CS 464 Introduction to Machine Learning

HW1 Report

Berdan Akyürek

21600904

1.1)

= P(B or not head) * P(B or not head) * P(B or not head) * P(B or not head) * P(B or not head) * P(B or not head) * P(B or not head) * P(A and head)

= $P(\text{B or not head})^7$ * P(A and head)

P(B or not head) = P3*(1-P1)+P2

P(A and head) = P3*P1

Answer = $(P3-P1*P3+P2)^7*P3*P1$

1.2)

Binomial distribution. E[x] = n*p=10p=10(P3*P1+(1-P3)*P2)

1.3.a)

The results may be measured by repeating the experiment many times. For example, if we count the number of heads that occur when he says heads and divide it by the total number of times Oliver predicts head, we should find 0.95.

1.3.b)

P(oliver predits head) = 0.99

so P(oliver predicts head 8 times in a row) = $(0.99)^8$

1.3.c)

P(oliver predicts not head | head comes)

= P(oliver predicts not head and head comes)/P(head comes)

P(oliver predicts not head and head comes) = 1/100*1/100

P(head) = 99/100*95/100 + 1/100 * 1/100

P(oliver predicts not head and head comes)/P(head) = P(oliver predicts not head | head comes) = 1/(99*95+1) = 1/9406 ≈ 0.0001

2.1)

When a high number of dimensions is the case, it is better to use Manhattan distance against Euclidian is preferable. This problem has 8 features and 8 dimensions. So Manhattan distance is used.

2.2)

Some features may be irrelevant. If the dataset contains an unbalanced number of these irrelevant features by chance, using these irrelevant features may reduce accuracy.

2.3)

Code is uploaded with the report.

2.4)

Training takes a lot more time compared to validation. This is because the validation set is way smaller compared to training set and it is easier to process it.

3.1)

Requested training is done and code is uploaded with the report.

3.2)

Probabilities of spam and normal should be estimated for each word. This makes 2*vocabulary length.

3.3.a)

Requested training is done and code is uploaded with the report.

3.3.b)

There is no time complexity difference. However running time should differ. If we use fewer features, it takes less time to process it.

3.4)

Since Bernoulli does not work properly, it is not possible to make good comparisons.

0 19

0 81

this is an example confusion matrix for Bernoulli for 100 features. It leads to 81% accuracy however it is possible to see there are no negative predictions done. This high accuracy is due to unbalance of the dataset.
However multinomial leads to 73% accuracy. This looks worse than Bernoulli but since other predictions are also done, multinomial works better.