CS446 Introduction to Machine Learning (Fall 2013)
University of Illinois at Urbana-Champaign
http://courses.engr.illinois.edu/cs446

LECTURE 19: PROBABILISTIC MODELS (I)

Prof. Julia Hockenmaier
juliahmr@illinois.edu

Probabilistic models

Three different applications (in CS446):

- Classification

  Generative models (Naïve Bayes)
  Discriminative models (Logistic Regression, aka Maximum Entropy/loglinear models)

- Clustering

  k-means clustering with the EM algorithm

- Sequence labeling (e.g. POS-tagging)

  Generative: Hidden Markov models, supervised/unsupervised

  (CS546) Discriminative models: Conditional Random Fields

(Anticipated) Schedule

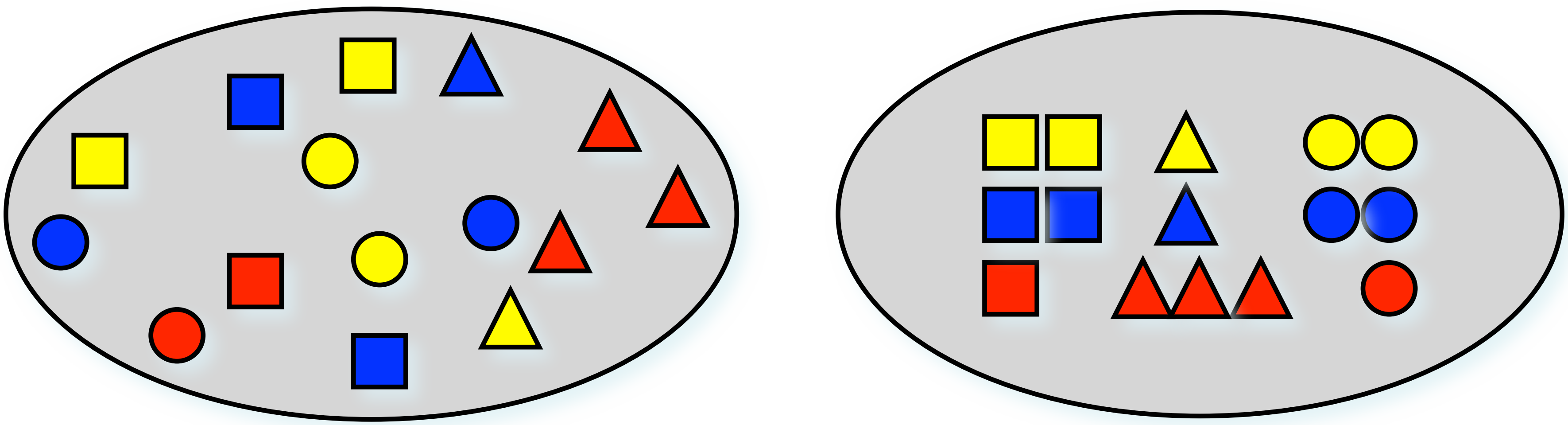
Today:
- Brief probability review
- The Naïve Bayes classifier

Thursday:
- Logistic Regression

Next week and afterwards:
- Graphical models
- k-means clustering
- HMMs

Probability review I: Events

What is the probability of…?



P(■) = 2/15
P(blue) = 5/15
P(blue | □) = 2/5

P(■) = 1/15
P(red) = 5/15
P(□) = 5/15

P(■ or ▲) = 2/15
P(△ | red) = 3/5

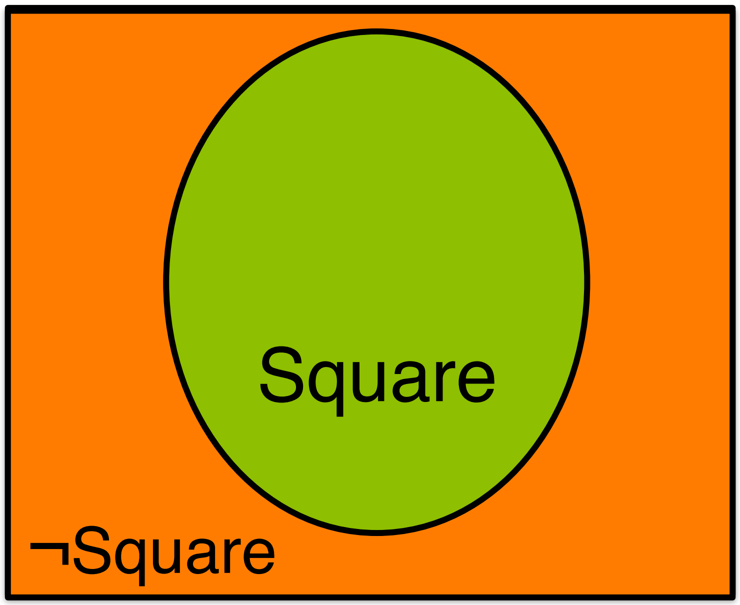
Some terminology…

Trial: e.g. picking a shape

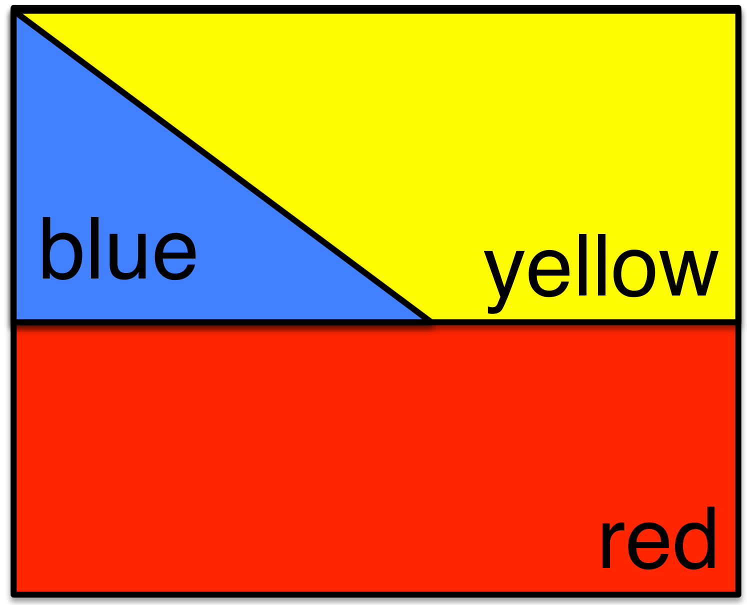
Sample space Ω: the set of all possible outcomes (e.g. all kinds of shapes)

Event $\omega \subseteq \Omega$: an actual outcome of a trial (a subset of Ω)

Atomic events

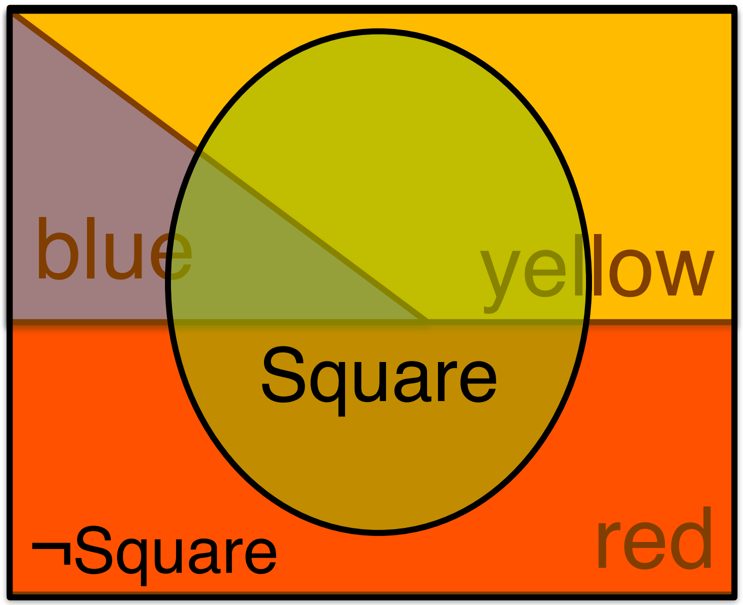


Boolean random variable *Square*



Categorical random variable *Color*

Complex events



What is the probability of…

…. a circle when drawing a red shape?

(# of red circles) / (# of red shapes)

Conditional probability P(A | B):

Probability of one event (*circle*) given another (*red)*

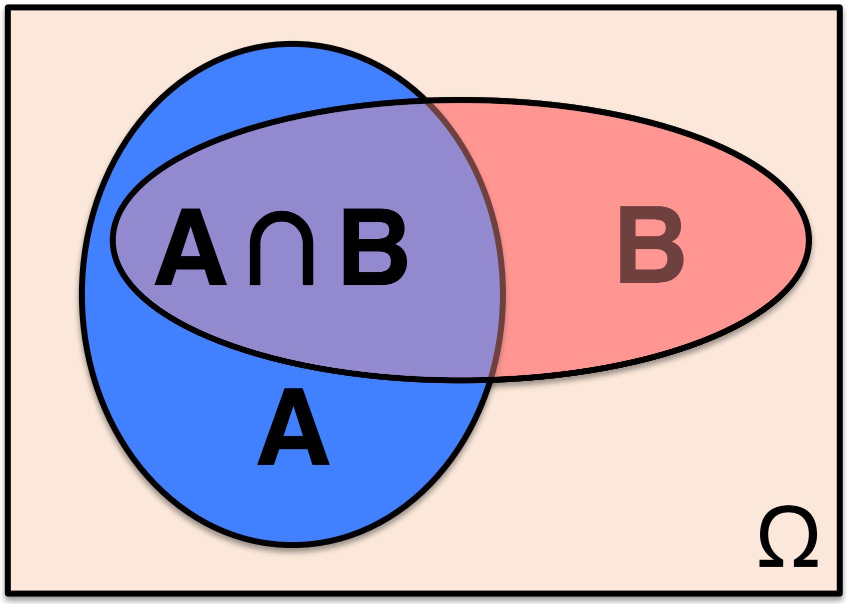
…. drawing a red circle?

(# of red circles) / (# of all shapes in bag)

Joint probability P(A, B):

Probability of two events (*red* and *circle*) occurring together

Laws of probability



$$P(\Omega) = 1$$

$$\forall A \subseteq \Omega: \quad 0 \leq P(A) \leq 1$$

$$\forall A,B \subseteq \Omega: P(A \cap B) \leq P(A)$$

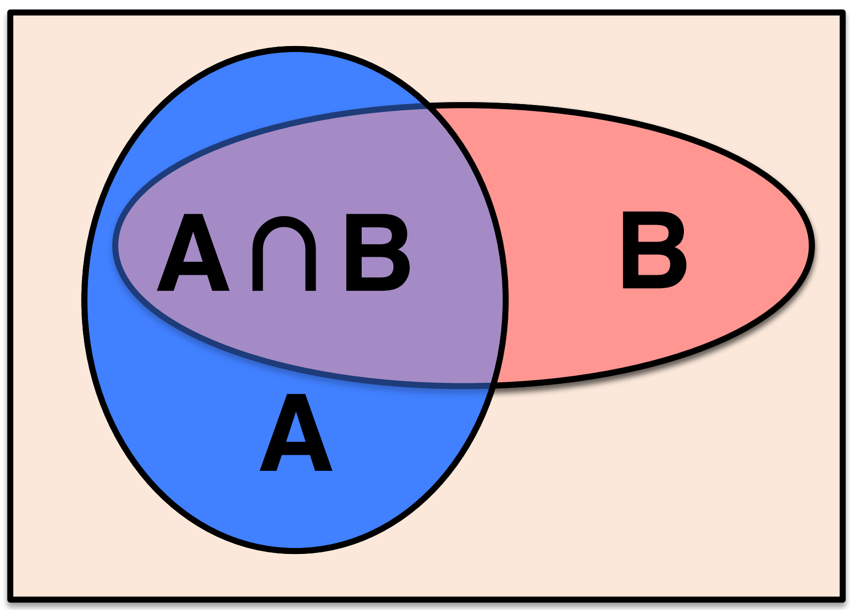
$$\forall A,B \subseteq \Omega: P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Joint probability P(A, B)

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

If A and B are Boolean:

$P(A,B) = P(A \wedge B)$



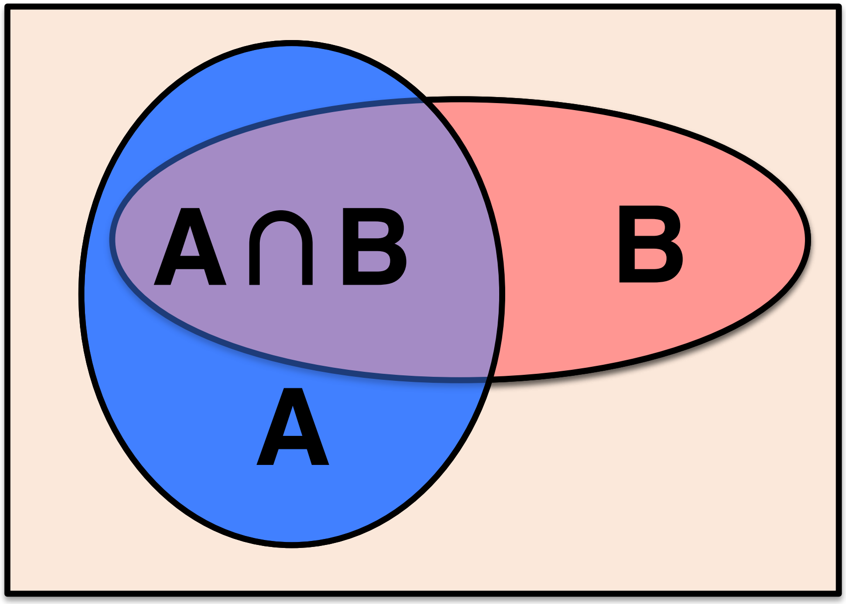
Conditional probability P(A|B)

Bayes rule:

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

Product rule

$$P(A,B) = P(A \mid B)P(B)$$

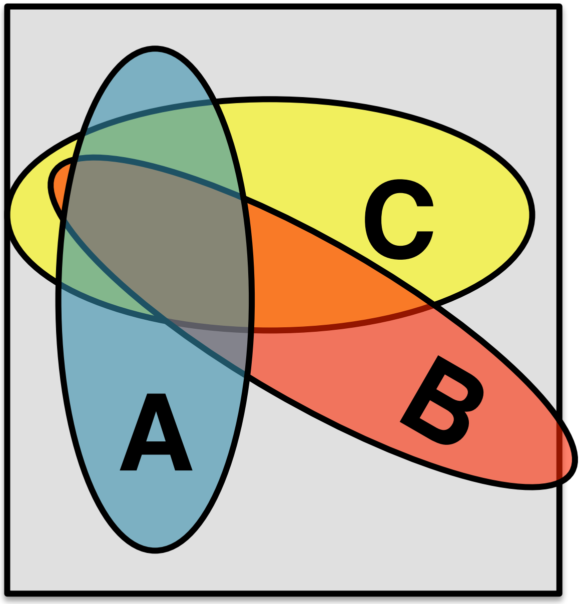


The chain rule

Extends the product rule to multiple variables:

$$
\begin{aligned}
P(X_1, \ldots, X_n) = \; & P(X_1) \\
& \times P(X_2 \mid X_1) \\
& \times P(X_3 \mid X_{1..2}) \\
& \times \ldots \\
& \times P(X_i \mid X_{1..i-1}) \\
& \times \ldots \\
& \times P(X_n \mid X_{1..n-1})
\end{aligned}
$$

Conditional probability



Probability of A given B:

$P(A \mid B)$

Probability of A and B given C:

$P(A,B \mid C)$

Probability of A given B and C:

$P(A \mid B, C)$

Conditional probabilities of events

If A and B are events (A: 'red'), (B:'triangle'), $P(A \mid B)$ indicates the probability of event A given event B:

$$P(\text{red} \mid \text{triangle}) = 0.5$$

Probability review II: Random variables and distributions

Random variables

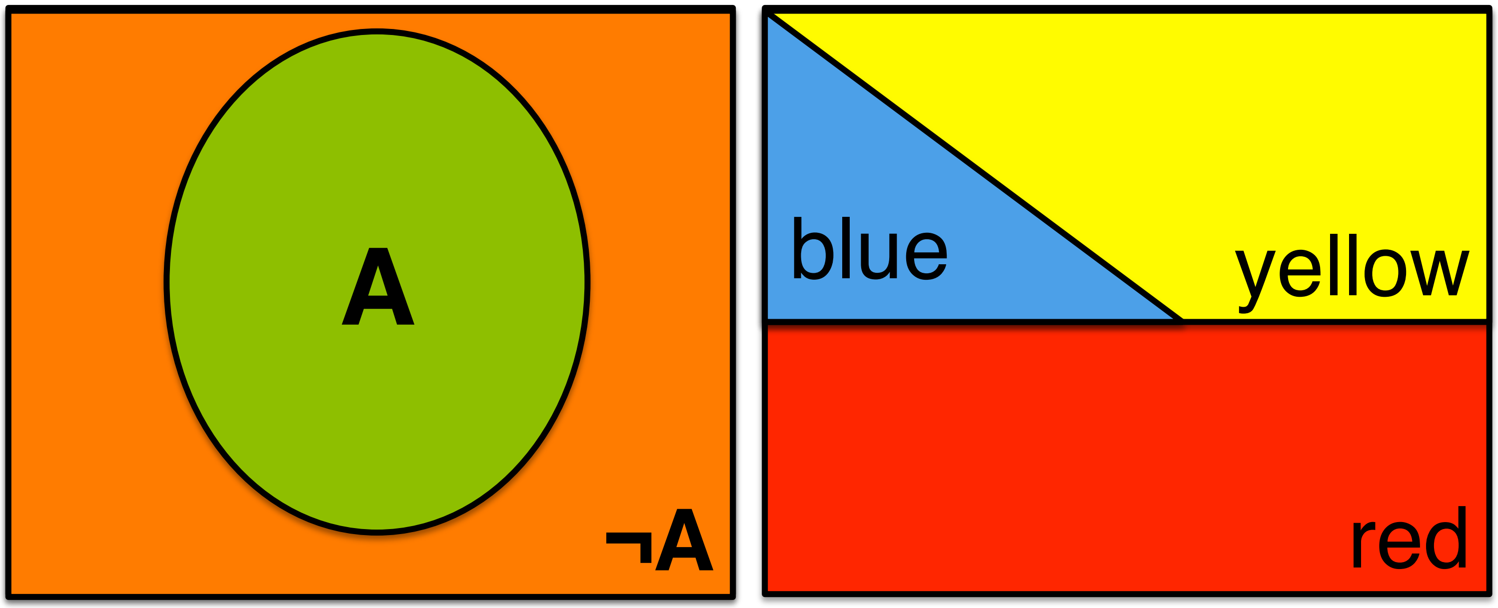
A function which maps every element in the sample space to some value.

Boolean random variables: heads or tails?

Categorical random variables: color, shape

Continuous random variables: size, height,...

Discrete random variables



The possible outcomes of discrete random variables (=atomic events) partition the sample space

Distributions of R.V.s

Each R.V. X is associated with a distribution P(X)

- Discrete distributions $P(X = x)$
  are defined by a *probability mass function f*:
  $f(X = x) = P(X = x)$ with $\sum_x P(X = x) = 1$
- Continuous distribution $P(X \in A)$
  are defined by a *probability density function f*:
  $P(X \in A) = \int_A f(X{=}x)dx$ with $\int_\Omega f(X{=}x)dx = 1$

Important note re. notation:

We often abbreviate $P(X = x)$ or $P(X \in A)$ as $P(X)$

Examples of distributions

You should have seen at least some of the following *parametric* families of distributions:

Discrete: Bernoulli, Binomial, Categorical, Multinomial, Poisson, Geometric,...

Continuous: Gaussian (Normal), Beta, Dirichlet, χ^2, ...

Parametric = each individual distribution is specified by a particular set of parameters, e.g.:

N(0, 0.5): Gaussian with 0 mean and 0.5 variance

Coin tossing

Bernoulli distribution:

Probability of success (*head*) in single yes/no trial

The probability of *head* is p.

The probability of *tail* is $1-p$.

Binomial distribution:

Prob. of k heads in n independent yes/no trials

$$P(k \text{ heads, } n-k \text{ tails}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Rolling a die

Categorical distribution:

Prob. of getting one of K outcomes in a single trial

The probability of outcome c_i is p_i $(\sum p_i = 1)$

Multinomial distribution:

Prob. of observing each possible outcome c_i exactly x_i times in a sequence of n yes/no trials

$$P(X_1 = x_i, \ldots, X_N = x_N) = \frac{n!}{x_1! \cdots x_N!} p_1^{x_1} \cdots p_N^{x_N} \quad \text{if} \sum_{i=1}^{N} x_i = n$$

The parameters of a distribution

How many numbers do we need to specify a distribution?

Bernoulli distribution:

1 parameter (two, but one is implied)

Categorical distribution:

N-1 parameters (N, but one is implied)

Probability review III: Joint and conditional distributions

Conditional distributions P(A|B)

If A and B are random variables
(A:'Color', B:'Shape'),
P(A | B) is a set of distributions

If B is discrete, there is one distribution
for each distinct value of B:

P(Color | Shape) =

P(Color = red | Shape = 'triangle') = 3/5
P(Color = blue | Shape = 'triangle') = 1/5
P(Color = yellow | Shape = 'triangle') = 1/5,
P(Color = red | Shape = 'square') = 1/5,….

Parameters of conditional distributions

P(A |B) corresponds to K_B distributions; one for each possible value *b* of *B*.

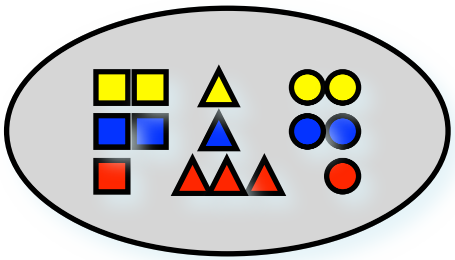
P(A |B=b) has K_A parameters, one for each possible value of *A*
(minus one implied parameter)

P(A |B) has $K_A \times K_B$ (or $(K_A - 1) \times K_B$) parameters

Joint distribution $P(X_1,\ldots,X_i,\ldots,X_n)$

The joint distribution of n discrete R.V.s $X_1 \ldots X_i \ldots X_n$ with K_i possible outcomes each is a distribution with $K_1 \times \ldots \times K_i \times \ldots \times K_n$ parameters:

P(Color = red, Shape = triangle)	= 3/15
P(Color = red, Shape = square)	= 1/15
P(Color = red, Shape = circle)	= 1/15
P(Color = yellow, Shape = triangle)	= 1/15
P(Color = yellow, Shape = square)	= 2/15
P(Color = yellow, Shape = circle)	= 2/15
P(Color = blue, Shape = triangle)	= 1/15
P(Color = blue, Shape = square)	= 2/15
P(Color = blue, Shape = circle)	= 2/15



Probability review IV: (Conditional) Independence

Independence $(X \perp Y)$

Two random variables X and Y are independent (written as $X \perp Y$) if

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

If X and Y are independent: $P(X|Y) = P(X)$:

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

X,Y are independent

Independence ($X \perp Y$)

Showing that X and Y are independent
= Showing that $P(X,Y) = P(X) \times P(Y)$

You have to show that for all possible outcomes x of X and y of Y,

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

N.B.: true independence is rare

Conditional Independence ($X \perp Y \mid Z$)

Two random variables X and Y are conditionally independent given a third random variable Z
(written as $X \perp Y \mid Z$)

if $P(X,Y \mid Z) = P(X \mid Z) \times P(Y \mid Z)$

Independence assumptions

In probabilistic modeling, we often **assume** random variables X, Y, Z are independent

Therefore we can **factor** the joint distribution:
$P(X,Y,Z) = P(X) \times P(Y) \times P(Z)$

How many parameters do we need to know to specify P(X,Y, Z) if we assume independence?

Only $K_X + K_Y + K_Z$

Graphical models

Graphical models

Graphical models are a **notation for probability models.**

Each random variable X
is represented as a node:

$$P(X) = X$$

Arrows represent dependencies:

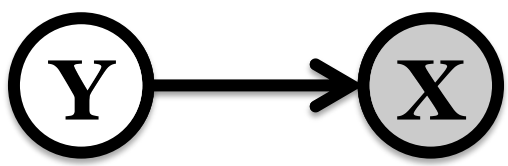
$$P(Y)\,P(X \mid Y) = Y \longrightarrow X$$

Graphical models

Shaded nodes represent observed variables

White nodes represent hidden variables

$P(Y)\, P(X \mid Y)$ with Y hidden and X observed:

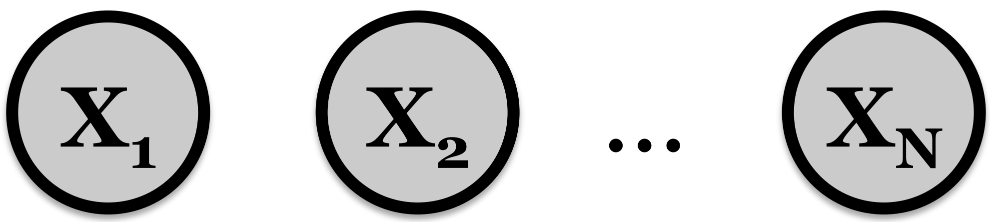


Probabilistic models for classification

Probabilistic classification

Each item is defined by a set of features.

– Each **feature X_i** is **one** (typically **observed) random variable:**



– The **class label Y** is a **hidden random variable**

Probabilistic classification

Task: Return the *most likely* class y^* for the item $\mathbf{x} = (x_1 \ldots x_n)$:

$$y^* = \text{argmax}_y \, P(\, y \mid \mathbf{x})$$
$$= \text{argmax}_y \, P(\, y \mid x_1 \ldots x_n)$$

Probabilistic classification

Discriminative (Conditional) model:

Model $P(Y \mid X_1 \ldots X_n)$ directly

Generative (Joint) model:

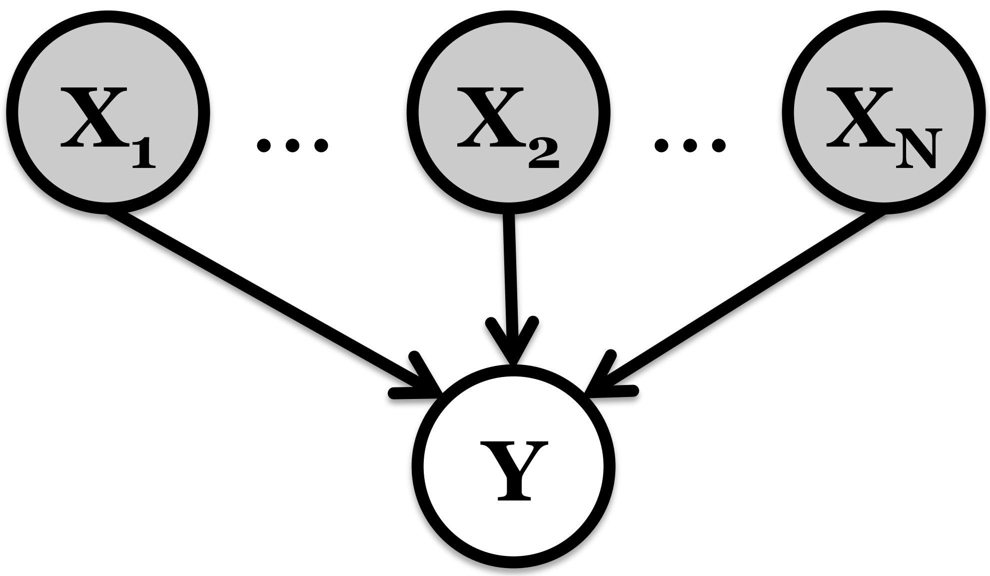
Model $P(Y, X_1 \ldots X_n)$ and use Bayes Rule:

$$
\begin{aligned}
y^* &= \operatorname{argmax}_y P(y \mid x_1 \ldots x_n) \\
&= \operatorname{argmax}_y P(y, x_1 \ldots x_n)/P(x_1 \ldots x_n) \\
&= \operatorname{argmax}_y P(y, x_1 \ldots x_n) \text{ (since } x_1 \ldots x_n \text{ is given)}
\end{aligned}
$$

Probabilistic classification

Discriminative (Conditional) model:

Model $P(Y \mid X_1 \ldots X_n)$ directly



Probabilistic classification

Generative (Joint) model:

Model $P(Y, \mathbf{X})$ and use Bayes Rule:

$$
\begin{aligned}
y^* &= \text{argmax}_y\, P(Y \mid \mathbf{X}) \\
&= \text{argmax}_y\, P(y, \mathbf{X})/P(\mathbf{X}) \\
&= \text{argmax}_y\, P(y, \mathbf{X}) \text{ (since } \mathbf{X}\mathbf{=}\mathbf{x} \text{ is given)}
\end{aligned}
$$

Naïve Bayes classifier

Naïve Bayes classifier

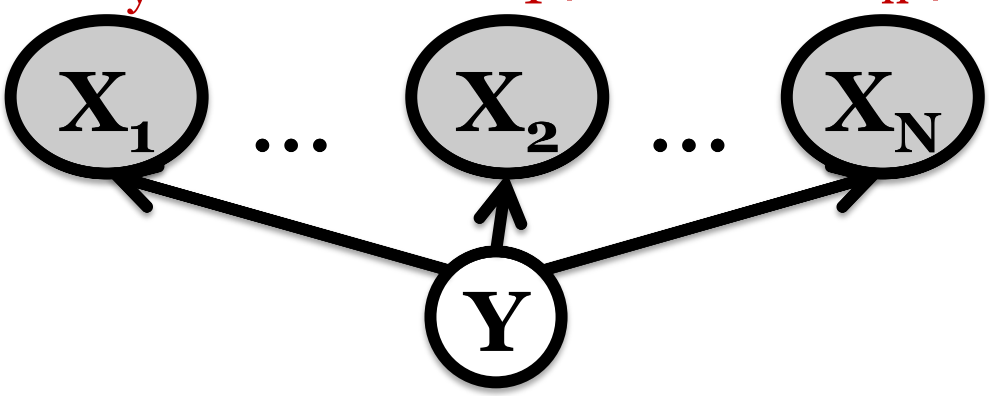
Generative (Joint) model: Assume features are conditionally independent given the class label

$y^* = \text{argmax}_y\ P(Y, X_1 \ldots X_n)$

$\quad = \text{argmax}_y\ P(Y)\ P(X_1 \ldots X_n \mid Y)$

(Independence assumption)

$\quad = \text{argmax}_y\ P(Y)\ P(X_1 \mid Y) \ldots P(X_n \mid Y)$



Naïve Bayes

$\text{argmax}_Y\ P(Y|\ X_1 \ldots X_n) =$

$= \text{argmax}_Y\ P(X_1 \ldots X_n \mid Y)\ P(Y)$

$= \text{argmax}_Y \prod_j P(X_j \mid Y)\ P(Y)$

We need to estimate:

– the categorical distribution P(Y)

– for each attribute X_j and class y, $P(X_j \mid y)$

Supervised learning of a Naïve Bayes classifier

If we have a set of N labeled training items:

- the multinomial $P(Y=y) = freq(y)/N$
- for each attribute X_j and class y:
  $P(X_j = x| y) = freq(X_j = x, y)/freq(y)$

  $freq(y)$ = the number of items with class c

  $freq(x, y)$ = the number of items with attribute value $X_j = x$ and class y.

Learning = parameter estimation

- Probabilistic models (e.g. Naïve Bayes classifiers) are defined by probability distributions
- In a probabilistic model, learning means usually estimating the parameters of the model's distributions
- N.B.: Learning the structure of a probabilistic model (i.e. which variables depend on each other) is a much harder task

Using a Naïve Bayes classifier

Given an item $\mathbf{x} = (x_1 \ldots x_n)$,
return y^* with

$$y^* = \text{argmax}_y \, P(Y = y) \prod_i P(X_i = x_i \mid Y = y)$$