Reminders: August 31, 2015

- Assignment 1 is due on September 16
- Thought questions 1 are due on September 9
    - Chapters 1 and 2
- My office hours are today
    - Feel free to also email me for an appointment

Previous questions

- Can discrete and continuous distributions be unified?
  - CDFs and characteristic functions were mentioned
  - discretization was also mentioned
- Why do we focus on pmfs and pdfs?
  - i.e., when CDFs and characteristic functions are more general?
  - for our purposes, pdfs will be key (this will become more clear)
- How do we formally write down a proof?

Proof Example

Using only the definition of a sigma field, prove that a sigma field $\mathcal{F}$ is closed under set difference: $A_1, A_2 \in \mathcal{F} \implies A_1 \backslash A_2 \in \mathcal{F}$

Proof: First, note that $\mathcal{F}$ is closed under intersection, because

1. $A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$ by DeMorgan's laws;
2. $A_1^c, A_2^c \in \mathcal{F}$ because $\mathcal{F}$ is closed under complementation;
3. $A_1^c \cup A_2^c \in \mathcal{F}$ because $\mathcal{F}$ is closed under union;
4. finally $(A_1^c \cup A_2^c)^c \in \mathcal{F}$ by closure under complementation.

First rewrite set difference as

$$A_1 \setminus A_2 = (A_1 \cap A_2)^c \cap A_1$$

Then from the above argument, $A_1 \cap A_2 \in \mathcal{F}$, and by closure under complementation $(A_1 \cap A_2)^c \in \mathcal{F}$ and finally we can use closure under intersection to obtain $(A_1 \cap A_2)^c \cap A_1 \in \mathcal{F}$.

Independence of Events

$(\Omega, \mathcal{F}, P)$ = a probability space

Events A and B are **independent** if:

$$P(A \cap B) = P(A) \cdot P(B)$$

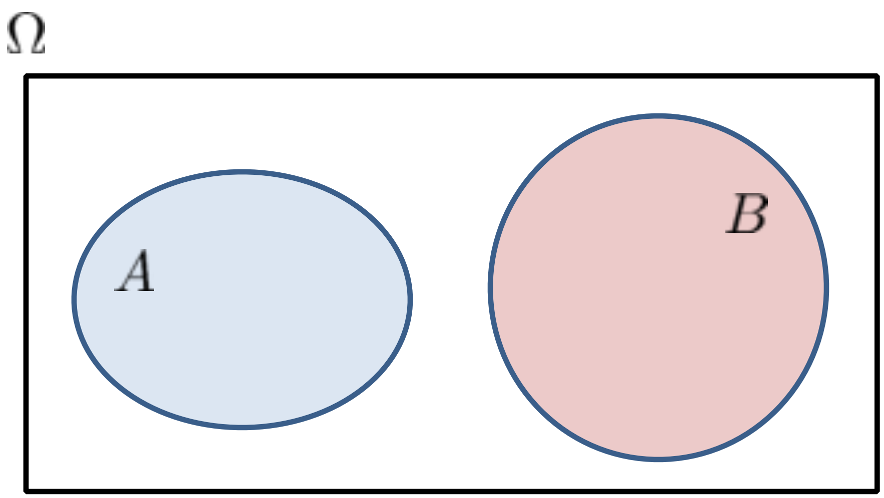
Events A and B are **conditionally independent** given C if:

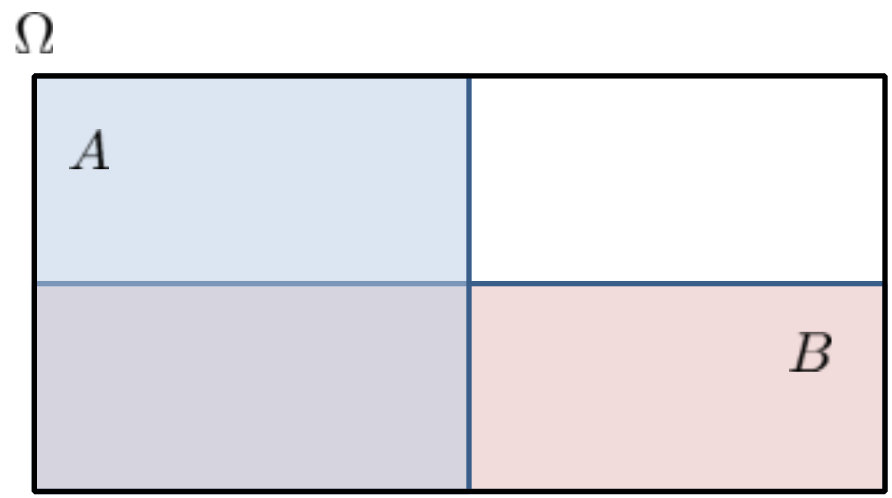
$$P(A \cap B | C) = P(A|C) \cdot P(B|C)$$

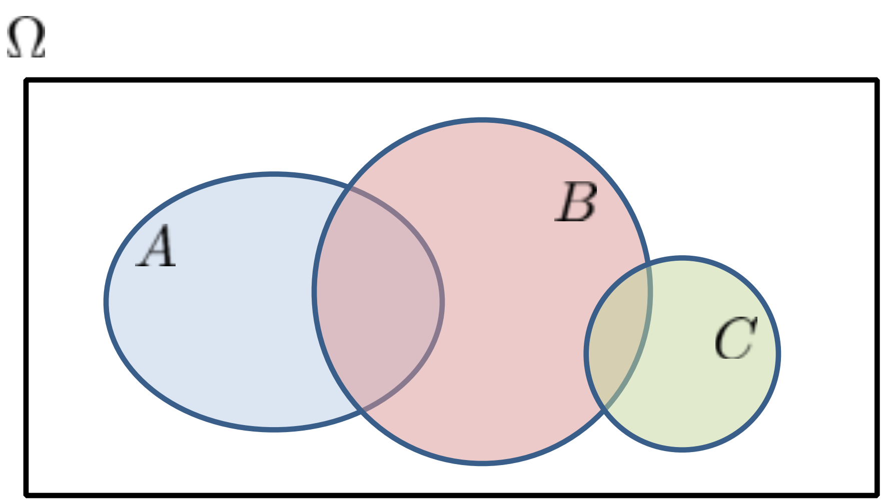
What if we had multiple events?

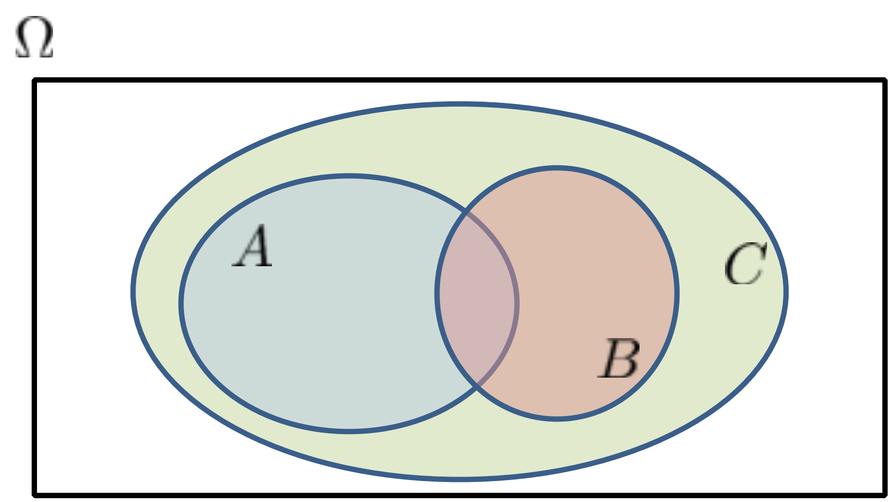
Independence Examples

$(\Omega, \mathcal{F}, P)$ = a probability space









Conditional Independence Examples

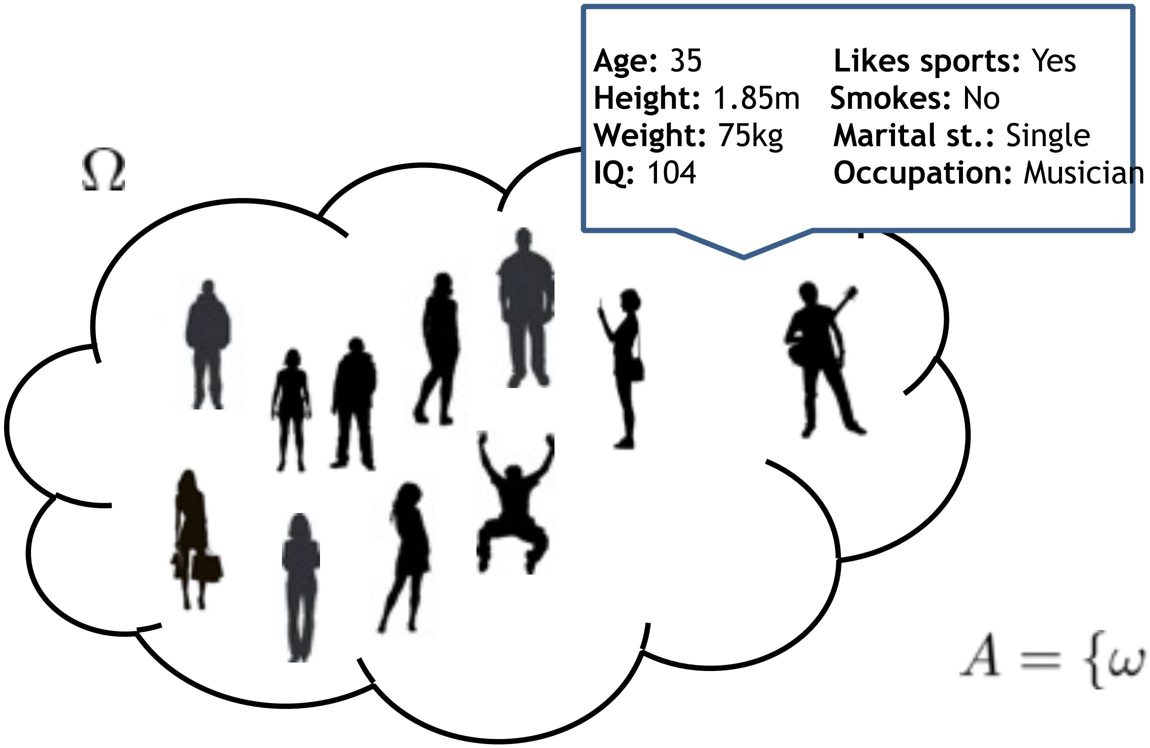
- Let Omega = {1,2,3,4,5,6} (die roll)
- Let A = {3,5}, B = {2,3}, C = {3,4}
- Are A and B conditionally independent given C?
- Recall: P(A | C) = P(A int C)/P(C)
- Recall: CI only if P(A int B | C) = P(A | C) P(B | C)

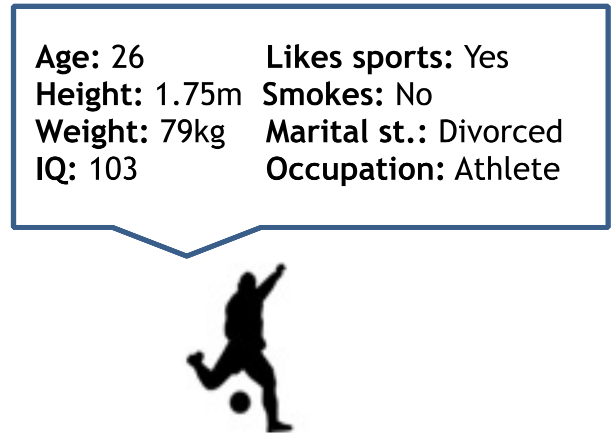
- P(B|C) = P(B int C)/P(C) = (1/6)/(1/3) = 1/2
- P(A|C) = 1/2
- P(A int B | C) = P({3} | C) = 1/2
- What if A = {1,2}?

Random Variables

$$(\Omega, \mathcal{F}, P)$$

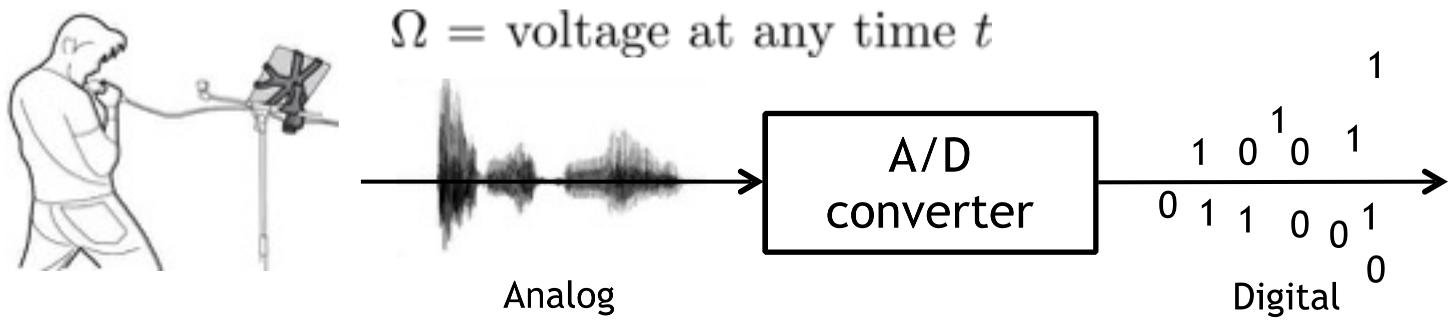
Ω





$$A = \{\omega \in \Omega : Musician(\omega) = \text{yes}\}$$

Ω = voltage at any time t



We instinctively create this transformation

Assume Ω is a set of people.

Compute the probability that a randomly selected person $\omega \in \Omega$ has a cold.

Define event $A = \{\omega \in \Omega : \text{Disease}(\omega) = \text{cold}\}$.

Disease is our new random variable, $P(Disease = cold)$

Disease is a function that maps outcome space to new outcome space $\{\text{cold}, \text{not cold}\}$

RANDOM VARIABLES

Example: three consecutive (fair) coin tosses
X = the number of heads in the first toss
Y = the number of heads in all three tosses
Find the probability spaces after the transformations.

Where is the probability space $(\Omega, \mathcal{F}, P)$?

Where is the randomness?

$$\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$$

$$\mathcal{F} = \mathcal{P}(\Omega)$$

$$P = ?$$

$$P(\Omega) = 1$$

$$P(\{\text{HHH}, \text{TTT}\}) = \tfrac{2}{8}$$

$$\vdots$$

Random Variables

$X : \Omega \to \{0, 1\}$
$Y : \Omega \to \{0, 1, 2, 3\}$

ω	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$X(\omega)$	1	1	1	1	0	0	0	0
$Y(\omega)$	3	2	2	1	2	1	1	0

What are the probability spaces $(\Omega_X, \mathcal{F}_X, P_X)$ and $(\Omega_Y, \mathcal{F}_Y, P_Y)$?

Where does the randomness come from?

Random Variable: Formal Definition

$(\Omega, \mathcal{F}, P)$ = a probability space

Random variable:

1. $X : \Omega \to \Omega_X$
2. $\forall A \in \mathcal{B}(\Omega_X)$ it holds that $\{\omega : X(\omega) \in A\} \in \mathcal{F}$

It follows that:

$$P_X(A) = P(\{\omega : X(\omega) \in A\})$$

Discrete Random Variable

$(\Omega, \mathcal{F}, P)$ = a discrete probability space

Probability mass function (pmf):

$$\begin{aligned} p_X(x) &= P_X(\{x\}) \\ &= P(\{\omega : X(\omega) = x\}) \end{aligned} \qquad \forall x \in \Omega_X$$

The probability of an event A:

$$P_X(A) = \sum_{x \in A} p_X(x) \qquad \forall A \subseteq \Omega_X$$

$P(\{\omega : X(\omega) \in A\})$

Continuous Random Variable

Cumulative distribution function (cdf):

$$\begin{aligned} F_X(t) &= P_X\left(\{x : x \leq t\}\right) \\ &= P_X\left((-\infty, t]\right) \\ &= P\left(X \leq t\right) \\ &= P\left(\{\omega : X(\omega) \leq t\}\right) \end{aligned}$$

Probability density function (pdf), if it exists:

$$p_X\left(x\right) = \left.\frac{dF_X\left(t\right)}{dt}\right|_{t=x}$$

Continuous Random Variable

If the probability density function (pdf) exists:

$$F_X(t) = \int_{-\infty}^{t} p_X(x)\,dx$$

The probability of an event $A = (a, b]$:

$$\begin{aligned} P_X((a,b]) &= \int_a^b p_X(x)\,dx \\ &= F_X(b) - F_X(a) \end{aligned}$$

$P(a < X \leq b)$

Joint and Marginal Distributions

$(\Omega, \mathcal{F}, P)$ = a discrete probability space

Joint probability distribution:

$$
\begin{aligned}
p_{XY}(x, y) &= P(X = x, Y = y) \\
&= P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\})
\end{aligned}
$$

Extend to k-D vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$

Marginal probability distribution:

$$
p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} p_{\boldsymbol{X}}(x_1, \ldots, x_k)
$$

Joint and Marginal Distributions

$$(\Omega, \mathcal{F}, P) = \left(\mathbb{R}^k, \mathcal{B}(\mathbb{R})^k, P_{\boldsymbol{X}}\right) = \text{a continuous probability space}$$

Joint probability distribution:

$$\begin{aligned} F_{\boldsymbol{X}}(\boldsymbol{t}) &= P_{\boldsymbol{X}}\left(\{\boldsymbol{x} : x_i \leq t_i, i = 1 \ldots k\}\right) \\ &= P\left(X_1 \leq t_1, X_2 \leq t_2 \ldots\right) \end{aligned}$$

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \left.\frac{\partial^k}{\partial t_1 \cdots \partial t_k} F_{\boldsymbol{X}}(t_1, \ldots t_k)\right|_{\boldsymbol{t}=\boldsymbol{x}} \qquad \text{(if it exists)}$$

Marginal probability distribution:

$$p_{X_i}(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_k} p_{\boldsymbol{X}}(\boldsymbol{x})\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_k$$

Conditional Distributions

Conditional probability distribution:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

The probability of an event A, given that $X = x$, is:

$$P_{Y|X}(Y \in A | X = x) = \begin{cases} \sum_{y \in A} p_{Y|X}(y|x) & Y : \text{discrete} \\ \int_{y \in A} p_{Y|X}(y|x) dy & Y : \text{continuous} \end{cases}$$

Chain Rule

Conditional probability distribution:

$$p(x_k|x_1,\dots,x_{k-1}) = \frac{p(x_1,\dots,x_k)}{p(x_1,\dots,x_{k-1})}$$

This leads to:

$$p(x_1,\dots,x_k) = p(x_1)\prod_{l=2}^{k} p(x_l|x_1,\dots,x_{l-1})$$

Independence of Random Variables

X and Y are **independent** if:

$$p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$$

X and Y are **conditionally independent** given Z if:

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) \cdot p_{Y|Z}(y|z)$$

What if we had k random variables?

Conditional Independence Examples

- Let Z = bias of a coin (say outcomes are 0.3, 0.5, 0.8 with associated probabilities 0.7, 0.2, 0.1)
- Let X and Y be independent flips of the coin
- Are X and Y independent?
- Are X and Y conditionally independent, given Z?

EXPECTATIONS

$(\Omega_X, \mathcal{B}(\Omega_X), P_X)$ = a probability space

Consider a function $f : \Omega_X \to \mathbb{C}$

$$E_x\left[f(x)\right] = \begin{cases} \sum_{x \in \Omega_X} f(x) p_X(x) & X : \text{discrete} \\ \\ \int_{\Omega_X} f(x) p_X(x) dx & X : \text{continuous} \end{cases}$$

Expectations You Know About

$f(x)$	Symbol	Name
x	$E[X]$	Mean
$(x - E[X])^2$	$V[X]$	Variance
x^k	$E[X^k]$	k-th moment; $k \in \mathbb{N}$
$(x - E[X])^k$	$E[(x - E[X])^k]$	k-th central moment; $k \in \mathbb{N}$
e^{tx}	$M_X(t)$	Moment generating function
e^{itx}	$\varphi_X(t)$	Characteristic function
$\log \frac{1}{p_X(x)}$	$H(X)$	(Differential) entropy
$\log \frac{p_X(x)}{q(x)}$	$D(p_X \| q)$	Kullback-Leibler divergence
$\left(\frac{\partial}{\partial\theta} \log p_X(x\|\theta)\right)^2$	$\mathcal{I}(\theta)$	Fisher information

Conditional Expectations

Consider a function $f : \Omega_Y \to \mathbb{C}$

$$E_y\left[f(y)|x\right] = \begin{cases} \sum_{y \in \Omega_Y} f(y) p_{Y|X}(y|x) & Y : \text{discrete} \\ \int_{\Omega_Y} f(y) p_{Y|X}(y|x) dy & Y : \text{continuous} \end{cases}$$

$$E\left[Y|x\right] = \sum y p_{Y|X}(y|x)$$

$$E\left[Y|x\right] = \int y p_{Y|X}(y|x) dy$$

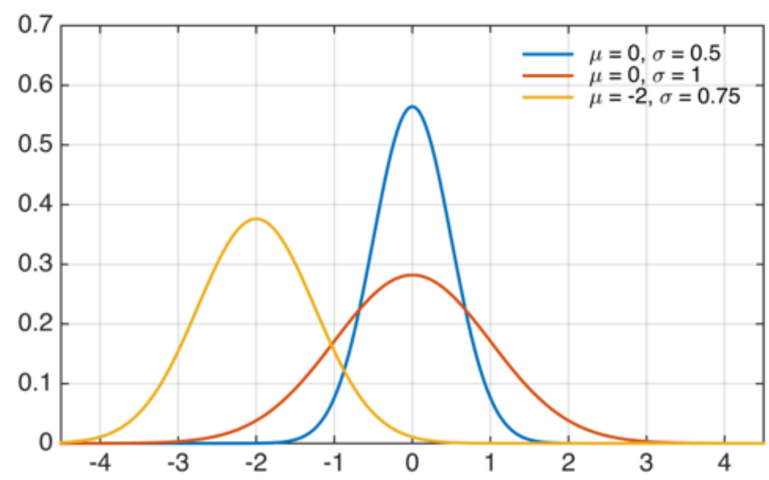
$\Leftarrow$ Regression function!

EXERCISE: RVS, PDFS AND UNCERTAINTY

- In ML, common strategy to assume trying to learn a deterministic function, from noisy measurements
- Denoised "truth": y = f(x)
- Noisy observation: f(x) + noise
  - one common assumption is the noise N is a Gaussian RV
  - E[f(x) + noise] = f(x) + E[noise] = f(x)
- For a sample x of RV X:

$$N \sim \mathcal{N}(0, \sigma^2)$$

$$Y = f(x) + N \sim \mathcal{N}(f(x), \sigma^2)$$



Expectations for Two Variables

Consider a function $f : \mathbb{R}^2 \to \mathbb{C}$

$$E_{x,y}\left[f(x,y)\right] = \begin{cases} \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f(x,y) p_{XY}(x,y) & X, Y : \text{discrete} \\ \int_{\Omega_X} \int_{\Omega_Y} f(x,y) p_{XY}(x,y) dx dy & X, Y : \text{continuous} \end{cases}$$

Expectations You Know About

$f(x,y)$	Symbol	Name
$(x - E[X])(y - E[Y])$	$\text{cov}(X,Y)$	Covariance
$\frac{(x-E[X])(y-E[Y])}{\sqrt{V[X]V[Y]}}$	$\text{corr}(X,Y)$	Correlation
$\log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$	$I(X;Y)$	Mutual information
$\log \frac{1}{p_{XY}(x,y)}$	$H(X,Y)$	Joint entropy
$\log \frac{1}{p_{X\|Y}(x\|y)}$	$H(X\|Y)$	Conditional entropy

MIXTURES OF DISTRIBUTIONS

Mixture model:

A set of m probability distributions, $\{p_i(x)\}_{i=1}^m$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ and non-negative and

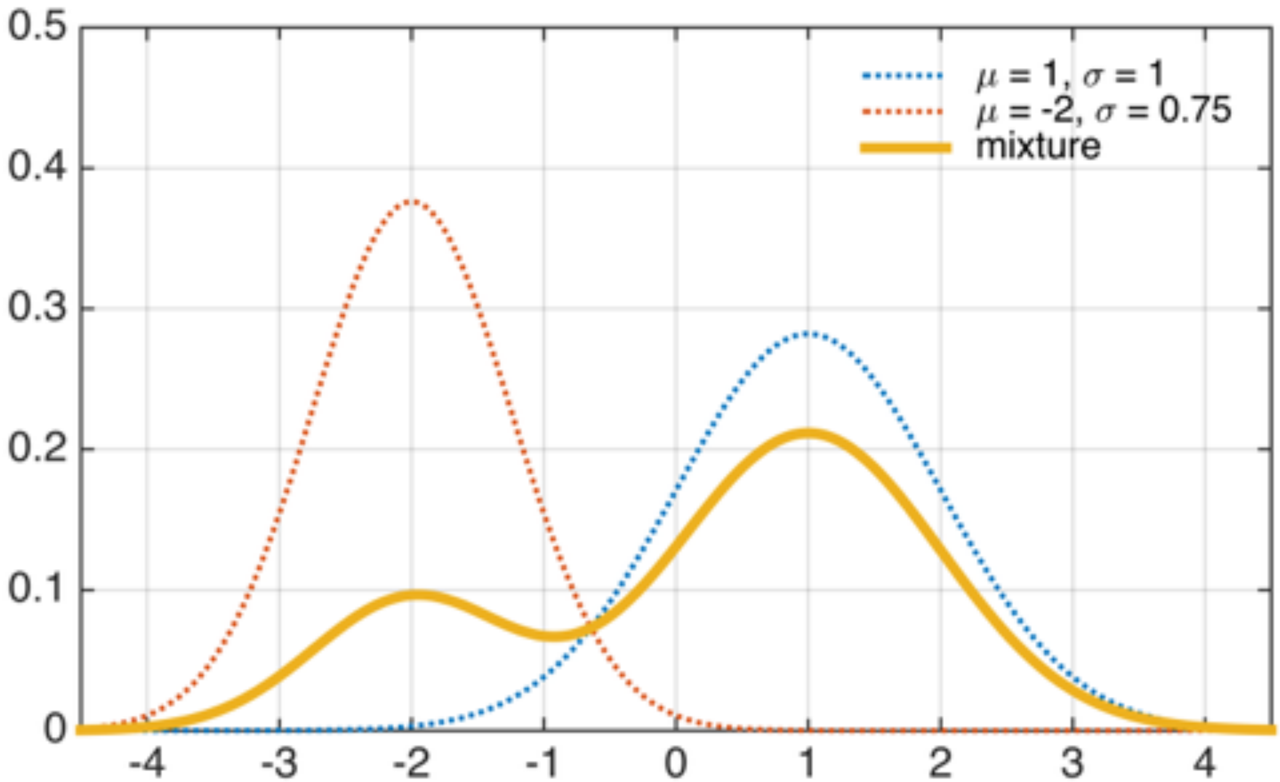
$$\textstyle\sum_{i=1}^{m} w_i = 1$$

Mixtures of Gaussians

Mixture of $m = 2$ Gaussian distributions:

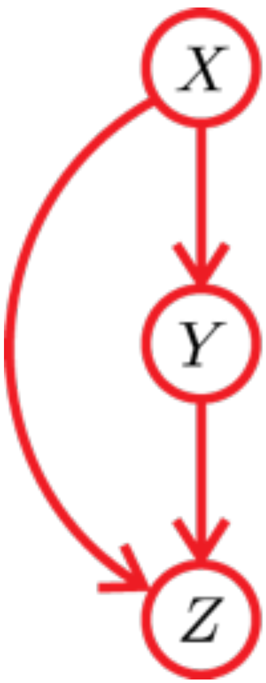
$w_1 = 0.75,\ w_2 = 0.25$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$



Graphical Representations

Bayesian Network: $p(\boldsymbol{x}) = \prod_{i=1}^{k} p\left(x_i | \boldsymbol{x}_{\text{Parents}(X_i)}\right)$



$P(X=1)$
0.3

X	$P(Y=1\|X)$
0	0.5
1	0.9

X	Y	$P(Z=1\|X,Y)$
0	0	0.3
0	1	0.1
1	0	0.7
1	1	0.4

Factorization:

$$p(x,y,z) = p(x)p(y|x)p(z|x,y)$$

Graphical Representations

Bayesian Network: $p(\boldsymbol{x}) = \prod_{i=1}^{k} p\left(x_i | \boldsymbol{x}_{\text{Parents}(X_i)}\right)$

$P(X=1)$
0.3

Y	$P(Z=1\|Y)$
0	0.2
1	0.7

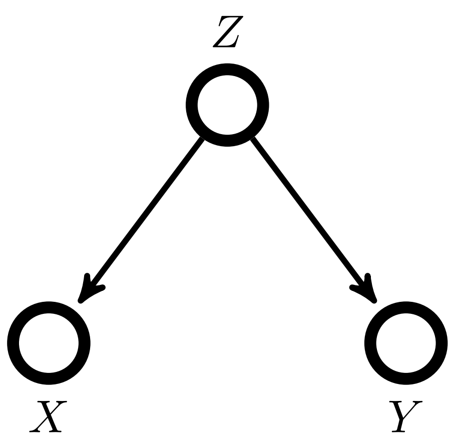
X	$P(Y=1\|X)$
0	0.5
1	0.9

Factorization:

$$p(x,y,z) = p(x)p(y|x)p(z|y)$$

Graphical Representations: conditional independ.

Bayesian Network: $p(\boldsymbol{x}) = \prod_{i=1}^{k} p\left(x_i | \boldsymbol{x}_{\text{Parents}(X_i)}\right)$

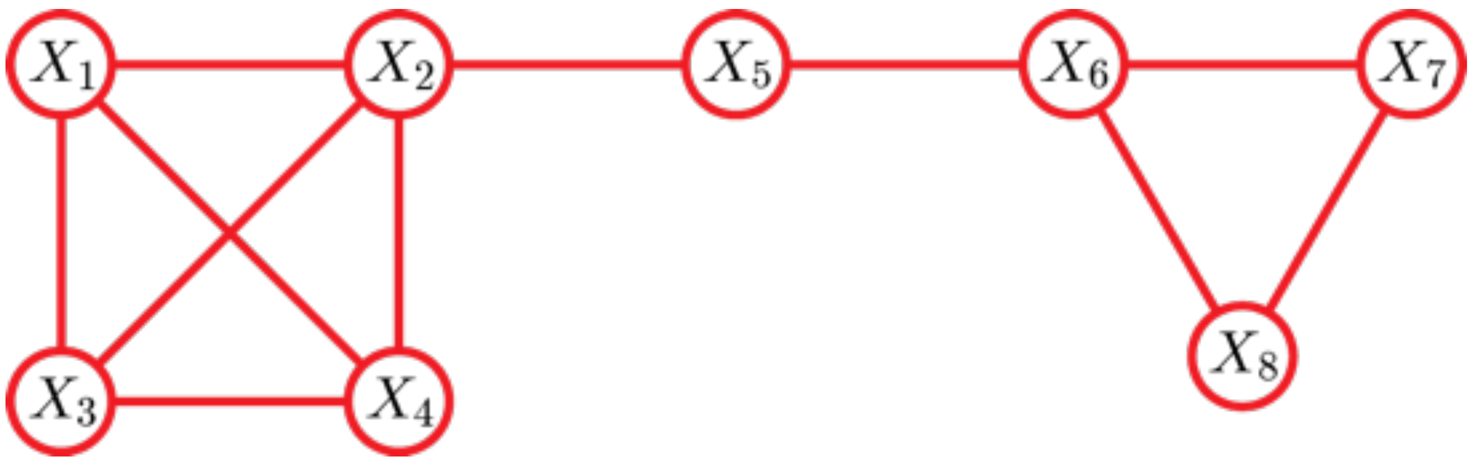


Factorization:

$$p(x, y|z) = p(x|z)p(y|z)$$

Graphical Representations

Markov Network: $p\left(x_i|\boldsymbol{x}_{-i}\right) = p\left(x_i|\boldsymbol{x}_{N(X_i)}\right)$



Factorization:

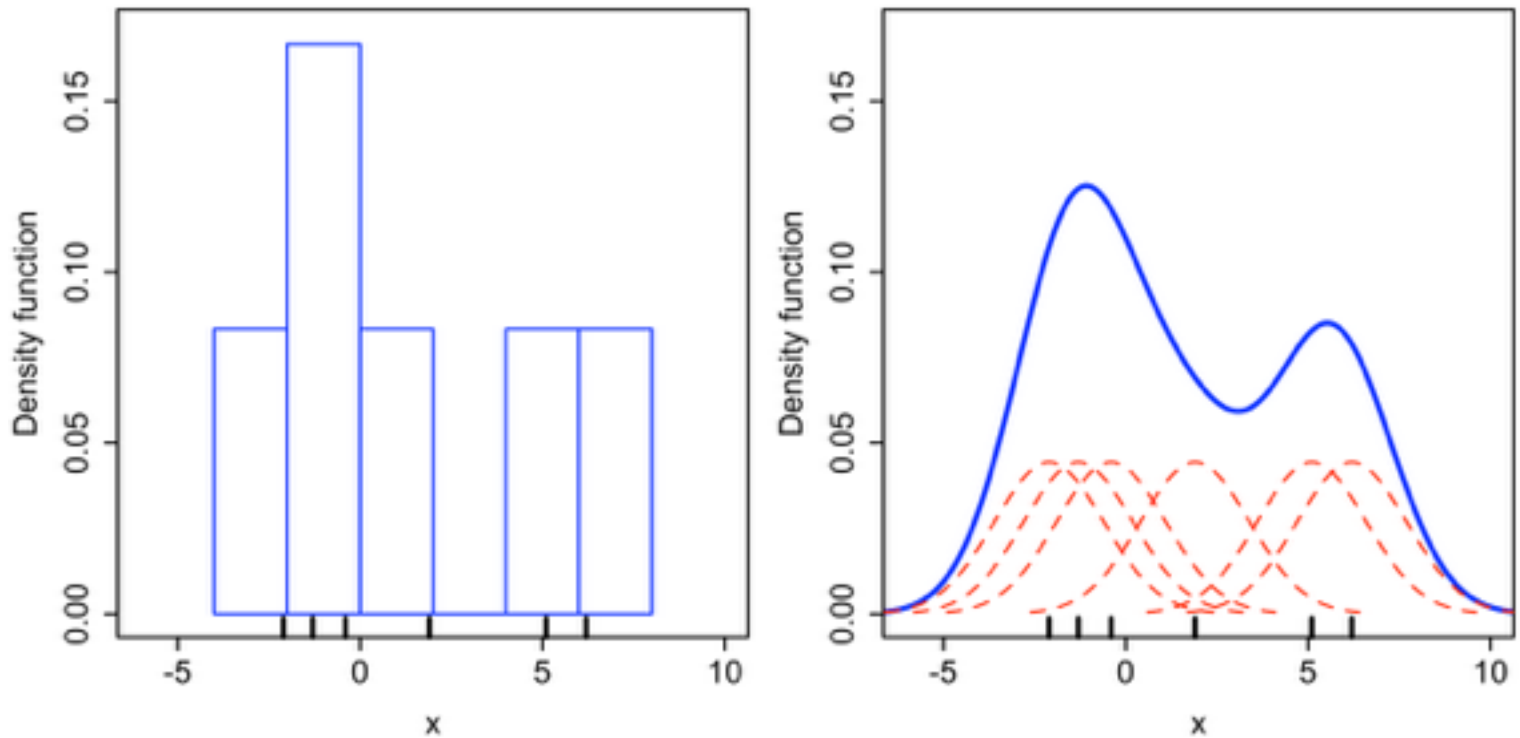
$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C)$$

SUMMARY: PARAMETRIC MODELS

- We will consider many parametric models in machine learning
- To model the data, we pick a parametric class and do parameter estimation (next)
- Given a model, we can make statements about our data
  - predict target given inputs (conditional probs)
  - find underlying structure of data
  - find explanatory variables
  - …

NON-PARAMETRIC MODELS

- Do not assume knowledge of distribution
  - might not even assume pdf exists (e.g., for more see work on kernel embedding of distributions)
- Often accomplished using kernels
  - we'll discuss this more later



Next: Parameter Estimation

- For a given model type, we want to determine the "best" modeling parameters
- Parameter estimation deals with finding model parameters, informed by the observed data
- These model parameters can be themselves parametrized in different ways
  - for supervised learning
  - for unsupervised learning
  - with augmented representations
  - …