## Introduction to Probability and Statistics



Instructor: Chris Callison-Burch TAs: Mitchell Stern, Justin Chiu

Website: mt-class.org/penn



#### Last time ...

Formulate a *model* of pairs of sentences.
 *Learn* an instance of the model from *data*.
 Use it to *infer* translations of new inputs.

# Why Probability?

- Probability formalizes ...
  - the concept of models
  - the concept of **data**
  - the concept of *learning*
  - the concept of *inference* (prediction)



Probability is expectation founded upon partial knowledge.

# $p(x \mid \text{partial knowledge})$

"Partial knowledge" is an apt description of what we know about language and translation!

## Probability Models

- Key components of a probability model
  - The space of events ( $\Omega$  or S)
  - The assumptions about conditional independence / dependence among events
  - Functions assigning probability (density) to events
  - We will assume discrete distributions.

#### **Events and Random Variables**

A random variable is a function from a random event from a set of possible outcomes ( $\Omega$ ) and a probability distribution (p), a function from outcomes to probabilities.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$X(\omega) = \omega$$

$$\rho_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6\\ 0 & \text{otherwise} \end{cases}$$

#### **Events and Random Variables**

A random variable is a function from a random event from a set of possible outcomes ( $\Omega$ ) and a probability distribution (p), a function from outcomes to probabilities.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$Y(\omega) = \begin{cases} 0 & \text{if } \omega \in \{2, 4, 6\} \\ 1 & \text{otherwise} \end{cases}$$

$$\rho_Y(y) = \begin{cases} \frac{1}{2} & \text{if } y = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

#### What is our event space?

# What are our random variables?

## Probability Distributions

A probability distribution  $(p_X)$  assigns probabilities to the values of a random variable (X).

There are a couple of philosophically different ways to define probabilities, but we will give only the invariants in terms of **random variables**.

$$\sum_{x \in \mathcal{X}} \rho_X(x) = 1$$
$$\rho_X(x) \ge 0 \quad \forall x \in \mathcal{X}$$

Probability distributions of a random variable may be specified in a number of ways.

## Specifying Distributions

- Engineering/mathematical convenience
- Important techniques in this course
  - Probability mass functions
    - Tables ("stupid multinomials")
    - Log-linear parameterizations (maximum entropy, random field, multinomial logistic regression)
  - Construct random variables from other r.v.'s with known distributions

## Sampling Notation



## Sampling Notation

# $x = 4 \times z + 1.7$ $y \sim \text{Distribution}(\theta)$ $\int Distribution$ Parameter

## Sampling Notation

# $y \sim \text{Distribution}(\boldsymbol{\theta})$ $y' = y \times x$

#### Multivariate r.v.'s

Probability theory is particularly useful because it lets us reason about (cor)related and dependent events.

A joint probability distribution is a probability distribution over r.v.'s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

 $\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = 1 \qquad \rho_Z \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) \ge 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$ 

$$\Omega = \{1,2,3,4,5,6\}$$
 $X(\omega) = \omega$ 

$$\begin{split} \Omega &= \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ &(2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ &(3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ &(4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ &(5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ &(6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \} \end{split}$$



$$X(\omega) = \omega_1 \quad Y(\omega) = \omega_2$$
$$\rho_{X,Y}(x,y) = \begin{cases} \frac{1}{36} & \text{if } (x,y) \in \Omega\\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
 $X(\omega) = \omega$ 

$$\begin{split} \Omega &= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),\} \end{split}$$



$$X(\omega) = \omega_1 \quad Y(\omega) = \omega_2$$
$$\rho_{X,Y}(x,y) = \begin{cases} \frac{x+y}{252} & \text{if } (x,y) \in \Omega\\ 0 & \text{otherwise} \end{cases}$$

Marginal Probability  

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

$$p(X = x) = \sum_{y'=\mathcal{Y}} p(X = x, Y = y')$$

$$p(Y = y) = \sum_{x'=\mathcal{X}} p(X = x', Y = y)$$

$$\begin{split} \Omega &= \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \quad p(X=4) = \sum_{y' \in [1,6]} p(X=4,Y=y')\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),\} \end{split}$$

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{1}{36} & \text{if } (x,y) \in \Omega\\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \} \end{cases}$$

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{x+y}{252} & \text{if } (x,y) \in \Omega\\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \} \end{cases}$$

#### Conditional Probability

The **conditional probability** of one random variable given another is defined as follows:

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{\text{joint probability}}{\text{marginal}}$$

Given that  $p(y) \neq 0$ 

Conditional probability distributions are useful for specifying joint distributions since:

$$p(x \mid y)p(y) = p(x, y) = p(y \mid x)p(x)$$
  
Why might this be useful?

#### Conditional Probability Distributions

A conditional probability distribution is a probability distribution over r.v.'s X and Y with the form  $\rho_{X|Y=y}(x)$ .

$$\sum_{x \in \mathcal{X}} \rho_{X|Y=y}(x) = \mathbf{1} \, \forall y \in \mathcal{Y}$$

#### Chain rule

The chain rule is derived from a repeated application of the definition of conditional probability:

Use as many times as necessary!





#### Independence

Two random variables are independent iff p(X=x,Y=y)=p(X=x)p(Y=y)

Equivalently, (use def. of cond. prob to prove)  $p(X = x \mid Y = y) = p(X = x)$ 

Equivalently again:

$$p(Y = y \mid X = x) = p(Y = y)$$

"Knowing about X doesn't tell me about Y"

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{1}{36} & \text{if } (x,y) \in \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \} \end{cases}$$

 $\Omega$ 

$$\rho_{X,Y}(x,y) = \begin{cases} \frac{x+y}{252} & \text{if } (x,y) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), \} \end{cases}$$

#### Independence

Independence has practical benefits. Think about how many parameters you need for a naive parameterization of  $\rho_{X,Y}(x, y)$  vs  $\rho_X(x)$  and  $\rho_Y(y)$ 

O(xy) vs O(x+y)

#### Conditional Independence

Two equivalent statements of conditional independence:

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

and:

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

"If I know B, then C doesn't tell me about A"

#### **Conditional Independence**

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

"If I know B, then C doesn't tell me about A"

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

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

Do we need more parameters or fewer parameters in conditional independence?

#### Independence

- Some variables are independent In Nature
  - How do we know?
- Some variables we pretend are independent for computational convenience
  - Examples?
- Assuming independence is equivalent to letting our model "forget" something that happened in its past
  - What should we forget in language?

## A Word About Data

- When we formulate our models there will be two kinds of random variables: observed and latent
  - Observed: words, sentences(?), parallel corpora, web pages, formatting...
  - Latent: parameters, syntax, "meaning", word alignments, translation dictionaries...

![](_page_29_Figure_0.jpeg)

#### Observed

Garcia and associates .

Garcia y asociados . Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados . his associates are not strong .

sus asociados no son fuertes . Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry.

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos . the company has three groups .

> la empresa tiene tres grupos . its groups are in Europe .

sus grupos estan en Europa . the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine.

los grupos no venden zanzanina .

the small groups are not modern.

los grupos pequenos no son modernos.

#### Hidden

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong.

sus asociados no son fuertes .

Garcia has a company also.

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry.

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos.

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe.

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes.

the groups do not sell zanzanine.

los grupos no venden zanzanina.

the small groups are not modern.

los grupos pequenos no son modernos.

## Learning

- Let's say we have formulated a model of a phenomenon
  - Made independence assumptions
  - Figured out what kinds of parameters we want
- Let's say we have collected data we assume to be generated by this model
  - E.g. some parallel data

What do we do now?

#### Parameter Estimation

#### • Inputs

- Given a model with unspecified parameters
- Given some data
- Goal: learn model parameters
- How?
  - Find parameters that make the model make predictions that look like the data do
  - What do we mean "look like the data?"
    - Probability (other options: accuracy, moment matching)

#### Strategies

#### • Maximum likelihood estimation

• What is the *probability* of generating the data?

#### • Accuracy

- Using an auxiliary similarity function, find parameters that maximize the (expected?) accuracy of data
- Bayesian techniques

![](_page_35_Picture_0.jpeg)

![](_page_35_Picture_1.jpeg)

p(heads) = 1 - p(heads)

#### p(heads)?

![](_page_36_Picture_1.jpeg)

#### $p(data) = p(heads)^7 \times p(taip(s))^3 ads)^3$

![](_page_37_Figure_0.jpeg)

p(heads)

![](_page_38_Figure_0.jpeg)

#### Optimization

- For the most part, we will be working with maximum likelihood estimation
- The general recipe is:
  - Come up with an expression of the likelihood of your probability model, as a function of data and the model parameters
  - Set the parameters to maximize the likelihood
  - This optimization is generally difficult
    - You must respect any constraints on the parameters (>0, sum to I, etc)
    - There may not be analytical solutions (log-linear models)

## Probability lets us

Formulate a *model* of pairs of sentences.
 *Learn* an instance of the model from *data*.
 Use it to *infer* translations of new inputs.

## Key Concepts

- Joint probabilities
- Marginal probabilities
- Conditional probabilities
- Chain rule
- Bayes' rule
- Independence
- Latent versus observed variables
- Maximum likelihood estimation

## Supplemental Reading

 If this was unfamiliar to you, then please read Chapter 3 from the textbook "Statistical Machine Translation" by Philipp Koehn

![](_page_42_Picture_2.jpeg)

#### Announcements

- HW 0 has been posted on the web site.
- It's a setup assignment to make sure that you can upload results, have them scored, and that they correctly appear on the leaderboard