EID 403: Machine Learning

Assignment-2

Sections 4B1 and 4B11

Due Date: December 13, 2020

1. Using Bayes Theorem to construct an E-Mail Spam detector using Natural Language Processing. Assuming that out of 100 e-mails in my inbox, 30% of emails are spam and 70% are desired e-mails.

   The word ‘offer’ frequently exists in spam e-mails. But, 10% of the desired e-mails contain the word ‘offer’.

   What is the probability of a new e-mail to be spam if it contains the word ‘offer’?
2. Consider the following training dataset:

Example Number	**Color**	**Type**	**Origin**	**Stolen?**
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Use Naïve Bayesian Classifier to predict if a ‘Red Domestic SUV’ may be stolen.

Hint:

$$Prediction = arg\max_{v_j \epsilon V} P(v_j)\ \pi P(A_i|v_j)$$

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

where

n=number of training examples

nc=number of examples for which v=v_j and a=a_i

p=a priori estimate of P(a_i|v_j)

m=the equivalent sample size

3. Consider the dataset below:

X1=Acid Durability (in seconds)	X2=Strength (in Kg/square meter)	Y=Classification
7	7	Bad

7	4	Bad
3	4	Good
1	4	Good

This dataset is formed using survey from the people(last attribute) as well as objective tests (first two attributes) to classify whether a given tissue paper is good or bad.

The factory now produces a new tissue paper with X1=3 and X2=7. Without again going for another round of surveys, can you predict its classification using k-Nearest Neighbour Classifier (assume k=3)