CS 70 - Foundations Of Applied Computer Science

Carter Kruse

Reading Assignment 5

CS 70 - Chapter 9 (Probability)

Overview: Probability theory gives us the mathematical tools to model and deal with uncertainty. Probability can be thought of as the proportion of times an event is likely to occur, or it can be thought of as the degree of belief in a certain event.

Basic Definitions & Axioms

The basic notion in probability is that of a random experiment, which is a repeatable procedure with a fixed set possible outcomes.

Sample Space: The set of all possible outcomes of an experiment is called the sample space and is often denoted Ω.

Event Space: Often we are not interested in a single outcome, but whether some subset of outcomes matches a certain criterion. An event is a subset of the sample space Ω. The events space is the set of all possible subsets of Ω, which we denote $\mathcal{A}$.

Basic Set Operations: Since events are sets, we can apply the usual set operations to them, and reason about the results visually using Venn diagrams.

- The *union* of two events $A \cup B$ is the event that either A or B (or both) occur).
- The *intersection* $A \cap B$ is the event that A and B occur.
- The *complement* $A^C = \Omega \backslash A$ of an event A is the event that A does not occur.
- If A is a subset of B, $A \subseteq B$, then the events in A are contained in B and event A is said to *imply* event B.
- Two events A and B are disjoint when they do not overlap, meaning their intersection is the empty set: $A \cap B = \varnothing$.

Probability Function: To model a random experiment, we must specify the probability of each event in our experiment. A probability measure, $\mathbb{P} : \mathcal{A} \to [0, 1]$ is a function that assigns to each event $A \in \mathcal{A}$ a number between 0 and 1 measuring the probability or degree of belief that the event will occur. A probability measure must further satisfy:

- $\mathbb{P}(\Omega) = 1$
- For any collection of disjoint events, $A_1, A_2, \ldots, A_n$:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

Theorem: Properties Of Probability: For events A and B, we have

- $\mathbb{P}(\varnothing) = 0$
- $\mathbb{P}(A) = 1 - \mathbb{P}\left(A^C\right)$
- If $B \subseteq C$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Probability Space: The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a probability space.

When we are modeling an experiment, it is our responsibility to carefully specify the sample space Ω, the set of events $\mathcal{A}$, and a probability function $\mathbb{P}$ that accurately reflect the nature of the experiment.

Conditional Probability: How does the probability of an event A change when we know something about another event B? Well, given that B occurs, then A will occur only if $A \cap B$ occurs and we are interested in how likely this is relative to just B occurring. The conditional probability of A given that some event B has occurred is written $\mathbb{P}(A|B)$ and defined as:

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

We divide by $\mathbb{P}(B)$ because our sample space is effectively restricted to B when we only consider the cases where B has occurred. Consequently, the conditional probability is only defined if $\mathbb{P}(B) > 0$; we cannot condition on something that never occurs.

Independence: The events A and B are independent if knowledge of B has no effect on A (and vice versa): $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(B|A) = \mathbb{P}(B)$. Inserting this into the definition of conditional probability, we see that for independent events, the probability of their intersection is the product of their individual probabilities:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

Chain/Product Rule:

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

Bayes' Rule: Sometimes we know the conditional probability $\mathbb{P}(A|B)$ but we'd like to know $\mathbb{P}(B|A)$ (or vice versa). Thus, we move between such "inverse" conditional probabilities. By rearranging terms, we obtain a fundamental result from probability theory:

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

Random Variables

Sometimes it is not straightforward or even necessary to precisely define the sample space of an experiment. It may be easier to model the observations of the random experiment and provide a way to reason about it numerically. This is where random variables come in.

Formally, a random variable is a function $X : \Omega \to \mathbb{R}$ that assigns numbers in $\mathbb{R}$ to each outcome in a sample space Ω.

Probability Distribution: A probability distribution defines how likely a random variable is to take on any of its possible states. How we describe a probability distribution depends on whether the variable is continuous or discrete.

Cumulative Distribution Function: Both discrete and continuous random variables can be described by a cumulative distribution function (cdf) $F : \mathbb{R} \to [0, 1]$ which gives the probability that the random variable will be at most a certain value x:

$$F(x) = \mathbb{P}(X \leq x)$$

Theorem: The cdf F of a random variable X satisfies the following properties:

- $0 \leq F(X) \leq 1$ for any $x \in \mathbb{R}$
- F is monotonically increasing; if $x \leq y$, then $F(x) \leq F(y)$

Probability Mass Function Function: Alternatively, we can fully describe the probability distribution of a discrete random variable X by its probability mass function (pmf), which is a function $p : \mathbb{R} \to [0, 1]$ that simply returns the probability that the variable takes on any particular value x:

$$p(x) = \mathbb{P}(X = x)$$

The sum of the pmf over all possible states x is one:

$$\sum_x p(x) = 1$$

Probability Density Function: We can describe the probability distribution of continuous random variables using a probability density function (pdf) rather than a pmf. A pdf is a function $p : \mathbb{R} \to [0, \infty)$ whose anti-derivative is the cdf

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(u)\ du$$

From this definition, the pdf must have a total integral of 1:

$$\int_{\mathbb{R}} p(x)\ dx = 1$$

Conversely, the pdf p is the derivative of the cdf F:

$$p(x) = \frac{d}{dx} F(x)$$

Note that we do not require $p(x) \leq 1$ for a pdf. The values returned by the pdf are *not probabilities* themselves, since they can exceed 1. In fact, the probability that a continuous random variable takes on any particular value x is generally 0. Rather, we can think of the density as determining the probability of a random variable falling within a small ϵ-sized region around x.

More generally, the probability that a continuous random variable falls between two values can be obtained by integrating its pdf between those values.

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\ dx$$

The most common distribution over the real numbers is the normal distribution (Gaussian distribution). It is a continuous, bell-shaped distribution parameterized by its mean μ (which gives the location of its peak and center of mass) and its variance σ^2 (which defines the spread of the distribution about the mean).

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

We write $X\ \mathcal{N}\left(x, \mu, \sigma^2\right)$ to denote that X is normally distributed with mean μ and variance σ^2. If $\mu = 1$ and $\sigma = 1$, the distribution is known as the standard normal distribution.

Multidimensional Random Variables

Joint Distributions: So far we have talked about univariate distributions (those involving a single random variable). Probability distributions can be defined over many variables simultaneously. This is called a joint distribution. For instance, if we have two discrete random variables X and Y, we can write

$$p(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x \cap Y = y)$$

to indicate the probability of X taking on the value x, and Y taking on the value y simultaneously. The joint probability is therefore the probability of the intersection of the events $X = x$ and $Y = y$. If the random variables are continuous, then $p(x, y)$ denotes the joint probability density.

Random Vectors & Multivariate Distributions Another way to interpret the joint distribution is using random vectors. When we have multiple random variables $X_1, X_2, \ldots X_n$, we can either think of them as distinct random variables, each defined over the real numbers $\mathbb{R}$, or we can stack them together and interpret them as the coefficients of a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$.

We can write $p(\boldsymbol{x})$ to refer to the multivariate probability (density) of the random vector $\boldsymbol{X}$ taking on the value $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, which is just the joint distribution of its coefficients: $p(\boldsymbol{x}) = p(x_1, x_2)$.

Marginal Distributions: If we have a joint distribution over two (or more) random variables X and Y, it is possible to obtain a marginal distribution just over one of the variables by "summing out" (or "integrating out") the other variable(s):

$$p(x) = \begin{cases} \sum_y p(x, y) & \text{if } Y \text{ is discrete} \\ \int_{\mathbb{R}} p(x, y)\, dy & \text{if } Y \text{ is continuous} \end{cases}$$

Summary Statistics

While the pmf and pdf completely characterize the distribution of discrete and continuous random variables, it is often useful to summarize this information with just a few numbers. Summary statistics provide a way to achieve this.

Expected Value: Intuitively, the expected value, expectation, or mean of a random variable tells us what value the random variable takes, on average.

More generally, the expected value is an average of the values that the random variable takes on, weighted by the probability (density) of those values occurring:

$$\mathbb{E}[X] = \begin{cases} \sum_i x_i p(x_i) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x p(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

Theorem (Properties Of Expected Values): For any random variables X and Y, and constants α and β, we have

- $\mathbb{E}[\alpha + X] = \alpha + \mathbb{E}[X]$
- $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$
- $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$

In essence, we can pull multiplicative and additive factors out, and the expected value of of a linear combination is the linear combination of the expected values. If the random variables X and Y are independent,

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Variance: Intuitively, variance expresses how spread apart the values are from the center.

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

In other words, it is the average squared distance of X from its mean. Using the linearity of expected values, we can also write this as

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Theorem (Properties Of Variance): For any random variable X and constant α, we have

- $\mathbb{V}[X] \geq 0$
- $\mathbb{V}[\alpha + X] = \mathbb{V}[X]$
- $\mathbb{V}[\alpha X] = \alpha^2 \mathbb{V}[X]$

Basically, multiplicative factors get squared when we pull them out, and additive factors disappear (adding a constant to a random variable shifts all values as well as the mean, so the average squared distance from the mean remains the same).

If two (or more) random variables X and Y are independent, then the variance of their sum is the sum of their variances.

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

Standard Deviation: While variance is a useful statistic for gauging how widely spread a distribution is, it has the downside that its units are the square of the units of the random variable (because of the squaring). We can instead express the spread of a distribution using the standard deviation—typically denoted σ - which is simply the square root of the variance:

$$\sigma[X] = \sqrt{\mathbb{V}[X]}$$

The standard deviation has the same units as the random variable. Also, for standard deviation, multiplicative factors remain the same when we pull them out.

$$\sigma[\alpha X] = \sqrt{\mathbb{V}[\alpha X]} = \sqrt{\alpha^2 \mathbb{V}[X]} = \alpha \sigma[X]$$

Covariance & Correlation: Covariance measures how much two random variables are linearly related to one another, and is defined as

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

where the outer expectation must be taken over the joint distribution of X and Y.

Unlike variance, covariance may be negative or positive. If the covariance is positive then the two variables tend to take high values simultaneously and low values simultaneously. When the covariance is negative, the variables show opposing tendencies: higher values of one variable occur with lower values of the other.

A large absolute value of covariance can mean a strong relationship, but it can also just mean that one (or both) variables have a large expected value. Normalizing the covariance by dividing by the standard deviations of the random variables gives the correlation

$$Corr[X, Y] = \frac{Cov[X, Y]}{\sigma[X]\,\sigma[Y]} \in [-1, 1]$$

which is always between -1 and 1.

If $Cov[X, Y] = 0$ (which also implies that $Corr[X, Y] = 0$, we say that the variables X and Y are uncorrelated, otherwise they are correlated. Covariance and (in)dependence are related, but distinct concepts. If two variables are independent then they are uncorrelated, but the converse is not true in general: two variables can be dependent but still have zero covariance.

Covariance Matrix: See Textbook

CS 70 - Chapter 10 (MLE & MAP Data Fitting)

Overview: In the prior least-squares and data fitting chapters, we posed over-constrained problems as minimizing a quadratic error function. We will now see how to treat the problem of data fitting from a probabilistic/statistical perspective.

In probabilistic data fitting, we are given some data which we assume is generated by some underlying, parameterized probability distribution. We wish to find the parameters of the distribution that best fit this data.

Maximum Likelihood Estimation (MLE): Find the parameters of a probability distribution that best "explains" the data we have observed.

Suppose we have a vector of data values $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. In MLE we assume that these are observations of a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ which has an unknown probability distribution: $p(y_1, y_2, \ldots, y_n|\theta) = p(y|\boldsymbol{\theta})$. The exact form of this distribution is determined by a set of parameters $\boldsymbol{\theta}$ (for instance, maybe we assume that y_1 was drawn from a normal distribution with an unknown mean and variance). Our goal is to determine the parameters that best explain our observations.

$$\hat{\boldsymbol{\theta}}_{MLE} = argmax_\theta \left[\mathcal{L}(\boldsymbol{\theta})\right]$$

where $\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood function telling us how likely we are to observe the data $\boldsymbol{y}$ given that the parameter of the distribution is $\boldsymbol{\theta}$.

The high-dimensional distribution $p(\boldsymbol{y}|\boldsymbol{\theta})$ is difficult to deal with, so we often assume that the random variables are *independently and identically distributed* (iid). This allows us to write

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

Products are tricky, so it is often convenient to now take the log, which turns the product into a sum and gives us the log-likelihood.

$$\log \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{\theta})$$

We can now maximize the log-likelihood

$$\hat{\boldsymbol{\theta}}_{MLE} = argmax_\theta \left[\sum_{i=1}^{n} \log p(y_i|\boldsymbol{\theta})\right]$$

which is equivalent to maximizing the likelihood since the log is a monotonically increasing function.

Maximum a Posteriori (MAP) Estimation: While MLE maximizes the likelihood $p(\boldsymbol{Y}|\boldsymbol{\theta})$, in a *maximum a posteriori estimation* (MAP), we assume the parameters themselves are also a random variable, and we instead maximize the *posterior* $p(\boldsymbol{theta}|\boldsymbol{Y})$. We can use Bayes' theorem to understand how these two conditional distributions are related:

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{p(\boldsymbol{\theta})\, p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y})}$$

Computing the normalization constant (evidence) in the denominator is often intractable since the distribution is very high dimensional, but luckily, since it is a constant, it just scales the result but does not change what value $\boldsymbol{\theta}$ maximizes the posterior. We can therefore maximize

$$\hat{\boldsymbol{\theta}}_{MAP} = argmax_\theta \left[p(\boldsymbol{\theta})\, p(\boldsymbol{Y}|\boldsymbol{\theta})\right]$$

In essence, *MAP estimation allows us to incorporate prior knowledge*, $p(\boldsymbol{\theta})$ about which values of $\boldsymbol{\theta}$ are more probable.

If we assume the observations are iid, then we can again split the high-dimensional $p(\boldsymbol{Y}|\boldsymbol{\theta})$ into the product of 1-D marginals and taking the log gives:

$$\hat{\boldsymbol{\theta}}_{MAP} = argmax_{\theta} \left[\log p(\boldsymbol{\theta}) + \sum_{i=1}^{n} \log p(y_i|\boldsymbol{\theta}) \right]$$