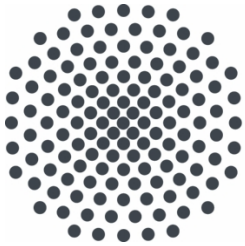

Universität Stuttgart

Machine Learning

Assignment 2

Jugal Yadav (3510720)
study programme: M.Sc. Computer Science

Suganth Natarajan (3522323)
study programme: M.Sc. Autonome Systeme

Abdul Rehman (3440146)
study programme: M.Sc. INFOTECH

11. Mai 2021

1 Solution for Task 1: Simple Bayes

1.1

Given : Box 1 has 4 apples and 10 oranges ; Box 2 has 6 apples and 8 oranges.

$P(apple) = P(Box1ischoosen)P(appleinbox1) + P(Box2ischoosen)P(appleinbox2)$
$P(apple) = (1/2)(4/14) + (1/2)(6/14) = 10/28$

$P(Box1|apples) = [P(apples|Box1)P(Box1)]/[P(apples)]$
$P(Box1|apples) = [(4/14)(1/2)]/(10/28) = 4/10$

1.2

We need to compute as to what is the probablity of the yellow MM coming from a 1994 bag. But there can be 2 scenarios here - bag1 could be of 1994 or bag2. Let's assume that the probablity of a bag being of 1994 or 1996 is evenly distributed. i.e. $P(S_1) = P(S_2) = 1/2$
Thus now what we need to compute is;
$P(S_1|(yellow, green)) = [P(yellow, green)P(S_1)]/P(yellow, green)$
$P(S_1|(yellow, green)) = [P(yellow, green)P(S_1)]/\sum_{S1,S2} P(yellow, green|Si)P(S_i)$

But, $P(yellow, green|S1)P(S_1) = (30/100)(24/100) = (72/1000)$ and,
$P(yellow, green|S2)P(S_2) = (16/100)(20/100) = (32/1000)$

Therefore, $P(S_1|(yellow, green)) = [(72/1000)(1/2)]/[(72/1000)(1/2)+(32/1000)(1/2)]$
$P(S_1|(yellow, green)) = (72/104) = (9/13)$

2 Solution for Task 2: Fake News Classification with Naive Bayes

Please look into the Jupyter Notebook.

3 Solution for Task 3: : kNN for Text Classification

We can formulate the following kNN problem as follows;
Consider we have set of documents as the training sample. Each of this document

has n number of words. Thus we build up a corpus for every document. The cost function which we could use here is Jaccard similarity or Dice coefficient which measures the similarity between two documents. Now whenever any unseen data (document) appears this similarity measure is calculated and the decision rule which can be used to predict the final output to the unseen document could be default with maximum number of votes or based on weighted rules - like probablity distribution, cost function etc.

Example:

d1 = It is a nice weather
d2 = I do not like this windy weather
d3 = Rain in spring is not what I expect !

Let the class label be *positie* or *negative*
Thus d1 has a label *positive* and d2,d3 has a label *negative*

Now corpus is as follows;
$d1 = \{It, is, a, nice, weather\}$
$d2 = \{I, do, not, like, this, windy, weather\}$
$d3 = \{Rain, in, spring, is, not, what, I, expect\}$

Now consider the unseen data d4 = It is a beautiful weather
Corpus $d4 = \{It, is, a, beautiful, weather\}$

Now $J = [|A \cap B|]/[A \cup B]$ and $D = 2[|A \cap B|]/[|A| + |B|]$
Let's take Jaccard similarity measure (J).
$J(d1, d4) = 4/6$
$J(d2, d4) = 1/11$
$J(d3, d4) = 1/12$

Amongst all $J(d1, d4)$ has the highest value which refers to most similar. The decision rule which we will use here is based on distance(similarity) metric. Thus the output label given to d4 will be *negative*.

The disadvantage here is that this method discussed above does not take into account the word frequency which also has a vital role in influencing the predicted output.

4 Solution for Task 4: kNN in High-Dimensional Feature Spaces

In high dimensional feature spaces, almost all the data samples have similar distances with respect to each other. Thus it makes the usage of any distance metric - like euclidean or cosine, irrelevant for its usage.
One way to prevent this problem is to bring down the dimensionality of our feature space using regularization techniques.