CMU 10-701, Coding Assignment 2

Joseph Datz

October 2020

The way the problem is presented leads us to the *Multinomial Distribution*:

$$\prod_{k=1}^{K} \mu_k^{x_k}$$

Where u_k is a probability and x_k is a vector of $K-1$ zeroes and one 1. Note that this implicitly has the constraint that

$$\sum_{k=1}^{K} \mu_k = 1$$

So that this satisfies the definition of a probability distribution. Under the Maximum Likelihood (MLE) approach to estimate parameters, the Likelihood function takes the form:

$$P(D|\mu) = \prod_{k=1}^{K} \mu_k^{m_k}$$

Where the m_k's are the counts of the observed words and μ is the vector of all probabilities μ_k. To find the maximum likelihood estimate, we will take the log of this function and use Lagrange Multipliers with our implicit constraint. As an example, we will use the ith partial derivative μ_i to apply to the other partial derivatives:

$$\frac{\partial \ln P(\mu|D)}{\partial \mu_i} + \lambda \frac{\partial g}{\partial \mu_i} = \frac{m_i}{\mu_i} + \lambda,$$

$$\mu_i = \frac{m_i}{\lambda}$$

The above equation for μ_i is found when the partial derivative is set equal to 0. This can placed into our constraint equation to get:

$$\sum_{k=1}^{K} \mu_k = 1 \Leftrightarrow -\frac{1}{\lambda} \sum_{k=1}^{K} m_k = 1 \Leftrightarrow \lambda = -\sum_{k=1}^{K} m_k$$

This gets us the final equation for a parameter μ_i:

	Actual Class																			
Predicted Class	249	0	0	0	0	1	0	0	1	0	0	2	0	3	3	24	2	3	4	26
	0	286	13	14	9	22	4	1	1	0	1	11	8	6	10	1	2	0	0	0
	1	33	204	57	19	21	4	2	3	0	0	12	5	10	8	3	1	0	5	3
	0	11	30	277	20	1	10	2	1	0	1	4	32	1	2	0	0	0	0	0
	0	17	13	30	269	0	12	2	2	0	0	3	21	8	4	0	1	0	1	0
	0	54	16	6	3	285	1	1	3	0	0	5	3	6	4	0	1	1	1	0
	0	7	5	32	16	1	270	17	8	1	2	0	7	4	6	0	2	1	2	1
	0	3	1	2	0	0	14	331	17	0	0	1	13	0	4	2	0	0	6	1
	0	1	0	1	0	0	2	27	360	0	0	0	3	1	0	0	1	1	0	0
	0	0	0	1	1	0	2	1	2	352	17	0	1	3	3	5	2	1	5	1
	2	0	1	0	0	0	2	1	2	4	383	0	0	0	0	1	2	0	1	0
	0	3	0	3	4	1	0	0	0	1	1	362	2	2	2	0	9	0	5	0
	3	20	4	25	7	4	8	11	6	0	0	21	264	9	7	1	3	0	0	0
	5	7	0	3	0	0	3	5	4	1	0	1	8	320	8	7	6	5	8	2
	0	8	0	1	0	3	1	0	1	0	1	4	6	5	343	3	2	1	12	1
	11	2	0	0	0	2	1	0	0	0	0	0	0	2	0	362	0	1	2	15
	1	1	0	0	0	1	1	2	1	1	0	4	0	5	2	1	303	5	23	13
	12	1	0	1	0	0	1	2	0	2	0	2	1	0	0	6	3	326	18	1
	6	1	0	0	1	1	0	0	0	0	0	5	0	10	6	2	63	6	196	13
	39	3	0	0	0	0	0	0	1	1	0	1	0	2	6	27	10	3	7	151

Table 1: The Confusion Matrix for the 20 newsgroups.

$$\mu_i = \frac{m_i}{\sum_{k=1}^{K} m_k}$$

Which can interchanged with any other parameter. As for the Maximum a Posteriori (MAP) estimate, we start with both the Likelihood function and a conjugate prior choice in the from of the *Dirichlet Distribution*:

$$P(\mu|D) \propto P(D|\mu)P(\mu) = \prod_{k=1}^{K} \mu_k^{m_k} * \prod_{k=1}^{K} \mu_k^{\alpha} = \prod_{k=1}^{K} \mu_k^{m_k+\alpha}$$

In general α is a vector instead of a singular value and shifted by 1 to give the appearance of $\alpha - 1$ in the formula, but to be consistent with the problem's premise there is a single scalar α and an implicitly added 1. This can be substituted directly into the MLE estimate formula:

$$\mu_i = \frac{\alpha + m_i}{K\alpha + \sum_{k=1}^{K} m_k}$$

To give us our final MAP estimate for a given parameter μ_i.

Question 3.1. In this setting, each of the Random Variables X_i can take 1 of 50,000 parameter values. This makes the amount of data necessary to accurately calculate the probability distribution for each Random Variable very high; significantly higher than the 1000 documents provided.

Question 3.2. See Table 1 for the Confusion Matrix. With the given choice of α, the best accuracy achievable for the model was 78.5%.

Question 3.3. Using the confusion matrix, it does appear that there are some classes where the model had an easier time than others - classes 2, 3, 4, and 5 see some interchange between each other; classes 1 and 20 are mistaken for each other, etc. The answer for this is likely that the subjects in the newsgroups rely on a similar subset of words as part of their regular language.

Question 3.4. A graph of different accuracy rates for different choices of the parameter α is provided below. I do not see the degree of accuracy change that is in the original solution's plot.