Capturing All Possibilities in d Dimensions

Suppose you have two binary features (i.e. your feature can only be one of two values, such as "0" or "1", "yes" or "no"). Now, what is the size of your feature space or equivalently, how many different combinations of features are possible?

When you think you know the answer, click the image to reveal the solution

Click to reveal the answer

Answer: $2^2 = 4$

Now consider instead of two binary features, imagine you have twenty binary features. What is the size of your feature space?

Click to reveal the answer

Answer: $2^{20} = 111{,}048{,}576$

With only twenty features, you already have a large input space, as it grows exponentially with the number of dimensions. In real data sets, we often have more than just twenty binary features (often we have thousands of features) and collecting all the data in the feature space is too expensive and not feasible. Soon you could obtain more possible inputs than there are electrons in the entire universe — certainly at this stage you would rarely (effectively never) observe the same input more than once.

So how do you overcome the curse of dimensionality? How can we apply this methodology if it fails to work on common data sets in practice? To get around this problem, we will have to make a “naive” assumption to simplify our setting: the Naive Bayes assumption.

◂ Previous

Next ▸