Definitions and Inequalities

Basically probability ideas today, because that is a part of the subject, part of deep learning as we get there. You will have seen the sample mean, the average of the data. And you will know about the expected mean.

What's the expected mean? So this is the expectation of the value x

$$E[x] = x_1p_1 + x_2p_2 + \cdots + x_np_n$$

The symbol E is like everywhere. It gives a handy shorthand. For example, the variance is the expected value of $(x - E[x])^2$

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

$m = E[x]$, and in general, of course, this covariance matrix we could express with that E natation. But let us just stetch it as far as. What would be the expected value of any function of x?

$$E[f(x)] = p_1f(x_1) + \cdots + p_nf(x_n)$$

We weight those by the probabilities that they happen. And if that $f(x) = (x - m)^2$, then we get what we expect. So we just want to keep going with variance.

$$Variance = p_1(x_1 - m)^2 + \cdots + p_n(x_n - m)^2$$

And everybody should know a second expression, a second way to do that the sum. If we just write out those squares and combine them a little differently, we get a second expression which is really useful, often a little faster to compute.

$$Var = p_1({x_1}^2 - 2x_1m + m^2) + \cdots + p_n({x_n}^2 - 2x_nm + m^2)$$

We take that sum. So we get

$$Var = p_1{x_1}^2 + \cdots + p_n{x_n}^2 - 2(p_1x_1m + \cdots + p_nx_nm) + m^2$$

$$Var = E[x^2] - 2m^2 + m^2 = E[x^2] - (E[x])^2$$

So you have the anther handy way to compute the variance. OK, let's go a little deeper with something here.

There are two great inequalities in statistics. And the first one is due to Markov -- Markov's inequality. It comes out easily, in fact, too easily.

Markov inequality, Markov was a great Russian mathematician, about 1900. And we will see Markov chains and Markov processes. It applies to non-negative events -- applies when all $x_i \geq 0$. So it doesn't apply to something like Gaussian, because there, the Gaussian, the outputs go all the way from $-\infty$ to $+\infty$. It does apply to a lot of important ones and simple ones. We will see a proof for the finite probability and there will be a similar proof, similar discussion every where here for continuous probability.

It's natural to want to estimate the probability that

$$P[x \geq a]$$

Gets some idea of what' s the probability of $x \geq a$. This is certainly a number between 0 and 1. That number is going to ger smaller as a increases, because wea re going to be asking for more. If I take a to b, say twice the mean, can we estimate what that probability could be. And that's what Markov has done. He says the probability of that is at most to the mean divided by a

$$P[x \geq a] \leq \frac{E[x]}{a}$$

And as we expect, as a increases, the probability goes down. So that is a pretty simple estimate to get this probability just in terms of the number a, which has to come in, because it's part of the question, and the mean. Let's see an example as $a = 3$.

We want to show that the probability of x being greater or equal to 3. We don't have many facts to work with. So if we write those down, we should see the reason. So we know that the mean is $E[x]$. And we are going to take $E[x] = 1$. And we are asking for what is the chance that x will be bigger than 3. And we will get an estimate of $1/3$.

$$\text{Markov Inequality}$$

$$P[x \geq 3] \leq \frac{1}{3}$$

If we write down what we know, we will see that. We know the definition of the mean. We know that $x_1 p_1 + \cdots + x_n p_n = 1$, and what's it that we want to prove? We want to know the probability if being greater of equal 3. So what's the probability that the result will be ≥ 3 ?. It's $p_3 + p_4 + \cdots + p_n$, and we want to show it is $\leq 1/3$. What we liked about this elementary approach is that we have stated these facts.

If we take a special case that if we take $x_1 = 1, x_2 = 2, x_3 = 3, \cdots, x_n = n$. So that satisfies my condition that the $x_i \geq 0$. And we know $p_{sum} = 1$, and we also know $p_i \geq 0$. My idea is substitute p_3 here using $p_3 + p_4 + \cdots + p_n$ and we get

$$p_1 + 2p_2 + 3(p_3 + p_4 + \cdots + p_n) + p_4 + 2p_5 + 3p_6 + \cdots + (n-3)a_n = 1$$

What is Markov telling me about that $p_3 + p_4 + \cdots + p_n$? Less or equal to $1/3$. We want to prove $3(p_3 + p_4 + \cdots + p_n) \leq 1$. But suppose it was greater than 1. Do you see the problem? All the other number are ≥ 0. And the total things adds to 1. So that piece has to be ≤ 1. That's it. So a lot of talking there. Simple idea. You will see a more conventional proof in the Prof. Gilbert Strang's notes. But they are somehow more mysterious.

Chebyshev is the other great Russian probabilist of the time. And he gets his inequality.

So Chebyshev was interested in the probability that $(x - m)$ is greater equal than a -- the probability of being sort of a distance away from mean. So again, as a increases, we are asking more. And the probability will drop. And the question is can we estimate this? So this is a different estimate. But it's similar question. And what Chebyshev's answer for this?

$$\text{Chebyshev Inequality}$$

$$(P[(x - m)] \geq a) \leq \frac{\sigma^2}{a^2}$$

So that's Chebyshev. And we just take time to do these two in this lecture. Because they involve analysis. They are basic tools. They are sort of the first thing you think of if you are trying to estimate a probability. Does it fit Markov? And Markov only applies when $x_i \geq 0$. In Chebyshev we are not concerned about the size of x_i. And we are taking a distance from m. So we are obviously in the world of variances. We are distances from m. And the proof of $Chebyshev$ comes directly from $Markov$. So we are going to apply Markov to y. this will be a new output. And it will be $y_i = |x_i - m|^2$.

In order to apply Markov, we need to figure out what is the mean of y.

$$E[y] = \sum p_i y_i = \sum p_i (x_i - m)^2 = \sigma^2$$

You are supposed to recognize it. Do you see that now Chebyshev is looking like Markov?

So those are two basic inequalities. Now, the other topic that we wanted to deal with was covariance, covariance matrix. You have to get comfortable with what's the covariance.

Covariance matrix will be $m \times m$ when we have m experiments at once. Let's take $m = 2$. You will see everything for $m = 2$. So we are expecting to get 2×2 matrix. And what are we staring with? We start we are doing two experiments at once. So we have two outputs, x_i, y_i. For example we are flipping two coins.

coin1	**x= 0 or 1**	**p=1/2**
coin2	**y= 0 or 1**	p=1/2

There is a connection between the output, if we glue the coins together, then the two outputs are the same. This is a model question that brings out the main point of covariance. If we flipped two coins separately, quite independently, then we don't know more about each other. But is the two coins are glued together, then head will come up for both coins, we only have two possibilities. It will be heads heads / tails tails. Let's write down those two different scenarios.

A tensor if a three-way structure.

Covariance matrix -- V.

P_{ij} is the probability that x is x_i and y is y_j both happen. Let's see a example to keep in mind.

Suppose we are looking at age and height. x is the age of the sample, the person. And y is the height. We want to know so what fraction have a certain age and a certain height.

What is the meaning of $\sum_{i=1}^{n} P_{ij}$? The answer is $\sum_{i=1}^{n} P_{ij} = P_j$. Those would be called the marginals of the joint probability. So we can complete the definition of covariance matrix.

$$V = \sum_{x_i, y_j} P_{ij} \begin{bmatrix} x_i - m_x \\ y_j - m_y \end{bmatrix} \begin{bmatrix} x_i - m_x & y_j - m_y \end{bmatrix}$$

this the two experiments. A 2×2 covariance matrix. And let's see what would be the $1, 1$ entry in that matrix. We get the standard variance of the x experiment.

$$V = \begin{bmatrix} {\sigma_x}^2 & \sigma_{xy} \\ \sigma_{xy} & {\sigma_y}^2 \end{bmatrix}$$

in the independent coins, unglued coins,$\sigma_{xy} = 0$. So you have a diagonal matrix. But if the two coins are glued together, then the matrix will be singular.

$${\sigma_{xy}}^2 = {\sigma_x}^2 {\sigma_y}^2$$

So, this matrix is positive semidefinite always. Because it's column times row, we know that's positive semidefinite. And it's multiplied by numbers greater or equal 0 -- P_{ij}. So it's a combination of rank 1 positive semidefinite. So it's positive semidefinite or positive definite.

In between, coins that were partly glued, partly independent, but not completely independent experiments, then σ_{xy} would be small and smaller than the diagonal elements.