1: hello everyone, today we will talk about two important terms in machine learning, which are “design matrix” and “response vector”. Our background story will be the “Cockcroft-Gault formula”, which is a formula in nephrology that is widely used.

2: This formula was invented by the Canadian physicians Dr. Donald W. Cockcroft and Dr. Henry M. Gault, and was published in 1976 in an article named “prediction of creatinine clearance from serum creatinine”. It helps us in estimating the rate of urinary excretion of a plasma protein called creatinine, which is important in evaluating kidney function. The formula does so by using creatinine concentration in the plasma and the patient’s age. It is useful, since measuring creatinine urinary excretion rate directly requires a 24-hrs urine collection while measuring it’s concentration in the plasma only requires a single blood-test.

3: some disclosures before we begin.

The first one is a mathematical assumption I made, and it’s that throughout the video we will assume that the formula predicts perfectly the rate of urinary excretion of creatinine. It means that the value predicted of the formula is identical to the one we would get if we measured its value directly. This is how I generated the data that we will later use.

The second disclosure that I have to make is that the thinking process of Dr. Cockcroft and Dr. Gault might have been different than the one presented here. However, pretending it actually was like that would be fun, so I decided we could allow ourselves to imagine it was so anyway.

4: Now we could start to describe the experiment. In their experiment they collected data about 249 patients, and the data that they chose to collect about each patient was age, serum creatinine concentration and urinary creatinine excretion. Each of these properties had a dedicated column in a table; in which each row represented a patient.

Then, they arranged this data in a mathematical structure called a matrix. A matrix is simply a table of numbers, and in the context of machine learning the matrix of the sampled data is called **a design matrix**. In this matrix, each row represents a patient, and each column represents a feature, which mean a property that we use to make a prediction.  
The last property, which is the one that they wanted to later “predict”, was represented as a sequence of numbers which is called **a response vector**.

Since all data collected is real numbers, and there are m samples and d features, we say that the design belongs to the field of real numbers of dimensions m times d. similarly, the response vector belongs to the field of real numbers of dimension m.

The next thing they did, being all curious about the relationship of these three magnitudes, is to observe them in a plot. So let’s do it too!

vs code 1: Here we can get an impression about the mentioned relationship. In this plot, each sample is represented as a point, and each axis represents a different feature, or the response value. This type of plot is called a **scatter plot**, since it represents the samples as scattered dots in space. We see that urinary creatinine excretion decreases with age and increases with serum creatinine concentration. In order to predict urinary creatinine excretion rate they had to find an equation that describes this relationship. However, this function seems to be curved, so it would be hard to do. At this point they must have had quite a headache – how could they possibly “flatten” this curve?

5: Given the sampled data, as described here, the equation of a fitting plane is of the form coefficient1 (or weight) times serum creatinine concentration plus coefficient2 times age. So, they came up with an elegant idea– let’s create new features based on the old ones. and so they did.

The first new feature is the inverse of the second old feature. The second new feature is the ratio between the first old feature and the second old feature. These are entirely new features; however, they don’t require sampling new data, which is excellent.

6: Accordingly, they got now a new design matrix. Again, each column represents a feature and each row represents a sample, and the matrix is in R m times d. The response vector stays the same.

vs code 2 without plane: They must have been really happy when re-observing the data, they saw it lied on a single plain. Here’s the plane: vs code 2 with plane.

7: the existence of such a plane means that the given equation can be solved. Using a mathematical method called “singular value decomposition”, which is very interesting in itself; we find that the coefficients that describe the plane are as follows. Finally, replacing w1 and w2 with the corresponding values gave the well-known formula.

. Remember that I told you we assumed that the formula predicts perfectly the rate of creatinine urinary excretion? That’s because, as I mentioned earlier, the data was \*drawn from\* a distribution based on that equation, so obviously the equation describes it.

That’s all for today people. Thank you for watching, and you are welcome to like the video or leave a comment. bye!