Homework 1 Xiaowei Zhao UNI: xz2767

Problem 1

(1) The prior probability for a random email to be spam is $P(spam) = \frac{3}{5}$

The prior probability for a random email to be ham is $P(ham) = \frac{2}{5}$

(2) $P(word|class)$

| word \ class | spam | ham |
|---|---|---|
| buy | $\frac{1}{12}$ | 0 |
| car | $\frac{1}{12}$ | $\frac{1}{7}$ |
| Nigeria | $\frac{1}{6}$ | $\frac{1}{7}$ |
| profit | $\frac{1}{6}$ | 0 |
| money | $\frac{1}{12}$ | $\frac{1}{7}$ |
| home | $\frac{1}{12}$ | $\frac{2}{7}$ |
| bank | $\frac{1}{6}$ | $\frac{1}{7}$ |
| check | $\frac{1}{12}$ | 0 |
| wire | $\frac{1}{12}$ | 0 |
| fly | 0 | $\frac{1}{7}$ |

(3) $y^* = \arg\max_y P(y) \prod_i P(x_i|y)$

0. Nigeria $y^* = \arg\max_{class} P(class) \prod_i P(Word_i|class)$

When class = spam, $y^* = P(spam)\,P(Nigerial|spam) = \frac{3}{5} \cdot \frac{1}{6} = \frac{1}{10}$

When class = ham, $y^* = P(ham)\,P(Nigerial|ham) = \frac{2}{5} \cdot \frac{1}{7} = \frac{2}{35}$

Since $\frac{1}{10} > \frac{2}{35}$, the class should be spam.

o Nigeria home

When class = spam, $y_x = p(\text{spam})\, p(\text{Nigeria}|\text{spam})\, p(\text{home}|\text{spam}) = \frac{3}{5} \cdot \frac{1}{6} \cdot \frac{1}{12} = \frac{1}{120}$

When class = ham, $y_x = p(\text{ham})\, p(\text{Nigeria}|\text{ham})\, p(\text{home}|\text{ham}) = \frac{2}{5} \cdot \frac{1}{7} \cdot \frac{2}{7} = \frac{4}{245}$

Since $\frac{1}{120} < \frac{4}{245}$, the class should be ham

o home bank money

When class = spam, $y_x = p(\text{spam})\, p(\text{home}|\text{spam})\, p(\text{bank}|\text{spam})\, p(\text{money}|\text{spam})$

$$= \frac{3}{5} \times \frac{1}{12} \times \frac{1}{6} \times \frac{1}{12} = \frac{1}{1440}$$

When class = ham, $y_x = p(\text{ham})\, p(\text{home}|\text{ham})\, p(\text{bank}|\text{ham})\, p(\text{money}|\text{ham})$

$$= \frac{2}{5} \times \frac{2}{7} \times \frac{1}{7} \times \frac{1}{7} = \frac{4}{1715}$$

Since $\frac{1}{1440} < \frac{4}{1715}$, the class should be ham.

Problem 2.

Base case: when $n=1$, $\sum_{w_1} p(w_1) = \sum_{w_1} p(w_1|\text{start}) = 1$

Introduction assumption: let's assume $\sum_{w_1, w_2 \cdots w_n} p(w_1, w_2, \cdots w_n) = \sum_{w_1, w_2 \cdots w_n} p(w_1|\text{start})\, p(w_2|w_1) \cdots p(w_n|w_{n-1}) = 1$

Introduction steps: For the n+1-th step,

$$\sum_{w_1, \cdots w_n, w_{n+1}} p(w_1, w_2, \cdots w_n, w_{n+1}) = \sum_{w_1, \cdots w_n, w_{n+1}} p(w_1|\text{start}) \cdots p(w_n|w_{n-1})\, p(w_{n+1}|w_n)$$

$$= \sum_{w_{n+1}} \sum_{w_1 \cdots w_n} p(w_1|\text{start}) \cdots p(w_n|w_{n-1})\, p(w_{n+1}|w_n)$$

$$= \sum_{w_{n+1}} \sum_{w_1 \cdots w_n} p(w_{n+1}|w_n)$$

$$= \sum_{w_{n+1}} \cdot p(w_{n+1})$$

$$= 1$$

Conclusion: Hence, we can say if you sum up the probabilities of all sentence of length n under a bigram language model, this sum is exactly 1

## Result Table

| *Perpl* | Unigram (no smoothing) | Bigram (no smoothing) | Trigram (no smoothing) | Bigram (smoothed) | Bigram (Add-one smoothed) | Bigram (Katz backoff smoothed) |
|---|---|---|---|---|---|---|
| English | 49.69407533 | inf | inf | 36.2706049 | 93.37432686 | 39.77277337 |
| French | 74.65342427 | inf | inf | 46.204344 | 150.7925373 | 53.653989 |
| German | 30.63471311 | inf | inf | 22.426784 | 53.14784201 | 24.17793953 |
| | | | | | | |
| *Acc* | Unigram (no smoothing) | Bigram (no smoothing) | Trigram (no smoothing) | Bigram (smoothed) | Bigram (Add-one smoothed) | Bigram (Katz backoff smoothed) |
| English | 0.613333333 | 0.48 | 0.426666667 | 0.6 | 0.6 | 0.606666667 |
| French | 0.38 | 0.533333333 | 0.54 | 0.40666667 | 0.393333333 | 0.033333333 |
| German | 0.006666667 | 0.24 | 0.326666667 | 0.02 | 0.006666667 | 0.413333333 |
| All-three | 0 | 0.126666667 | 0.146666667 | 0.01333333 | 0 | 0.026666667 |

## Part 4

As we can see in above table, if the model is not smoothed, only the unigram one have finite perplexity, both bigram and trigram goes to infinity. Also, we can observe that the normal smoothing and katz backoff smoothing models have smaller total perplexity compared to Add-one smoothing method, which is not performed very well. As we learned in the class, the Additive Smoothing method is inaccurate in practice. Moreover, we can conclude the smoothing does remove unseen word (maybe add-one not).

In the analysis of accuracy, overall speaking, all of the accuracies of all-three classes using different n-gram model are not very good. However, if we look at the English / Not English, French / Not French, German / Not German classification, some are relatively good, such as 0.61 in Unigram (no smoothing) model for English, 0.53 in Bigram (no smoothing) model for French, 0.54 in Trigram (no smoothing) model for french and around 0.6 in both Bigram (smoothed) & Bigram (Katz backoff smoothed) models for English. We can say it is different from English to French and German. German has relatively lowest accuracy, the reason might be its grammar structure and the complexity of vocabulary are hard to learn. In the meanwhile, English is the most accurate one comparing to other two language, proving English is easier to learn.