Notes on Classification

95865 Recitation by Emaad Ahmed Manzoor, last updated 2018.02.16

1 Parameters vs. Hyperparameters

Consider a Gaussian mixture model. In this model, the *parameters* are the Gaussian mean and covariance, μ and Σ, for each of the mixture components; these parameters are *learned* from the data to maximize some objective function. The *hyperparameter* is k, the number of mixture components; this is picked by the user of the model and not learned from the data.

Another example is Naive Bayes. In this model, the parameters are the conditional probabilities that are learned from the data. If using Laplacian smoothing, the hyperparameter is α, the smoothing constant.

2 Laplace Smoothing

Consider a list of ham and spam emails as follows:

Email	**Class**
play sports today	Ham
went play sports	Ham
secret sports event	Ham
sport is today	Ham
sport costs money	Ham
offer is secret	Spam
click secret link	Spam
secret sports link	Spam

1. What is the size of the vocabulary $|V|$?
2. What is the probability that a random email is spam, $P(\text{spam})$?
3. What is the probability that a random email is ham, $P(\text{ham})$?
4. What is $P(\text{spam}|\text{today is secret})$?

The answer to Q4 is a problem: no matter how much evidence we have from the other words in the email, the probability of "today" cannot be conditioned away. To handle issues like this, we may resort to *Laplace smoothing*.

Recall that (via maximum-likelihood estimation) in Naive Bayes, the probability of a word was given by:

$$P(\text{word } i) = \frac{\text{count}(\text{word } i)}{\sum_{\text{word } j \in V} \text{count}(\text{word } j)} \tag{1}$$

With Laplace smoothing and hyperparameter α, this changes to:

$$P(\text{word } i) = \frac{\text{count}(\text{word } i) + \alpha}{\sum_{\text{word } j \in V} \text{count}(\text{word } j) + \alpha|V|} \tag{2}$$

Think of this as adding a *pseudo-count* of α to every word in your vocabulary; so no word can have a count $< \alpha$. This way, an unseen word has a count of α, and hence, a non-zero probability.

5. With $\alpha = 1$, what is $P(\text{spam}|\text{today})$?

3 Cross-Validation Demo

- *Notebook:* `https://gist.github.com/emaadmanzoor/0ba78a2920ea0858b54942eff8b08820`
- *Slides:* `https://speakerdeck.com/emaadmanzoor/introduction-to-classification`