CMSC 478 Lecture Notes

KMA Solaiman

September 29, 2023

1 Laplace Smoothing

Laplace smoothing, also known as *add-one smoothing* or *add-k smoothing*, is a technique used in statistics and machine learning to estimate the probabilities of events or outcomes in situations where some events may have zero observed occurrences in the data. It's a simple and commonly used method for dealing with sparse data or preventing zero probabilities.

Laplace Smoothing: In Laplace smoothing, a fixed value (typically denoted as α, where $\alpha > 0$) is added to the count of each event or outcome in the data. This is done to account for the possibility of unobserved events and to ensure that no probability estimate is exactly zero. The α value is often chosen to be 1, but it can be adjusted based on the specific problem and the amount of smoothing desired.
The smoothed probability of an event or outcome is calculated as:

$$P(\text{event}) = \frac{\text{Count(event)} + \alpha}{\text{Total Count} + \alpha \cdot \text{Number of Possible Events}}$$

where

- Count(event) is the number of times the event occurred in the data.
- Total Count is the total number of events in the data.
- α is the smoothing parameter (typically 1 for Laplace smoothing).
- Number of Possible Events is the total number of possible events or outcomes.

Example 1:
Suppose you have a simple text classification problem with two short documents:

- Document 1: "I like apples."
- Document 2: "I like bananas."

Without smoothing, if a word doesn't appear in a document, its probability is zero. Laplace smoothing helps avoid zero probabilities. Now, let's calculate the word probabilities for each word in the two documents for two cases: Laplace smoothing with $\alpha = 1$ and Laplace smoothing with $\alpha = 2$.

Case 1: Laplace Smoothing with $\alpha = 1$

In this case, you add 1 to the count of each word in your vocabulary. Let's calculate the probabilities for the word "apples" in Document 1:

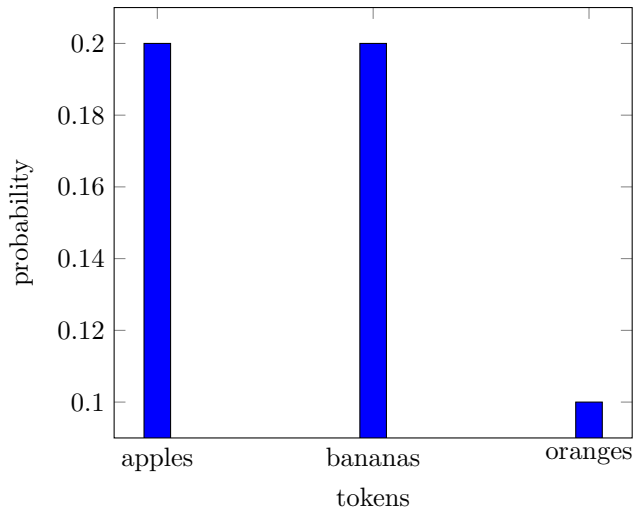
- Count of "apples" in Document 1 = 1
- Total words in Document 1 = 4
- Total unique words in your dataset = 6 (includes "I," "like," "apples," "bananas," plus two more)

The probability of "apples" in Document 1 with Laplace smoothing $\alpha = 1$ would be:

$$\frac{1+1}{4+6} = \frac{2}{10} = \frac{1}{5}$$

You want to calculate the probability of the word "oranges" occurring in these documents. In the original count, "oranges" doesn't appear in either document. The probability of "oranges" in Document 1 with Laplace smoothing $\alpha = 1$ would be:

$$\frac{0+1}{4+6} = \frac{1}{10}$$



In this simple example, you can see how Laplace smoothing assigns non-zero probabilities to words that didn't appear in the original data. It ensures that

no probability is zero and that you can make probabilistic predictions even for previously unseen events.

Case 2: Laplace Smoothing with $\alpha = 2$

In this case, you add 2 to the count of each word in your vocabulary. Let's calculate the probabilities for the word "apples" in Document 1:

- Count of "apples" in Document 1 = 1
- Total words in Document 1 = 4
- Total unique words in your dataset = 6

The probability of "apples" in Document 1 with Laplace smoothing $\alpha = 2$ would be:

$$\frac{1+2}{4+2\times 6} = \frac{3}{16}$$

So, the key difference is in the amount added to the counts. With $\alpha = 1$, you add 1 to each count, while with $\alpha = 2$, you add 2 to each count. This makes Laplace smoothing with $\alpha = 2$ "smoother" and spreads the probability mass more widely than Laplace smoothing with $\alpha = 1$. In practical terms, it makes the model more robust to rare events or unseen words.

Example 2 (from Stanford Notes):
Take the problem of estimating the mean of a multinomial random variable z taking values in $\{1, \ldots, k\}$. We can parameterize our multinomial with $\phi_j = p(z = j)$. Given a set of n independent observations $\{z^{(1)}, \ldots, z^{(n)}\}$, the maximum likelihood estimates are given by:

$$\phi_j = \frac{1}{n}\sum_{i=1}^{n} 1\{z(i) = j\} = \frac{n_j}{n}$$

As we saw previously, if we were to use these maximum likelihood estimates, then some of the ϕ_j's might end up as zero, which was a problem. To avoid this, we can use Laplace smoothing with $\alpha = 1$, which replaces the above estimate with:

$$\phi_j = \frac{\sum_{i=1}^{n} 1\{z(i) = j\} + 1}{n + 1 \times k} = \frac{n_j + 1}{n + 1 \times k}$$

since the number of possible events could be k.