L665 Machine Learning for NLP
2. Probability Theory

Damir Cavar

Indiana University

January 2018

Outline

Reading

See syllabus

- Maleki & Do: Review of Probability Theory
- Manning & Schuetze: Ch. 2
- Goodfellow et al.: Ch. 3

The following slides are based on Maleki & Do: *Review of Probability Theory*

Probability Theory

- Mathematical framework: representing uncertain statements
  - quantifying uncertainty
  - axioms for deriving new uncertain statements
- In AI we use Probability Theory:
- laws of probability tell us how AI systems should reason
- so we design our algorithms to compute or approximate various expressions derived using probability theory.
- Second, we can use probability and statistics to theoretically analyze the behavior of proposed AI systems

Axioms of Probability

Sample Space Ω

- The set of all the outcomes of a random experiment.
- Each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.

Axioms of Probability

Set of Events or **Event Space** $\mathcal{F}$

- A set whose elements $A \in \mathcal{F}$ (called events) are subsets of Ω.
- $A \subseteq \Omega$ is a collection of possible outcomes of an experiment.
- $\mathcal{F}$ should satisfy three properties:
  - $\emptyset \in \mathcal{F}$
  - $A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}$
  - $A_1, A_2, \dots \in \mathcal{F} \implies \cup_i A_i \in \mathcal{F}$

Axioms of Probability

Probability Measure

- A function $P : \mathcal{F} \rightarrow R$ that satisfies the following properties,
  - $P(A) \geq 0$, for all $A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

Axioms of Probability

Example

- Tossing six-sided die.
- Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Different event spaces:
  1. simple event space $\mathcal{F} = \{\emptyset, \Omega\}$
  2. another event space could be all subsets of $\mathcal{F}$
- for 1., the unique probability measure is: $P(\emptyset) = 0, P(\Omega) = 1$
- for 2., one possibility would be to assign the probability of each set to $\frac{i}{6}$, with $i =$ number of elements in the particular set, i.e. $P(\{1, 2, 3, 4\}) = \frac{4}{6}$ or $P(\{1, 2, 3\}) = \frac{3}{6}$

Axioms of Probability

Properties

- $A \subseteq B \implies P(A) \leq P(B)$
- $P(A \cap B) \leq min(P(A), P(B))$
- (Union Bound) $P(A\ cupB) \leq P(A) + P(B)$
- $P(\Omega \setminus A) = 1 - P(A)$
- (Law of Total Probability) if $A_1, \ldots, A_k$ is a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_i) = 1$

Conditional Probability

With B, and event with non-zero probability:

- conditional probability of any event A given B is:

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \tag{1}$$

- $P(A|B)$ is the probability of event A after observing the event B.
- Two events are independent, if and only if

$$P(A \cap B) = P(A)P(B), or\ P(A|B) = P(A) \tag{2}$$

- Independence: the occurrence of B has no impact on the probability of A occurring.

Random Variables

Experiment:

- flipping 10 coins, asking for the number of coins coming up heads
- sample space Ω contains sequences of length 10 of heads and tails, e.g. $w_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$
- Practically: we are not necessarily interested in obtaining any particular sequence of heads and tails, but rather
- real-valued functions of outcomes
  - number of heads among 10 coin flips
  - length of the longest run of tails
- these functions are known as: **Random Variables**

Random Variables

Experiment:

- random variable X is a function $X : \Omega \to \mathbb{R}$.
- random variables denoted using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied)
- value that a random variable may take denoted using lower case letters x

Random Variables

Example: experiment above

- $X(\omega)$ is the number of heads that occur in the sequence of tosses ω.
- only 10 coins are tossed, $X(\omega)$ can take only a finite number of values
- so it is known as a discrete random variable.
- Here, the probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) \coloneqq P(\omega : X(\omega) = k) \tag{3}$$

Random Variables

Example:

- Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay.
- In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a continuous random variable.
- We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as

$$P(a \leq X \leq b) := P(\omega : a \leq X(\omega) \leq b) \tag{4}$$

Cumulative Distribution Functions

Functions:

- to specify the probability measures used when dealing with random variables
- alternative functions (CDFs, PDFs, and PMFs) are specified
- probability measures governing an experiment immediately follow from these functions

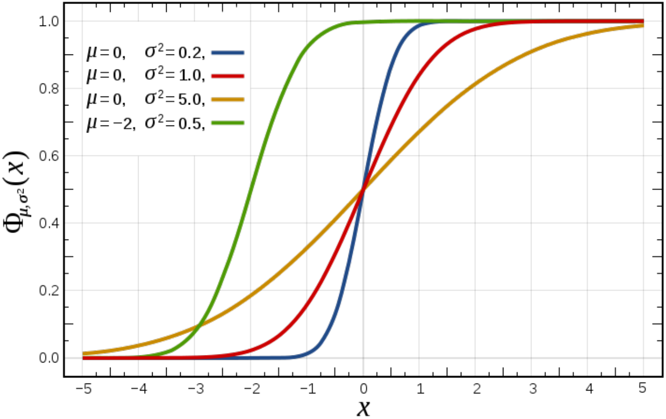
Cumulative Distribution Functions

CDF:

- for a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x
- a function $F_X : \mathbb{R} \rightarrow [0, 1]$ which specifies a probability measure as, $F_X(x) \triangleq P(X \leq x)$.
- by using this function one can calculate the probability of any event in $\mathcal{F}$ (i.e. the Event Space)

Cumulative Distribution Functions

Sample CDF functions (Wikipedia)



Cumulative Distribution Functions

CDF Properties:

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- $x \leq y \implies F_X(x) \leq F_X(y)$

Probability Mass Functions

PMF:

- X, a random variable that takes on a finite set of possible values (i.e. X is a discrete random variable)
- representing the probability measure associated with X by directly specifying the probability of each value that the random variable can assume
- that is: a PMF is a function that gives the probability that a discrete random variable is exactly equal to some value

Probability Mass Functions

PMF:

- A function $p_X : \Omega \rightarrow \mathbb{R}$ such that $p_X(x) \triangleq P(X = x)$
- For discrete random variables: $Val(X)$ is the set of possible values that the random variable X may assume
- Example: if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of a coin, then $Val(X) = \{0, 1, 2, \ldots, 10\}$.

Probability Mass Functions

PMF Properties:

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in Val(X)} p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = P(X \in A)$
  (A a set of elements in the Event Space $\mathcal{F}$)

Rates at which quantities change

The differential represents the principal part of the change in a function $y = f(x)$ with respect to changes in the independent variable.
The differential dy is defined by

$$dy = f'(x)\,dx \tag{5}$$

where $f'(x)$ is the derivative of f with respect to x, and dx is an additional real variable (so that dy is a function of x and dx). The notation is such that the equation

$$dy = \frac{dy}{dx}\,dx \tag{6}$$

Calculus: Differentiation

Rates at which quantities change

Constant function: $f(x) = c$

- Derivative: $f'(x) = 0$
- Example: $f(x) = -10$, then $f'(x) = 0$

Power rule: $f(x) = x^r$ with r a constant from $\mathbb{R}$

- Derivative: $f'(x) = rx^{r-1}$
- Example: $f(x) = x^{-2}$, then $f'(x) = -2x^{-3} = \frac{-2}{x^3}$

Calculus: Differentiation

Function multiplied by a constant: $f(x) = cg(x)$

- Derivative: $f'(x) = cg'(x)$
- Example: $f(x) = 3x^3$, for $c = 3$ and $g(x) = x^3$,
  $f'(x) = cg'(x) = 3(3x^2) = 9x^2$

Sum of functions: $f(x) = g(x) + h(x)$

- Derivative: $f'(x) = g'(x) + h'(x)$
- Example: $f(x) = x^2 + 4$, for $g(x) = x^2$ and $h(x) = 4$,
  $f'(x) = g'(x) + h'(x) = 2x + 0 = 2x$

Difference of functions: $f(x) = g(x) - h(x)$

- Derivative: $f'(x) = g'(x) - h'(x)$
- Example: $f(x) = x^3 - x^{-2}$, for $g(x) = x^3$ and $h(x) = x^{-2}$,
  $f'(x) = g'(x) - h'(x) = 3x - (-2x^{-3}) = 3x^2 + 2x^-3$

Calculus: Differentiation

Product of two functions: $f(x) = g(x)h(x)$

- Derivative: $f'(x) = g(x)h'(x) + h(x)g'(x)$
- Example: $f(x) = (x^2 - 2x)(x - 2)$, for $g(x) = x^2 - 2x$ and $h(x) = x - 2$,
  $f'(x) = g(x)h'(x) + h(x)g'(x) = (x^2 - 2x)1 + (x - 2)(2x - 2) = x^2 - 2x + 2x^2 - 6x + 4 = 3x^2 - 8x + 4$

Quotient of two functions: $f(x) = \frac{g(x)}{h(x)}$

- Derivative: $f'(x) = \frac{(h(x)g'(x) - g(x)h'(x))}{h(x)^2}$
- Example: $f(x) = \frac{x-2}{x+1}$, for $g(x) = x - 2$ and $h(x) = x + 1$,
  $f'(x) = \frac{h(x)g'(x) - g(x)h'(x)}{h(x)^2} = \frac{(x+1)1 - (x-2)1}{(x+1)^2} = \frac{3}{(x+1)^2}$

Probability Density Functions

- For some continuous random variables, the CDF $F_X(x)$ is differentiable everywhere.
- That is, for a function to be differentiable means that for one real variable a derivative exists at each point in its domain.
- In these cases: the PDF is the derivative of the CDF

$$f_X(x) \triangleq \frac{dF_X(x)}{dx} \tag{7}$$

The PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere).
According to the properties of differentiation, for very small Δx,

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x \tag{8}$$

Probability Density Functions

PDF:

- density of a continuous random variable
- is a function
  value at any given sample (or point) in the sample space (the set of possible values taken by the random variable)
  interpreted as providing a relative likelihood that the value of the random variable would equal that sample
- the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value
- probability given by the integral of this variable's PDF over that range

Expectation

The expectation of $g(X)$: a "weighted average" of the values that $g(x)$ can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$.

- X a discrete random variable with PMF $p_X(x)$ and an arbitrary function $g : \mathbb{R} \to \mathbb{R}$.
- $g(X)$ can be considered a random variable, and we define the expectation or expected value of $g(X)$ as $E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x)$
- If X is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as, $E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx$

As a special case of the above, note that the expectation, $E[X]$ of a random variable itself is found by letting $g(x) = x$; this is also known as the mean of the random variable X.

Expectation

Example:

- six-sided dice
- expected value is the arithmetic mean of the value of a large number of experiments

In Python?

```
 % frame needs [fragile]
import random
random.seed()
count = 1000
sum([random.randrange(1,7) for x in range(count)])/count
```

Variance

For a random variable X: how concentrated the distribution of a random variable X is around its mean.

- Formal definition: the variance of a random variable X:

$$Var[X] \triangleq E[(X - E(X))^2] \tag{9}$$

We can derive the alternate expression:

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned} \tag{10}$$

- the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation

Two Random Variables

Considering more than one quantity during a random experiment.
Experiments with two variables:

- flip a coin ten times:
  - variable 1: $X(\omega)$ = the number of heads that come up
  - variable 2: $Y(\omega)$ = the length of the longest run of consecutive heads

Joint Cumulative Distribution Function

Given two random variables X and Y:

- **Possibility 1:** Consider each of them separately, we need only $F_X(x)$ and $F_Y(y)$
- **Possibility 2:** Considering the values that X and Y take simultaneously during outcomes of a random experiment, we need the **Joint Cumulative Distribution Function** (CDF) of X and Y:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \tag{11}$$

Knowing the Joint CDF, the probability of any event involving X and Y can be calculated.

Joint Cumulative Distribution Function

Joint CDF $F_{XY}(x, y)$ and the **Joint Distribution Functions** $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)dy \tag{12}$$

$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y)dx \tag{13}$$

$F_X(x)$ and $F_Y(y)$ are the **Marginal Cumulative Distribution Functions** of $F_{XY}(x, y)$.

Joint Cumulative Distribution Function

Properties:

- $0 \leq F_{XY}(x, y) \leq 1$
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$
- $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$

Joint and Marginal Probability Mass Functions

If X and Y are discrete random variables, then the **Joint Probability Mass Function** $p_{XY} : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y) \tag{14}$$

We assume, $0 \leq P_{XY}(x, y) \leq 1$ for all x, y, and $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1$.

How does the joint PMF over two variables relate to the probability mass function for each variable separately?

Joint and Marginal Probability Mass Functions

Next Steps

For Probability Theory:

- Maleki & Do: *Review of Probability Theory*
- Read Manning and Schuetze Chapter 2
- Read Deep Learning Book Chapter 3

For Linear Algebra:

- Read Deep Learning Book Chapter 2
- Kolter (and Do): Linear Algebra Review and References