Machine Learning Department
School of Computer Science
Carnegie Mellon University

The Probabilistic Approach to Learning from Data

Prob. Readings:
Lecture notes from 10-600
(See Piazza post for the pointers)

Murphy 2
Bishop 2
HTF --
Mitchell --

Matt Gormley
Lecture 4
January 30, 2016

Reminders

- Website schedule updated
- Background Exercises (Homework 1)
  - Released: Wed, Jan. 25
  - Due: Wed, Feb. 1 at 5:30pm
    (The deadline was extended!)
- Homework 2: Naive Bayes
  - Released: Wed, Feb. 1
  - Due: Mon, Feb. 13 at 5:30pm

Outline

- **Generating Data**
  - Natural (stochastic) data
  - Synthetic data
  - Why synthetic data?
  - Examples: Multinomial, Bernoulli, Gaussian
- **Data Likelihood**
  - Independent and Identically Distributed (i.i.d.)
  - Example: Dice Rolls
- **Learning from Data (Frequentist)**
  - Principle of Maximum Likelihood Estimation (MLE)
  - Optimization for MLE
  - Examples: 1D and 2D optimization
  - Example: MLE of Multinomial
  - Aside: Method of Langrange Multipliers
- **Learning from Data (Bayesian)**
  - *maximum a posteriori* (MAP) estimation
  - Optimization for MAP
  - Example: MAP of Bernoulli—Beta

Generating Data

Whiteboard

- Natural (stochastic) data
- Synthetic data
- Why synthetic data?
- Examples: Multinomial, Bernoulli, Gaussian

In-Class Exercise

1. With your neighbor, **write a function** which returns **samples from a Categorical**
   - Assume access to the `rand()` function
   - Function signature should be: `categorical_sample(phi)` where phi is the array of parameters
   - Make your implementation as **efficient** as possible!
2. What is the **expected runtime** of your function?

Data Likelihood

Whiteboard

- Independent and Identically Distributed (i.i.d.)
- Example: Dice Rolls

Learning from Data (Frequentist)

Whiteboard

- Principle of Maximum Likelihood Estimation (MLE)
- Optimization for MLE
- Examples: 1D and 2D optimization
- Example: MLE of Multinomial
- Aside: Method of Langrange Multipliers

Learning from Data (Bayesian)

Whiteboard

- *maximum a posteriori* (MAP) estimation
- Optimization for MAP
- Example: MAP of Bernoulli—Beta

Takeaways

- One view of what ML is trying to accomplish is **function approximation**
- The principle of **maximum likelihood estimation** provides an alternate view of learning

- **Synthetic data** can help **debug** ML algorithms
- Probability distributions can be used to **model** real data that occurs in the world
(don't worry we'll make our distributions more interesting soon!)

The remaining slides are **extra slides** for your reference.

Since they are **background material** they were not (explicitly) covered in class.

Outline of Extra Slides

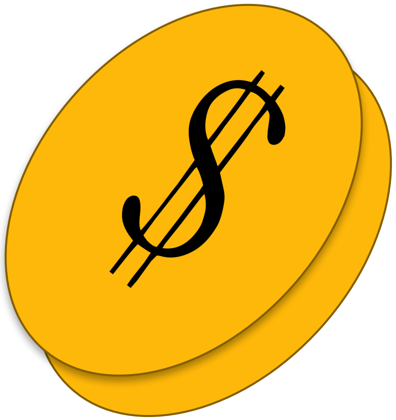
- **Probability Theory**
  - Sample space, Outcomes, Events
  - Kolmogorov's Axioms of Probability
- **Random Variables**
  - Random variables, Probability mass function (pmf), Probability density function (pdf), Cumulative distribution function (cdf)
  - Examples
  - Notation
  - Expectation and Variance
  - Joint, conditional, marginal probabilities
  - Independence
  - Bayes' Rule
- **Common Probability Distributions**
  - Beta, Dirichlet, etc.

PROBABILITY THEORY

Probability Theory: Definitions

Example 1: Flipping a coin

Sample Space	Ω	{Heads, Tails}
Outcome	$\omega \in \Omega$	Example: Heads
Event	$E \subseteq \Omega$	Example: {Heads}
Probability	$P(E)$	$P(\{\text{Heads}\}) = 0.5$ $P(\{\text{Tails}\}) = 0.5$



Probability Theory: Definitions

Probability provides a science for inference about interesting events

Sample Space	Ω	The set of all possible outcomes
Outcome	$\omega \in \Omega$	Possible result of an experiment
Event	$E \subseteq \Omega$	Any subset of the sample space
Probability	$P(E)$	The non-negative number assigned to each event in the sample space

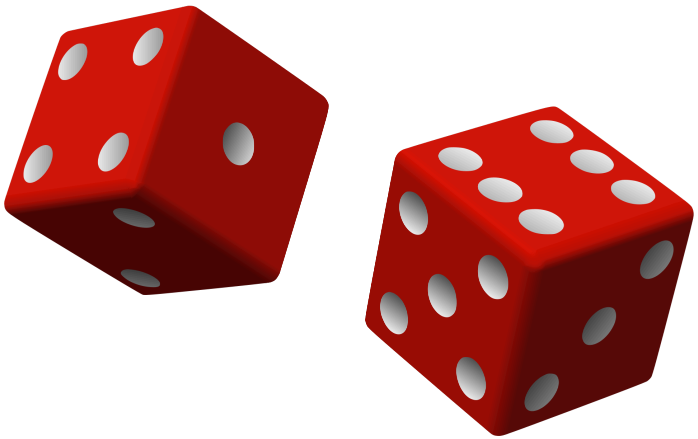
- Each outcome is unique
- Only one outcome can occur per experiment
- An outcome can be in multiple events
- An **elementary event** consists of exactly one outcome

Probability Theory: Definitions

Example

Example 2: Rolling a 6-sided die

Sample Space	Ω	{1,2,3,4,5,6}
Outcome	$\omega \in \Omega$	Example: 3
Event	$E \subseteq \Omega$	Example: {3} (the event "the die came up 3")
Probability	$P(E)$	$P(\{3\}) = 1/6$ $P(\{4\}) = 1/6$

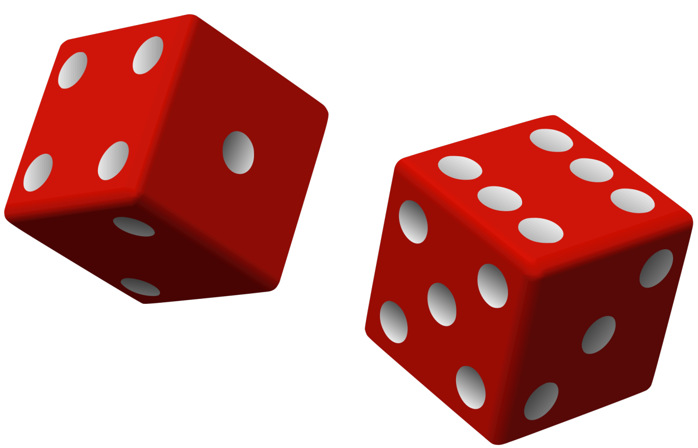


Probability Theory: Definitions

Example

Example 2: Rolling a 6-sided die

Sample Space	Ω	{1,2,3,4,5,6}
Outcome	$\omega \in \Omega$	Example: 3
Event	$E \subseteq \Omega$	Example: {2,4,6} (the event "the roll was even")
Probability	$P(E)$	$P(\{2,4,6\}) = 0.5$ $P(\{1,3,5\}) = 0.5$

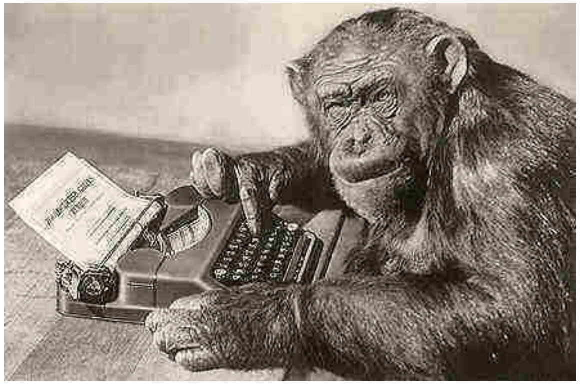


Example

Probability Theory: Definitions

Example 3: Timing how long it takes a monkey to reproduce Shakespeare

Sample Space	Ω	$[0, +\infty)$
Outcome	$\omega \in \Omega$	Example: 1,433,600 hours
Event	$E \subseteq \Omega$	Example: $[1, 6]$ hours
Probability	$P(E)$	$P([1,6]) = 0.000000000001$ $P([1{,}433{,}600, +\infty)) = 0.99$



Kolmogorov's Axioms

1. $P(E) \geq 0$, for all events E
2. $P(\Omega) = 1$
3. If $E_1, E_2, \ldots$ are disjoint, then
   $P(E_1 \text{ or } E_2 \text{ or } \ldots) = P(E_1) + P(E_2) + \ldots$

Kolmogorov's Axioms

1. $P(E) \geq 0$, for all events E
2. $P(\Omega) = 1$
3. If $E_1, E_2, \ldots$ are disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

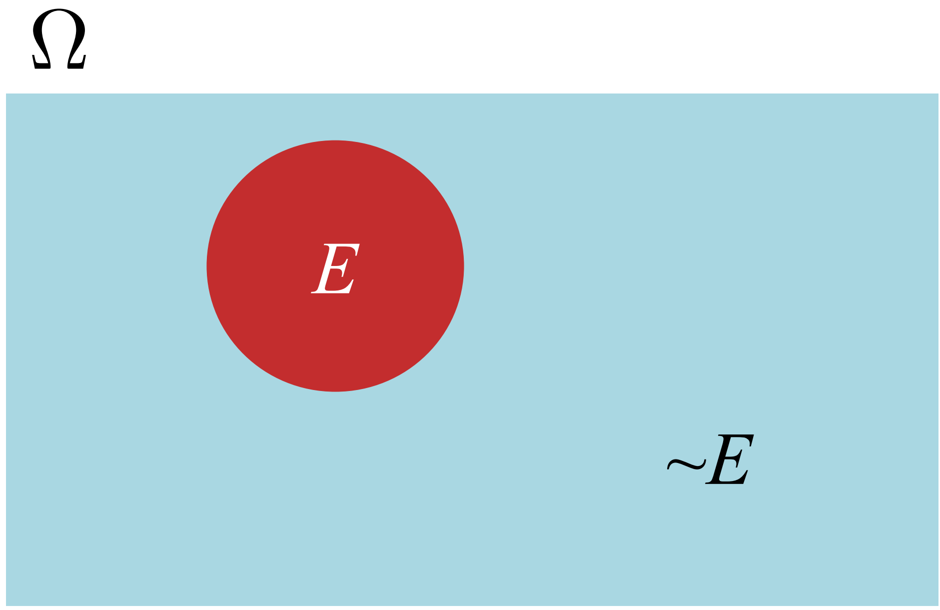
All of probability can be derived from just these!

In words:

1. Each event has non-negative probability.
2. The probability that *some* event will occur is one.
3. The probability of the union of many disjoint sets is the sum of their probabilities

Probability Theory: Definitions

- The **complement** of an event E, denoted $\sim E$, is the event that E does not occur.



RANDOM VARIABLES

Random Variables: Definitions

Random Variable	X (capital letters)	Def 1: Variable whose possible values are the outcomes of a random experiment
Value of a Random Variable	x (lowercase letters)	The value taken by a random variable

Random Variables: Definitions

Random Variable	X	Def 1: Variable whose possible values are the outcomes of a random experiment
Discrete Random Variable	X	Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})
Continuous Random Variable	X	Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range (3, 5))

Random Variables: Definitions

Random Variable	X	Def 1: Variable whose possible values are the outcomes of a random experiment Def 2: A measureable function from the sample space to the real numbers: $$X : \Omega \to E$$
Discrete Random Variable	X	Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})
Continuous Random Variable	X	Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range (3, 5))

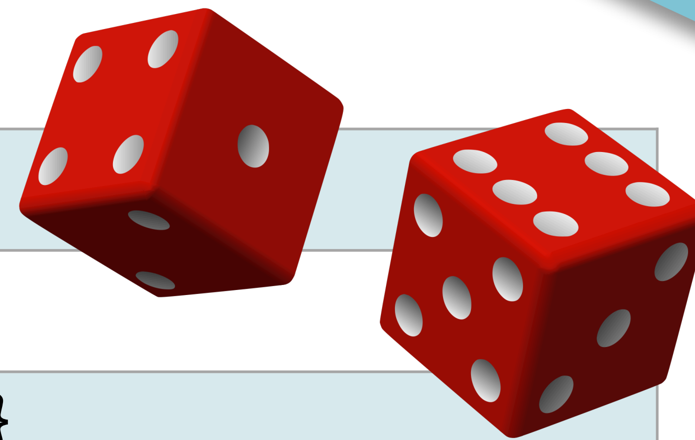
Random Variables: Definitions

Discrete Random Variable	X	Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})
Probability mass function (pmf)	$p(x)$	Function giving the probability that discrete r.v. X takes value x. $$p(x) := P(X = x)$$

Example

Random Variables: Definitions

Example 2: Rolling a 6-sided die

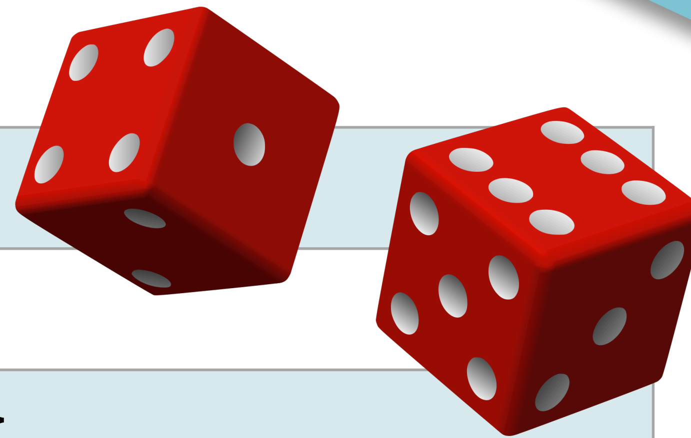


Sample Space	Ω	{1,2,3,4,5,6}
Outcome	$\omega \in \Omega$	Example: 3
Event	$E \subseteq \Omega$	Example: {3} (the event "the die came up 3")
Probability	$P(E)$	$P(\{3\}) = 1/6$ $P(\{4\}) = 1/6$

Example

Random Variables: Definitions

Example 2: Rolling a 6-sided die

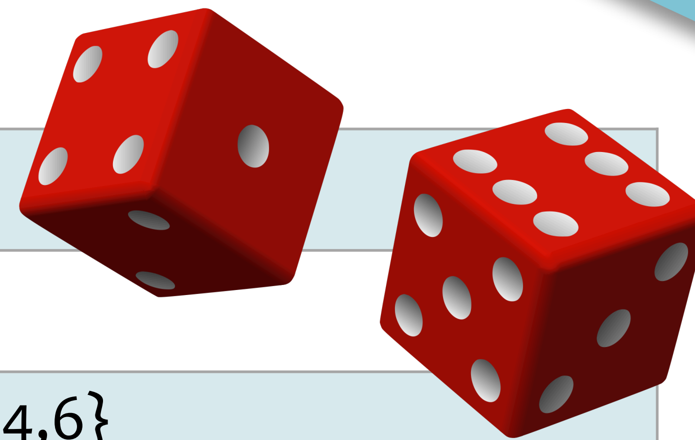


Sample Space	Ω	{1,2,3,4,5,6}
Outcome	$\omega \in \Omega$	Example: 3
Event	$E \subseteq \Omega$	Example: {3} (the event "the die came up 3")
Probability	$P(E)$	$P(\{3\}) = 1/6$ $P(\{4\}) = 1/6$
Discrete Random Variable	X	Example: The value on the top face of the die.
Prob. Mass Function (pmf)	$p(x)$	p(3) = 1/6 p(4) = 1/6

Example

Random Variables: Definitions

Example 2: Rolling a 6-sided die



Sample Space	Ω	{1,2,3,4,5,6}
Outcome	$\omega \in \Omega$	Example: 3
Event	$E \subseteq \Omega$	Example: {2,4,6} (the event "the roll was even")
Probability	$P(E)$	$P(\{2,4,6\}) = 0.5$ $P(\{1,3,5\}) = 0.5$
Discrete Random Variable	X	Example: 1 if the die landed on an even number and 0 otherwise
Prob. Mass Function (pmf)	$p(x)$	p(1) = 0.5 p(0) = 0.5

Random Variables: Definitions

Discrete Random Variable	X	Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})
Probability mass function (pmf)	$p(x)$	Function giving the probability that discrete r.v. X takes value x. $$p(x) := P(X = x)$$

Random Variables: Definitions

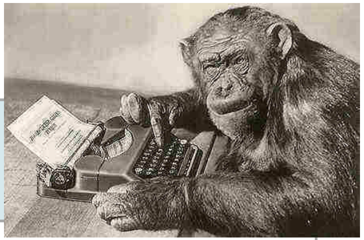
Continuous Random Variable	X	Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range (3, 5))
Probability density function (pdf)	$f(x)$	Function the returns a nonnegative real indicating the relative likelihood that a continuous r.v. X takes value x

- For any continuous random variable: $P(X = x) = 0$
- Non-zero probabilities are only available to intervals:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Random Variables: Definitions

Example 3: Timing how long it takes a monkey to reproduce Shakespeare

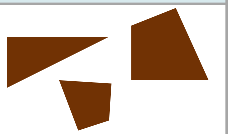


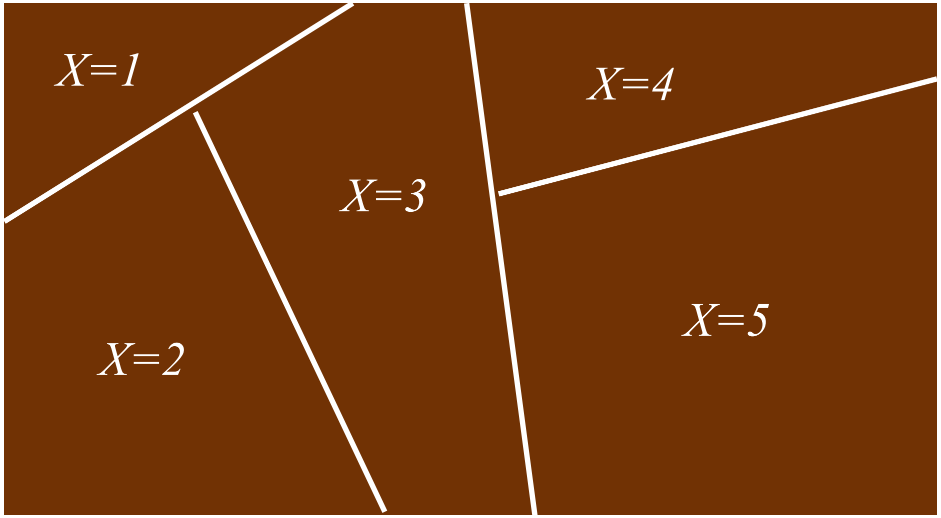
Sample Space	Ω	[0, +∞)
Outcome	$\omega \in \Omega$	Example: 1,433,600 hours
Event	$E \subseteq \Omega$	Example: [1, 6] hours
Probability	$P(E)$	*P*([1,6]) = *0.000000000001* *P*([1,433,600, +∞)) = *0.99*
Continuous Random Var.	X	Example: Represents time to reproduce (not an interval!)
Prob. Density Function	$f(x)$	Example: Gamma distribution

Example

Random Variables: Definitions

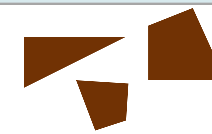
“Region”-valued Random Variables

Sample Space	Ω	{1,2,3,4,5}
Events	x	The sub-regions 1, 2, 3, 4, or 5 
Discrete Random Variable	X	Represents a random selection of a sub-region
Prob. Mass Fn.	$P(X=x)$	Proportional to size of sub-region



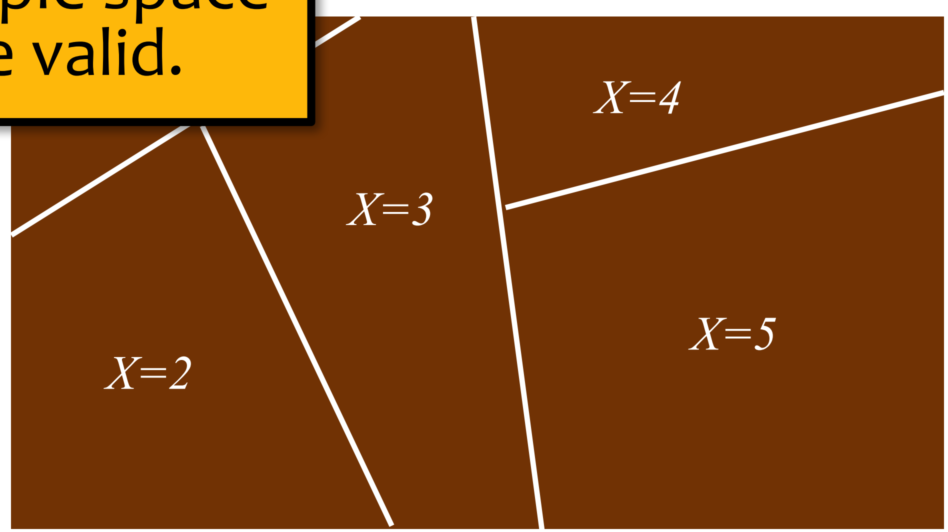
Random Variables: Definitions

"Region"-valued Random Variables

Sample Space	Ω	All points in the region: 
Events	x	The sub-regions 1, 2, 3, 4, or 5 
D… de…		Represents a random selection of a sub-region
Pr…		Proportional to size of sub-region

Recall that an event is any subset of the sample space.

So both definitions of the sample space here are valid.



Example

Random Variables: Definitions

String-valued Random Variables

Sample Space	Ω	All Korean sentences (an infinitely large set)
Event	x	Translation of an English sentence into Korean (i.e. elementary events)
Discrete Random Variable	X	Represents a translation
Probability	$P(X=x)$	Given by a model

English:	machine learning requires probability and statistics
Korean:	$P(X=$ 기계 학습은 확률과 통계를 필요 $)$
	$P(X=$ 머신 러닝은 확률 통계를 필요 $)$
	$P(X=$ 머신 러닝은 확률 통계를 이 필요합니다 $)$
	…

Random Variables: Definitions

Cumulative distribution function	$F(x)$	Function that returns the probability that a random variable X is less than or equal to x: $F(x) = P(X \leq x)$

- For **discrete** random variables:

$$F(x) = P(X \leq x) = \sum_{x' < x} P(X = x') = \sum_{x' < x} p(x')$$

- For **continuous** random variables:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x')dx'$$

Random Variables and Events

Question: Something seems wrong…

- We defined *P(E)* (the capital 'P') as a function mapping *events* to probabilities
- So why do we write $P(X=x)$?
- A good guess: $X=x$ is an event…

Random Variable	Def 2: A measureable function from the sample space to the real numbers: $$X : \Omega \to \mathbb{R}$$

Answer: $P(X=x)$ is just shorthand!

These sets are events!

Example 1:

$$P(X = x) \equiv P(\{\omega \in \Omega : X(\omega) = x\})$$

Example 2:

$$P(X \leq 7) \equiv P(\{\omega \in \Omega : X(\omega) \leq 7\})$$

Notational Shortcuts

A convenient shorthand:

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

$\Rightarrow$ For all values of a and b:

$$P(A=a|B=b) = \frac{P(A=a,B=b)}{P(B=b)}$$

Notational Shortcuts

But then how do we tell *P(E)* apart from *P(X)* ?

Event

Random Variable

Instead of writing:

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

We should write:

$$P_{A|B}(A|B) = \frac{P_{A,B}(A,B)}{P_B(B)}$$

…but only probability theory textbooks go to such lengths.

Expectation and Variance

The **expected value** of X is $E[X]$. Also called the mean.

- Discrete random variables:

  Suppose X can take any value in the set $\mathcal{X}$.

$$E[X] = \sum_{x \in \mathcal{X}} x p(x)$$

- Continuous random variables:

$$E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

Expectation and Variance

The **variance** of *X* is *Var(X)*.

$$Var(X) = E[(X - E[X])^2]$$

$$\mu = E[X]$$

- Discrete random variables:

$$Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x)$$

- Continuous random variables:

$$Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Joint probability

Marginal probability

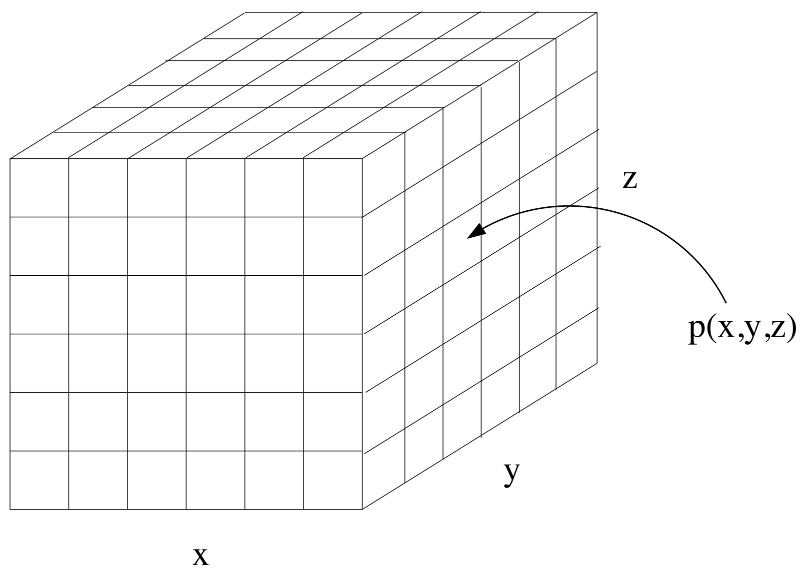
Conditional probability

MULTIPLE RANDOM VARIABLES

Joint Probability

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write

$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$



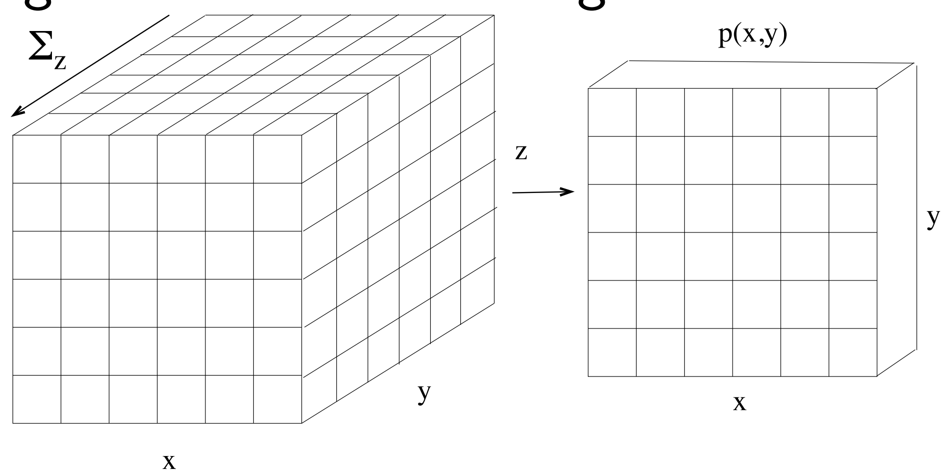
Slide from Sam Roweis (MLSS, 2005)

Marginal Probabilities

- We can "sum out" part of a joint distribution to get the *marginal distribution* of a subset of variables:

$$p(x) = \sum_y p(x, y)$$

- This is like adding slices of the table together.



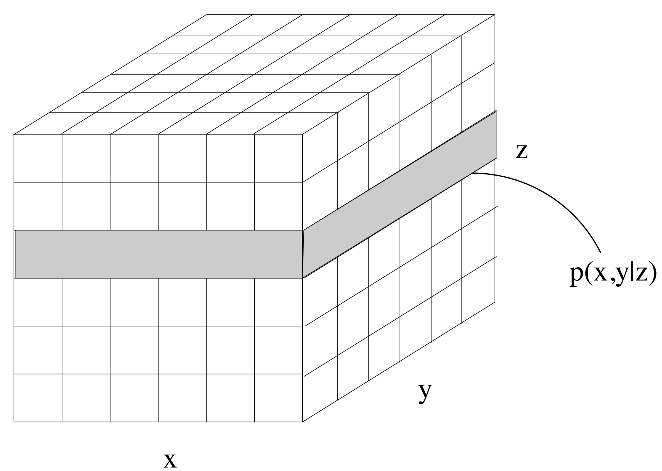
- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

Slide from Sam Roweis (MLSS, 2005)

Conditional Probability

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.

$$\boxed{p(x|y) = p(x,y)/p(y)}$$

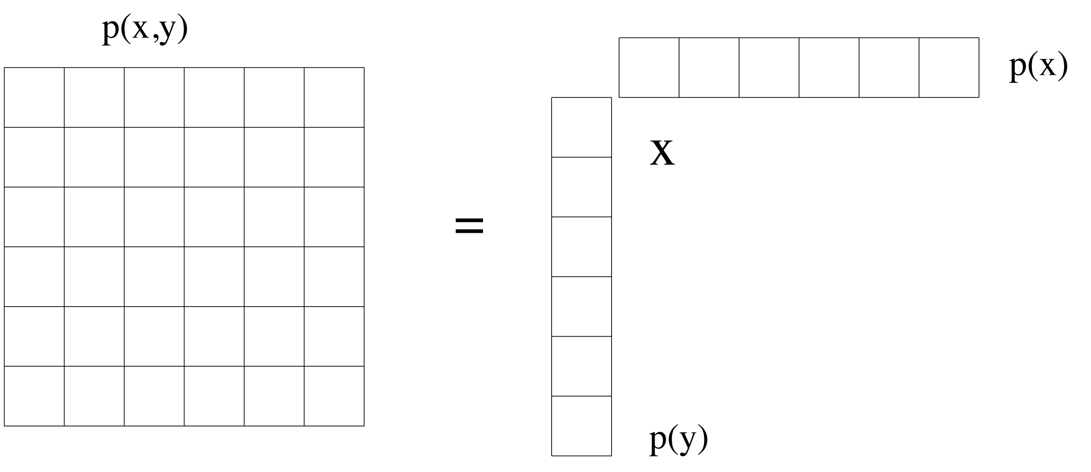


Slide from Sam Roweis (MLSS, 2005)

Independence and Conditional Independence

- Two variables are independent iff their joint factors:

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



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x, y|z) = p(x|z)p(y|z) \qquad \forall z$$

Slide from Sam Roweis (MLSS, 2005)

MLE AND MAP

MLE

What does maximizing likelihood accomplish?

- There is only a finite amount of probability mass (i.e. sum-to-one constraint)
- MLE tries to allocate **as much** probability mass **as possible** to the things we have observed…

  …**at the expense** of the things we have **not** observed

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Maximum Likelihood Estimate (MLE)

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Background: MLE

Example: MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ.
- Compute second derivative and check that it is concave down at λ^{MLE}.

Background: MLE

Example: MLE of Exponential Distribution

- First write down log-likelihood of sample.

$$\ell(\lambda) = \sum_{i=1}^{N} \log f(x^{(i)}) \tag{1}$$

$$= \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)})) \tag{2}$$

$$= \sum_{i=1}^{N} \log(\lambda) + -\lambda x^{(i)} \tag{3}$$

$$= N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \tag{4}$$

Background: MLE

Example: MLE of Exponential Distribution

- Compute first derivative, set to zero, solve for λ.

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \quad (1)$$

$$= \frac{N}{\lambda} - \sum_{i=1}^{N} x^{(i)} = 0 \quad (2)$$

$$\Rightarrow \lambda^{\text{MLE}} = \frac{N}{\sum_{i=1}^{N} x^{(i)}} \quad (3)$$

Background: MLE

Example: MLE of Exponential Distribution

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim$ Exponential(λ) for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ.
- Compute second derivative and check that it is concave down at λ^{MLE}.

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Maximum Likelihood Estimate (MLE)

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum *a posteriori* (MAP) estimate

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})\underbrace{p(\boldsymbol{\theta})}_{\text{Prior}}$$

COMMON PROBABILITY DISTRIBUTIONS

Common Probability Distributions

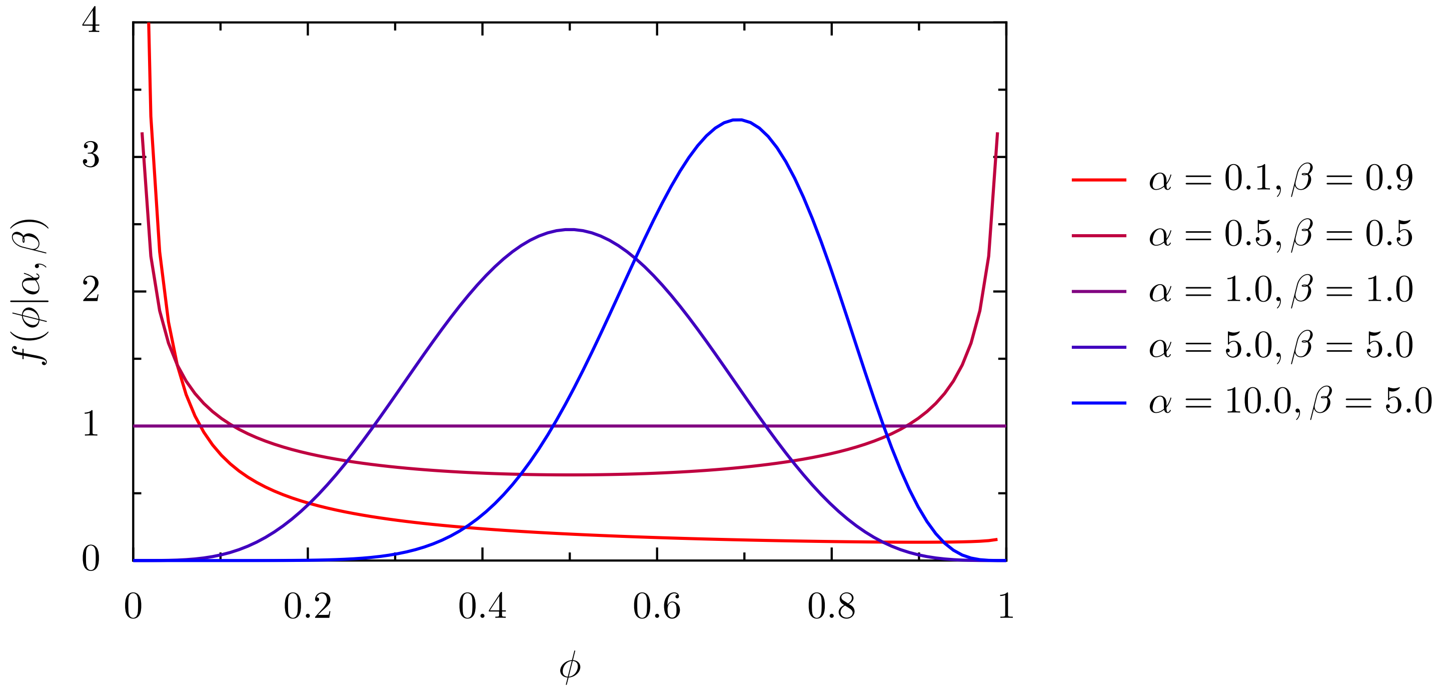
- For Discrete Random Variables:
  - Bernoulli
  - Binomial
  - Multinomial
  - Categorical
  - Poisson
- For Continuous Random Variables:
  - Exponential
  - Gamma
  - Beta
  - Dirichlet
  - Laplace
  - Gaussian (1D)
  - Multivariate Gaussian

Common Probability Distributions

Beta Distribution

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

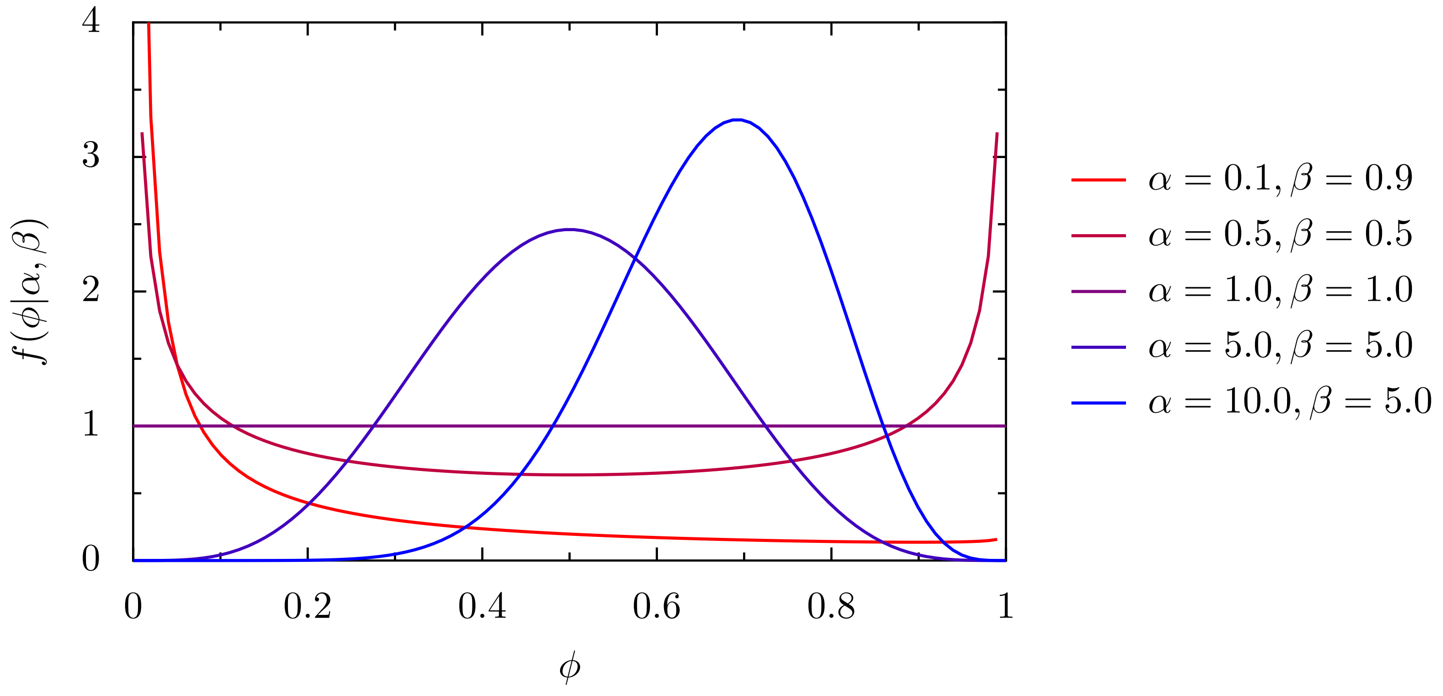


Common Probability Distributions

Dirichlet Distribution

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$



Common Probability Distributions

Dirichlet Distribution

probability density function:

$$p(\vec{\phi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \quad \text{where } B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$