Lecture 18 (Using Naive Bayes)

For computing $\underset{y}{argmax}$ we need to ensure that the numerator doesn't underflow and hence we need to take log before computing the probability

1 Smoothening of the Parameters (Laplace Smoothing)

1. If a certain combination never happens, $\theta_{jl|y=b} = 0$ and if a new data comes with $x_j = l$, the prediction will be $\sim b$
2. The smoothening is done by adding 1 in the numerator and L in the denominator (multiplied by a factor of α for better smoothening)

2 Text Classification

1. We are given a vocabulary of words $\{w_i\}_{i=1}^{|V|}$
2. We have to classify the documents into different classes

2.1 Binomial Model

1. $x_j = \{0, 1\}$ and $y = \{1, 2, \ldots, r\}$
2. The limitation is that the counts don't affect the probabilities

2.2 Multinoulli Model

This model will be terrible since a small change in number of occurences of w_i will change the prediction

2.3 Alternate Multinoulli Model

1. Feature is the word at position j
2. This can now be modelled using multinoulli distribution
3. The problem is that the number of parameters is $O(|V| \cdot \max n \cdot r)$ (each document can have different number of words) which is very huge, this reuqires a very large test data

2.4 Simplifying Assumption Model - Bag of Words

1. We modify the above multinomial distribution so that $\theta_{jl|y=k} = \theta_{j'l|y=k}$
2. The number of parameters are now $O(|V| \cdot r)$
3. We lose the information of ordering of the words but this somewhat prevents overfitting
4. $\theta_{l|y=k} = \frac{\sum_{i=1}^{m}(y_i = k)(n_i)_l}{\sum_{i=1}^{m}(y_i = k)n_i}$ ($(n_i)_l$ is the number of times w_l appears in x_i)
5. We need to add the smoothing factor

3 MAP Estimate

1. We find $\underset{\theta}{argmax} P(\theta|D)$
2. This has a good relation with Laplace smoothing