CS 464 Introduction to Machine Learning

HW1 Report

Berdan Akyürek
21600904

1)

1.1) S = {1 word novel, 2 word novel, 1 word poetry, 2 word poetry, 3 word poetry, 1 word story, 2 word story, 3 word story}

1.2) A = {1 word novel, 2 word poem, 3 word poem }

1.3)

A probability always must be in interval [0,1].

P(S) = 1 where S is sample space.

For disjoint events, probability of the union of events is the same as the sum of their probabilities separately. With other words, where A,B,C… are disjoint, $P(A \cup B \cup C...) = P(A) + P(B) + P(C)...$

1.4) This estimation can be disproved easily. Because according to first line, P(3 word story or 2 word novel) = P(3 word story) + P(2 word novel) = 0.045 Since they are disjoint events. According to second line,

P(3 word story or 2 word novel or 2 word poetry) = P(3 word story) + P(2 word novel) + P(2 word poetry) = 0.045 + P(2 word poetry) = 0.11 since they are disjoint events. So,

P(2 word poetry) = 0.11 - 0.045 = 0.065

According to the third line P(2 word poetry or 3 word story) = P(2 word poetry) + P(3 word story) = 0.06 since they are disjoint events. According to the results of the previous line, P(2 word poetry) = 0.065. So,

0.065 + P(3 word story) = 0.06. So, P(3 word story) = -0.005 which is impossible. Because a probability always must be in interval [0,1] and -0.005 is not in this interval. So, Donald"s estimates are wrong.

2)

2.1)

Binomial distribution. $P(Y < 3|X=10) = P(Y<3 \cap X=10)/P(X=10) =$ $(20^{10}e^{-20}/10!)*(C(10,0)(0.3)^0(0.7)^{10}+C(10,1)(0.3)^1(0.7)^9+C(10,2)(0.3)^2(0.7)^8)/(20^{10}e^{-20}/10!) = (C(10,0)(0.3)^0(0.7)^{10}+C(10,1)(0.3)^1(0.7)^9+C(10,2)(0.3)^2(0.7)^8) =$ $((0.3)^0(0.7)^{10}+10(0.3)^1(0.7)^9+45(0.3)^2(0.7)^8) \approx 0.3827827864$

2.2) $20^2e^{-20}/2!*(0.7)^2 = 200e^{-20}(0.7)^2= 98e^{-20}$

2.3) $E(Y) = E(E(Y|X)) = E(0.3k) = 0.3*20 = 6$

3)

3.1) In the dataset, there are 4085 emails, where 2911 of them are spams and 1174 of them are not. With other words, around 71% of the emails are spam in training dataset. But to be balanced, there should be almost the same amount of spam and normal emails. There is a significant difference between them in this case. This makes the dataset skewed. An unbalanced dataset can lead to bias because the algorithm will more tend to choose the one with more data. To solve this, it is possible to reduce the number of spams in the both datasets until both datasets are balanced.

3.2)

```
-----------------------------
Results
-----------------------------
Accuracy rate: 73.20441988950276
Confusion matrix:
-------
475   5
286 320
-------
Number of wrong predictions: 291
```

According to results, it is possible to say the program makes more mistakes in normal emails. The reason behind this may be the unbalanced structure of the dataset.

3.3)

```
Results
-----------------------------
Accuracy rate: 78.82136279926335
Confusion matrix:
-------
537   6
224 319
-------
Number of wrong predictions: 230
```

\

In this case, results are more accurate when alpha value is 1 instead of zero. This change helps the algorithm to distinguish spams but there was not a significant difference on distinguishing normal mails.

3.4)

```
----------------------------
Results
----------------------------
Accuracy rate: 82.59668508287292
Confusion matrix:
-------
595  23
166 302
-------
Number of wrong predictions: 189
```

When this situation is compared to the multinomial model, we see a significant difference between results. It is possible to say that, when we ignore multiple occurences of a word while detecting spams, we get more accurate results. Also it is possible to say that, by ignoring multiple occurences, algorithms can detect spams easier while it detects normal emails harder.