Formalize the Curse of Dimensionality

Even in a very limited 3-dimensional case such as the previous spam email example, you notice that we cannot simply fit a Binomial Distribution to predict the label for each input on the data: **this approach is only possible if there are many training inputs with identical feature vectors of a test input**. You could try to collect more data to make sure you capture every possible feature vector in training, but consider the feasibility of such an approach as the dimensionality of the data set increases . As you include additional features in the classifier, such as other words that might indicate spam or metadata about the email itself, the possibility of two inputs matching exactly becomes more and more unlikely. Could you imagine exactly matching two inputs when you have hundreds or even thousands of different features?

When you fit a Binomial distribution to estimate the probability of discrete labels y (i.e. spam or not spam) given an input $\mathbf{x}$, i.e. $\mathrm{P}(y|\mathbf{x})$, you use the following formula:

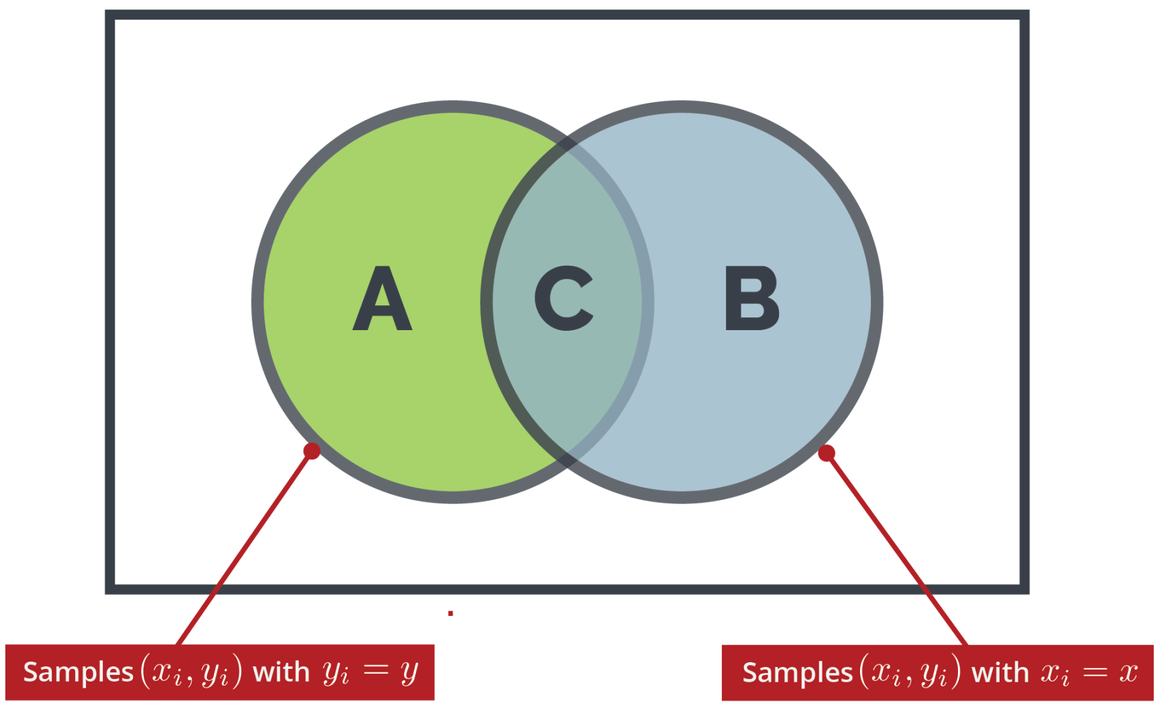
$$\mathrm{P}(y|\mathbf{x}) = \frac{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x} \wedge y_i = y)}{\sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x})}$$

The numerator of the conditional probability expression is counting the number of times y_i takes on the value y and the remaining features $\mathbf{x}_i$ take on the value $\mathbf{x}$ (this is indicated by the symbol $\wedge$).

☆ Key Points

Fitting a Binomial or Bernoulli distribution on the data with MLE is only useful if there are many training inputs with feature vectors identical to a test input.

Adding additional data to compensate for the lack of identical training and test points is not feasible in high dimensional spaces.



Take a look at the Venn diagram. When you estimate the probability of a particular label given an input, $\mathrm{P}(y|\mathbf{x})$, what you're effectively estimating is:

$$\mathrm{P}(y|\mathbf{x}) = \frac{|C|}{|B|}$$

In high dimensional spaces, your data points become more "spread out" and it becomes less likely to observe any given feature value; $|B| \to 0$ and $|C| \to 0$. But what is a high dimensional space? How quickly does this approach break down as you add more features to the vector?

◂ Previous

Next ▸