Probability basics

January 15, 2019

The purpose of this handout is to refresh your mind of some of the basics of probability you have covered in the course Foundations of Data Science and provide a quick reference to some basic tool/notions from probability introduced in the first part of this course.

Probability Spaces

A Probability Space is a triple $(\Omega, \Sigma, \mathbf{P})$ where each component is defined as follows:

- The *Sample Space* Ω contains all the possible *outcomes* $\omega_1, \omega_2, \dots$ of the experiment.
- The *Event Space* Σ is the power-set[1] of Ω and contains all *events* which are combinations of outcomes (subsets of Ω).
- The *Probability Measure* $\mathbf{P}$ is a function from Σ to $\mathbb{R}$ satisfying
  (i) $0 \leq \mathbf{P}[\,\mathcal{E}\,] \leq 1$, for all $\mathcal{E} \in \Sigma$
  (ii) $\mathbf{P}[\,\Omega\,] = 1$
  (iii) If $\mathcal{E}_1, \mathcal{E}_2, \dots \in \Sigma$ are pairwise disjoint ($\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ for all $i \neq j$) then

$$\mathbf{P}\left[\bigcup_{i=1}^{\infty} \mathcal{E}_i\right] = \sum_{i=1}^{\infty} \mathbf{P}[\,\mathcal{E}_i\,].$$

We will now provide some examples to hopefully help illuminate these definitions.

Example (Examples of Sample Spaces)**.**

- *Flipping a single coin* $\omega = H$ *or* $\omega = T$ *thus* $\Omega = \{H, T\}$ *or equivalently* $\{0, 1\}$.
  Two coins $\Omega = \{HH, HT, TH, TT\}$ *or equivalently* $\{0, 1\}^2$.
  Any number of coins $\Omega = \{H, T, HH, HT, TH, TT, HHH, \cdots\}$ *or equivalently* $\{0, 1\}^{\mathbb{N}}$.
- *Drawing a card from a deck* $\Omega = \{A\clubsuit, 2\clubsuit, \dots, K\clubsuit, A\diamondsuit, \dots\}$ *or*
  poker starting hands $\Omega = \{\{A\spadesuit, J\heartsuit\}, \{2\clubsuit, 7\diamondsuit\}, \dots\}$.
- *Uniformly selecting a point on a unit circle* $\omega \in [0, 2\pi)$ *and* $\Omega = [0, 2\pi)$.

Example (Examples of Events)**.**

- *Flipping a single coin and nothing happening:* $E = \emptyset \in \Sigma$ *(Coin landing on it's side!?!?).*
  At least one head from two coin flips: $E = \{HH, HT, TH\} \in \Sigma$.
  An even number of heads from any number of coins $E = \{HH, HHTT, HTHT, \dots\}$.
- *Drawing two aces "bullets" in poker* $E = \{\{A\clubsuit, A\spadesuit\}, \{A\heartsuit, A\spadesuit\}, \dots\}$.

[1] We will be working with discrete samples spaces so it serves us take Σ as the powerset of Ω. The general situation is more delicate and Σ only needs to be a sub-set of the power set called a sigma-algebra, this means it must satisfy the following properties:

- $\emptyset \in \Sigma$ and $\Omega \in \Sigma$.
- if $E_1, E_2, \dots \in \Sigma$ then $E_1 \cup E_2 \cup \dots \in \Sigma$.
- if $E \in \Sigma$ then $\Omega \backslash E \in \Sigma$.

- *Let X be a random variable on Ω (see lower down the sheet), $a \in \mathbb{R}$ and $\mathcal{E} = \{X \leq a\} = \{\omega \in \Omega : x(\omega) \leq a\}$. Then $\mathcal{E} \in \Sigma$ is an event.*
- *Uniformly selecting a point from the first or third quadrant of the circle $E = [0, \pi/2) \cup [\pi, 3\pi/2)$.*

Example (Examples of probability measures)**.**

- *Fair coin:* $\mathbf{P}[\, H \,] = 1/2$*, biased coin:* $\mathbf{P}[\, H \,] = \beta \in [0, 1]$.
- $\mathbf{P}[$ *even number of heads from any number of fair coins* $] = 1/2$.
- *If E is drawing two aces in poker then* $\mathbf{P}[\, E \,] = \binom{4}{2}/\binom{52}{2} = \frac{4}{52} \cdot \frac{3}{51}$.
- *The uniform probability measure: Given a finite discrete sample space Ω the uniform measure places weight $1/|\Omega|$ each outcome $\omega \in \Omega$. Thus if $\mathcal{E} \in \Sigma$ then*

$$\mathbf{P}[\, \mathcal{E} \,] = \sum_{\omega \in \mathcal{E}} \mathbf{P}[\, \{\omega\} \,] = \frac{|\mathcal{E}|}{|\Omega|}.$$

This measure arises frequently in discrete probability.

Properties and inequalities for Probability Measures

Events $A_1, A_2, \ldots, A_n \in \Sigma$ are *independent* (with respect to $\mathbf{P}$) if

$$\mathbf{P}[\, A_1 \cap A_2 \cap \cdots \cap A_n \,] = \mathbf{P}[\, A_1 \,]\, \mathbf{P}[\, A_2 \,] \cdots \mathbf{P}[\, A_n \,].$$

The events $A_1, A_2, \ldots, A_n \in \Sigma$ are *pairwise independent* (with respect to $\mathbf{P}$) if for any $i \neq j$, $\mathbf{P}[\, A_i \cap A_j \,] = \mathbf{P}[\, A_i \,]\, \mathbf{P}[\, A_j \,]$.

Two events $A, B \in \Sigma$ are *disjoint* if

$$A \cap B = \emptyset.$$

The Union bound is very useful for bounding the probability of unions of events but it may give a very poor bound if the events have a large overlap.

Theorem (Union Bound/Boole's inequality)**.** *For any events $\mathcal{E}_1, \ldots, \mathcal{E}_n \in \Sigma$ the following holds*

$$\mathbf{P}[\, \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_n \,] \leq \mathbf{P}[\, \mathcal{E}_1 \,] + \cdots + \mathbf{P}[\, \mathcal{E}_n \,]$$

with equality if the events are disjoint.

The following theorem non assessed shows us how to calculate the probability of a union exactly.

Theorem (Inclusion-Exclusion)**.** *For any events $\mathcal{E}_1, \ldots, \mathcal{E}_n \in \Sigma$ the following holds*

$$\begin{aligned}\mathbf{P}[\, \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_n \,] = \mathbf{P}[\, \mathcal{E}_1 \,] + \cdots + \mathbf{P}[\, \mathcal{E}_n \,] - \sum_{1 \leq i_1 < i_2 \leq n} \mathbf{P}[\, \mathcal{E}_{i_1} \cap \mathcal{E}_{i_2} \,] \\ + \sum_{1 \leq i_1 < i_2 < i_3 \leq n} \mathbf{P}[\, \mathcal{E}_{i_1} \cap \mathcal{E}_{i_2} \cap \mathcal{E}_{i_3} \,] - \cdots - (-1)^n \mathbf{P}[\, \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_n \,].\end{aligned}$$

Finally the Bonferroni inequalities non assessed let us interpolate between the crude but easy to apply/calculate union bound and the exact but often costly/painful to compute inclusion-exclusion expression.

Theorem (Bonferroni inequalities)**.** *For any events $\mathcal{E}_1, \ldots, \mathcal{E}_n \in \Sigma$ let*

$$S_k := \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbf{P}[\, \mathcal{E}_{i_1} \cap \cdots \cap \mathcal{E}_{i_k} \,].$$

Then for any odd $k \in \{1, \ldots, n\}$

$$\mathbf{P}[\, \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_n \,] \leq \sum_{i=1}^{k} (-1)^{i-1} S_i.$$

For any even $k \in \{2, \ldots, n\}$

$$\mathbf{P}[\, \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_n \,] \geq \sum_{i=1}^{k} (-1)^{i-1} S_i$$

With equality if $k = n$.

Conditional Probability Measures

For $A \in \Sigma$ of $(\Omega, \Sigma, \mathbf{P})$ define the conditional probability measure $\mathbf{P}[\,\cdot|A\,]$ by

$$\mathbf{P}[\,B|A\,]\,\mathbf{P}[\,A\,] = \mathbf{P}[\,A \cap B\,] \quad \text{for alll } B \in \Sigma.$$

The measure $\mathbf{P}[\,B|A\,]$ is only defined when $\mathbf{P}[\,A\,] > 0$. If A, B are independent then

$$\mathbf{P}[\,A|B\,] = \mathbf{P}[\,A\,].$$

Theorem (Bayes Theorem)**.** *For any events A and B, for which $\mathbf{P}[\,A\,] > 0$ and $\mathbf{P}[\,B\,] > 0$, we have*

$$\mathbf{P}[\,B|A\,] = \frac{\mathbf{P}[\,A|B\,]\,\mathbf{P}[\,B\,]}{\mathbf{P}[\,A\,]}.$$

Theorem (Law of Total Probability)**.** *For any collection of disjoint events B_i such that $\bigcup_{i=0}^{\infty} B_i = \Omega$ and any event $A \in \Sigma$, we have*

$$\mathbf{P}[\,A\,] = \sum_{i=0}^{\infty} \mathbf{P}[\,A|B_i\,]\,\mathbf{P}[\,B_i\,].$$

Random Variables

A *random variable* X on $(\Omega, \Sigma, \mathbf{P})$ is a function $X : \Omega \to \mathbb{R}$ mapping outcomes to real numbers. Random variables are the "observables" in our experiment.

We say the random variables $X_1, \ldots, X_n$ are *independent* if, for any $x_1, \ldots, x_n \in \Omega$, the following holds

$$\mathbf{P}[\,X_1 = x_1, \ldots, X_n = x_n\,] = \mathbf{P}[\,X_1 = x_1\,] \cdots \mathbf{P}[\,X_n = x_n\,].$$

Otherwise we say that the random variables are dependent.

Example (Examples of random variables)**.**

- *The indicator random variable of $\mathcal{E} \in \Sigma$:* $\mathbf{1}_{\mathcal{E}}(\omega) = \begin{cases} 1 & \text{if } \omega \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$
- *Let $\omega = \{a_i\}_{i=0}^{\infty}$ be a sequence of random variables, then $M_n(\omega) = \max_{i \le n} a_i$ is a random variable.*
- *In a round of Texas hold 'em Poker against one opponent the total amounts X_0, X_1, X_2, X_3 which I bet at the four occasions when I have the option/obligation to bet (blind, flop, turn, river) are random variables. This is as they are functions on the state space (card in my hand, opponents hand and on the table). These random variables depend highly on the corresponding random variables of my opponents - their bet amounts Y_0, Y_1, Y_2, Y_4. For example, if my opponent makes a large bet at some stage (say $Y_2 = 100$) and I am holding average cards I may bet nothing ($X_2 = 0$, fold) whereas if they bet a small amount (say $Y_2 = 5$) I also bet the same ($X_2 = 5$). This is an example where although the sequence X_i are random variables (functions of Ω) they are highly dependent on another sequence of random variables (the Y_i) and themselves.*

The function $f(x) = \mathbf{P}[\,X = x\,] = \mathbf{P}[\,\{\omega \in \Omega : X(\omega) = x\}\,]$ is known as the *probability density function* of X and this gives us the distribution of X.

Moments

For a discrete random variable X the *Expectation* $\mathbf{E}[\,X\,]$ is defined as

$$\mathbf{E}[\,X\,] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbf{P}[\,\{\omega\}\,],$$

and this is known as the first moment and we will often use the symbol μ to denote expectation.

For $k \ge 1$ the k^{th} moment $\mathbf{E}\big[\,X^k\,\big]$ is

$$\mathbf{E}\big[\,X^k\,\big] = \sum_{\omega \in \Omega} X(\omega)^k \cdot \mathbf{P}[\,\{\omega\}\,].$$

In the special case where X is non-negative integer valued, for $k \geq 1$ we have

$$\mathbf{E}[\, X^k \,] = \sum_{i=0}^{\infty} i^k \cdot \mathbf{P}[\, X = i \,].$$

Properties of Expectation:

- If $X_1, \ldots, X_n$ are independent then

$$\mathbf{E}[\, X_1 \cdots X_n \,] = \mathbf{E}[\, X_1 \,] \cdots \mathbf{E}[\, X_n \,].$$

- $\mathbf{E}[\,\cdot\,]$ is linear: For any random variables X, Y and $a, b \in \mathbb{R}$

$$\mathbf{E}[\, aX + bY \,] = a\mathbf{E}[\, X \,] + b\mathbf{E}[\, Y \,]$$

- If $X = c$ is a constant (this is still a random variable) then $\mathbf{E}[\, X \,] = c$ and thus for any random variable Y, $\mathbf{E}[\, \mathbf{E}[\, Y \,] \,] = \mathbf{E}[\, Y \,]$.

The *Variance* $\mathbf{Var}\,[\, X \,]$ of a random variable X is the centred second moment of X and is given by

$$\mathbf{Var}\,[\, X \,] = \mathbf{E}\Big[\, (X - \mathbf{E}[\, X \,])^2 \,\Big] = \mathbf{E}[\, X^2 \,] - \mathbf{E}[\, X \,]^2 .$$

We will often use the notation σ^2 for the variance and $\sigma = \sqrt{\mathbf{Var}\,[\, X \,]}$ is known as the standard deviation.

The Covariance $\mathbf{Cov}\,[\, X, Y \,]$ between two random variables X, Y is given by

$$\mathbf{Cov}\,[\, X, Y \,] = \mathbf{E}[\, (X - \mathbf{E}[\, Y \,])\,(Y - \mathbf{E}[\, Y \,]) \,] = \mathbf{E}[\, XY \,] - \mathbf{E}[\, X \,]\,\mathbf{E}[\, Y \,].$$

The covariance gives us an indication of the correlation between two random variables.

Properties of Variance/Covariance:

- The variance is non-negative: for any random variable X, $\mathbf{Var}\,[\, X \,] \geq 0$ with equality if and only if $X = c$ is a constant.
- The variance is shift invariant: for any random variable X and $a \in \mathbb{R}$, $\mathbf{Var}\,[\, X + a \,] = \mathbf{Var}\,[\, X \,]$.
- For any random variable X and $a \in \mathbb{R}$, $\mathbf{Var}\,[\, aX \,] = a^2\mathbf{Var}\,[\, X \,]$.
- For any random variables X, Y and $a, b \in \mathbb{R}$

$$\mathbf{Var}\,[\, aX + bY \,] = a^2\mathbf{Var}\,[\, X \,] + b^2\mathbf{Var}\,[\, Y \,] + 2ab\mathbf{Cov}\,[\, X, Y \,].$$

- If X, Y are independent then $\mathbf{Cov}\,[\, X, Y \,] = \mathbf{E}[\, XY \,] - \mathbf{E}[\, X \,]\,\mathbf{E}[\, Y \,] = 0$ note that $\mathbf{Cov}\,[\, X, Y \,] = 0$ does not imply independence.
- For any random variables X, Y the Cauchy-Schwartz inequality implies that

$$|\mathbf{Cov}\,[\, X, Y \,]\,| \leq \sqrt{\mathbf{Var}\,[\, X \,]\,\mathbf{Var}\,[\, Y \,]}.$$

Example (Examples of moments)**.**

- *For the indicator random variable of $\mathcal{E} \in \Sigma$ we have*

$$\mathbf{E}[\, \mathbf{1}_{\mathcal{E}} \,] = 0 \cdot \mathbf{P}[\, \mathcal{E}^c \,] + 1 \cdot \mathbf{P}[\, \mathcal{E} \,] = \mathbf{P}[\, \mathcal{E} \,].$$

- *If $X_1, \ldots, X_n$ take values in $\{0, 1\}$ and $X = \max_i X_i$ then*

$$\mathbf{E}[\, X \,] = 1 - \mathbf{P}[\, X_1 = 0, \ldots, X_n = 0 \,].$$

- *If X is the value of one fair die roll then*

$$\mathbf{E}[\, X \,] = \sum_{i=1}^{6} \frac{i}{6} = \frac{7}{2}, \qquad \mathbf{E}[\, X^2 \,] = \sum_{i=1}^{6} \frac{i^2}{6} = \frac{91}{6} \quad \textit{and} \quad \mathbf{Var}\,[\, X \,] = \sum_{i=1}^{6} \left(i - \frac{7}{2}\right)^2 \frac{1}{6} = \frac{35}{12}.$$

Probability distributions

Bernoulli (Ber(p)): Think of Ber(p) as being a (biased) coin flip or random bit. It takes values in $\{0, 1\}$ has parameter p and its density is given by

$$\mathbf{P}[\,\mathrm{Ber}(p) = 1\,] = p, \qquad \mathbf{P}[\,\mathrm{Ber}(p) = 0\,] = 1 - p.$$

The expectation is given by $\mu = p$ and variance is $\sigma^2 = p(1-p)$.

Geometric Geo(p): This is the number of p-coins you need to flip before getting heads (or Ber(p) random variables sampled until success). It takes values in $\mathbb{N}_+$ and its density is given by

$$\mathbf{P}[\,\mathrm{Geo}(p) = k\,] = p\,(1-p)^{k-1}\,, \qquad \mathbf{P}[\,\mathrm{Geo}(p) \geq k\,] = (1-p)^{k-1}\,.$$

The expectation is given by $\mu = 1/p$ and variance is $\sigma^2 = (1-p)/p^2$.

Exponential Exp(λ): This is the continuous analogue of the Geometric distribution. It takes values in $[0, \infty)$ and its density is given by

$$f(x) = \lambda e^{-\lambda x}, \qquad \mathbf{P}[\,\mathrm{Exp}(\lambda) \geq k\,] = e^{-\lambda x}.$$

The expectation is given by $\mu = 1/\lambda$ and variance is $\sigma^2 = 1/\lambda^2$.

Will will elaborate slightly more on what we mean when we say the Exponential is the continuous analogue of the Geometric. The probability distribution of some random variable X is *memoryless* if for any two elements s, t in the range of X, we have

$$\mathbf{P}[\,X > t + s \mid X > t\,] = \mathbf{P}[\,X > s\,]\,.$$

To see that Geo(p) is memoryless we have the following by Bayes Theorem

$$\mathbf{P}[\,X > t + s \mid X > t\,] = \frac{\mathbf{P}[\,X > t + s, X > t\,]}{\mathbf{P}[\,X > t\,]} = \frac{\mathbf{P}[\,X > t + s\,]}{\mathbf{P}[\,X > t\,]} = \frac{(1-p)^{t+s}}{(1-p)^t} = (1-p)^s = \mathbf{P}[\,X > s\,]\,.$$

Similarly one can show that $\exp \lambda$ is memoryless. It can also be shown that the only discrete memoryless distribution is the Geometric and the only memoryless continuous distribution is the Exponential.

Binomial Bin(n, p): This is the sum of n independent Bernoulli random variable with parameter p. It takes values in $\{0, \ldots, n\}$ and its density is given by

$$\mathbf{P}[\,\mathrm{Bin}(n,p) = k\,] = \binom{n}{k} p^k\,(1-p)^{n-k}\,.$$

The expectation is given by $\mu = np$ and variance is $\sigma^2 = np(1-p)$.

Poisson Poi(λ): If the time between events is Exp(), then the total number of events in time t is distributed Poi(λt) so this counts "the maximum number of exponential $\exp \lambda$ random variables we can sum and have a total less than t". It takes values in $\mathbb{N}$ and its density is given by

$$\mathbf{P}[\,\mathrm{Poi}(\lambda) = k\,] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

The expectation is given by $\mu = \lambda$ and variance is $\sigma^2 = \lambda$.

Normal/Gaussian If $X \sim \mathcal{N}(\mu, \sigma^2)$ is Normal then X is a continuous random variable taking values in the entire real line, and

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

The expectation is given by μ and the variance is σ^2. If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \rho^2)$ are independent, then for any $a, b \in \mathbb{R}$

- $aX + bY \sim \mathcal{N}(a\mu + b\nu, a^2\sigma^2 + b^2\rho^2)$
- $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.