# Chapter 13 - The Statistics of Machine Learning

## Means, variances, and standard deviations

### 13.1 - Introduction

In an ideal world, our measurements would be perfect and free of noise, and the relationship between our inputs and the outcomes we want to predict would be 1-to-1 mapping. Early cancer research was hopeful this would be true for our genetics. The presence of a specific gene would mean you will develop, provided you lived long enough, and in its absence, you would not. Sadly, things weren't that clear cut, as they seldom are. Yes, we know that specific genes increase the likelihood of certain cancers, like the BRCA1 and BRCA2 genes and breast cancer, but it isn't a given, far from it. This uncertainty isn't limited to cancer. It is pretty much a feature of our universe. First, our measurements will be noisy. Cell counts can be off, family histories incomplete, and biopsy samples can be contaminated. But even with perfect measurements, we can't possibly control everything in our studies, and the systems we study are quite noisy. Therefore, we need statistics in machine learning. This chapter is a quick run-through of some of the most relevant statistical concepts you will encounter in machine learning.

### 13.2 - The mean, variance, standard deviation, and covariance

Four metrics form the backbone of statistics. A good understanding and intuition for these values and what they represent go a long way in understanding the probabilistic nature of almost all machine learning algorithms and data science.

### 13.3 - The mean

The mean or average is a value familiar to most people, even in the absence of any exposure to statistics. But even a value as essential as the mean has some interesting nuances. There are data for which calculating the mean makes relatively little sense. It can be highly uninformative or even misleading. In addition, there isn't just one mean. Depending on the nature of your data and the problem you are using the data to solve, you might need a different definition of the mean. We are familiar with the arithmetic mean, the sum of all elements divided by their count. However, in some instances, the geometric or harmonic mean, calculated entirely differently, provides a better metric of what an average across data points represents. The algorithms we discuss here are all based on the arithmetic mean. Hence, we only define that specific one here, as a complete discussion about the differences between the various means is beyond the scope of this course.

|  |  |  |
| --- | --- | --- |
|  |  | (13.1) |
|  | **Equation 13.1** - The arithmetic mean is defined as the sum of all data points divided by their count |  |

It often surprises people that there is more than one way to define a mean. Similarly, the arithmetic mean becomes misleading, for that isn't distributed symmetrically around some central value. Many processes give rise to exponentially decaying distributions. For example, consider all the words used in the English language. If we look at the relative frequencies of these words, they are not equally likely to occur. That much makes perfect sense. We don't say xylophone all that often. But it is a bit more dramatic than that. If we pick a billion words at random from randomly selected English texts and count how often any of those words occur, we find that for a billion random words, 25% of those words are in the top 15 of the most frequent words. Given that English roughly has 10000 words give or take, it means that 0.15% of all possible words account for approximately 25% of all the words in any randomly selected English text. That is a significant skew. Similar, heavily skewed numbers are found elsewhere, for example, in the size of cities, corporations, and incomes, to name a few. It is a phenomenon called Zipf's law, named after the person, associated with popularizing its broad range of applications. But what does this have to do with the mean? The mean is quite misleading in heavily skewed, long-tailed distributions. When we compute it on such exponentially decaying data, we find that it is usually a value much higher than seems reasonable. This overestimation occurs because a few large data points completely dominate the larger number of relatively small data points. It thus biases the mean towards a mean that isn't representative of most data points, as most data points are relatively small in an exponentially decaying distribution. In this case, an alternative value is more appropriate, called the median. This metric finds the point where 50% of the data will be smaller and 50% will be larger. And in our example, this will be to the right of the mean. The mean had the most points to the right because it was skewed toward the large data points. On the other hand, the median sits right, with an equal amount of data to the left and right. This property makes the median more representative across many data points, not just some.

|  |  |  |
| --- | --- | --- |
|  |  | (13.2) |
|  |  | (13.3) |
|  | **Equations 13.2** and **13.3** - Population (13.2) and sample (13.3) variance as the sum of the squared difference between each data point and the mean across data points, normalized by the number of data points *n* (population variance) or n-1 (sample variance). |  |

### 13.4 - Variance

Whereas the mean represents our data's central tendency, variance provides insight into the width of our data's spread around the mean. It is tempting to say that high variance means higher variability or noise in our measurements. Still, you can only make this statement when comparing it to the variance of a distribution that measures the same quantity in the same units but under different conditions. For example, we could compare the distribution of at what age men were diagnosed with cancer with the same distribution for women. We could find that the mean age for this diagnosis is roughly the same for men and women and that the variance in men is much larger. This observation is an interpretable result since we are comparing the same type of measurement (onset of cancer) in the same units (age). If the men have more variance in their cancer onset age, it means that cancer risk for men is more similar across their lifespan. In contrast, women seem to have a narrower age range during which their risk for cancer is quite significant compared to before and after this range. We compute variance as the summed squared difference of each data point and the mean across all data points. The bigger the overall difference between data points and­­ the mean, the higher the variance. The squaring serves to do two things. First, we don't care if our current data point is smaller or larger than the mean. We are interested in the absolute difference, not whether this is a negative or positive difference. The second effect of squaring the difference is that it grows quadratically as they become more and more distinct from the mean. When we are off by 1 unit between the data point and mean, the variance for that point is still 1 (12). However, if it is off by as much as 10, the variance is 100 (102). This property has some desirable properties that come into play later when we need to fit function to the data points.

### 13.5 - Population versus sample variance

The following is a necessary complicating factor in computing variance. I would bet that of all concepts that students find hard to grasp from an introduction to statistics, as outlined here, is the understanding of why we compute variance in two different ways, yielding two different values. The actual variance across all possible data points is called the population variance. We use the word 'population' as statistics began by mainly being applied to human traits and characteristics. In that perspective, the population variance of BMI is the variance we would observe should we somehow be able to collect data for the entire population. Collecting the data for each person isn't very practical, so we must settle for a measure of variance based on sampling the whole population. We compute the sample variance almost identically, with a tiny difference. The population variance is the summed squared difference between each data point and the mean across all data points, normalized by the number of data points. That is reasonable because that's how you usually compute an average or mean of something. Sum all the measurements and divide by the number of measurements. However, we can mathematically prove that if you do this only for a small sample of those data points (the population), you tend to underestimate the population variance. To correct this, we do not divide the sum by n but by the sum of n-1. I know it sounds a little peculiar. But the math is solid. It is also, unfortunately, beyond the scope of this course. But we can urge you to find some explanations elsewhere.

|  |  |  |
| --- | --- | --- |
|  |  | (13.4) |
|  |  | (13.5) |
|  | **Equations 13.4** and **13.5** - Population (13.4) and sample (13.5) variance as the sum of the squared difference between each data point and the mean across data points, normalized by the number of data points *n* (population variance) or n-1 (sample variance). |  |

### 13.6 - Standard Deviation

Since we squared the difference between each point and the mean, the resulting variance metric is no longer in the same units as the original data. For example, take IQ. We can set the mean IQ across the population to exactly 100 since we do not express it in units of some physical property. That means the scores we assign to different intelligence scores are up to us. So, for ease of understanding, scores computed from IQ tests are adjusted until, for a large enough sample, the mean score is indeed 100. But even without this link to a physical quantity, we still have interpretable units of IQ. We can ask how many people have an IQ higher than 140 or above 160. But we can't do that directly with variance, as we express it in IQ units, but rather the square of IQ units. Therefore, it is very common to take the square root of the variance, reversing the effect of the squaring operation used when we computed it and bringing it back to the spread of observed IQs around the mean expressed in terms of a unit of IQ. The metric is called the standard deviation, and we denote it with the same Greek letter sigma, but this time without the symbol for squaring, for the above reasons.

### 13.7 - Covariance

Whereas variance (and standard deviation) gives us a sense of how data spread around its mean, covariance provides a metric for how two variables (x and y) vary in unison from data point to data point. For each data point (x,y), we compute the difference between x and its mean and y and its mean. Instead of squaring them each, we take the product of (x-x) and (y-y). Here is the intuition, if both points are to the left of their respective means, both differences are negative, so their product is positive.

|  |  |  |
| --- | --- | --- |
|  |  | (13.6) |
|  | **Equation 13.6** - Population (top) and sample (bottom) covariance, an indicator of how much two variables x and y fluctuate in unison across data points. |  |

If they were to the right of the means, both differences would be positive, and their product would be positive. When the points fall on opposite sides of their means, the resulting product will always be negative (because one will always be positive, guaranteeing the other will be negative by definition). So, you can see that if two variables x and y vary in unison, finding themselves mainly on the same side of their means, most values will be positive, and their sum will be highly positive. If they are so an anti-correlated behavior, always finding themselves on opposite sides of their means, the resulting values and therefore sum across values will be highly negative. Finally, if x and y operate independently and don't consistently find themselves on the same or opposite side, the sum of their products will start to average towards 0. Therefore, a significant positive covariance tells us that x and y rise and fall together, a positive correlational relationship. Highly negative covariance still tells us x and y covary quite a bit. Still, they show a negative correlational relationship: if x goes up, chances are y goes down, and vice versa. Finally, small to no covariance means no coupling between x and y in their behavior. Instead, they vary independently.

### 13.8 - Probability Distributions

When we observe a particular data point in our data, we will find some probability of observing that point relative to other points. For example, the probability of someone having an IQ close to the population average is relatively high. But the probability of the IQ is either very low or very high is lower. We assume we can explain these probabilities across different values by some underlying distribution that describes the probability of all possible values. Knowing the type of distribution and being able to parametrize its shape and form helps us 1) understand our data, 2) what models we can and should not use, 3) what we can assume are outliers, and 4) how much predictive power we can expect for a particular variable or combination of variables.

### 13.9 - The Normal Distribution

The most common distribution that describes your data will be the normal or gaussian distribution. It is symmetric and centered around a mean (μ), the first moment or parameter of interest. The second is the width of the distribution, or variance (σ2), which tells us the spread of values we can expect to encounter (Equations 5 and 6). Visual inspection of these empirical (i.e., measured) distributions can reveal problems early on that should be mitigated. For example, some of our data might violate the assumptions of what we think is the underlying type of distribution. Such violations might have serious consequences when training our models.

|  |  |  |
| --- | --- | --- |
|  | Chart, line chart  Description automatically generated |  |
|  | **Figure 13.1** - The normal distribution, as constructed with equation 13.8. The distribution is sometimes also referred to as the bell curve. |  |
|  |  | (13.7) |
|  | **Equation 13.7** - Mathematical shorthand that the data in matrix X comes from a normal distribution parameterized by mean μ and variance σ2 |  |
|  |  |  |
|  |  | (13.8) |
|  | **Equation 13.8** - Probability density function of the normal distribution given by mean μ and variance σ2 |  |

In addition, it allows us to determine the statistical dependencies between variables. Input variables that are highly correlated are redundant to our model and can also introduce serious statistical ambiguities that lead to failures in training our model. Possible violations include asymmetries (skew) or multimodal distributions, while we expect symmetrical and unimodal ones.

|  |  |  |
| --- | --- | --- |
|  | Chart, line chart  Description automatically generated |  |
|  | **Figure 13.2** - The distribution of IQ for a range of x (20-180), and a mean μ = 100 and variance σ2 = 225, standard deviation σ= 15. |  |

### 13.10 - Joint distributions – IQ and temperature

To work through an understanding of what to look for, let us start with a toy data set. Imagine we are interested in predicting someone's IQ from the temperature (in Fahrenheit) on the day they were born. Per definition, we set the mean IQ score to 100, with a variance σ2 = 225, see Figure 13.2. As seen above, we can convert the variance into standard deviation, computed by taking the square root, as in Equations 13.5 and 13.6. A variance σ2 = 225 yields a standard deviation of 15.

|  |  |  |
| --- | --- | --- |
|  | Chart, line chart  Description automatically generated |  |
|  | **Figure 13.3** - The distribution of temperature for a range of x (0-140), and a mean μ = 70 and variance σ2 = 900, standard deviation σ= 30. |  |

These parameters inform us that ~100% of the people have an IQ score between 0 and 200, and 98% percent have an IQ between 70 and 130. And 68% of people have an IQ between 85 and 115. Taking the square root converts the variance into a metric that has the same units as the data. A standard deviation of 15 means 15 IQ points. As for temperature, I set the mean temperature at 70 degrees Fahrenheit, with a variance of 900, which equals a standard deviation of 30 degrees, producing a reasonable distribution for a city like New York City. The distributions in Figures 13.2 and 13.3 represent the probabilities p(temperature) and p(IQ). What if we look at the probability of observing a specific pair of temperature and IQ? How often do we see (temperature = 50° and IQ = 120)? The result is called a joint distribution, and it can reveal critical insights into the dependence or independence of the two variables of interest. When we say dependence, we mean covariance: the probability of observing a value for variable A is partly a function of observing a certain value for variable B, and vice versa. This last statement is important because covariance is agnostic to cause and effect. High covariance between A and B can mean that A causes B, B causes A, both exert an influence on each other, or A and B simply covary or correlative because both are driven by a third process C. Even if we were working with real data, we have no good reason to believe that the temperature at somebody's birthday would correlate with their IQ. We can estimate the joint distribution and plot it as a 2D surface, where the peaks indicate higher probabilities.

|  |  |  |
| --- | --- | --- |
|  | Chart, surface chart  Description automatically generated |  |
|  | **Figure 13.4** - The joint probability distribution is symmetric without any dependence on the temperature on the day someone is born and their IQ. At each point, the joint probability p(IQ,temperature) equals the product of p(IQ) and p(temperature). This equality is telling, as it proves there is no p(IQ) dependence on p(temperature) or vice versa. |  |

When we do so, we observe a single peak at (temperature 70, IQ = 100). This observation should not be surprising because those are the two means of the distributions for each variable in isolation. Furthermore, the peak is symmetrical along both axes. If you were to squash the probabilities of one axis together, you compute a marginal distribution by taking the sum across all slices along an axis. There are two such marginals in our case, one for IQ (x) and one for temperature (y), see Equations 13.9 and 13.10.

|  |  |  |
| --- | --- | --- |
|  |  | (13.9) |
|  |  | (13.10) |
|  | **Equations 13.9** and **13.10** - The marginals p(x) and p(y) of the joint distribution p(x, y) |  |

The marginals (the two curves plotted alongside the surface) look like the distributions for temperature and IQ in isolation (see Figures 13.2 and 13.3). Furthermore, and this is the crucial bit, should we take the outer product of the two marginals, we end up with the same surface as the joint distribution. This relationship is perhaps a little hard to wrap your head around at first, so let us unpack this a bit. The outer product of two vectors *x* and *y* is a matrix where each element is the product of an element in x and an element y, as in Equation 13.11.

|  |  |  |
| --- | --- | --- |
|  |  | (13.11) |
|  | **Equations 13.11** - The outer product of the two marginals p(x) and p(y) creates the joint distribution p(x,y) if p(x) and p(y) are independent. |  |

Suppose we find that the joint distribution empirically differs significantly from the one constructed from the outer product of the two marginals of that empirically established joint distribution. In that case, we have evidence of covariance between p(x) and p(y). If x was some type of input data we were hoping to use to predict some outcome y, this is good news. The intuition is that by taking the outer product between the two marginals, we are essentially creating the joint distribution under the condition that p(x) and p(y) are indeed independent, as we compute each element in joint distribution as the product between some 𝑝(,𝑥-𝑖.) and 𝑝(,𝑦-𝑗.).

### 13.11 - Joint distributions – IQ and maternal IQ

Let's pick another example. Let's look at a data set in which we collected some number of individuals' IQs and their mother's IQs. This time, we suspect a correlation: we know that IQ is at least in part hereditary. In addition, your mother's IQ could also shape your intellectual development through childrearing. Both should yield a positive correlation, if not an outright causal effect. When we generate some example data with an artificially introduced dependence shows an oriented cloud of points: a positive correlation between your IQ and your mother's IQ.

|  |  |
| --- | --- |
| Chart, surface chart  Description automatically generated |  |
| **Figure 13.5** - In the presence of an actual dependence, with clear covariance between p(x) and p(y), the distribution is no longer symmetrical along either axis. Instead, the distribution appears elongated and rotated relative to the axes of p(x) and p(y). A ridge of higher probability cuts across along a line where p(x) = p(y). To put that in words: the most common observation in this joint distribution is those in which a person's IQ is like their mother's. People with a low IQ are more likely to have mothers with a low IQ, people with a high IQ will be more likely to have mothers with a high IQ, and so on. |  |

With sufficient data, we can again estimate the underlying distribution. Notice that the contour lines appear rotated 45 degrees this time. This rotation reflects a person's IQ's covariance or statistical dependence on their mother's IQ. Given the correlation, we expect that observing an IQ of 140 in someone has a probability different than observing an IQ of 140 in someone whose mother also had an IQ of 140. In probability theory, we expressed this as . You should read as follows: two events have a probability of happening. However, the probability of observing both is not the same as the probability of them occurring in isolation. And this is indeed what we observe.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (A)  Chart, surface chart  Description automatically generated | (B)Chart, surface chart  Description automatically generated | (C)Chart, engineering drawing  Description automatically generated with medium confidence |  |
|  | **Figure 13.6** - (A) An empirical joint distribution created by generating random samples from two independent normal distributions. (B) The joint probability distribution is created from the marginals obtained from (A). (C) The difference between (A) and (B). Since (A) was generated from two independent variables p(x) and p(y), the resulting joint distribution p(x,y) equals p(x)p(y). Computing the joint distribution from the marginals yields the same result, as it computes the joint distribution as the outer product of p(x) and p(y), which will yield the same distribution as when p(x) and p(y) are independent. Therefore, their difference is 0 at all pairs p(xi,yi). | | |  |

The probability of observing someone with an IQ of 140 knowing that their mother has the same IQ is greater than observing someone with an IQ of 140 in general. Notice, however, that the marginals look perfectly symmetrical around their mean (IQ=100). If we were to use these marginals to create the joint distribution where p(x,y) = p(x)\*p(y), the result would be symmetrical, as well as if there was no dependence between p(x) and p(y). Any significant difference between the empirical joint distribution and the one created from the marginals will tell us that there is covariance between p(x) and p(y). We cannot claim p(x) influences p(y), p(y) influences p(x) or some unknown p(z) influences both (correlation does not imply causation). All we can say is that they show significant covariance.

### 13.12 - Dependencies

Dependence or covariance between data is both a blessing and a curse. It is a curse if you observe it between the columns of your input data. On the other hand, dependency is a blessing when we see it happen between input and output. Because this time, the dependence indicates the predictive power of the input in explaining the outcome. If two variables correlate, how do we decide which one is predictive of the output?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (A)  Chart, surface chart  Description automatically generated | (B) Chart, surface chart  Description automatically generated | (C) Chart, surface chart  Description automatically generated |  |
|  | **Figure 13.7** - (A) An empirical joint distribution created by generating random samples from two normal distributions that covary (x ~ y). (B) We create the joint probability distribution from the marginals obtained from (A). (C) The difference between (A) and (B). Since (A) was generated from two variables p(x) and p(y), that exhibit a dependency, the resulting joint distribution p(x,y) does not equal p(x)p(y). Computing the joint distribution from the marginals yields the same result as before (Figure 6B), as it again computes the joint distribution as the outer product of p(x) and p(y), which will yield the same distribution as if p(x) and p(y) are independent. Now, the empirical joint probability distribution is significantly different than what we would expect if p(x) and p(y) were independent, as is evident when we plot the difference between (A) and (B) in (C) | | |  |

### 13.13 - Other statistical distributions

Visualizing the empirical distributions of whatever data set you work with is never a bad idea. Distributions capture the probability of the different values our variables take on. Specifically, the normal distribution is valid for real-valued and continuous observations. We typically use binomial and multinomial distributions for categorical observations and probabilities. Categorical values have no meaningful mean and variation; probabilities have explicit bounds (0 and 1) and additive properties that violate the normal distribution. Machine learning models typically rely on estimating the parameters of the underlying distributions, like the mean, standard deviation, kurtosis, and skew. If our data violates these underlying assumptions (e.g., the input data is not normally distributed with a mean and variance), at best our model will fail to produce any predictive power. At worst, its output is misleading and biased, which can have serious consequences if we use the model for actual decision-making. In those situations, understanding what process gave rise to your data can guide you in picking a more appropriate probability distribution, like the exponential, Bernoulli, multinomial, binomial, gamma, beta, and Poisson, to name but a few.

### 13.14 - Conclusions

Understanding what kind of distribution might be underlying your data is worth the time and effort. The normal distribution can accurately describe not everything in the world that varies continuously. Our data could be skewed, making it non-symmetrical around its mean. Our data might contain multiple peaks of increased probability, making it multimodal rather than the unimodal nature of the normal distribution. And some data we observe in the world might be characterized more closely by a slew of other distributions

### 13.15 – Demo Notebooks

Notebook 013.00 has example code that walks through how to compute means, variances, standard deviations, and covariances in python. This code is for educational purposes only, as most numerical packages will have functions to compute these metrics quickly and efficiently. I urge you to use those in any kind of actual project. Notebook 013.02 mainly generates the data and the figures for the joint distributions discussed in sections 13.3 to 13.7.

### 13.16 - References

1. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
2. Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (4th ed.). WW Norton.
3. Rieke, F., etc., Warland, D., van Steveninck, R. de R., & Bialek, W. (1997). Spikes: Exploring the neural code. MIT Press.