N} p(x_i;\theta)$$



Question How do we find the MLE?

Answer Remember calculus…

Is $\mathcal{L}_N(\theta)$ convex?

Yes → Unique, closed-form solution →

Approach

- Compute derivative $\frac{d\mathcal{L}_N}{d\theta}$
- Set to zero and solve

$$\frac{d\mathcal{L}_N}{d\theta} = 0 \Rightarrow \hat{\theta}^{\mathrm{MLE}}$$

No → Gradient-based optimization →

Still have to compute derivative…

Maximum Likelihood

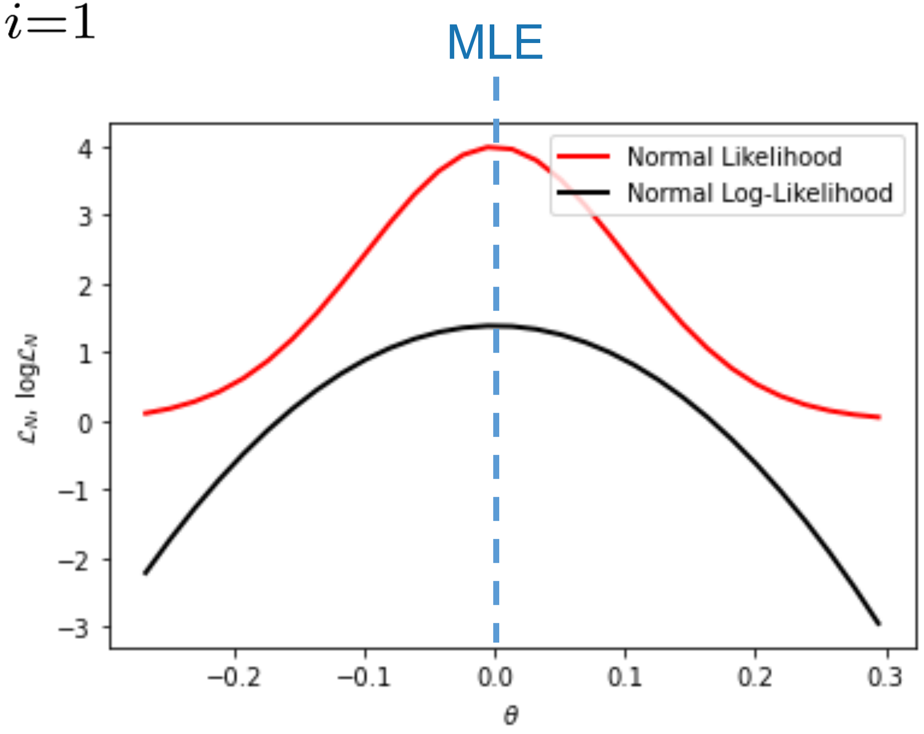
Maximizing log-likelihood makes the math easier (as we will see) and doesn't change the answer (logarithm is an increasing function)

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \log p(x_i;\theta)$$

Derivative is a linear operator so,

$$\frac{d}{d\theta} \log \mathcal{L}_N(\theta) = \underbrace{\sum_{i=1}^{N} \frac{d}{d\theta} \log p(x_i;\theta)}_{\text{One term per data point; Can be computed in parallel (big data)}}$$

One term per data point
Can be computed in parallel
(big data)

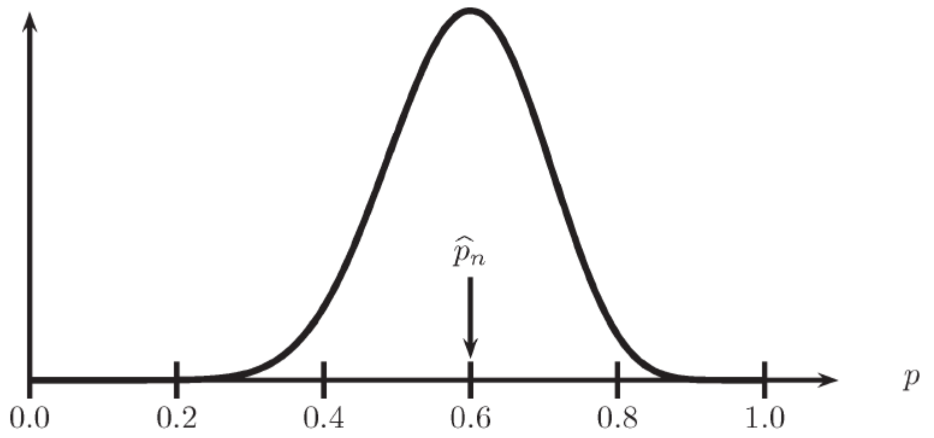


Maximum Likelihood

[Source: Wasserman, L. 2004]

Example Suppose we have N coin tosses with $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ but we don't know the coin bias p. The likelihood function is,

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^S(1-p)^{n-S}$$



Likelihood function for Bernoulli with n=20 and $\sum_i x_i = 12$ *heads*

where $S = \sum_i x_i$. The log-likelihood is,

$$\log \mathcal{L}_n(p) = S \log p + (n-S)\log(1-p)$$

Set the derivative of $\log \mathcal{L}_n(p)$ to zero and solve,

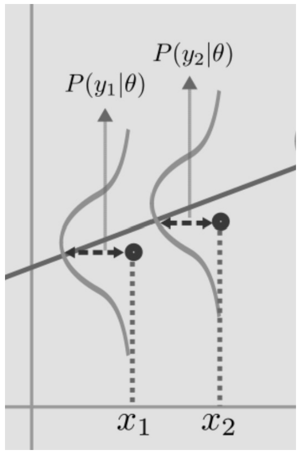
$$\hat{p}^{\text{MLE}} = S/n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

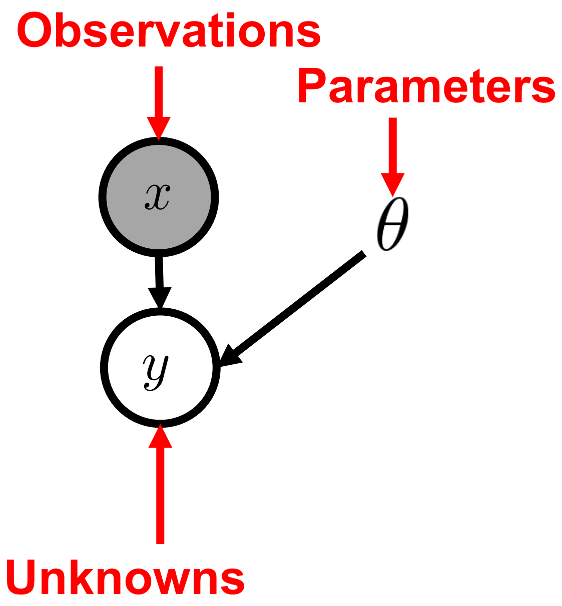
Maximum likelihood is equivalent to sample mean in Bernoulli

Discriminative vs Generative modeling

Discriminative model:

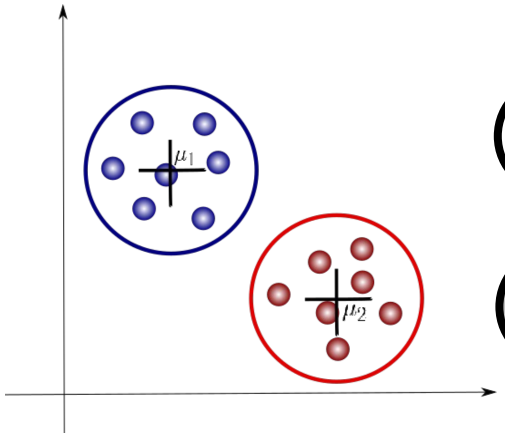
- Only models $P(\, y \mid x, \theta \,)$ -- i.e. *doesn't model data x*
- Recall linear regression: $y \mid x; \theta \sim N(x^\top \theta, \sigma^2)$
- Logistic regression: $y \mid x; \theta \sim \mathrm{Bernoulli}(\sigma(x^\top \theta))$

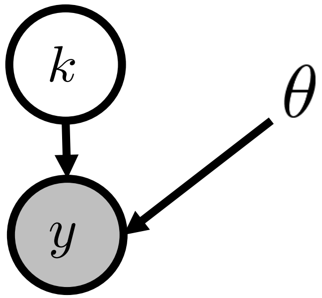




Generative model:

- Models everything including data: $P(k, y) = P(k)P(y \mid k, \theta)$
- e.g., Gaussian mixture model (GMM)
  - $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1}^{K}$
  - $k \sim \mathrm{Categorical}(\pi)$ (*hidden*), i.e. $P(k = l) = \pi_l$
  - $y \mid k \sim N(\mu_k, \Sigma_k)$

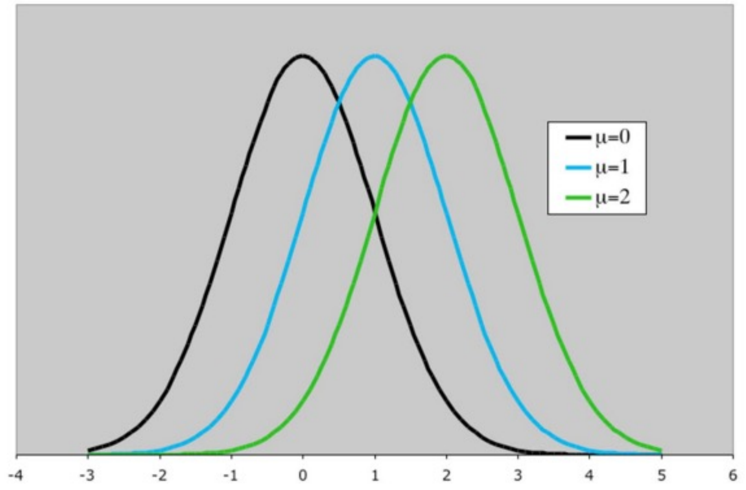




Barbershop Example

Suppose you go to a barbershop at every last Friday of the month. You want to be able to predict the waiting time. You have collected 12 data points (i.e., how long it took to be served) from the last year: $S = \{x_1, \ldots, x_{12}\}$

- 1. Modeling assumption: $x_i \sim$ Gaussian distribution $N(\mu, 1)$
  - $p(x; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$
  - Observation: this distribution has mean μ

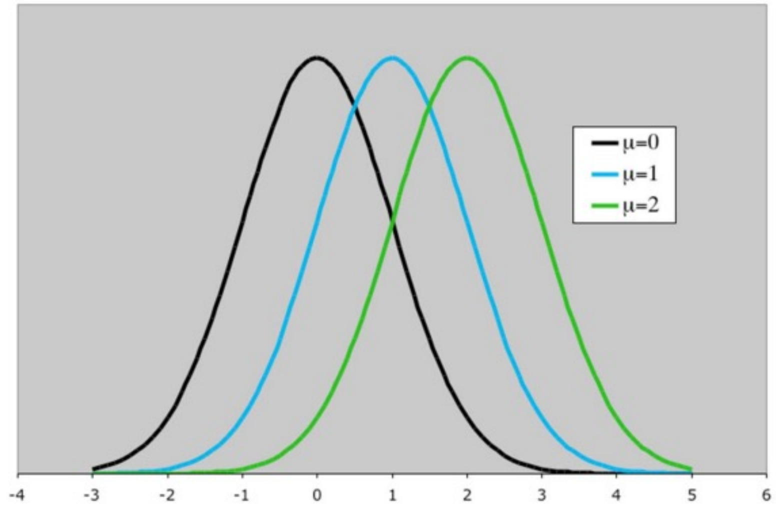


- 2. Find the MLE $\hat{\mu}$ from data S
  - (2.1) write down the neg. log likelihood of the sample

  $$L_n(\mu) = -\ln P(x_1, \ldots, x_n; \mu) = 12 \ln\sqrt{2\pi} + \frac{1}{2}\sum_{i=1}^{12}(x_i - \mu)^2$$

Is this a <u>generative</u> or <u>discriminative</u> model?

Generative model: basic example I (cont'd)



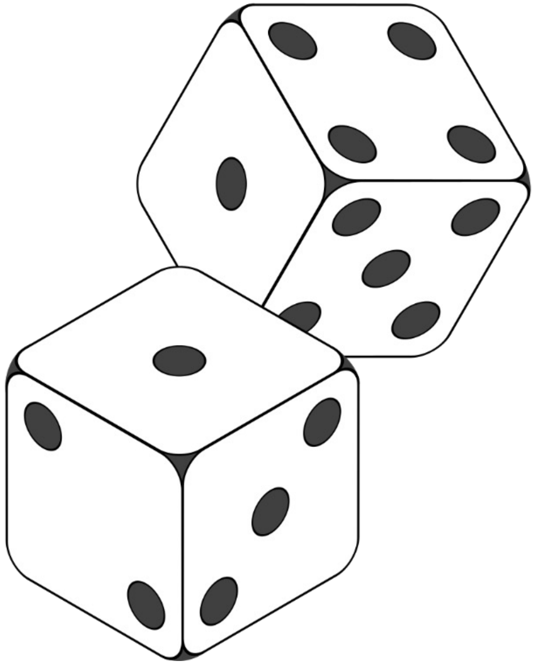
2. Find the MLE $\hat{\mu}$ from data S
   - (2.2) compute the first derivative, set it to 0, solve for λ (be sure to check convexity)

     $L_n'(\mu) = \sum_{i=1}^{12}(x_i - \mu) = 0 \Rightarrow \mu = \frac{x_1 + \cdots x_{12}}{12}$ Sample Mean

3. The learned model $N(\hat{\mu}, 1)$ is yours!
   - Simple prediction: e.g., predict the next wait time by $\mathbb{E}_{X \sim N(\hat{\mu},1)}[X]$
   - which is $\hat{\mu} = \frac{x_1 + \cdots x_{12}}{12}$

4. (Optional: Model Checking) Generate some data... Does it look realistic?

(Aside) Categorical Distribution



Distribution on integer-valued RV $X \in \{1, \ldots, K\}$

$$p(X) = \prod_{k=1}^{K} \pi_k^{\mathbf{I}(X=k)} \quad \text{or} \quad p(X) = \sum_{k=1}^{K} \mathbf{I}(X = k) \cdot \pi_k$$

with parameter $p(X = k) = \pi_k$ *and Kroenecker delta:*

$$\mathbf{I}(X = k) = \begin{cases} 1, & \text{If } X = k \\ 0, & \text{Otherwise} \end{cases}$$

Can also represent X as *one-hot* binary vector,

$X \in \{0, 1\}^K$ where $\sum_{k=1}^{K} X_k = 1$ then $p(X) = \prod_{k=1}^{K} \pi_k^{X_k}$

Basic Example II

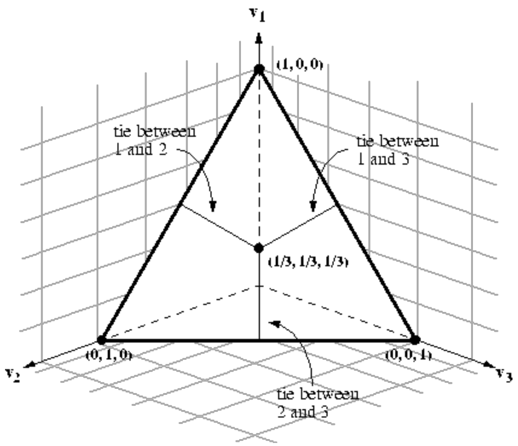
Data $S = \{y_i\}_{i=1}^{n}$, where $y_i \in \{1, \dots, C\}$

Generative Story

$y \sim \text{Categorical}(\pi)$, where $\pi = (\pi_1, \dots, \pi_C) \in \Delta^{C-1}$ ($\pi_c \geq 0$ and $\pi_1 + \cdots + \pi_C = 1$)

e.g. y_i = the color of i-th ball drawn randomly from a bin (with replacement)

$p(y; \pi) = \pi_y \left(= \prod_{c=1}^{C} \pi_c^{I(y=c)}\right)$



Training

(2.1) $L_n(\pi) = -\ln P(y_1, \dots, y_n; \pi) = \sum_{i=1}^{n} -\ln \pi_{y_i} = -\sum_{c=1}^{C} n_c \ln \pi_c$,

where $n_c = \#\{i : y_i = c\} = \sum_{i=1}^{n} I(y_i = c)$

Basic Example II (Cont'd)

Training

(2.2) $\text{minimize}_{\pi \in \Delta^{C-1}} L_n(\pi) := -\sum_{c=1}^{C} n_c \ln \pi_c$

Constrained maximization problem; solve by Lagrange multipliers

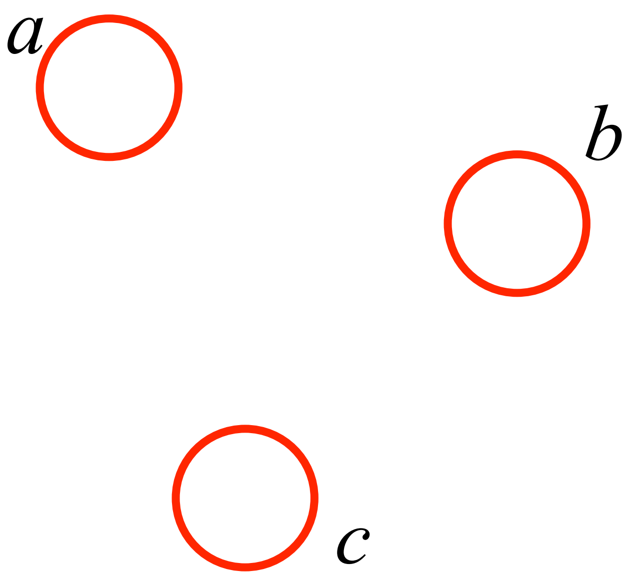
$$\frac{\partial}{\partial \pi}\left(-\sum_{c=1}^{C} n_c \ln \pi_c - \lambda\left(\sum_{c=1}^{C} \pi_c - 1\right)\right) = -\frac{n_c}{\pi_c} - \lambda = 0 \Rightarrow \pi_c = -\frac{n_c}{\lambda}$$

Combined with the constraint that $\pi_1 + \cdots + \pi_C = 1 \Rightarrow \hat{\pi}_c = \frac{n_c}{n}$, for all c

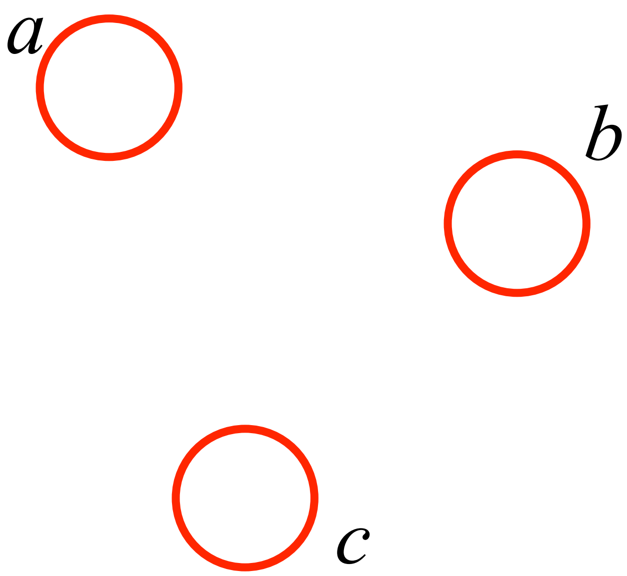
Test predict label $\text{argmax}_c P(y = c; \hat{\pi}) = \text{argmax}_c \hat{\pi}_c$

- Probability Refresher
- Probabilistic Graphical Models
- Naïve Bayes

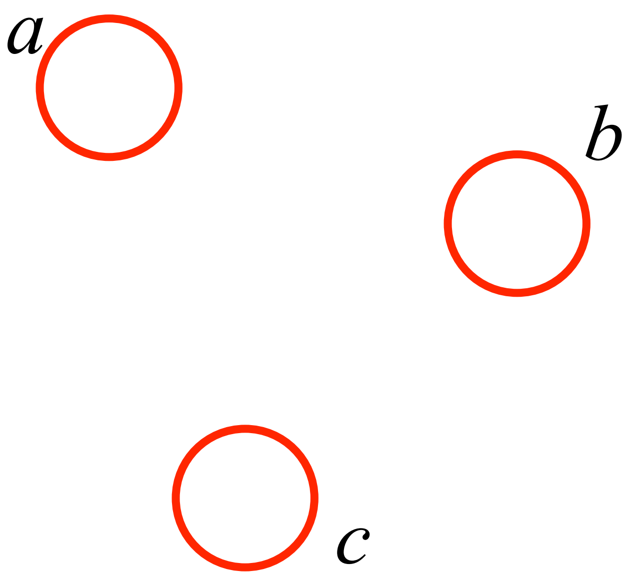
What is the joint factorization?



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

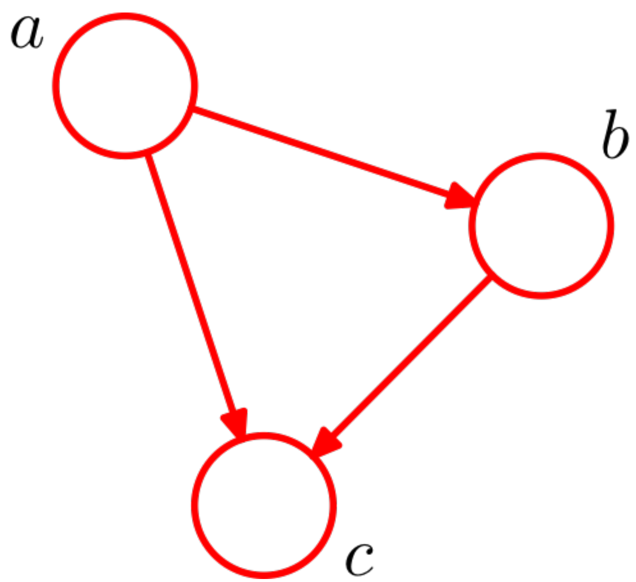


Are a and b independent ($a \perp b$)?



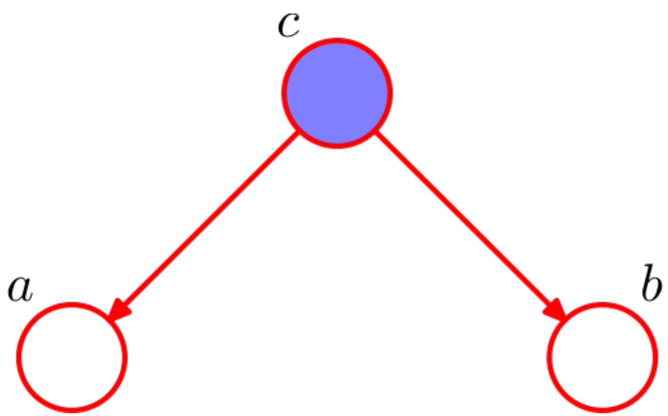
$$\mathbf{p(a,b,c) = p(a)p(b)p(c)}$$

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



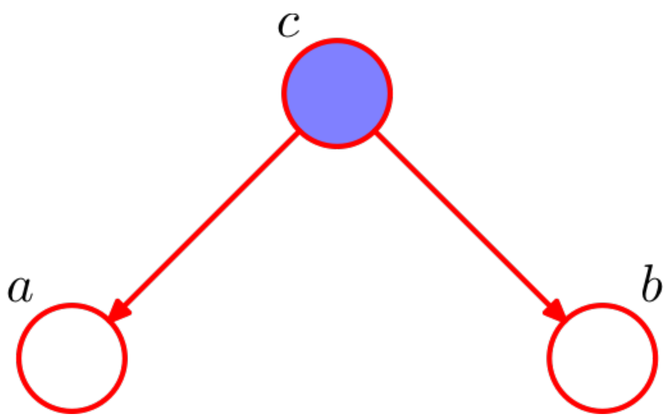
Note there are **no conditional independencies**

Case one where c is observed



Is $a \perp b \mid c$?

Case one where c is observed



$$a \perp b \mid c$$

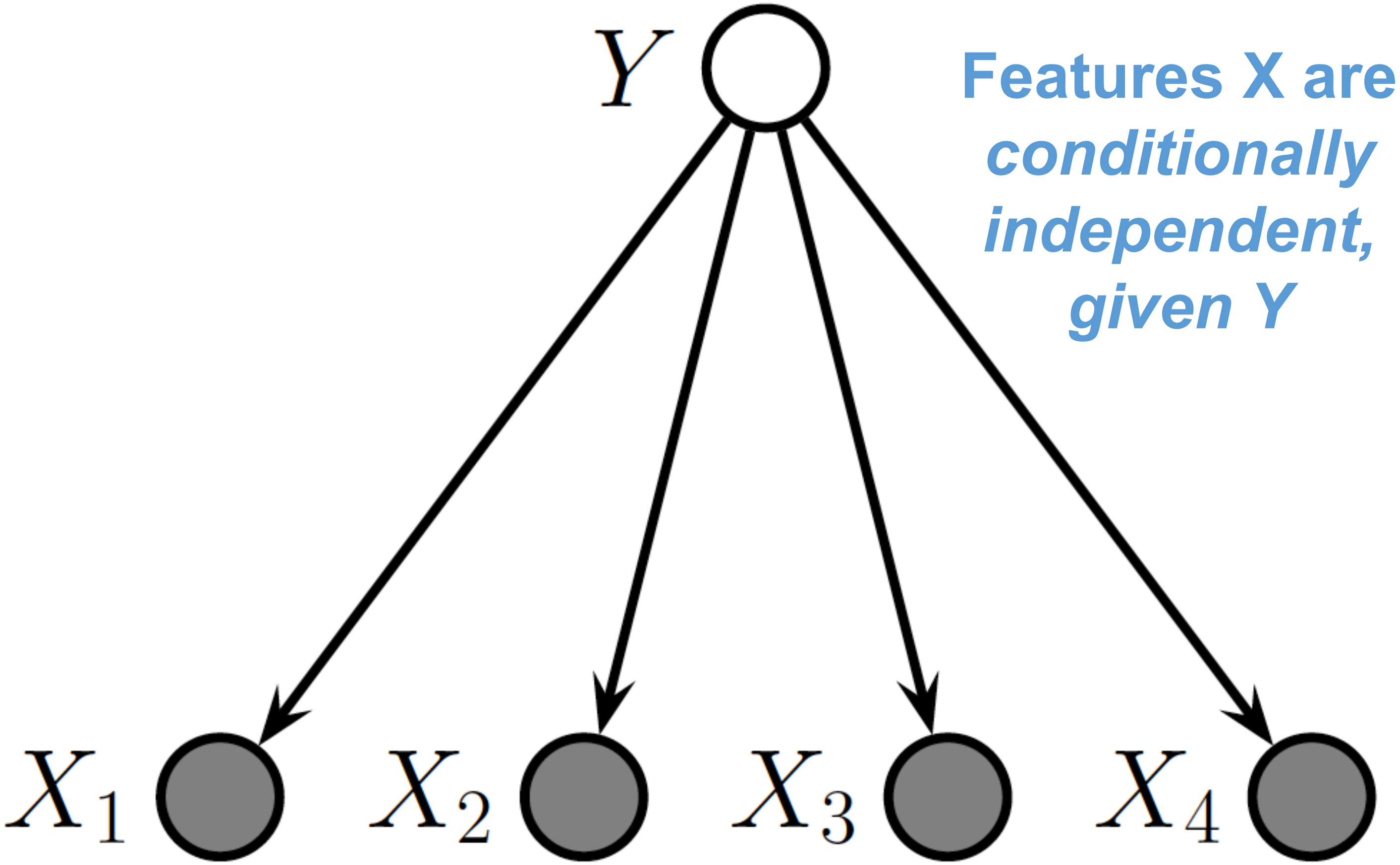
$$p\left(a,b,c\right) = p(c)p\left(a \middle| c\right)p\left(b \middle| c\right) \qquad \text{(what the graph represents in general)}$$

$$p\left(a,b \middle| c\right) = p\left(a \middle| c\right)p\left(b \middle| c\right) \qquad \text{(with } c \text{ observed)}$$

This is the definition of $a \perp b \middle| c$

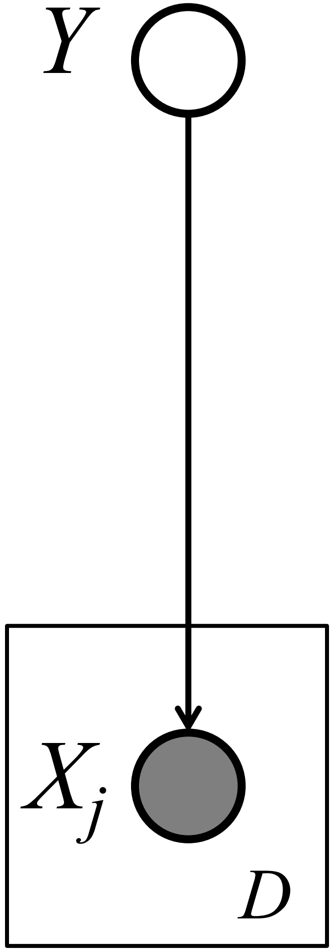
Shading & Plate Notation

***Convention:** Shaded nodes are observed, open nodes are latent/hidden/unobserved*



Features X are conditionally independent, given Y

$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y)$$



Plates denote replication of random variables

Naïve Bayes for supervised learning

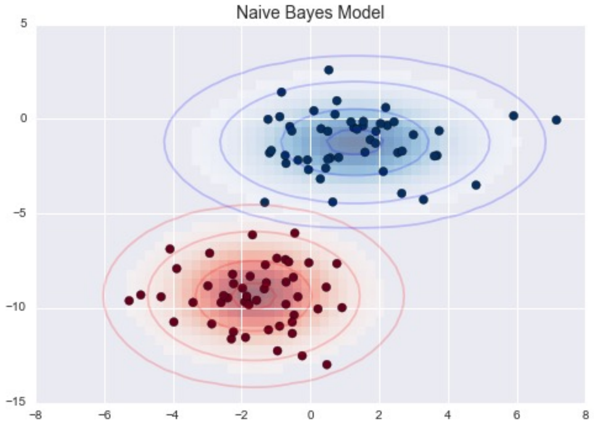
- Motivation: supervised learning for classification
- high-dimensional $x = (x(1), \dots, x(F))$, modeling $P(x \mid y)$ can be tricky
- In general, $P(x \mid y) = P(x(1) \mid y) \cdot P(x(2) \mid x(1), y) \cdot \dots \cdot P(x(F) \mid x(1), \dots, x(F-1), y)$

- A modeling assumption: $x(1), \dots, x(F)$ are conditionally independent given y

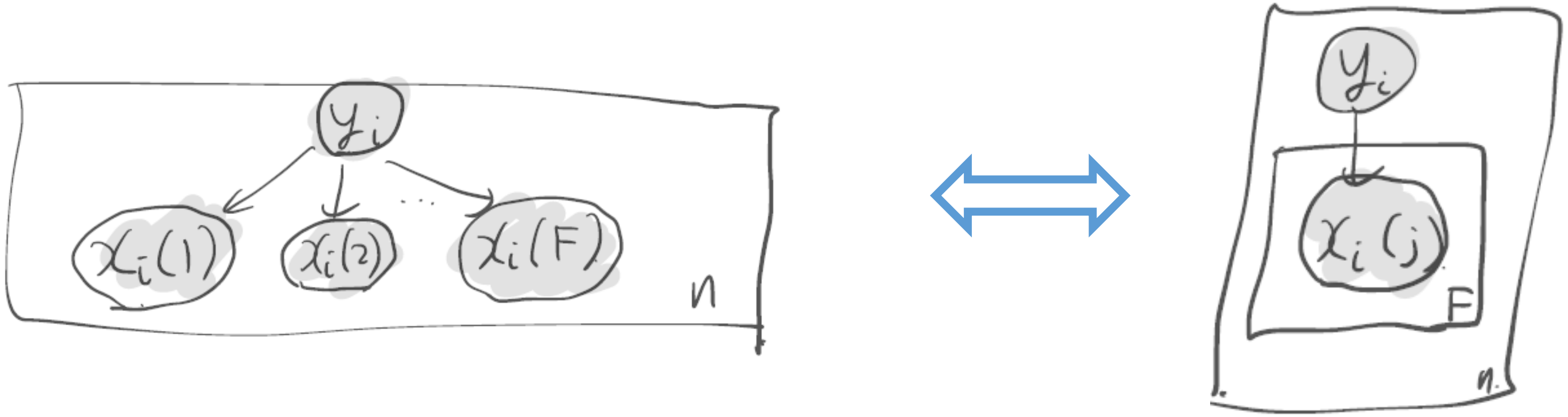
  i.e. for all i

$$x(i) \perp\!\!\!\perp \big(x(1), \dots, x(i-1), x(i+1), \dots, x(F)\big) \mid y$$

  (Conditional independence notation: $A \perp\!\!\!\perp B \mid C$)



- Equivalently $P(x \mid y) = P(x(1) \mid y) \cdot \dots P(x(F) \mid y)$



Recall : Class Preference Prediction

Define the labeled training dataset $S = \{(x_i, y_i)\}_{i=1}^{m}$

Features

Feature Values

Labels

Data Point

Rating	Easy?	AI?	Sys?	Thy?	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

To make this a <u>binary</u> classification we set
"Liked" = {+2,+1,0}
"Nah" = {-1,-2}

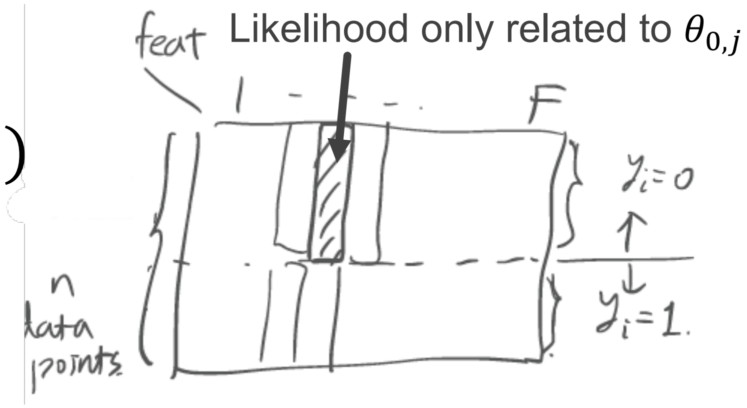
Naïve Bayes: binary-valued features

Training Data $S = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \{0,1\}^F$ $y_i \in \{0,1\}$

Generative Story

$y \sim \text{Bernoulli}(\pi)$; for all $j \in [F]$, $x(j) \mid y = c \sim \text{Bernoulli}(\theta_{c,j})$

#parameters = $1 + 2F$



Training (denote by $\theta = \{\theta_{c,j}\}$)

$$\max_{\pi,\theta} \sum_{i=1}^{n} \ln P(x_i, y_i; \pi, \theta) = \sum_{i=1}^{n} \ln P(y_i; \pi) + \sum_{i=1}^{n} \ln P(x_i \mid y_i; \theta)$$

$$= \max_{\pi} \sum_{i=1}^{n} \ln P(y_i; \pi) + \max_{\{\theta_{0,j}\}} \sum_{i:y_i=0} \ln P(x_i \mid y_i; \theta) + \max_{\{\theta_{1,j}\}} \sum_{i:y_i=1} \ln P(x_i \mid y_i; \theta)$$

Key observation: optimal π, optimal $\{\theta_{0,j}\}$, optimal $\{\theta_{1,j}\}$ can be found separately

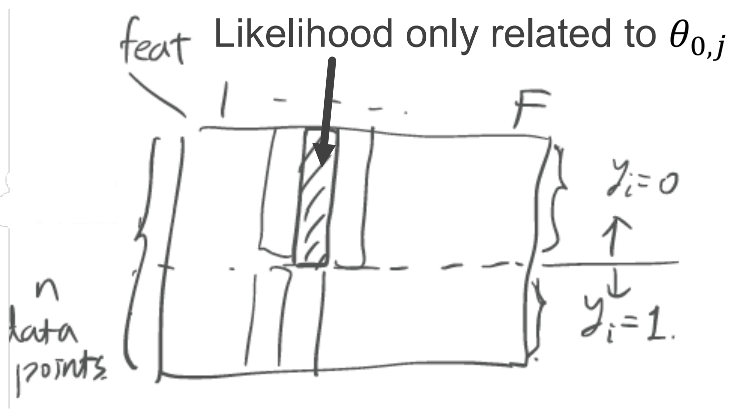
Optimal π: $\max_{\pi} \sum_{i=1}^{n} \ln P(y_i; \pi) = \max_{\pi} n_0 \ln(1-\pi) + n_1 \ln(\pi)$ => $\hat{\pi} = \frac{n_1}{n}$

Naïve Bayes: binary-valued features (cont'd)

By the Naïve Bayes modeling assumption,

$$\max_{\{\theta_{0,j}\}} \sum_{i:y_i=0} \ln P(x_i \mid y_i; \theta) = \max_{\{\theta_{0,j}\}} \sum_{j=1}^{F} \sum_{i:y_i=0} \ln P\big(x_i(j) \mid y_i; \theta_{0,j}\big)$$

$$= \sum_{j=1}^{F} \max_{\theta_{0,j}} \sum_{i:y_i=0} \ln P\big(x_i(j) \mid y_i; \theta_{0,j}\big)$$



Again, can optimize each $\theta_{0,j}$ separately,

- Optimal $\theta_{0,j}$: $\max_{\theta_{0,j}} \sum_{i:y_i=0,\, x_i(j)=1} \ln \theta_{0,j} + \sum_{i:y_i=0,\, x_i(j)=0} \ln (1-\theta_{0,j})$

$$\hat{\theta}_{0,j} = \frac{\#\{i: y_i=0, x_i(j)=1\}}{\#\{i: y_i=0\}}; \quad \text{similarly,} \quad \hat{\theta}_{1,j} = \frac{\#\{i: y_i=1, x_i(j)=1\}}{\#\{i: y_i=1\}}$$

Naïve Bayes: binary-valued features (cont'd)

Test Given $\hat{\pi}$, $\{\hat{\theta}_{c,j}\}$, Bayes optimal classifier

$$\hat{f}_{BO}(x) = \text{argmax}_y \, P(x, y; \hat{\pi}, \{\hat{\theta}_{c,j}\}) = \text{argmax}_y \log P(x, y; \hat{\pi}, \{\hat{\theta}_{c,j}\})$$

- $\log P(x, y = 0; \pi, \{\theta_{c,j}\}) = \ln(1-\pi) + \sum_{j=1}^{F} \ln P\big(x(j) \mid y; \theta_{0,j}\big)$

$$= \ln(1-\pi) + \sum_{j=1}^{F} \ln\big(1-\theta_{0,j}\big) I(x(j) = 0) + \ln\big(\theta_{0,j}\big) I(x(j) = 1)$$

$$= \ln(1-\pi) + \sum_{j=1}^{F} \ln(1-\theta_{0,j}) + \sum_{j=1}^{F} x(j) \ln \frac{\theta_{0,j}}{1-\theta_{0,j}}$$

- Similarly, $\log P(x, y = 1; \pi, \{\theta_{c,j}\}) = \ln(\pi) + \sum_{j=1}^{F} \ln(1-\theta_{1,j}) + \sum_{j=1}^{F} x(j) \ln \frac{\theta_{1,j}}{1-\theta_{1,j}}$
- Therefore, $\hat{f}_{BO}(x) = 1 \Leftrightarrow \underbrace{\ln\left(\frac{\pi}{1-\pi}\right) + \sum_{j=1}^{F} \ln\left(\frac{1-\theta_{1,j}}{1-\theta_{0,j}}\right)}_{b} + \sum_{j=1}^{F} x(j) \underbrace{\left(\ln \frac{\theta_{1,j}}{1-\theta_{1,j}} - \ln \frac{\theta_{0,j}}{1-\theta_{0,j}}\right)}_{w(j)} \geq 0$
- I.e. Bayes classifier is *linear*

Naïve Bayes: Discrete (Categorical-valued) features

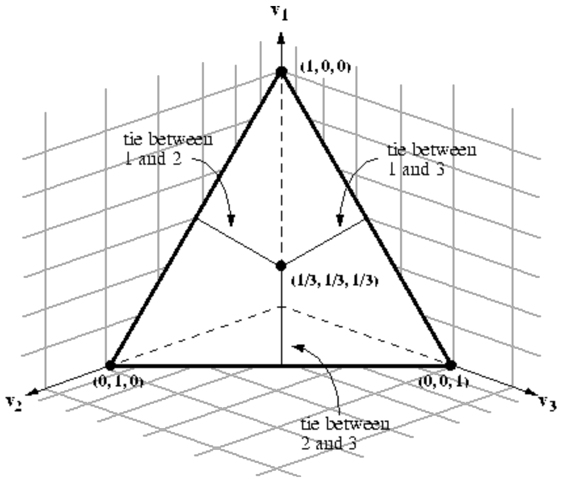
Data $S = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in [W]^F$ $y_i \in \{0,1\}$

Generative story

$y \sim \text{Bernoulli}(\pi)$; for all $j \in [F]$, $x(j) \mid y = c \sim \text{Categorical}(\theta_c)$ $(\theta_c \in \Delta^{W-1})$

#parameters = $1 + 2W$

Note: in this example, θ_c *shared across all features*!

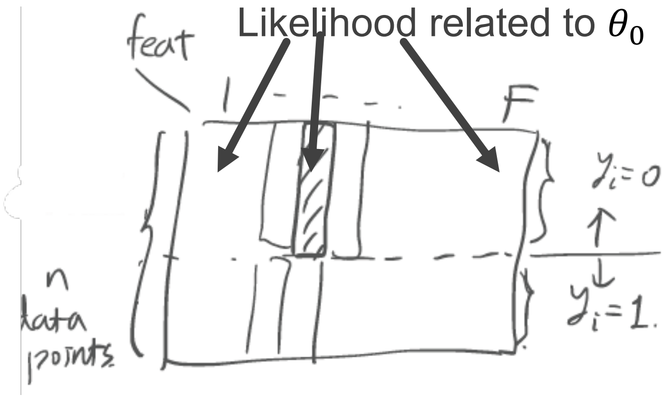


Training

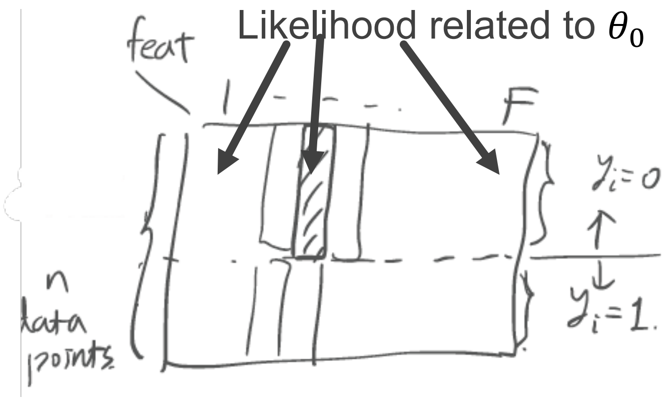
Similar to previous example, optimal π, optimal θ_0, optimal θ_1 can be found separately,

by maximizing the respective part of the likelihood function (exercise)

Optimal π same as previous example



Naïve Bayes: Discrete features (cont'd)



Training

Optimal θ_c:

$$\max_{\theta_0} \sum_{i:y_i=0} \ln P(x_i \mid y_i; \theta_0) = \max_{\theta_0} \sum_{j=1}^{F} \sum_{i:y_i=0} \ln P(x_i(j) \mid y_i; \theta_0)$$

$$= \max_{\theta_0} \sum_{w=1}^{W} \sum_{j=1}^{F} \sum_{i:y_i=0} I(x_i(j) = w) \ln \theta_{0,w}$$

$$= \max_{\theta_0} \sum_{w=1}^{W} \ln \theta_{0,w} \, \#\{(i,j): y_i = 0, x_i(j) = w\}$$

=> $\hat{\theta}_{c,w} = \frac{\#\{(i,j):\, y_i=c, x_i(j)=w\}}{\#\{i:\, y_i=c\} \times F}$

Exercise: how to extend this to variable-length x_i's (e.g. for text classification)?

Test

Bayes optimal classification rule with $(\hat{\pi}, \hat{\theta}_0, \hat{\theta}_1)$ (exercise)

Summary

Fundamental rules of Probability:

- Law of total probability: $p(Y) = \sum_x p(Y, X = x)$
- Probability chain rule: $p(X \mid Y) = \frac{p(X,Y)}{p(Y)}$
- Conditional probability: $p(X, Y) = p(Y)p(X \mid Y)$

Independence of Random Variables:

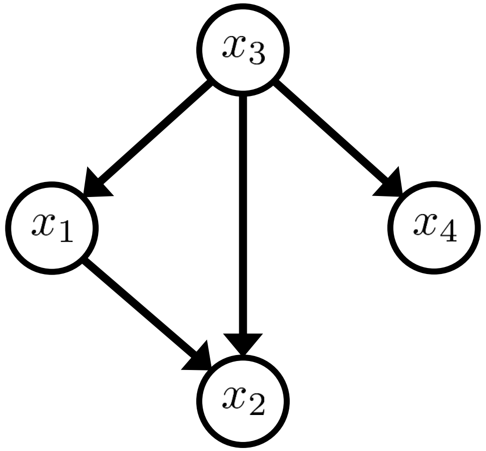
- Two RVs are independent if: $p(X = x, Y = y) = p(X = x)p(Y = y)$
- Or: $p(X \mid Y) = p(X)$
- They are *conditionally independent* if:

$$p(X = x, Y = y \mid Z = z) = p(X = x \mid Z = z)p(Y = y \mid Z = z)$$

- Or: $p(X \mid Y, Z) = p(X \mid Z)$

Summary

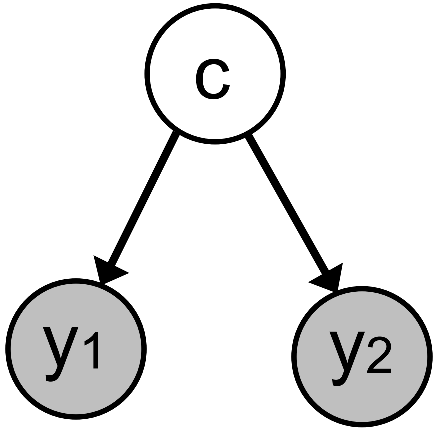
A Bayes Network expresses a unique probability factorization:



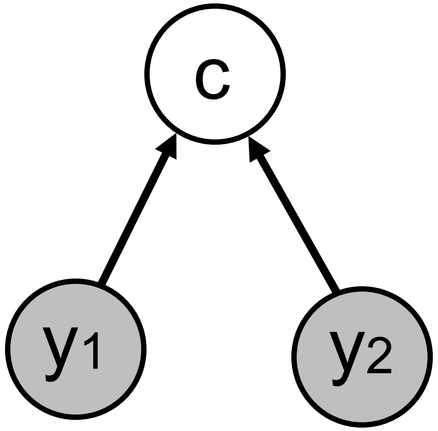
$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid x_{\mathrm{Pa}(s)})$$

Parents of node *s*

Inference is performed by Bayes' rule (posterior distribution):

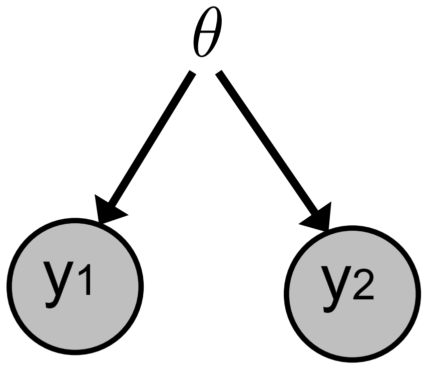


$$p(c \mid y_1, y_2) = \frac{p(c)p(y_1 \mid c)p(y_2 \mid c)}{p(y_1, y_2)}$$



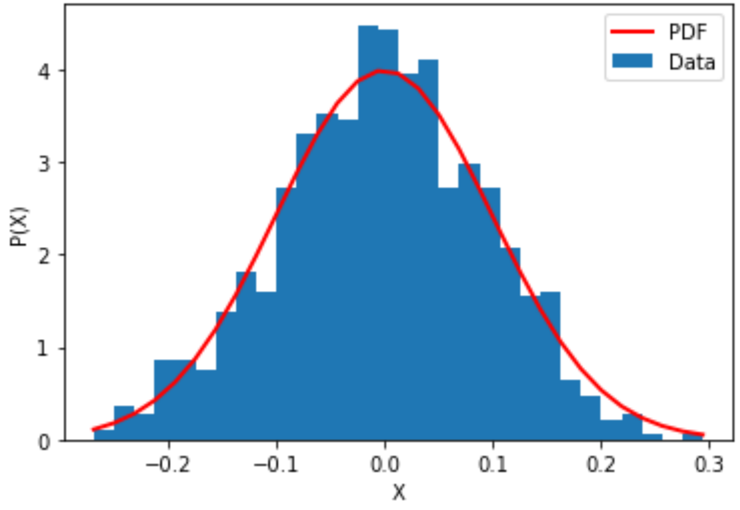
Summary

Hyperparameters must be estimated (e.g. Maximum Likelihood):

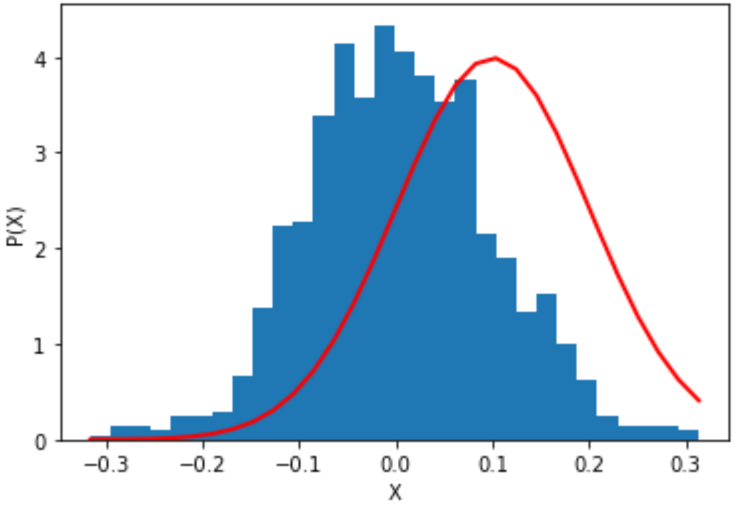


$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log p(y_1, \ldots, y_n \mid \theta)$$

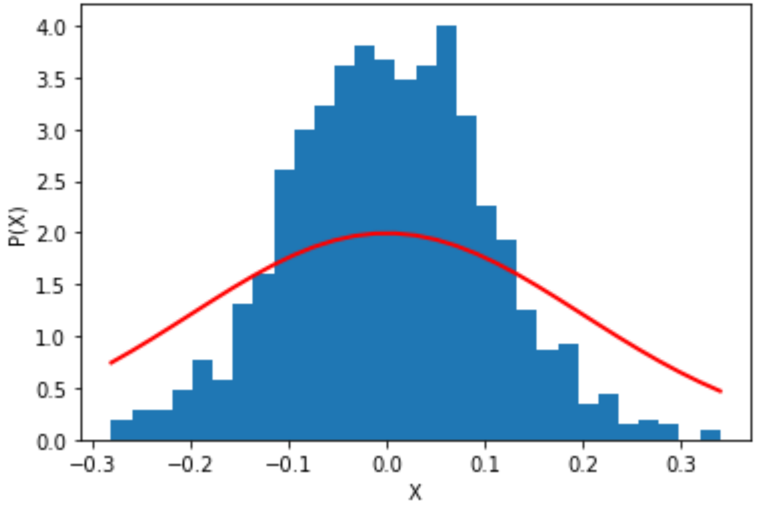
High Likelihood



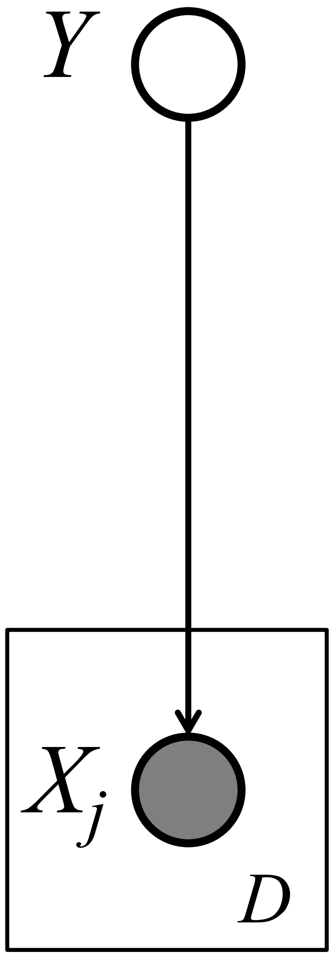
Low Likelihood (mean)



Low Likelihood (variance)



Summary



Naïve Bayes classifier assumes features are *conditionally independent* given class Y:

$$x(j) \perp\!\!\!\perp \big(x(1), \dots, x(j-1), x(j+1), \dots, x(D)\big) \mid y$$

Joint distribution factorizes as:

$$p(x, y) = p(y) \prod_{\{j=1\}}^{D} p(x(j) \mid y)$$

Allows easier fitting of hyperparameters for *class conditional distributions* (they can be fit independently of each other)