COSC343: Artificial Intelligence

Lecture 4: Probability Theory: introduction

Lech Szymanski

Dept. of Computer Science, University of Otago

In today's lecture

- Mathematical framework for dealing with uncertainty
- Probability distributions
- Conditional probability
- Independence
- Expectation

Probability Theory

- Fundamental concept underlying all machine learning is **uncertainty**
- Probability theory = mathematical framework for quantification and manipulation of uncertainty
- What's the *best* action to take, when the outcome is uncertain?

Defining a sample space

A **sample space** is a model of 'all possible ways the world can be'.

- Formally, it's the space of all possible values of the input and outputs to the function
- Each of these defines one dimension of the samples space
- Each possible combination is called a **sample point**

Formally, a **probability model** assigns a probability to each sample point in a sample space.

- Each probability is between 0 and 1 inclusive
- Probabilities for all points in the space sum to 1

Examples of sample spaces

Coin toss

Tails	Heads

X

Dice roll

1	2	3	4	5	6

X

Double Dice roll

X_2						
	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)

X_1

Probability distribution

Imagine we roll a single die. Our sample space has a single **random variable** (call it X), which has 6 possible values.

n	1	2	3	4	5	6
	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$	$P(X=5)$	$P(X=6)$

We can estimate the probability at each point by generating a training set of N die rolls and using **relative frequencies** of events in this set

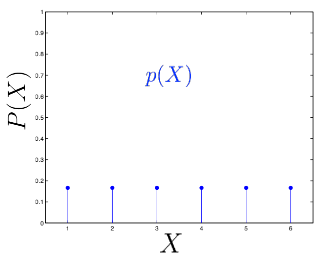
$$P(X=n) = \frac{\text{count}(X=n)}{N}$$

A simple probability model

A probability model induces a **probability distribution** for each possible value of the random variable.

- This distribution is a function, whose domain is all possible value for the random vairable, which returns probability for each possible value
- The distribution must sum to 1

n	1	2	3	4	5	6
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



$$p(X) = \begin{cases} \frac{1}{6} & ,\{X \mid X \in \mathbb{Z} \wedge 1 \leq X \geq 6\} \\ 0 & ,\text{otherwise} \end{cases}$$

- Discrete uniform distribution – countable number of events and each event is equally likely

Joint distribution

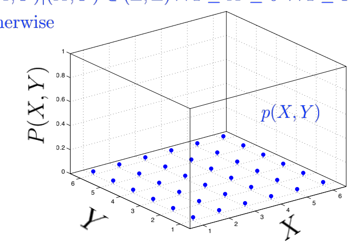
A distribution function over two, or more, random variables is called a **joint distribution**

- E.g. Double dice roll

X

Y	1	2	3	4	5	6
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

$$p(X,Y) = \begin{cases} \frac{1}{36} & ,\{(X,Y) \mid (X,Y) \in (\mathbb{Z},\mathbb{Z}) \wedge 1 \leq X \geq 6 \;\wedge 1 \leq Y\} \\ 0 & ,\text{otherwise} \end{cases}$$



- Discrete uniform distribution – countable number of events and each event is equally likely

Some terminology

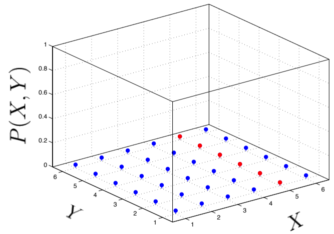
- An **event** is any subset of points in a sample space.
- The probability of an event E is the sum of probabilities of each sample point it contains.

$$P(E) = \sum_{\{n \in E\}} P(X = n)$$

Events

X

Y	1	2	3	4	5	6
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$



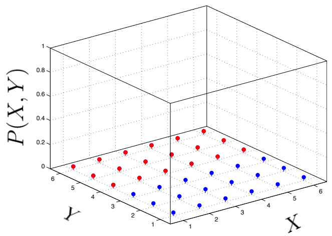
$p(X,Y)$

- Double dice roll
- What's $P(X=5)$?

Events

X

Y	1	2	3	4	5	6
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$



$p(X,Y)$

- Double dice roll
- What's $P(Y \geq 4)$?

A simple medical example

Consider a medical scenario, with 3 Boolean variables

- *cavity* (does the patient have a cavity or not?)
- *toothache* (does the patient have a toothache or not?)
- *catch* (does the dentist's probe catch on the patient's tooth?)

Here's an example probability model: the joint probability distribution $p(Toothache, Cavity, Catch)$

"not"

	toothache		¬*toothache*	
	catch	¬*catch*	*catch*	¬*catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬*cavity*	**.016**	**.064**	**.144**	**.576**

Inference from a joint distribution

Given a full joint distribution, we can compute the probability of any event simply by summing the probabilities of the relevant sample points.

E.g. how to calcualte $P(toothache)$?

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.06 = 0.2$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬cavity	**.016**	**.064**	**.144**	**.576**

Inference from a joint distribution

Given a full joint distribution, we can compute the probability of any event simply by summing the probabilities of the relevant sample points.

"or"

E.g. how to calcualte $P(cavity \vee toothache)$?

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬cavity	**.016**	**.064**	**.144**	**.576**

Inference from a joint distribution

Given a full joint distribution, we can compute the probability of any event simply by summing the probabilities of the relevant sample points.

E.g. how to calcualte $P(cavity \vee toothache)$?

$$\begin{aligned} P(cavity \vee toothache) &= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 \\ &= 0.28 \end{aligned}$$

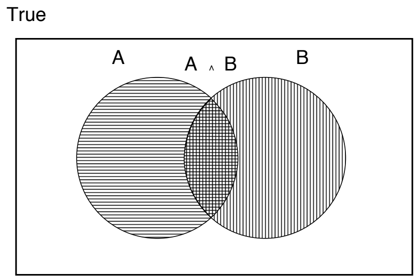
	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬cavity	**.016**	**.064**	**.144**	**.576**

Set-theoretic relationships in probability

Given a full joint distribution, we can compute the probability of any event simply by summing the probabilities of the relevant sample points.

"or" "and"

For instance: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

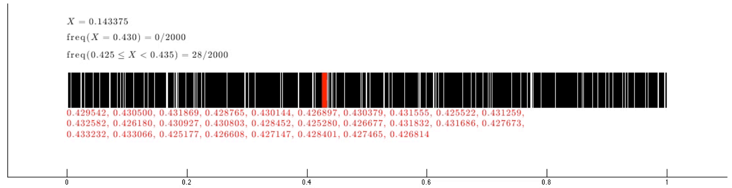


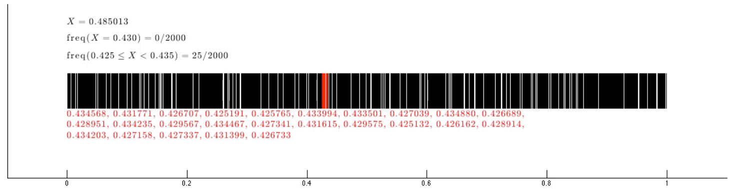
Continuous random variables

The sample spaces we've seen so far have been built from descrete random variables. But you can build probability models using **continuous random variables** too.

- E.g. we can define a random variable *Temperature*, whose domain is the real numbers.
- In the real domain (even if it's bounded) domain there is an infinite number of samples. Probability of continuous random variable hitting a specific value is 0.
- However, we can talk about probability of value being in certain range.

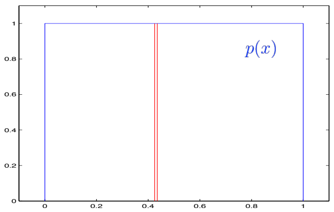
Continuous random variables





Probability density function

- For continuous variables, probability distributions are contintuos, and are referred to as **probability denstity functions**
- E.g. here's a function which gives uniform probability for values between 0 and 1



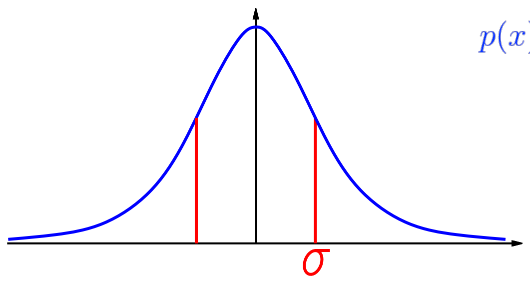
$$p(x) = \begin{cases} 1 & , \{x | x \in \mathbb{R} \wedge 0 \leq x \leq 1\} \\ 0 & , \text{otherise} \end{cases}$$

- This funcitonion is a density; itengrates to 1. So:

$$P(0.425 \leq x < 0.435) = \int_{0.425}^{0.435} p(x)dx = 0.01$$

Gaussian distribution

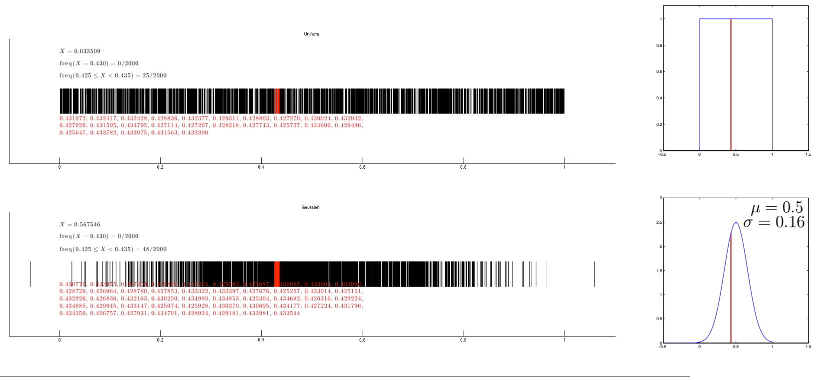
- A particularly useful probability function for continuous variables is the **Guassian** function (often referred to as the **normal** distribution)



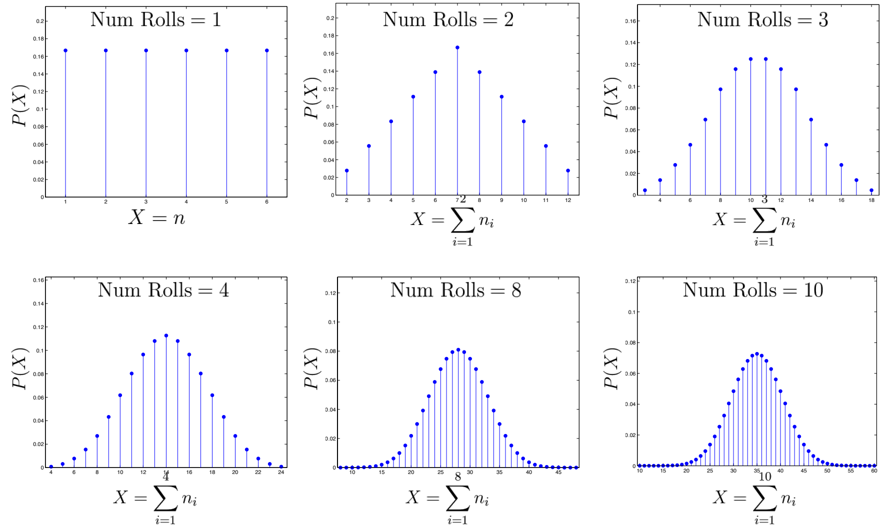
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Lots of real-world variables have this distribution

Gaussian distribution



Central Limit Theorem



Expectation

- Probability weighted value of all possible values of a function dependent on a random variable
- “Average” result expected

Discrete distribution

$$E\left[g(x)\right] = \sum_i p(x_i)g(x_i)$$

Continuous distribution

$$E\left[g(x)\right] = \int p(x)g(x)dx$$

Mean and variance

- The expected value of the random variable itself

$$\mu = E\left[x\right]$$

- The expected value of the squared deviation of random variable from its mean (measures the spread of a probability distribution).

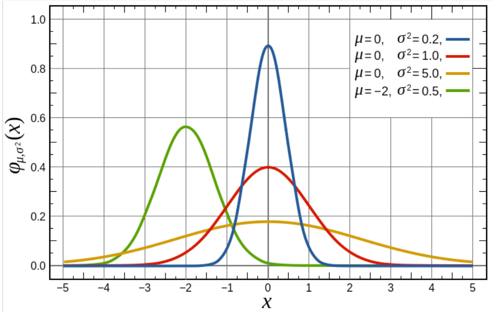
$$\sigma^2 = E\left[(x-\mu)^2\right]$$

An exampe: mean and variance of the normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[x] = \mu$$

$$E[(x-\mu)^2] = \sigma^2$$



https://en.wikipedia.org/wiki/Normal_distribution

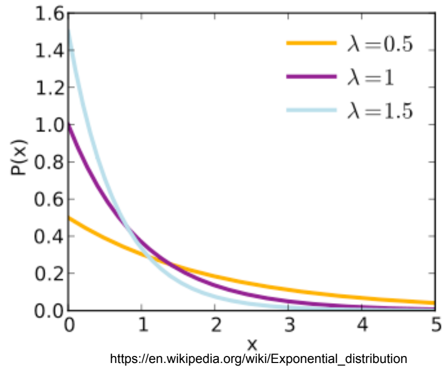
- Guassian distribution is completely parametrised by its meand and variance
- σ - standard deviation

An exampe: mean and variance of the exponential distribution

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

$$\mu = E[x] = \lambda^{-1}$$

$$E[(x-\mu)^2] = \lambda^{-2}$$



https://en.wikipedia.org/wiki/Exponential_distribution

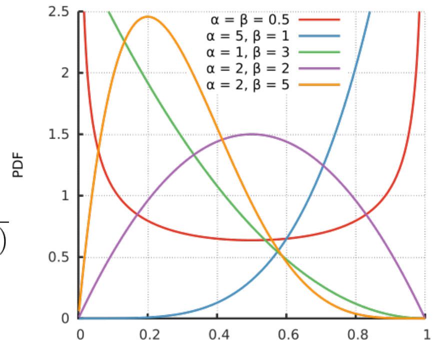
- Mean is the point splitting the probability density, such that are under curve is exactly 0.5 on either side

An example: mean and variance of the Beta distribution

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}$$

$$\mu = E[x] = \frac{\alpha}{\alpha+\beta}$$

$$E[(x-\mu)^2] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



https://en.wikipedia.org/wiki/Beta_distribution

Summary and reading

Probability theory is the foudnation for many learning algorithms.

- Key concepts: samples space, random variable, probability distribution, probability density, expectation

Reading for the lecture: AIMA Chapter 13 Sections 1-2

Reading for next lecture: AIMA Chapter 13 Section 3-6