Probability Review for Machine Learning

Roger Grosse, Rahul Krishnan, and Guodong Zhang [1]

University of Toronto

[1] Slides adapted from Erdogdu and Zemel

Motivation

Uncertainty arises through:

- Noisy measurements
- Variability between samples
- Finite size of data sets

Probability provides a consistent framework for the quantification and manipulation of uncertainty.

Sample Space

Sample space Ω is the set of all possible outcomes of an experiment.

Observations $\omega \in \Omega$ are points in the space also called sample outcomes, realizations, or elements.

Events $E \subset \Omega$ are subsets of the sample space.

In this experiment we flip a coin twice:

Sample space All outcomes $\Omega = \{HH, HT, TH, TT\}$

Observation $\omega = HT$ valid sample since $\omega \in \Omega$

Event Both flips same $E = \{HH, TT\}$ valid event since $E \subset \Omega$

Probability

The probability of an event E, $P(E)$, satisfies three axioms:

1: $P(E) \geq 0$ for every E

2: $P(\Omega) = 1$

3: If $E_1, E_2, \ldots$ are disjoint then

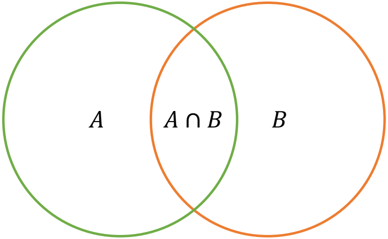
$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

Joint and Conditional Probabilities

Joint Probability of A and B is denoted $P(A, B)$.

Conditional Probability of A given B is denoted $P(A|B)$.

Joint: $p(A, B) = p(A \cap B)$



Conditional: $p(A|B) = \frac{p(A \cap B)}{p(B)}$

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

Conditional Example

Probability of passing the midterm is 60% and probability of passing both the final and the midterm is 45%.
What is the probability of passing the final given the student passed the midterm?

$$\begin{aligned} P(F|M) &= P(M,F)/P(M) \\ &= 0.45/0.60 \\ &= 0.75 \end{aligned}$$

Independence

Events A and B are independent if $P(A, B) = P(A)P(B)$.

- Independent: A: first toss is HEAD; B: second toss is HEAD;

$$P(A, B) = 0.5 \times 0.5 = P(A)P(B)$$

- Not Independent: A: first toss is HEAD; B: first toss is HEAD;

$$P(A, B) = 0.5 \neq P(A)P(B)$$

Independence

Events A and B are conditionally independent given C if

$$P(A, B|C) = P(B|C)P(A|C)$$

Consider two coins[2]: A regular coin and a coin which always outputs heads.

$$\begin{aligned} A &= \text{The first toss is heads;} \\ B &= \text{The second toss is heads;} \\ C &= \text{The regular coin is used} \\ D &= \text{The biased coin is used} \end{aligned}$$

Then A and B are conditionally independent given C and given D.

[2] `www.probabilitycourse.com/chapter1/1_4_4_conditional_independence.php`

Independence

Events A and B are conditionally independent given C if

$$P(A, B|C) = P(B|C)P(A|C)$$

Consider a coin which outputs heads if the first toss was heads, and tails otherwise.

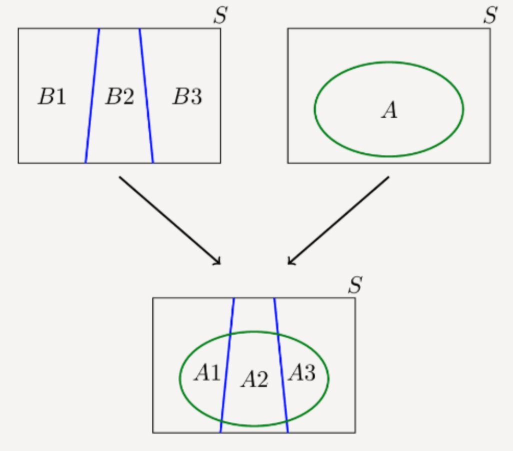
$$\begin{aligned} A &= \text{The first toss is heads;} \\ B &= \text{The second toss is heads;} \\ E &= \text{The eventually biased coin is used} \end{aligned}$$

Then A and B are conditionally dependent given E.

Marginalization and Law of Total Probability

Law of Total Probability [3]

$$P(A) = \sum_B P(A, B) = \sum_B P(A|B)P(B)$$



[3] www.probabilitycourse.com/chapter1/1_4_2_total_probability.php

Bayes' Rule

Bayes' Rule:

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

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$Posterior \propto Likelihood \times Prior$$

Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

This depends on the prior probability of the disease:

- $P(T = 1|D = 1) = 0.95$ (likelihood)
- $P(T = 1|D = 0) = 0.10$ (likelihood)
- $P(D = 1) = 0.1$ (prior)

So $P(D = 1|T = 1) = ?$

Bayes' Example

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

$P(T = 1|D = 1) = 0.95$ (true positive)

$P(T = 1|D = 0) = 0.10$ (false positive)

$P(D = 1) = 0.1$ (prior)

So $P(D = 1|T = 1) =?$
Use Bayes' Rule:

$$\begin{aligned} P(T = 1) &= P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0) \\ &= 0.95 * 0.1 + 0.1 * 0.90 = 0.185 \end{aligned}$$

$$P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)} = \frac{0.95 * 0.1}{P(T = 1)} = 0.51$$

Random Variable

How do we connect sample spaces and events to data?
A random variable is a mapping which assigns a real number $X(\omega)$ to each observed outcome $\omega \in \Omega$

For example, let's flip a coin 10 times. $X(\omega)$ counts the number of Heads we observe in our sequence. If $\omega = HHTHTHHTHT$ then $X(\omega) = 6$. We often shorten this and refer to the random variable X.

Expectations

From our example, we see that X does not have a fixed value, but rather a distribution of values it can take. It is natural to ask questions about this distribution, such as “What is the average number of heads in 10 coin tosses?”
This average value is called the expectation and denoted as $E[X]$. It is defined as

$$E[x] = \sum_{a \in \mathcal{A}} P[X = a] \times a$$

where $\mathcal{A}$ represents the set of all possible values $X(w)$ can take.

Expectation Practice

- What is the expected value of a fair die?
- X = value of roll

$$\begin{aligned} E[X] &= \sum_{a \in \{1,2,3,4,5,6\}} \frac{1}{6} a \\ &= \frac{1}{6} \sum_{a=1}^{6} a \\ &= \frac{21}{6} = \frac{7}{2} \end{aligned}$$

Linearity of expectations

There are two powerful properties regarding expectations.

1. $E[X + Y] = E[X] + E[Y]$.
   This holds even if the random variables are dependent.
2. $E[cX] = cE[X]$, where c is a constant.

Note we cannot say anything in general about $E[XY]$.

Expectation Practice

What is the expected value of the sum of two dice?
X_1 = value of roll 1
X_2 = value of roll 2

$$E[X_1 + X_2] = E[X_1] + E[X_2] = \frac{7}{2} + \frac{7}{2} = 7$$

(compare this to computing $2 \times \frac{1}{36} + 3 \times \frac{2}{36} + \dots$)

Expectation Practice 2

Suppose there are n students in class, and they each complete an assignment. We hand back assignments randomly. What is the expected number of students that receive the correct assignment? When $n = 3$? In general?

$X =$ Number of students that get their assignment back

$X_i =$ Student i gets their assignment back

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \ldots + X_n] \\ &= E[X_1] + E[X_2] + \ldots + E[X_n] \\ &= \frac{1}{n} \times n = 1 \end{aligned}$$

Variances

Knowing the expectation can only tell us so much. We have another quantity used to describe how far off we are from the expected value. It is defined as follows for a random variable X with $E[X] = \mu$:

$$\mathrm{Var}[x] = E[(X - \mu)^2]$$

The variance can be simplified as:

$$\begin{aligned} E[(X - \mu)^2] &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - E[2\mu X] + E[\mu^2] \\ &= E[X^2] - 2\mu E[X] + E[\mu^2] \\ &= E[X^2] - \mu^2 \end{aligned}$$

Variance Properties

Constants get squared:

$$\operatorname{Var}[cX] = c^2 \operatorname{Var}[X]$$

For independent random variables X and Y, we have

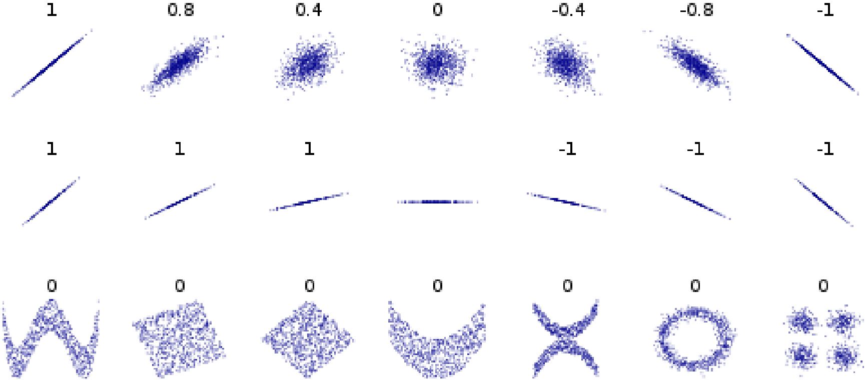
$$E[XY] = E[X]E[Y]$$

and

$$\operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$$

The quantity we encounter during the proof $E[XY] - E[X]E[Y]$ is called the covariance. It is 0 when X and Y are independent. Q: can it be 0 when X and Y are not independent?

Variance Properties



Variance Practice

Consider a particle that starts at position 0. At each time step, the particle moves one step to the left or one step to the right with equal probability. What is the variance of the particle at time step n?
$X = X_1 + X_2 + \ldots + X_n$

Each X_i is 1 or -1 with equal probability.

$$\mathrm{Var}(X_i) = 1$$
$$\mathrm{Var}(X) = \sum \mathrm{Var}(X_i) = n$$

The expected squared distance from 0 is n.

Discrete and Continuous Random Variables

Discrete Random Variables

- Takes countably many values, e.g., number of heads
- Distribution defined by probability mass function (PMF)
- Marginalization: $p(x) = \sum_y p(x,y)$

Continuous Random Variables

- Takes uncountably many values, e.g., time to complete task
- Distribution defined by probability density function (PDF)
- Marginalization: $p(x) = \int_y p(x,y)dy$

I.I.D.

Random variables are said to be independent and identically distributed (i.i.d.) if they are sampled from the same probability distribution and are mutually independent.
This is a common assumption for observations. For example, coin flips are assumed to be i.i.d.

Probability Distribution Statistics

Mean: First Moment, μ

$$E[x] = \sum_{i=1}^{\infty} x_i p(x_i) \qquad \text{(univariate discrete r.v.)}$$

$$E[x] = \int_{-\infty}^{\infty} x p(x) dx \qquad \text{(univariate continuous r.v.)}$$

Variance: Second (central) Moment, σ^2

$$\begin{aligned}\mathrm{Var}[x] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \\ &= E[(x-\mu)^2] \\ &= E[x^2] - E[x]^2\end{aligned}$$