CSE 6363 - Machine Learning

Probability Theory

Alex Dillhoff

University of Texas at Arlington

Probability Theory

What is probability theory?

A consistent framework for the quantification and manipulation of **uncertainty**.

Probability Theory

We will cover some probabilistic methods for machine learning in this course.

A brief review of probability theory and some fundamental distributions wouldn’t hurt!

Example: Mixing Cookies

Let’s start with an example…

There are two cookie jars

- A blue jar with 8 oatmeal raisin cookies
- A red jar with 10 chocolate chip cookies

Example: Mixing Cookies

However, some monsters took 2 chocolate chip cookies and put them in the blue jar

and placed 1 oatmeal raisin cookie in the red jar.

Example: Mixing Cookies

The blue jar now has 2 chocolate chip and 7 oatmeal raisin.

The red jar has 8 chocolate chip and 1 oatmeal raisin.

Example: Mixing Cookies

Let's further say that we happen to pick the red jar 80% of the time and the blue jar 20% of the time.

We can start to formulate about these events via probability by assigning these actions to **random variables**.

Example: Mixing Cookies

We have two types of things: cookies and jars. Let's assign them to variables.

- J - The type of jar, either blue b or red r.
- C - The type of cookie, either oatmeal o or chocolate chip c.

Example: Mixing Cookies

We can present the probabilities of picking each type of jar using a functional notation:

- $p(J = b) = 0.2$
- $p(J = r) = 0.8$

Notice that their sum is 1. This makes sense considering they are the only 2 jars in our problem.

Example: Mixing Cookies

We can also define the probability of picking each cookie.

- $p(C = o)$
- $p(C = c)$

However, this probability is based on which jar is picked.

Example: Mixing Cookies

We can also define the probability of picking each cookie.

	Chocolate Chip	Oatmeal Raisin
Blue Jar	$\frac{2}{9} = 0.222$	$\frac{7}{9} = 0.778$
Red Jar	$\frac{8}{9} = 0.889$	$\frac{1}{9} = 0.111$

Example: Mixing Cookies

Given these quantities, we can ask slightly more complication questions.

What is the probability that I will select the red jar AND take a chocolate chip cookie?

This is expressed as a **joint probability distribution**, $p(J = r, C = c)$.

Example: Mixing Cookies

$p(J = r, C = c)$ is defined based on

- the prior probability of picking the red jar,
- the conditional probability of picking the chocolate chip cookie **conditioned** on the red jar being picked.

Mathematically, $p(J = j, C = c) = p(C = c|J = r)p(J = r)$.

Example: Mixing Cookies

$$p(J = j, C = c) = p(C = c|J = r)p(J = r)$$

is also known as the **product rule**.

Example: Mixing Cookies

We have all of the quantities we need to answer this.

$$p(J = r) = 0.8 \text{ and } p(C = c|J = r) = 0.889$$

Thus,

$$p(J = r, C = c) = 0.8 * 0.889 = 0.711$$

Example: Mixing Cookies

If we knew nothing about the contents of the jar or the prior probabilities of selecting a specific jar, we could measure this probability empirically.

Example: Mixing Cookies

Conduct N trials of taking a cookie from one of the jars, recording it, and placing it back in the same jar.

Count the number of times we select a red jar AND a chocolate chip cookie.

Example: Mixing Cookies

If we wanted to measure the **conditional probability** $p(C = c|J = r)$...

Count the number of times a chocolate chip cookie is taken (and replaced) from the red jar and divide by N.

Example: Mixing Cookies

Calculating all joint probabilities produces a joint probability table:

	Chocolate Chip	Oatmeal Raisin
Red Jar	0.711	0.089
Blue Jar	0.044	0.156

Example: Mixing Cookies

Note that the sum of the sum of rows is equal to 1.

Likewise, the sum of the sum of columns is equal to 1.

Example: Mixing Cookies

The sum of the each column for a given row adds up to the prior probability of selecting a jar.

The sum of each row for a given column is the prior probability of selecting a cookie.

These are called **marginal probabilities**.

Example: Mixing Cookies

In general, the marginal probability can be computed by summing out the joint variables:

$$p(x_i) = \sum_j p(x_i, y_j)$$

Chain Rule

What if we want the joint probability over k variables?

$$p(a_1, \cdots, a_k)$$

Chain Rule

What if we want the joint probability over k variables?

$$p(a_1, \cdots, a_k) = p(a_1)p(a_2|a_1) \cdots p(a_k|a_1, \cdots, a_{k-1})$$

Bayes' Rule

Notationally, $p(X, Y)$ and $p(Y, X)$ would be written slightly differently, but they are equal.

Setting them equal to each other is the basis of the derivation of **Bayes' rule:**

$$\begin{aligned} p(X, Y) &= p(Y, X) \\ p(X|Y)p(Y) &= p(Y|X)p(X) \\ p(X|Y) &= \frac{p(Y|X)p(X)}{p(Y)} \end{aligned}$$

Bayes' Rule

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$$

This will come in handy in for many probabilistic methods later on.

Bayes' Rule

In the context of Bayes' rule, $p(X|Y)$ is referred to as the **posterior probability** of event X conditioned on the fact that we know event Y has occurred.

$p(X)$ is the **prior probability** of event X in the absence of any additional evidence.

Bayes' Rule Example: COVID-19 Testing

Murphy presents an excellent example of understanding probabilities via Bayes' rule.

- H is the infected state (1 for yes, 0 for no)
- Y is the test event (1 for positive, 0 for negative)

We want to find $p(H = h|Y = y)$: the probability of the state of infection given a test result.

Bayes' Rule Example: COVID-19 Testing

The **sensitivity** is $p(Y = 1|H = 1)$.

The probability of testing positive conditioned on actually being infected.

Bayes' Rule Example: COVID-19 Testing

The **false negative rate** is $1 - p(Y = 1|H = 1)$.

Also written as $p(Y = 0|H = 1)$.

Bayes’ Rule Example: COVID-19 Testing

The **specificity** is defined as $p(Y = 0|H = 0)$.

The probability of testing negative conditioned on no infection.

Bayes' Rule Example: COVID-19 Testing

The false positive rate is defined as $p(Y = 1|H = 0)$.

The probability of testing positive conditioned on no infection.

Also defined as $1 - p(Y = 0|H = 0)$.

Bayes’ Rule Example: COVID-19 Testing

Now that we have defined the likelihoods, we need the priors.

The **prevalence** of the disease in your area is $p(H = 1)$.

Bayes' Rule Example: COVID-19 Testing

Let's apply some values to these quantities. Suppose the likelihoods follow the table below.

	$Y = 0$	$Y = 1$
$H = 0$	0.975	0.025
$H = 1$	0.125	0.875

Additionally, we'll suppose the prevalence of infection $p(H = 1) = 0.05$.

Bayes' Rule Example: COVID-19 Testing

If you test positive, what is the probability that you are actually infected (true positive rate)?

$$
\begin{aligned}
p(H = 1|Y = 1) &= \frac{p(Y = 1|H = 1)p(H = 1)}{\sum_h p(Y = 1|H = h)p(H = h)} \\
&= \frac{0.875 \times 0.05}{0.875 \times 0.05 + 0.025 \times 0.95} \\
&= 0.648
\end{aligned}
$$

There is a 64.8% chance you are infected.

Bayes' Rule Example: COVID-19 Testing

If you test negative, what is the probability that you are actually infected?

Using the same data, we can calculate $p(H = 0|Y = 1) = .0067$, or 0.67%.

Independence

Two variables are **independent**, then

$$p(X, Y) = p(X)p(Y)$$

If two variables are conditionally independent given a third event, then

$$p(X, Y|Z) = P(X|Z)P(Y|Z)$$

Continuous Variables

In the introductory example, the events took on discrete values.

Most of the problems we will see in this course involve **continuous values**.

Continuous Variables

Consider a small differential of a random variable x, δx.

The probability density is $p(x)$.

Continuous Variables

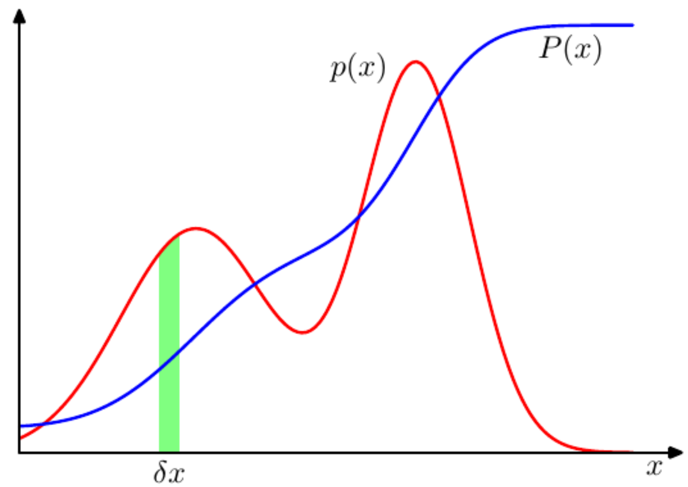

Figure 1: PDF $p(x)$ and CDF $P(x)$. Source: Bishop

Continuous Variables

With the small differential δx, the probability that x lies on some interval (a, b) is given by

$$p(a \leq x \leq b) = \int_a^b p(x)dx$$

Continuous Variables

The probability density must sum to 1 and cannot take a negative value.

$$p(x) \geq 0$$
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

However, it is possible to have a value greater than 1 as long as the integrals over any interval are $<=$ 1.

Continuous Variables

The cumulative distribution function $P(x)$ is the probability that x lies in the interval $(-\infty, z)$

$$P(z) = \int_{\infty}^{z} p(x)dx.$$

Note that the derivative of the cdf is equal to the pdf.

Continuous Variables

The **product rule** for continuous probability distributions takes on the same form as that of discrete distributions.

The **sum rule** is written in terms of integration:

$$p(x) = \int p(x, y) dy.$$

Moments of Distributions

A **moment** of a function describes a quantitative measurement related to its graph.

With respect to probability densities, the kth moment of $p(x)$ is defined as $\mathbb{E}[x^k]$.

The first moment is the **mean** of the distribution, the second moment is the **variance**, and the third moment is the **skewness**.

Expectation

The **expectation** of a function is the mean under a proability distribution $p(x)$

$$\mathbb{E}[f] = \sum_x p(x)f(x) \text{ and}$$
$$\mathbb{E}[f] = \int p(x)f(x)dx,$$

Expectation

Given a fair d6, for which each value is equally likely, $p(x) = \frac{1}{6}$, the expectation is

$$\begin{aligned}
\mathbb{E}[f] &= \sum_x p(x) f(x) \\
&= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\
&= 3.5
\end{aligned}$$

Expectation

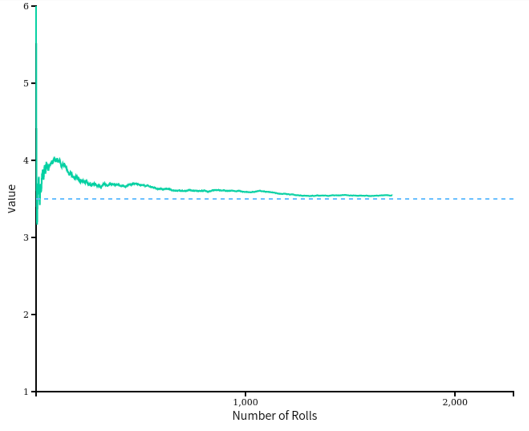

Figure 2: Expectation of rolling a d6 over 1800 trials converges to 3.5. Source: Seeing Theory

Variance

The **variance** of a function $f(x)$ under a probability distribution $p(x)$ measures how much variability is in $f(x)$ around the expected value $\mathbb{E}[f(x)]$

$$\begin{aligned} \mathrm{var}[f] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \end{aligned}$$

Variance

If we have a stack of 10 cards with values 1-10 and draw 1 (with replacement) over N trials, the variance is

$$\begin{aligned} \mathrm{var}[f] &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \\ &= 38.5 - 30.25 \\ &= 8.25 \end{aligned}$$

Variance

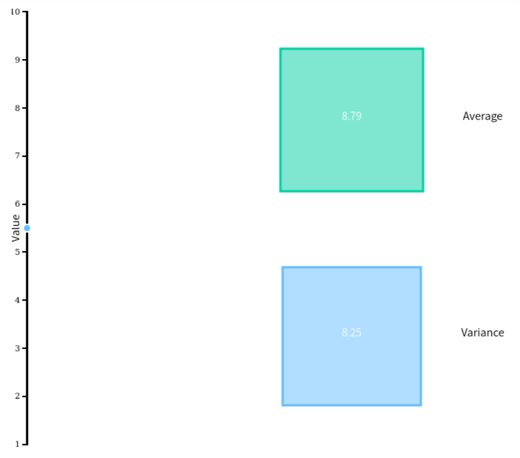

Figure 3: Variance of drawing a card (1-10) over N trials converges to 8.25. Source: Seeing Theory

Covariance

The **covariance** of two random variables x and y provides a measure of dependence between them.

$$\begin{aligned}\text{cov}[x, y] &= E_{x,y}[\{x - \mathbb{E}[x]\}\{y^T - \mathbb{E}[y^T]\}] \\ &= \mathbb{E}_{\mathrm{x,y}}[\mathrm{xy}^T] - \mathbb{E}[\mathrm{x}]\mathbb{E}[\mathrm{y}^T].\end{aligned}$$

Covariance

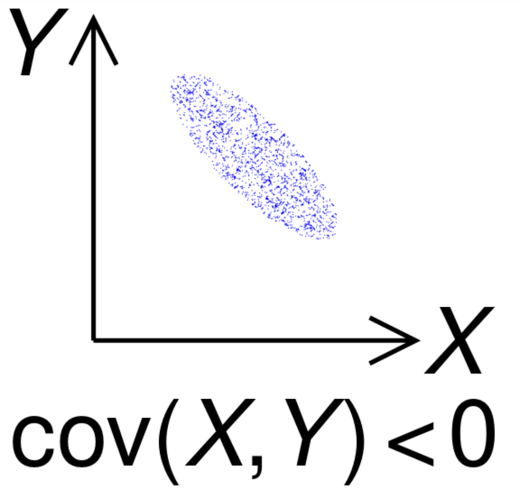

Figure 4: Plot of variables for with the covariance is negative. Source: Wikipedia

Covariance

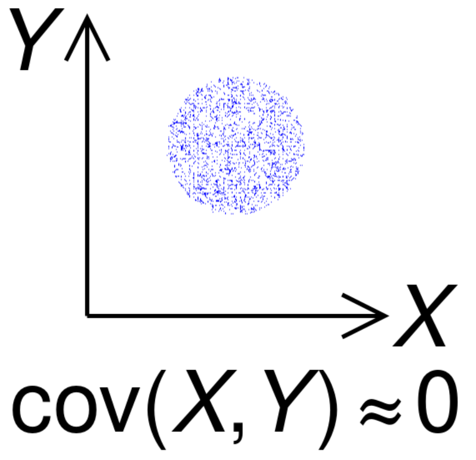

Figure 5: Plot of variables for with the covariance is approximately 0. Source: Wikipedia

Covariance

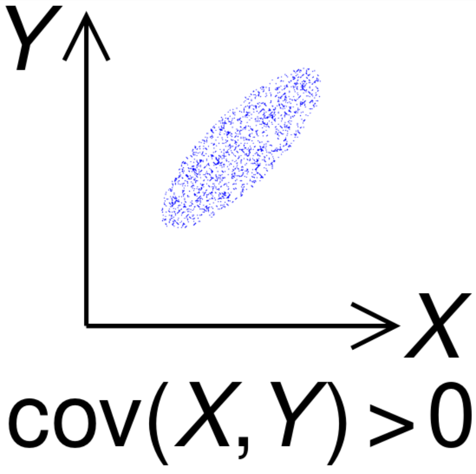

Figure 6: Plot of variables for with the covariance is positive. Source: Wikipedia

Correlation

The **correlation** between two random variables x and y relates to their covariance, but it is normalized to lie between -1 and 1.

$$\mathrm{corr}[\mathbf{x}, \mathbf{y}] = \frac{\mathrm{cov}[\mathbf{x}, \mathbf{y}]}{\sqrt{\mathrm{var}[\mathbf{x}]\mathrm{var}[\mathbf{y}]}}$$

Correlation

The correlation between two variables will equal 1 if there is a linear relationship between them.

We can then view the correlation as providing a measurement of linearity.

Correlation

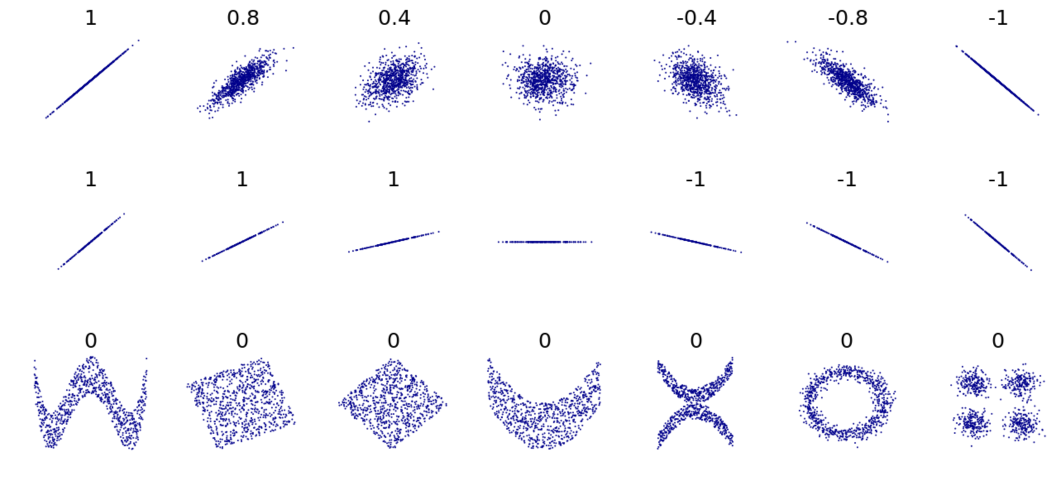

Figure 7: Sets of points with their correlation coefficients. Source: Wikipedia

Limits of Moments

When possible, it is always better to visualize the data.

An example of this is the **Anscombosaurus**, derived from the Anscombe's quartet.

The quartet consists of four datasets that have nearly identical summary statistics but are visually distinct.

A modern version, called the Datasaurus Dozen, consists of 12 datasets that have the same summary statistics but are visually distinct.

Anscombosaurus

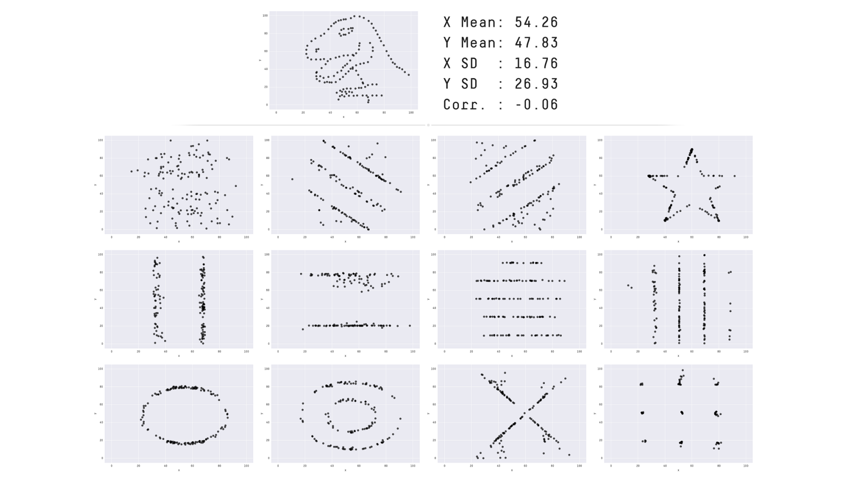

Figure 8: The Anscombosaurus. Source: Datasaurus Dozen