Math of Data Science: Lecture 4

Vlad Kobzar

APAM, Columbia

Sept 15, 2022

Course progress

- Previously
  - Diagonalization of square/symmetric matrices, projections
  - SVD and best k-rank approximation
- Today - probability basics

Probability - motivation

- Probabilistic and statistical methods will come in two forms:
  - *Probabilisitic models*: data is modeled by some unknown distribution; the problem would entail estimating that distribution
  - *Randomized algorithms* to sample from large datasets/train large parameter models, e.g., stochastic gradient descent for DL

Probability

- Probability measures the likelihood that an event will occur, quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).
- As with every mathematical model, it is not an exact copy of the physical world, but is designed to replicate some aspect of it.
- The simplest model involves phenomena where every outcome is equally likely, e.g. throwing a dice or flipping coins.
- The probability where every even is equally likely amounts to counting. You may have previously seen the theory of counting *combinatorial analysis*.

Sample space

- In probability theory, an *experiment* is any situation which has several possible outcomes, exactly one of which then happens.
- For example, flipping a pair of coins or order of finish in a 7-horse race.
- Probability tells us how to calculate the probabilities of different outcomes in an experiment.
- The *sample space* is the set of all possible outcomes often denoted by Ω. Conversely an *outcome* is always an element of Ω.
- Examples
  - Flipping a pair of coins: $\Omega = \{HH, HT, TH, TT\}$
  - Order of finish in a 7-horse race: $\Omega = \{\text{all orderings of } (1, 2, ..., 7)\}$;

Sample space

- An experiment is something that can happen in the real world.
- a sample space is a choice that we make about how to model that experiment.
- *Example*: Suppose that in a horse race, we care only about who wins, not the order of the other horses. Then we could use

  $$\Omega = \{\text{all orderings of } (1, 2, ..., 7)\}$$

  as before, or we could use

  $$\Omega = \{1, 2, ..., 7\} \text{ (possible numbers of the horse that wins)}$$

- The second is simpler. It is adequate to describe the outcomes only if we don't care who comes in second, third, etc.

Events

- Having chosen the sample space Ω, we will need to discuss different *events* that can occur. Formally specifying an event is equivalent to specifying the outcomes for which that event occurs.
- An *event* S is a subset of Ω.
- Two events are 'the same' if they consist of the same outcomes, even if they are described in two different ways.
  - 7-Horse race:
    $S = \{$all orderings of $(1, 2, ..., 7)$ starting with 3$\} = \{$horse 3 wins$\}$.
  - Flipping a pair of coins: $S = \{HH, HT\} = \{$first coin lands heads$\}$.

Probability space

- Probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of
  - A sample space Ω which contains all possible outcomes of an experiment
  - A set of events $\mathcal{F}$
  - A *probability measure* P that assigns probabilities to events in $\mathcal{F}$.
- The definition of probability measure captures simples ideas:
  - $0 \geq P(S) \leq 1$
  - *countable additivity*: $P(\cup S_i) = \sum_i P(S_i)$ is S_i are mutually exclusive ($P(S_i \cap S_j) = 0$)
  - $P(\emptyset) = 0$ and (unlike other measures, like mass) probability of the entire sample space is $P(\Omega) = 1$.
- There are also some technical requirements on $\mathcal{F}$ (called σ-algebra). We will just say that they formalizes a few simple ideas:
  - if we assign probability to an event then we must assign probability to its complement, and
  - if we assign probability to individual events then we must assign probability to their union, etc.

Random Variables

- For many experiments, we're not interested in every outcome, but rather the feature we really care about can be described by some numerical value
- Given a probability space $(\Omega, \mathcal{F}, P)$, what is a *random variable* X?

Random Variables

- (Informal) Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable X is a function $X : \Omega \to \mathbb{R}$
  - i.e., its value is determined by the outcome of the experiment.
- Then, for $S \subset \mathbb{R}$, can study the sample space given by the image of X.

$$P(X \in S) = P(\omega \in \Omega | X(\omega) \in S)$$

- For a fair coin flip, let X given by

$$X(H) = -1 \text{ and } X(T) = 1$$

- This is sometimes called *Radamacher* random variable

Cumulative distribution function (CDF)

- If X is a RV, then its *cumulative distribution function* (CDF) is the function

  $$F(a) = P\{X \leq a\}$$

  where $a \in \mathbb{R}$.

- Then (by the definition of probability, namely countable additivity)

  $$P\{a < X \leq b\} = F(b) - F(a)$$

Discrete distributions

- Random variables are intuitively simple
- But special complications arise for RVs that can take continuous range of possible values
- Let's start with the simple case
- A random variable X is *discrete* if there is a (finite or infinite) list of real numbers $x_1, x_2, ...$ s.t. X must always take a value from this list, i.e., the set of values is countable.
- Finite sets, and sets of whole and rational numbers are countable.
- The set of all real numbers, or even an interval on the real line, is not countable.

Probability mass function (PMF)

- If X is a discrete RV, then its *probability mass function* (PMF) is the function

$$p(x_i) = P(X = x_i)$$

  where $x_i \in \mathbb{R}$.

- For a set of real numbers A,

$$P(X \in A) = \sum_{\{i | x_i \in A \text{ and } p(x_i) > 0\}} p(x_i)$$

- in particular, the CDF of X is given by

$$F(a) = \sum_{\{i | x_i \leq a\}} p(x_i)$$

- For a discrete RV, if we know PMF, then we can work out the probability of any other event that can be described in terms of X

Example: binomial random variable

- Binom(n, p) counts the number k of "successes" ("heads" or H) in n independent coin flips where H occurs with prob. p.
- Given by the PMF

$$prob(k) = \binom{n}{k} p^k (1-p)^{k-n}$$

- To interpret this, there are

$$\frac{n!}{k!(n-k)!} =: \binom{n}{k}$$

  ways to choose k successes.
  - There are $(n-k)!$ permutations of the failures.
  - There are $k!$ permutations of successes.
  - $p^k(1-p)^{n-k}$ is the probability of getting any particular ordered sequence of k successes from n trials.

Expectation (expected value or mean)

- The idea of mean $E[X]$ is connected to *sample mean* given by

  $$\mu_n = \frac{X_1 + ... + X_n}{n}$$

  where $X_1, ...X_n$ are n random variables distributed identically to X, i.e., they are realizations (outcomes) of n experiments.
  - μ_n is also a random variable (but its particular realizations, i.e., numbers, are also sometimes called sample mean)
- For a discrete RV X with possible values $x_1, x_2, ...$, and PMF p, the *expectation* (or *expected value* or *mean*) is the number

  $$E[X] = \sum_i p(x_i)x_i = \sum_{\{i|\ p(x_i)>0\}} p(x_i)x_i$$

- When X only finitely many values $x_1, ..., x_n$.

  $$E[X] = p(x_1)x_1 + ... + p(x_n)x_n$$

- Since $p(x_1) + ... + p(x_n) = 1$, this is a weighted average of the possible values of X.
- The relation to the sample mean will be addressed on page 16

Why is expectation important?

- As with the sample mean, we can think of $E[X]$ as indicating where the values taken by X 'typically' lie (even though $E[X]$ may not actually equal any of the possible values of X)
- There are plenty of other quantities that can be used this way (such as 'median' and 'mode' in statistics).
- But the expectation has a better theory and more computational tools available, making it more useful to solve problems.
- For example, if the loss function depends on random inputs, its expectation is is a natural choice of the thing to minimize in machine learning problems
- Expectations turn out to be directly connected with long-run averages when we perform an experiment many independent times
- This will follow from the Law of Large Numbers (LLN).

Functions of random variables and linearity of expectation

- Suppose that X is a RV, and also that g is some function from real numbers to real numbers. Then we may define a new RV $g(X)$.
- Often, we know something about X, and want to turn that into information about $g(X)$: most obviously, its expectation.
- If X is discrete RV with possible values $x_1, x_2, ...$ and p is its PMF, then
  - $g(X)$ is discrete with possible values $g(x_1), g(x_2), ...$ (except that this list may contain repeats); and
  - $E[g(X)] = \sum_i p(x_i) g(x_i)$.
- *Linearity of expectation*: if a and b are constants, then

$$E[aX + b] = aE[X] + b$$

- This result generalizes to sums of several random variables
- Thus,

$$E_{X_1,...,X_n}[\mu_n] = E_{X_1,...,X_n} \frac{X_1 + ... + X_n}{n} = E_X[X]$$

i.e., the sample mean is an so-called *unbiased estimator* of the mean.

Variance

- If $X_1, X_2, ..., X_n$ are n random variables distributed identically to X and μ_n is their sample mean, then their *sample variance* is the random variable:

$$S_n^2 = \sum_{i=1}^{n} \frac{(X_i - \mu_n)^2}{n-1}$$

- It measures how 'spread out' the samples are around the mean.
- It is related to *variance* of X, which is a quantity (number)

$$\sigma^2(X) = E(X - E[X])^2$$

- Alternative formula, which can be derived by computing the expectation of a function of random variable:

$$\sigma^2(X) = E[X^2] - (E[X])^2$$

- We have

$$E_{X_1,\dots,X_n}[S_n^2] = \sigma^2$$

i.e., sample variance is an *unbiased estimator* of variance.

Continuous distributions

- Now, we want to model a random quantity, e.g., time when a train arrives
- A random variable X is *continuous* if there is an integrable function (*probability density function*) $p(x)$ on the real line such that

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

- The CDF of X is given by

$$F(a) = \int_{-\infty}^a p(x)dx$$

Example: uniform distribution

Unif(a, b) has the PDF

$$p(x) = \frac{1}{b-a}$$

and the CDF

$$F(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a}(x-a) & a \leq x \leq b \\ 1 & x > b \end{cases}$$

Expectation and variance

For a continuous RV X with pdf

$$E[X] = \int_{\mathbb{R}} p(x) x dx$$

and

$$\sigma^2(X) = E(X - E[X])^2 = \int_{\mathbb{R}} p(x)(x - E[X])^2 dx$$

LLN

- One of our basic intuitions about probability is this: If we perform an experiment independently many times, and E is an event that can happen for each performance of the experiment, then in the long-run average

$$\text{frequency of occurrence of } E \approx P(E).$$

- For instance, if 37% (not a real statistic) of US citizens have visible dandruff, and we randomly select a few thousand citizens (a large number, but much less than US population), then we expect about 37% of those sampled to have visible dandruff.
- So this is saying that, under these long-run average conditions, this 'frequency random variable' settles down, in some approximate sense, to the fixed value P(E).

LLN

- Instead of an event E, assume our basic experiment has a random variable X.
- Independent repeats of the experiment give independent copies of this random variable, say $X_1, X_2, ...$
- In general, a sequence of RVs $X_1, X_2, ...$ are independent and identically distributed ('i.i.d.') if (i) they are independent, and (ii) they all have the same distribution.
- (Weak Law of Large Numbers, 'WLLN'). In the situation above, for any $\epsilon > 0$, we have

$$P(|\mu_n - E[X]| \geq \epsilon) \to 0 \text{ as } n \to \infty$$

where μ_n is the sample mean defined previously

Sample mean

- Assume our basic experiment has a random variable X.
- Independent repeats of the experiment give independent copies of this random variable, say $X_1, X_2, \ldots$.
- For instance, if X is 1 with probability of p and 0 otherwise (Bernoulli(p) trials), then each X_i is the result of the i-th repeat of the flip.
- In the last lecture we defined the sample mean

$$\mu_n = \frac{1}{n} \sum_i X_i$$

- If each X_i is a Bernoulli(p) trial, then μ_n is the fraction of successes among the first n trials, i.e.,

$$\mu_n = \frac{1}{n} \text{binom}(n, p)$$

LLN

- *(Weak) Law of Large Numbers*: If

$$\mu_n = \frac{1}{n} \sum_i X_i$$

  where X_i's are i.i.d., for any $\epsilon > 0$, we have

$$P(|\mu_n - E[X]| \geq \epsilon) \to 0 \text{ as } n \to \infty$$

- A fundamental result that guarantees the convergence of the sample mean (moving or running average) to the mean.
- To justify this result we need to review independence of random variables and certain inequalities.

Inequalities

- So far we have spent a lot of course learning how to compute exactly with random variables.
- But there are also reasons to study estimates and inequalities concerning probabilities and random variables.
    - Sometimes we don't have enough information to compute a probability or expectation exactly
    - so we work out a range of possible values which are permitted given the information we do have.
    - Certain basic inequalities are "responsible" for the Limit Theorems, such the LLN and Central Limit Theorem, which describe the asymptotic behavior of large collections of RVs as the size of the collection tends to ∞.

Inequalities

- The most basic inequality: Let X be a RV such that $X \geq 0$: this means that the value taken by X is always non-negative, for every outcome of the experiment. Then

$$E[X] \geq 0$$

- REASON: $E[X]$s is a weighted average of the values taken by X.
- Immediate consequences:
  1. If $a < b$ are reals such that $a \leq X \leq b$, then

     $$a \leq E[X] \leq b$$

  2. (monotonicity of expectation) if X and Y are two RVs such that $X \geq Y$, then

     $$E[X] \geq E[Y]$$

Markov inequality

- Here is a slightly more subtle consequence of the monotonicity of expectation.
- If X is a non-negative RV, then for any $a > 0$ we have

$$P(X \geq a) \leq \frac{E[X]}{a}$$

- Let $\mathbb{1}_{X \geq a}$ be the indicator variable of the event $X \geq a$
- The inequality $X \geq a \cdot \mathbb{1}_{X \geq a}$ holds because

$$a \cdot \mathbb{1}_{X \geq a} = \begin{cases} a & \text{if } X \geq a \\ 0 & \text{if } X < a \end{cases}$$

- How we use the fact that X is a non-negative?
- Thus, $E[X] \geq aE[\mathbb{1}_{X \geq a}] = aP(X \geq a)$ where the last equality follows from the def'n of expectation
- So Markov inequality gives us an upper estimate on the probability that X takes a value above some threshold. But more often we want to estimate the probability that X takes a value far away from its expectation.

Chebyshev inequality

- If X is any random variable with $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then for any $\kappa > 0$ we have

  $$P\{|X - \mu| \geq \kappa\} \leq \frac{\sigma^2}{\kappa^2}$$

  IDEA: Apply Markov to $|X - \mu|^2$.
- Observe: Markov requires $X \geq 0$, but Chebyshev does not. If we let $\kappa = k\sigma$ for some positive integer k, then Chebyshev becomes

  $$P\{|X - \mu| \geq k\sigma\} \leq \frac{1}{k^2}$$
- 'The probability that X takes a value at least k standard deviations $(= \sigma)$ away from the mean $(= \mu)$ is at most $\frac{1}{k^2}$.'
- This justifies the idea that "the variance/standard deviation indicates how spread out a RV is".

Multivariate random variables

- Previously we've considered a single random variable attributable to an experiment (or a sequence of such random variables)
- We will now consider multiple random variables attributable to an experiment in a sample space Ω and with probability values given by P, let's say a pair X and Y, or a larger collection $X_1, ...X_n$ of potentially different random variables.
  - X and Y can represent different features of an experiment.
  - For example, if you flip two coins, X can represent the maximum of their values and Y can represent a minimum (we denote heads by 0 and tails by 1)

Multivariate random variables

- You can think of (X, Y), or a larger collection $(X_1, ...X_n)$ as a random vector.
- This is one place where probability meets multivariable calculus and linear algebra
- A general event defined in terms of X and Y is

  $$(X, Y) \in A$$

  for $A \subset \mathbb{R}^2$.
- The joint CDF is of X and Y is

  $$F(a, b) = P(X \leq a, Y \leq b)$$

  and for discrete RVs the joint PMF is

  $$p(x_i, y_i) = P(X = x_i, Y = y_i)$$

  and therefore

  $$F(a, b) = \sum_{x_i \leq a, y_i \leq b} p(x_i, y_i)$$

Multivariate random variables (jointly continuous)

- For continuous RVs with PDF $p(x, y)$, the joint CDF is

$$F(a, b) = \int_{x \leq a, y \leq b} p(x, y) dx dy$$

Independence and covariance

- If you flip two fair coins independently, then the probability of every outcome

  $$(0,0),(0,1),(1,0),(1,1)$$

  is $1/4$
- However if you glue them together, facing the same way, then

  $$(0,0),(1,1)$$

  have probability $1/2$ and the other outcomes (0,1),(1,0) have probability zero
- The notions of independence and covariance of random variables capture this distinction

Independence

- X and Y are *independent* if

$$F(a, b) = F_X(X \leq a)F_Y(Y \leq b)$$

- For discrete RVs, this implies that the joint PMF is

$$p(x_i, y_i) = p_X(x_i)p_Y(y_i)$$

  and for jointly continuous the joint PDF

$$p(x, y) = p_X(x)p_Y(y)$$

- So if we're told about two RVs separately and given their individuals PDFs or PMFs, and that they are independent, we can obtain the joint PDF/PMF right away.

Independence

- (Distribution of sum; discrete case) If X and Y are discrete with possible values $x_1, x_2, ..$ and $y_1, y_2, ...$, then

$$P(X+Y=z) = \sum_{i,j:x_i+y_j=z} p(x_i, y_j)$$

  for any real value z.

- If they are independent, the event $X+Y=z$ can be written as a union of disjoint events $(X=k, Y=z-k), 0 \leq k \leq z$, and the summation simplifies

$$\sum_{i,j:x_i+y_j=z} p(x_i, y_j) = \sum_{k=0}^{z} p_X(x_k) p_Y(y_{z-k})$$

- For a sum of independent continuous random variables, this type of an argument leads to the PDF of a sum given by a convolution of the individual PDFs.

Independence

- Let X, Y be independent binom(n, p) and binom(m, p).
- The independence allows us to treat X+Y as the number of successes in $n + m$ trials, which is binom$(n + m, p)$ by definition.

Independence and covariance

- If X and Y are independent, and g and h are any functions from reals to reals, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

- Not true without independence.
- Proof idea: split the 2D integral (or sum if the variables are discrete) into a product of 1D integrals (or sums).
- Now recall: the variance of a RV X is

$$\sigma^2(x) = E[(X - E[X])^2]$$

- It gives a useful measure of how 'spread out' X is.
- We can generalize it to two RVs X and Y.

Covariance

- Let X and Y be RVs and let $m_1 = E[X]$ and $m_2 = E[Y]$. The covariance of X and Y is

$$Cov(X, Y) = \sigma_{12} = E[(X - m_1)(Y - m_2)]$$

  provided that this expectation converges.
- Another notation we'll use (somewhat overloaded for consistency with Strang):
  - $Cov(X, Y) = \sigma_{12}$ for two scalar random variables X and Y or a random vector (X_1, X_2)
  - $Cov(X_i, X_j) = \sigma_{ij}$ for a random vector $X_1, ...X_n$
  - Also $Cov(X_i, Y_j) = \sigma_{ij}$ for a pair of random vectors $X_1, ...X_n$ and $Y_1, ...Y_m$
  - $Var(X) = \sigma^2 = \sigma_{11}$ and $Var(X_i) = \sigma_i^2 = \sigma_{ii}$

Covariance

- Some properties
  1. Symmetry: $\sigma_{12} = \sigma_{21}$
  2. Covariance generalizes variance: for $Var(X) = \sigma^2(X)$,
  $$\sigma_{11} = \sigma^2(X)$$
  3. Like variance, covariance has a useful alternative formula:
  $$\sigma_{12} = E[XY] - E[X]E[Y]$$
  4. So if X and Y are independent, by the result on a previous slide
  $$\sigma_{12} = 0$$

Covariance matirx

- The variance and covariance can be the previous results can be organized into a symmetric matrix

$$V = \begin{bmatrix} \sigma^2(X) & \sigma_{12} \\ \sigma_{12} & \sigma^2(Y) \end{bmatrix}$$

- If X and Y are independent, then $\sigma_{12} = 0$ as note previously
- For the coins glued together, facing the same way

$$\sigma_{12} = E[XY] - E[X]E[Y] = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$
$$\sigma^2(X) = \sigma^2(Y) = E[X^2] - E[X]^2 = \frac{1}{4}$$

Sample covariance

- Just like we can estimate the mean and variance by sample mean and sample covariance, we can estimate covariance matrix using sample covariance matrix.
- Let X be a d-dimensional random vector with mean $\overline{X}$.
- If we repeat the experiment N times, the sample covariance is a sum of rank-1 matrices

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(X_i - \overline{X})^T$$

Other properties of covariance

- What makes covariance really useful is how it transforms under sums and products:

  1. For any RVs X and Y, and any real value a, we have

     $$\sigma(aX, Y) = \sigma(X, aY) = a\sigma(X, Y) = a\sigma_{12}$$

     (so $Var(aX) = Cov(aX, aX) = a^2 Cov(X, X) = a^2 Var(X)$)
  2. For any RVs $X_1, ..., X_n$ and $Y_1, ... Y_m$, we have

     $$\sigma(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma(X_i, Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{ij}$$

     (so it behaves just like multiplying out a product of sums of numbers.)

Other properties of covariance

- In particular, if $m = n$ and $Y_i = X_i$ above, we get

$$\begin{aligned} Var(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j \leq n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} \sigma_i^2 + 2 \sum_{1 \leq i < j \leq n} \sigma_{ij} \end{aligned}$$

- If $X_1, ..., X_n$ are independent, then $\sigma(X_i, X_j) = 0$ whenever $i \neq j$, so in this case, we're left with

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma_i^2$$

Next steps

- Finish probability: LLN and central limit theorem
- Optimization review

References I

[1] Tim Austin, *Theory of Probability, unpublished lecture notes*, 2016

[2] Strang, Linear Algebra and Learning from Data, 2019

[3] Carlos Fernandez-Granda, *DS-GA 1013 / MATH-GA 2821 Mathematical Tools for Data Science, Lecture Notes*, 2020

[4] Carlos Fernandez-Granda, *Probability and Statistics for Data Science, Lecture Notes*, 2017 `https://cims.nyu.edu/~cfgranda/pages/stuff/probability_stats_for_DS.pdf`

[5] Bernstein, *Theory of Probability lecture notes*, http://www.cims.nyu.edu/ brettb/probSum2015/index.html

[6] Ross, *A First Course in Probability* (9th ed., 2014)

[7] Probability. In Wikipedia. Retrieved July 5, 2016, from https://en.wikipedia.org/wiki/Probability