Math Notes

Lessons Learned

Ladislav Šulák <laco.sulak@gmail.com>

April 28, 2021

Contents

1 General

Permutations

- You cannot repeat items that you already used. Equation:

$$\frac{n!}{(n-k)!} = n * n-1 * n-2 * n-3 * ... * n-k+1 \tag{1.1}$$

Combinations

- Order doesn't matter, and you can repeat already used items. Equation:

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{1}{k!} * \frac{n!}{(n-k)!} = \frac{n!}{k!(n-k)!} \tag{1.2}$$

Proof by Induction

- Induction is a way of proving something to be true. It is closely related to recursion.
- **Task**: Prove statement $P(k)$ is true for all $k >= b$.
  - **Base Case**: Prove the statement is true for $P(b)$. This is usually just a matter of plugging in numbers.
  - **Assumption**: Assume the statement is true for $P(n)$.
  - **Inductive Step**: Prove that if the statement is true for $P(n)$, then it's true for $P(n+l)$. This is like dominoes. If the first domino falls, and one domino always knocks over the next one, then all the dominoes must fall.
- **Example**: Let's use this to prove that there are 2^n subsets of an n-element set. Let $s = \{a_1, a_2, ..., a_n\}$ be the n-element set.
  - **Base case**: Prove there are 2^0 subsets of $\{\}$. This is true, since the only subset of $\{\}$ is $\{\}$ itself.
  - **Assume** that there are 2^n subsets of $\{a_1, a_2, ..., a_n\}$.
  - **Prove** that there are 2^{n+1} subsets of $\{a_1, a_2, ..., a_n + a_{n+1}\}$.

    Consider the subsets of $\{a_1, a_2, ..., a_n + a_{n+1}\}$. Exactly half will contain a_{n+1} and half will not. The subsets that do not contain a_{n+1} are just the subsets

of $\{a_1, a_2, ..., a_n\}$. We assumed there are 2^n of those. Since we have the same number of subsets with x as without x, there are 2^n subsets with a_{n+1}. Therefore, we have $2^n + 2^n$ subsets, which is 2^{n+1}.

- Many recursive algorithms can be proved valid with induction.

2 Algebra

2.1 Basics

- Algebraic operations
- Polynomial arithmetic (adding, subtracting, multiplication, division, factorization)
- Complex numbers
- Solving equations and inequalities
- Functions
- Sequences
- Trigonometry (unit circle, the Pythagorean identity, sinusoidal models, etc)

2.2 Linear Algebra

Linear algebra is a mathematical system for manipulating vectors in the spaces described by vectors. Or in another words, Linear algebra is the branch of mathematics concerning linear equations and functions, and their representations through matrices and vector spaces.

- **Vectors**
  - basic operations: $+$, $-$, $*$, and / (result is always vector of the same dimensions)
  - dot product (result of $a.b$ is a single number, not a vector)
  - size of vector ("general" Pythagoras theorem, the length of vector is also called as norm)
  - angle between vectors ($\frac{x.y}{||x||.||y||}$)
  - distance between vectors ($||x-y|| = \sqrt{(x-y, x-y)}$)
  - inner product
  - sub-spaces and the basis for a subspace
  - scalar projection (can be derived from cosine product of orthogonal triangle - if we want to project vector x on vector b, then it is $\frac{b.x}{||b||^2}$)

  - vector projection (result of scalar projection multiplied by the vector itself and divided by the size of the vector, so from the previous example: $\frac{b.x}{||b||^2}b$)
  - linear independence ($b_3 \neq a_1b_1 + a_2b_2$, for any a_1 or a_2 = algebraic understanding; geometric understanding is that b_3 does not lie in the plane spanned by b_1 and b_2.
  - basis (it is a set of vectors that are orthogonal to each other = are linearly independent, so they are not linear combinations of each other; and basis span the space)
  - changing basis of a vector (scalar projection of all axes)

- **Matrices**
  - solving equations with matrices (in an augmented matrix, each row represents one equation in the system and each column represents a variable or the constant terms)
  - basic operations: $+$,$-$, $*$, and / (remember that dimensions cannot be random)
  - determinant (scalar value denoted as $|A|$ that can be computed from the elements of a square matrix and encodes certain properties of the linear transformation described by the matrix)
  - matrix inverses (given 2x2 matrix A, inverse is $\frac{1}{|A|}$ multiplied by matrix that has switched scalars on diagonal and numbers on off-diagonal are multiplied by -1; if determinant is 0, then matrix is not invertible, and this is called a singular matrix)
  - matrix transposition
  - how matrices transform space (if we multiply matrix M with vector N, then the matrix just tells us where the basis vectors go; we can think of for example vector $\begin{bmatrix} 5 \\ 6 \end{bmatrix}$ as $5 * \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 6 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and calculate the original multiplication of $M * N$ with this)
  - matrix transformations (changing basis with a matrix for stretching, rotation, etc, but also combination and composition of matrix transformations; in fact, matrices are a way of representing linear transformations)

- **Alternate coordinate systems (bases)**
  - orthogonal complements
  - orthogonal projections
  - change of basis
  - orthonormal bases and the Gram-Schmidt process (method for orthonormalizing a set of vectors in an inner product space, most commonly the Euclidean space))

- **Eigenvalues and Eigenvectors**
  - They mean "characteristics". If we perform a certain linear transformation to a space some vectors from this space will not be changed at all, some will change their length, and some even their direction. Those vectors which do not change their original direction are eigenvectors, and their length changes are eigenvalues. Be careful, if a vector changes its direction by 180 degrees, that still means that it is eigenvector, but only its direction was reverted, thus its eigenvalue is -1.
  - Imagine a square space, and 3 vectors in that - one horizontal, one vertical, and one diagonal (between them). If you stretch the square vertically, then vertical vector will have bigger size, horizontal vector will not be changed at all, and diagonal vector will point to different direction (and will have different length). Vectors which direction was not changed by a given transformation are called **eigenvectors**. Because horizontal vector was not changed, it is eigenvector, and because the horizontal vector's length was unchanged, we say that it has a corresponding **eigenvalue** of 1 whereas, the vertical eigenvector doubled in length, so we say it has an eigenvalue of 2 (let's imagine that we stretched the square two times).
- **Project high dimensional data into lower dimensional space**
  - **Projection onto k-dimensional sub-spaces** (projection into 1-D is simpler, it is explained here as well). Consider a **n-dimensional vector space** V with the dot product at the inner product and a subspace U of V. With a basis vector $b_1, b_2, ..., b_k$ of U, we obtain the **orthogonal projection** of any vector $x \in V$ onto U via $\pi_U(x) = B\lambda$, where $B = (b_1|b_2|...|b_k) \in R^{nxk}$, and λ is the coordinate of $\pi_U(x)$ with respect to $b_1, b_2, ..., b_k$ of U, and can be calculated, in such multi-dimensional space as $\lambda = (B^TB)^{-1}B^Tx$ (and in 1-D space where we have just 1 basis vector b, as $\lambda = \frac{b^Tx}{b^Tb} = \frac{b^Tx}{||b||^2}$).
  - The projection matrix can be calculated as $P = B(B^TB)^{-1}B^T$ (and in 1-D space, it would be calculated as $P = \frac{bb^T}{b^Tb} = \frac{bb^T}{||b||^2}$), such that $\pi_U(x) = Px$ for all $x \in V$.
  - So the projected vector can be represented as a linear combination of the basis of the subspace, and the vector that connects the data point and its projection must be orthogonal to the subspace.
  - **For example**, let's have a vector $x = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$ and the subspace U spanned by the basis vectors $b_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $b_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$.

- The orthogonal projection was given as $\pi_u(x) = B\lambda$
- Basis B is calculated as a concatenation of all input basis, so $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$
- Lambda λ, which is a vector that contains coordinates of projection point with respect to bases, can be calculated as $\lambda = (B^T B)^{-1} B^T x$.
- So first, we can calculate $B^T x = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$
- Then, $B^T = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$
- Then, $B^T B = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}$ and its inverse is $(B^T B)^{-1} = \begin{bmatrix} 5/6 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$
- So, now we can calculate λ, either with the inverse matrix calculated from the previous step using equation $\lambda = (B^T B)^{-1} B^T x$, or with the following equation:$B^T B\lambda = B^T x$ (here we just eliminated inverse matrix $(B^T B)^{-1}$ so that only $B^T B$ is on the left side) - in either way, the result is $\lambda = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$
- And now, from lambda, we can calculate the projection of x onto space U: $\pi_U(x) = \lambda B = 5b_1 + (-3)b_2 = \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix}$
- The resulting projection matrix in this example is $P = 1/6 \begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix}$ and it can be seen that it is symmetric (projection matrices are always symmetrical).

– **PCA** - an algorithm that minimizes average reconstruction errors by orthogonal projections. It is quite old mathematical method, and it is detailed in my ML notes book.

3 Calculus

3.1 Limits and continuity

3.2 Differentiation

- **Multivariate Calculus** - the words multi-variable and multivariate are typically used interchangeably.
  - When you are differentiating some expression that has multiple variables, and you are differentiating the whole expression just by a single variable, then all other variables are considered to be constants (and constants differentiate to 0).
  - There are different symbols when you differentiate function with one variable (symbol is d), and function with many variables (this is called partial derivative, and the symbol is ∂).
  - Partial differentiation is essentially just taking a multi-dimensional problem and pretending that it's just a standard 1D problem when we consider each variable separately. So partial differentiation as just a simple extension of the single variable method.
  - Total derivative of a function is a sum of all possible partial derivatives (over all variables in a given function, so one partial derivation per variable).
  - For example, given a function $f(x, y, z) = sin(x)e^{yx^2}$
    - Then its total derivative is the following (where each variable x, y, z is some function of parameter t, but that should be known/given):
      $\frac{df(x,y,z)}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt}$
    - All partial derivatives:
      $\frac{\partial f}{\partial x} = cos(x)e^{yz^2}$
      $\frac{\partial f}{\partial y} = sin(x)e^{yz^2}z^2$
      $\frac{\partial f}{\partial z} = sin(x)e^{yz^2}2yz$
- **The sum rule, power rule, product rule,** and **chain rule.**
- **Partial and total derivative**
- **Dependent and independent variables**

  - Imagine a function that calculates a speed of vehicle (y-axis) and time (x-axis).
  - At some particular time, there is always just 1 speed.
  - On the contrary, one particular speed can happen in multiple times.
  - So, time is independent variable, and vehicle speed is dependent variable.

- **The Jacobian**
  - It is simply a row vector where each entry is the partial derivative of a given function $f$with respect to each one of its variables.
  - Once we have such vector, then we can put some values, calculate resulting vector with such numbers, and we can see the direction from that given point. We can do this with many points and we can then construct the whole space (with "regions" - higher / lower, that represents local or global minimums and maximums).
  - So Jacobian describes the gradient of a multi-variable system. And if you calculate it for a scalar valued multi-variable function, you get a row vector pointing up the direction of greater slope, with a length proportional to the local steepness.

- **The Hessian**
  - In many ways, the Hessian can be thought of as a simple extension of the Jacobian vector.
  - For the Jacobian, we collected together all of the first order derivatives of a function into a vector. In the Hessian, we're going to collect all of the second order derivatives together into a matrix.
  - It often makes life easier to find the Jacobian first and then differentiate its terms again to find the Hessian. But the Jacobian is vector, the Hessian is a square matrix. And the Hessian matrix is symmetrical across the leading diagonal, if a function is continuous, meaning that it has no sudden changes.
  - The power of the Hessian is, firstly, that if its determinant is positive, we know we are dealing with either a maximum or a minimum. Also, we can look on the first term, which is sitting at the top left-hand corner of the Hessian. If the number is also positive, we know we've got a minimum. Whereas, if it's negative, we've got a maximum.

- **Taylor Series**
  - Given some complicated function on the input, it is possible to build an approximation to it using a series of simpler functions.
  - But such approximation is only possible, if we know everything about the function at some point - the functions value, its first derivative, second derivative, third derivative, and so on.

  - Then we can use this information to reconstruct the function everywhere else. So, if I know everything about it at one place, I also know everything about it everywhere. However, this is only true for a certain type of function that we call well behaved, which means functions that are continuous and that you can differentiate as many times as you want.
  - By the way, **Maclaurin series** is a Taylor series expansion of a function about 0. So in another words, if the Taylor series is centered at zero, then we are talking about Maclauring series.
  - Maclaurin series says that if you know everything about a function at the point $x = 0$, then you can reconstruct everything about it everywhere. The Taylor series simply acknowledges that there is nothing special about the point $x = 0$. And so says that if you know everything about the function at any point, then you can reconstruct the function anywhere. So a small change, but an important one.

- **Newton Method**
  - Also known as **Newton-Raphson method**, is a way to quickly find a good approximation for the root of a real-valued function $f(x) = 0$. It uses the idea that a continuous and differentiable function can be approximated by a straight line tangent to it.[1]
  - It is iterative algorithm, and we can find a solution by following this equation (until we get a desired accuracy.):

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{3.1}$$

  where $\delta x = -\frac{f(x)}{f'(x)}$ is a step size of this iterative algorithm, and when $f'(x)$ is very small, this step can get big. In fact $f'(x)$ is exactly 0 at turning points of $f(x)$. **This is where Newton-Raphson behaves the worst since the step size is infinite**.
  - It's a really powerful way to solve an equation just by evaluating it and its gradient a few times.

- **Gradient Descent**

- **Lagrange Multipliers**
  - It is a strategy for finding the local maxima and minima of a function subject to equality constraints. So it can be seen as a technique to find a minimum of a function subject to a constraint, i.e. solutions lying on a particular curve (for example within a circle).
  - For example, calculate the minimum of function $f(x, y) = -exp(x - y^2 + xy)$ on the constraint $g(x, y) = cosh(y) + x - 2 = 0$

[1] https://brilliant.org/wiki/newton-raphson-method/

* At first, you should calculate 4 derivatives: $\frac{df}{dx}$, $\frac{df}{dy}$, $\frac{dg}{dx}$, and $\frac{dg}{dy}$.
* Then, calculate roots: $x = \frac{df}{dx} * \lambda\frac{dg}{dx}$, $y = \frac{df}{dy} * \lambda\frac{dg}{dy}$,and $-g(x, y)$ for some initial x, y, and λ.
* From resulting values, you should obtain minima or maxima.

3.3 Integration

4 Statistics and Probability

4.1 General

- A probability is a number that represents the likelihood of an uncertain event (and is between 0 and 1, inclusive).
- In statistics "**population**" refers to the total set of observations that can be made. For example, if we want to calculate average height of humans present on the earth, "population" will be the "total number of people actually present on the Earth".
- A **sample**, on the other hand, is a set of data collected/selected from a pre-defined procedure. For our example above, it will be a small group of people selected randomly from some parts of the Earth.
- When "population" is infinitely large it is improbable to validate any hypothesis by calculating the mean value or test parameters on the entire population. In such cases, a population is assumed to be of some type of a **distribution**.
- **Types of matrix / vector multiplication** are below. We got basically:
  - matrix multiplication (dot / inner product)
  - outer product
  - element-wise multiplication

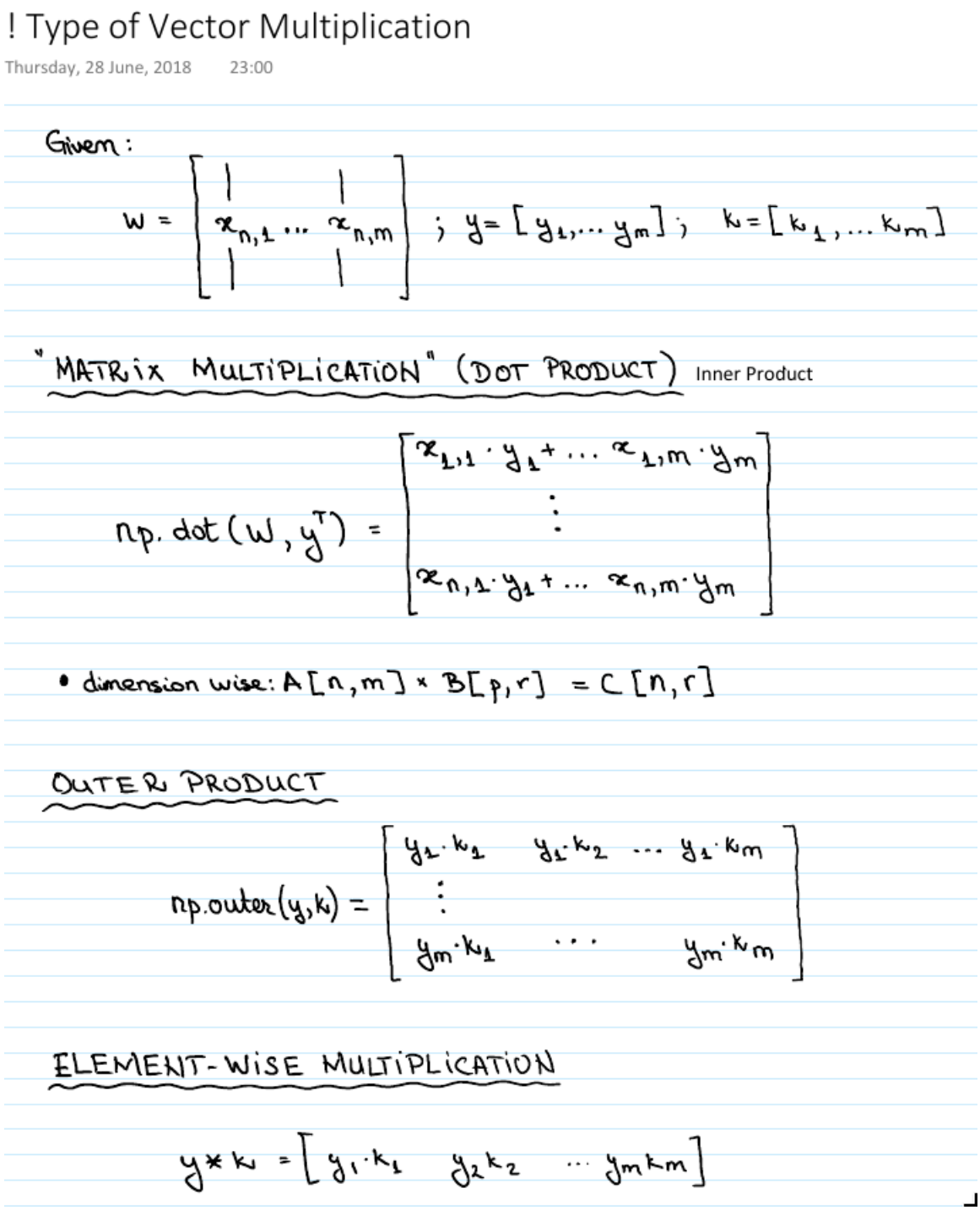

Figure 4.1: Types of vector and matrix multiplication by examples.

- **Variance**
  - It is used to characterize the variability or spread of data points in a dataset.
  - It is a statistic that is used to measure deviation in a probability distribution. Deviation is the tendency of outcomes to differ from the expected value.
  - So we can describe the concentration of data points around the mean value (=expected value) with variance.
  - If we would multiply each sample in a dataset by 2, its variance would be 4 times bigger. If we would just increment each value, variance would be the same.
  - Variance can be calculated as follows:

$$\sigma^2 = Var[X] = E[(X-\mu)^2] = \sum_x (x-\mu)^2 p(x) \tag{4.1}$$

    where X is a numerical discrete random variable with distribution $p(x)$ and expected value $\mu = E(X)$.
  - Note that from the definition, the variance is always non-negative, and if the variance is equal to zero, then the random variable X takes a single constant value, which is its expected value μ.

- **Standard deviation**
  - The standard deviation of a random variable X, denotedσ, is the square root of the variance:

$$\sigma(X) = \sqrt{Var[X]} \tag{4.2}$$

- **Covariance**
  - The covariance generalizes the concept of variance to multiple random variables.
  - Instead of measuring the fluctuation of a single random variable, the covariance measures the fluctuation of two variables with each other.
  - For example, imagine linear decreasing function (a simple line). So if the x value of a data point increases, then on average, the y value decreases. So that x and y are negatively correlated. This correlation can be captured by extending the notion of the variance to what is called the covariance of the data.
  - We can construct **covariance matrix**, in which variances are on the diagonal and cross-covariances on the off-diagonal.

- We can calculate covariance of random variables X and Y as follows:[1]

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{4.3}$$

- When dealing with a large number of random variables X_i it makes sense to consider a **covariance matrix** whose m, n-th entry is $Cov(X_m, X_n)$. Since $Cov(X, Y) = Cov(Y, X)$, the covariance matrix is always symmetric.

- **Expected value**[2]
  - It is the theoretical mean value of a numerical experiment over many repetitions of the experiment.
  - So it is a measure of central tendency; a value for which the results will tend to.
  - When a probability distribution is normal, a plurality of the outcomes will be close to the expected value.
  - **Expected value of discrete random variable**
    - Let X be a discrete random variable. Then the expected value of X, denoted as $E[X]$ or μ, is:

$$E[X] = \mu = \sum_x xP(X = x) \tag{4.4}$$

    - An example: A stack of cards contains one card labeled with 1, two cards labeled with 2, three cards labeled with 3, and four cards labeled with 4. If the stack is shuffled and a card is drawn, what is the expected value of the card drawn?

      Solution: So, there are $1 + 2 + 3 + 4 = 10$ cards. Let X be our random variable that represents the value of the card drawn:

      $P(X = 1) = \frac{1}{10}$
      $P(X = 2) = \frac{2}{10}$
      $P(X = 3) = \frac{3}{10}$
      $P(X = 4) = \frac{4}{10}$

      And this gives us expected value $E[X] = 1 * \frac{1}{10} + 2 * \frac{2}{10} + 3 * \frac{3}{10} + 4 * \frac{4}{10} = \frac{30}{10} = 3$. So the expected value of the card drawn is 3.
  - **Expected value of continuous random variable**

[1] https://brilliant.org/wiki/covariance
[2] https://brilliant.org/wiki/expected-value/

* Let X be a continuous random variable and $f(x)$ be a probability density function. Then the expected value of X, denoted as $E[X]$ or μ, is:

$$E[X] = \int_x xf(x)dx \tag{4.5}$$

* An example: Given the probability density function $f(x) = 3x^2$ defined on the interval $[0, 1]$, what is $E[X]$?

  By the definition above, $E[X] = \int_0^1 x3x^2dx = \int_0^1 3x^3dx = \left[\frac{3}{4}x^4\right]_0^1 = \frac{3}{4}$.

4.2 Central Limit Theorem

- CLT states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
- No matter what the shape of the original (parent) distribution, the sampling distribution of the mean approaches a normal distribution. A normal distribution is approached very quickly as n increases, and note that n is the **sample size for each mean** and not the number of samples. In a sampling distribution of the mean the number of samples is assumed to be infinite.

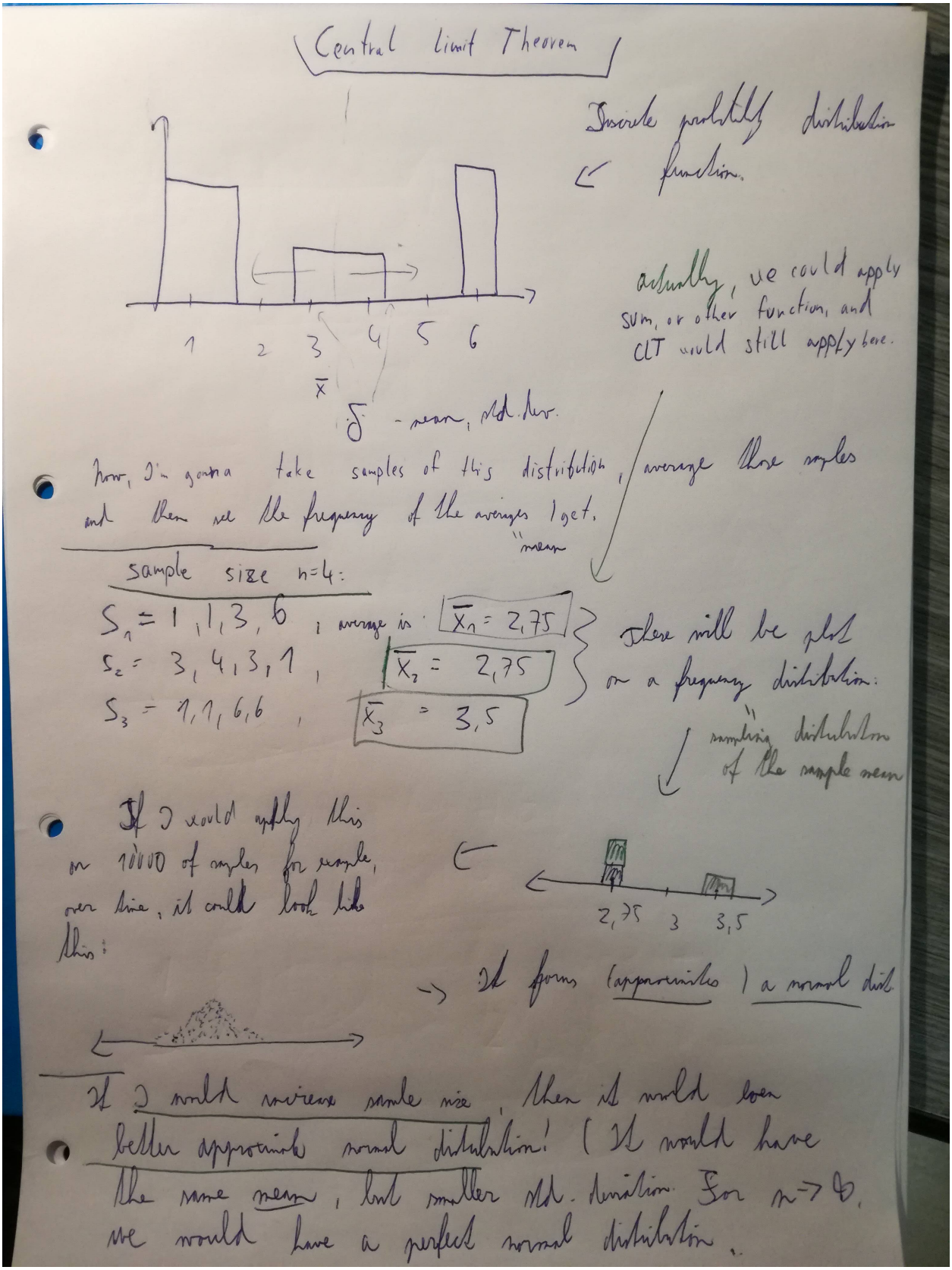

Figure 4.2: Central Limit Theorem explanation via example.

- So the probability distribution of the average of n independent, identically distributed (iid) random variables converges to the normal distribution for large n.[3] In fact, $n = 30$ is typically enough to observe convergence.
- The somewhat surprising strength of the theorem is that (under certain natural conditions) there is essentially no assumption on the probability distribution of the variables themselves; the theorem remains true no matter what the individual probability distributions are.

4.3 Confidence Interval

- In statistics, a confidence interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter. The interval has an associated confidence level, that quantifies the level of confidence that the deterministic parameter is captured by the interval.
- Confidence level is the probability that the value of a parameter falls within a specified range of values.
- More strictly speaking, the confidence level represents the frequency (i.e. the proportion) of possible confidence intervals that contain the true value of the unknown population parameter.
- In other words, if confidence intervals are constructed using a given confidence level from an infinite number of independent sample statistics, the proportion of those intervals that contain the true value of the parameter will be equal to the confidence level.[4]

4.4 Law of Large Numbers

- In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.
- The LLN is important because it guarantees stable long-term results for the averages of some random events.

[3] https://brilliant.org/wiki/normal-distribution/
[4] https://towardsdatascience.com/a-very-friendly-introduction-to-confidence-intervals-9add126e714

4.5 Statistical Significance

- An observed event is considered to be statistically significant when it is highly unlikely that the event happened by random chance.[5]
- More specifically, an observed event is statistically significant when its *p-value* falls below a certain threshold, called the level of significance. Passing this threshold and achieving statistical significance often marks a decision or conclusion to be drawn from the results of a study.
- A **p-value** is the probability that an event will happen that is as extreme as or more extreme than an observed event. This probability also comes with the assumption that extreme events occur with the same relative frequency as they do under normal circumstances. Put more simply, a p-value can be considered to be a measurement of how unusual an observed event is. The lower the p-value, the more unusual the event is. So p-values come from running experiments and comparing the results to what one would expect under normal circumstances.
- A challenge in interpreting data with statistics is that a result can always be attributed to random chance, even a result with an extremely low p-value. Applying a **level of significance** is a way to set a standard for when to stop attributing results to chance.
- A level of significance, denoted by α, is a numerical threshold that is compared to a p-value. When the p-value of an observed event passes below the level of significance, the observed event is considered to be statistically significant. Statistical significance often leads to a decision being made or a conclusion being drawn from the results of an experiment. The most commonly chosen level of significance is $\alpha = 0.05$.
- A smaller level of significance:
  - will ensure a more conservative interpretation of the results
  - is chosen when an incorrect conclusion can be harmful
  - often requires much more collection of data
- **p-values and statistical significance are used in hypothesis tests.** There are a multitude of different types of hypothesis tests, each with a different way to compute the p-value.

4.6 Statistical Tests

- **A null hypothesis**, proposes that no significant difference exists in a set of given observations. You got null and alternative hypothesis (negation of null hypothesis).

[5] https://brilliant.org/wiki/statistical-significance/

For rejecting a null hypothesis, a test statistic is calculated. This test-statistic is then compared with a critical value and if it is found to be greater than the critical value the hypothesis is rejected. To be more precise, the null hypothesis is rejected if the test statistic falls in the critical region. The critical values are the boundaries of the critical region. If the test is one-sided (like a $\chi 2$ test or a one-sided t-test) then there will be just one critical value, but in other cases (like a two-sided t-test) there will be two.

- A critical value is a point (or points) on the scale of the test statistic beyond which we reject the null hypothesis, and, is derived from the level of significance α of the test. Critical value can tell us, what is the probability of two sample means belonging to the same distribution. Higher, the critical value means lower the probability of two samples belonging to same distribution. The general critical value for a two-tailed test is 1.96, which is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean.
- Critical values can be used to do hypothesis testing in following way:
  1. Calculate test statistic
  2. Calculate critical values based on significance level alpha
  3. Compare test statistic with critical values.
- If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis.
- The determination of **distribution type** (e.g Poisson, discrete, binomial) is necessary to determine the critical value and test to be chosen to validate any hypothesis.
- As we know critical value is a point beyond which we reject the null hypothesis. **P-value** on the other hand is defined as the probability to the right of respective statistic (Z, T or chi).
- In **z-test**, the sample is assumed to be normally distributed. A z-score is calculated with population parameters such as "population mean" and "population standard deviation" and is used to validate a hypothesis that the sample drawn belongs to the same population.

$$z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where x is sample mean and μ is population mean. $\frac{\sigma}{\sqrt{n}}$ is population standard deviation.

- **t-test** is used to compare the mean of two given samples. Like a z-test, a t-test also assumes a normal distribution of the sample. A t-test is used when the population parameters (mean and standard deviation) are not known.

$$t = \frac{x_1 - x_2}{\frac{\sigma}{\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_2}}}$$

where x_1,n_1 are mean and size of sample 1 and x_2, n_2 are mean and size of sample 2.

- **ANOVA,** also known as analysis of variance, is used to compare multiple (three or more) samples with a single test.
- **Chi-square test**
  - It is used to compare categorical variables.
  - This refers to a class of statistical tests in which the sampling distribution is a chi-square distribution. Usually, the chi-squared test is used to test for independence between two data sets.[6]
  - The chi-squared statistic is defined by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.6}$$

    where O_i is the number of observations of type i, and E_i is the expected number of observations of type i.
  - Because of this approximation, a number of conditions need to hold in order for the test to be valid. Should they hold, the chi-squared test proceeds as follows:
    1. Calculate the chi-squared statistic χ^2, defined above.
    2. Determine the number of degrees of freedom *df* of the statistic. This depends on the particular expected distribution, but is usually $n - 1$ (where n is the number of categories).
    3. Select a confidence level, usually either 95% or 99%. See Section 4.3 for more information.
    4. Determine the critical value of the χ^2-distribution with *df* degrees of freedom and the confidence level chosen above. Essentially, this is defined as the value x at which the portion of the chi-squared distribution below x is at least the desired confidence level.
    5. Compare the chi-squared statistic to the critical value. If it is below the critical value, the null hypothesis is not rejected. If it is above the critical value, the null hypothesis is rejected, and the expected distribution is probably wrong.

[6] https://brilliant.org/wiki/chi-squared-test

- Intuitively, the test relies on the fact that if the expected distribution is indeed correct, the difference between the observed and expected distributions should approximate a multivariate normal distribution, which is approximated by a chi-squared distribution by the central limit theorem. If the chi-squared statistic is larger than the critical value, then it is unlikely to have occurred under this assumption, and thus the assumption is likely to be false.
- An example: Suppose that after 96 rolls of a die, the die has shown 24x 1s, 15x 2s, 14x 3s, 16x 4s, 14x 5s, and 13x 6s. Is the die unfair? This can be tabulated in the following table:

i	O_i	E_i	$O_i - E_i$	$\frac{(O_i-E_i)^2}{E_i}$
1	24	96/6=16	8	4
2	15	16	-1	0.0625
3	14	16	-2	0.25
4	16	16	0	0
5	14	16	-2	0.25
6	13	16	-3	0.5625

Table 4.1: Chi-square test example

so the chi-squared statistic is $4 + 0.0625 + 0.25 + 0 + 0.25 + 0.5625 = 5.1254$. The number of degrees of freedom is $df = 6 - 1 = 5$, and the chi-squared distribution with 5 degrees of freedom and 95% confidence level has critical value 11.07. Since the chi-squared statistic is less than the critical value, this observation does not provide enough information to reject the null hypothesis of fairness.

4.7 Density Estimation

- It can be said, that this belongs to unsupervised learning. Density estimation is a problem of modeling the **probability density function** (pdf) of unknown probability distribution from which the dataset has been drawn.
- It can be used for many applications, for example for **intrusion detection**.
- We can use **parametric** (for example a multivariate normal distribution - MVN), or **nonparametric model** (for example a kernel regression).
- Let $\{x_i\}_{i=1}^N$ be a one-dimensional dataset (a multi-dimensional case is similar), whose examples were drawn from a distribution with an unknown pdf f with $x_i \in \mathbb{R}$ for all $i = 1, ..., N$. We are interested in modeling the shape of f.

– Now consider using a kernel model of f, denotes as $\hat{f}_b$, which is given by:

$$\hat{f}_b(x) = \frac{1}{Nb}\sum_{i=1}^{N} k(\frac{x - x_i}{b}) \tag{4.7}$$

where b is a hyperparameter that controls the trade-off between bias and variance of our model and k is a kernel function. This can be, for example a Gaussian kernel:

$$k(z) = \frac{1}{\sqrt{2\pi}} exp(\frac{-z^2}{2}) \tag{4.8}$$

– We look for such a value of b that minimizes the difference between the real shape of f and the shape of our model $\hat{f}_b$. A reasonable choice of measure of this difference is called the **mean integrated squared error (MISE)**:

$$MISE(b) = E(\int_{\mathbb{R}} (\hat{f}_b(x) - f(x))^2 dx) \tag{4.9}$$

intuitively, it is the square difference between the real pdf f and our model of it $\hat{f}_b$. The integral $\int_{\mathbb{R}}$ replaces the summation $\sum_{i=1}^{N}$that is implemented in mean squared error, while the expectation operator E replaces the average $\frac{1}{N}$. Because our loss $(\hat{f}_b(x) - f(x))^2$ is a function with a continuous domain, we have to replace the summation with the integral. The expectation operator E means that we want b to be optimal for all possible realizations of our training set $\{x_i\}_{i=1}^{N}$. That is important, because $\hat{f}_b$ is defined on a finite sample of some probability distribution, while the real pdf f is defined on an infinite domain $\mathbb{R}$.

– Now we can rewrite the right-hand side term in equation. 4.9:
$E[\int_{\mathbb{R}} \hat{f}_b(x)^2 dx] - 2E[\int_{\mathbb{R}} \hat{f}_b(x) f(x) dx] + E[\int_{\mathbb{R}} f(x)^2 dx]$
where:

  * the first term: the unbiased estimator is given by $\int_{\mathbb{R}} \hat{f}_b(x)^2 dx$.
  * the second term: the unbiased estimator can be approximated by cross-validation $-\frac{2}{N}\sum_{i=1}^{N} \hat{f}_b^{(i)}(x_i)$, where $\hat{f}_b^{(i)}$ is a kernel model of f computed on our training set with the example x_i excluded. The term $\sum_{i=1}^{N} \hat{f}_b^{(i)}(x_i)$ is known in statistics as the leave one out estimate, a form of cross-validation in which each fold consists just of one example. It can be shown, that the leave one out estimate is a n unbiased estimator of $E(a)$ where $a = \int_{\mathbb{R}} \hat{f}_b(x) f(x) dx$ and a is expected value of the function $\hat{f}_b$, because f is a pdf.

* the third term is independent of b and thus can be ignored.

– Now, to find the optimal value for b, we minimize the cost defined as $\int_{\mathbb{R}} \hat{f}_b(x)^2 dx - \frac{2}{N}\sum_{i=1}^{N} \hat{f}_b^{(i)}(x_i)$ and we can find this value of b using grid search.

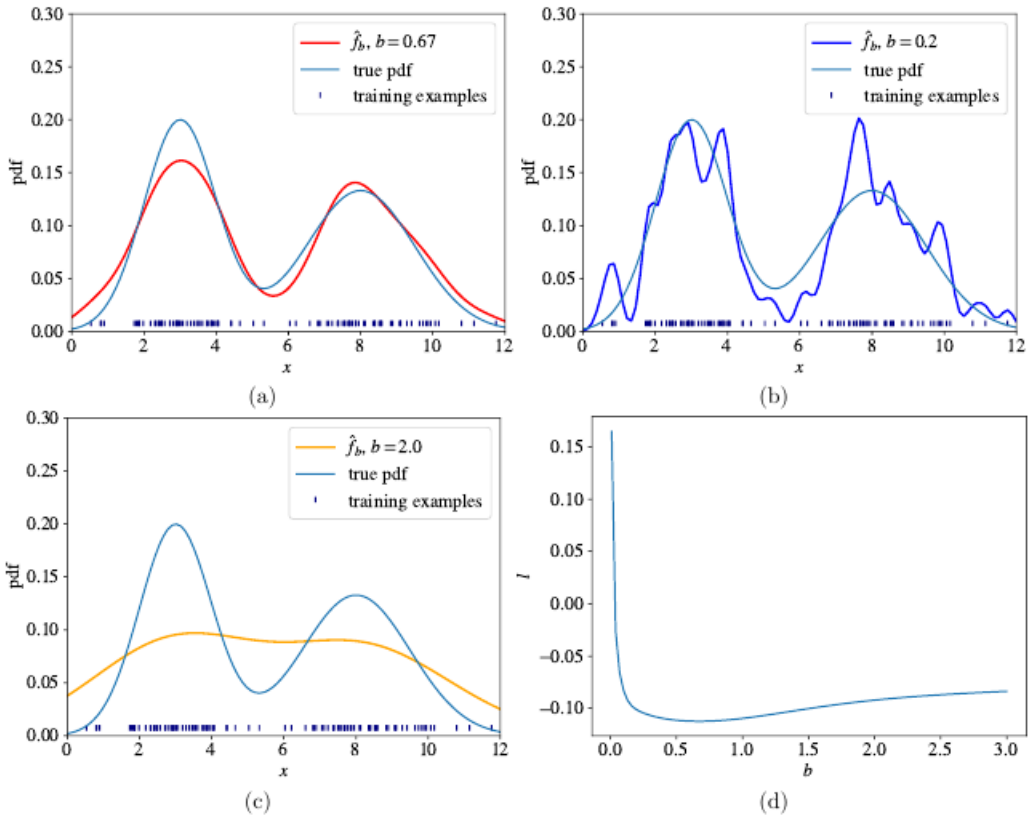

Figure 4.3: Kernel density estimation: (a) good fit; (b) overfitting; (c) underfitting; (d) the curve of grid search for the best value for b.

4.8 Frequentist Probability

- Historically, basic frequency probability theory dominated statistical analysis.
- It is an interpretation of probability; it defines an event's probability as the limit of its relative frequency in many trials. This interpretation supports the statistical needs of experimental scientists; probabilities can be found (in principle) by a repeatable objective process (and are thus ideally devoid of opinion).
- Frequentist probability has been misapplied in the past. Let's have a look on one

example, called *Monty Hall problem*.[7]

- The Monty Hall problem is a famous, seemingly paradoxical problem in conditional probability and reasoning using Bayes's theorem.
- Monty Hall is the host of a game show and gives a contestant the chance to choose 1 of 3 doors without knowing what is behind them. The catch is that one of the doors has a prize like a car, and the other two have goats.
- After the contestant picks a door, Monty then opens one of the doors that the contestant did not pick and reveals that this door has a goat behind it (Monty always knows where the goat is, and opens always door with a goat). Before the final reveal, Monty gives the contestant the chance to switch their choice of door.
- The frequency-probability-guided approach to looking at this choice is to think that because there are now only 2 doors left and 1 of them has a car and the other a goat, the chance of picking right is 50-50 and it doesn't matter if a contestant changes their door.
- This, however, is incorrect, and Bayesian thinking helps to illustrate why.
- A Bayesian probabilist will realize that Monty opening one door is additional evidence provided to the contestant (and this is important!). The Bayesian would realize that the contestant's initial guess had a 1/3 chance of being right, and a 2/3 chance of being wrong. Now that Monty has deliberately (and not randomly!) eliminated 1 wrong door and the 2/3 chance assigns itself to the unchosen and unopened door, staying with their door still has a 1/3 chance of being right, but switching has a 2/3 chance of being right.
- The Monty Hall problem isn't the only place where educated people become confused. Physicians and scientists have mistakenly used frequency probabilities when they should use Bayes' theorem (see Section 4.9) to report results and analyze clinical tests.

4.9 Bayes' Theorem and Conditional Probability

- Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence.
- Given a hypothesis H and evidence E, Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence $P(H)$ and

[7] https://brilliant.org/wiki/bayesian-theory-in-science-and-math/?subtopic=probability-2&chapter=conditional-probability
https://brilliant.org/wiki/monty-hall-problem/

the probability of the hypothesis after getting the evidence $P(H|E)$ is:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{4.10}$$

- This relates the probability of the hypothesis before getting the evidence $P(H)$, to the probability of the hypothesis after getting the evidence, $P(H \mid E)$. For this reason, $P(H)$ is called the prior probability, while $P(H \mid E)$ is called the posterior probability. The factor that relates to $\frac{P(E|H)}{P(E)}$ is called the likelihood ratio. Using these terms, Bayes' theorem can be rephrased as *"the posterior probability equals the prior probability times the likelihood ratio"*.

Deriving Bayes' Theorem

- Bayes' theorem centers on relating different conditional probabilities. A conditional probability is an expression of how probable one event is given that some other event occurred (a fixed value).
- For a joint probability distribution over events A and B, $P(A \cap B)$, the conditional probability of A given B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.11}$$

  For example, a joint probability is *"the probability that your left and right socks are both black"*, whereas a conditional probability is *"the probability that your left sock is black if you know that your right sock is black"*, since adding information alters probability.
- Note that $P(A \cap B)$ is the probability of both A and B occurring, which is the same as the probability of A occurring times the probability that B occurs given that A occurred: $P(B|A) * P(A)$
- Using the same reasoning, $P(A \cap B)$ is also the probability that B occurs times the probability that A occurs given that B occurs: $P(A|B) * P(B)$. The fact that these two expressions are equal leads to Bayes' Theorem.
- This result for dependent events and for Bayes' theorem are both valid when the events are independent. In these instances, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$, so the expressions simplify.

5 References

1. Mathematics for Machine Learning Specialization | Imperial College London (Coursera from 2017)

   https://www.coursera.org/specializations/mathematics-machine-learning

2. Khan Academy

   https://www.khanacademy.org/math

3. Brilliant

   https://brilliant.org/