Arlind Stafaj Data Mining 6930

**Homework 2**

1. **QUESTION 1:**

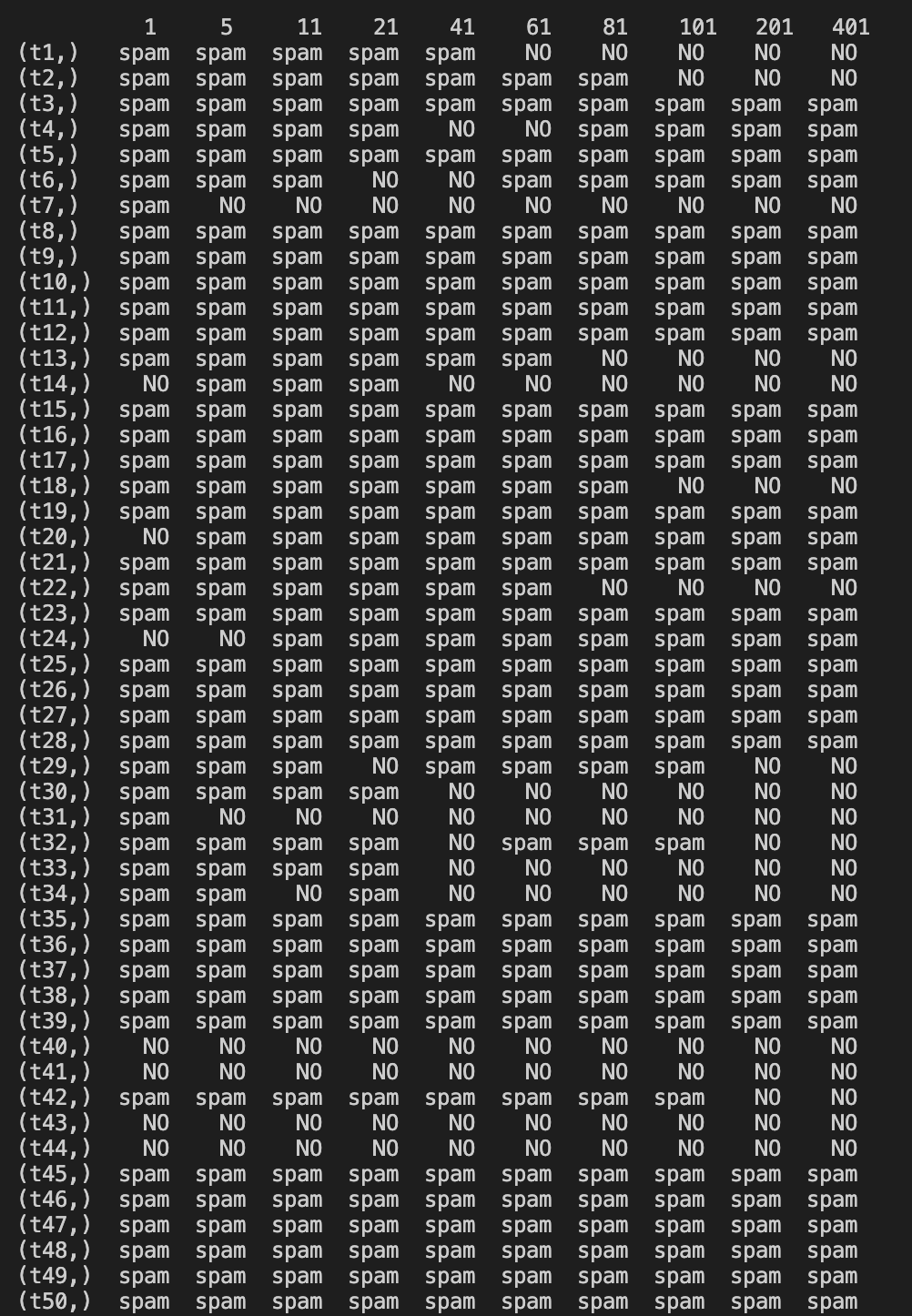
A close up of text on a black background

Description automatically generatedPart (a) & (b):

A screenshot of text

Description automatically generated

Part (c):



Part (d):

* With normalization the accuracy increases and the KNN algorithm is able to better predict the classifications of an input test data. An interesting feature was that for both normalized and non-normalized results, **k** = 11, was most accurate and the in both data sets as **k** grew very large accuracy started decreasing most likely due to surpassing the threshold of data available for a certain class.

Part (e):

* There are several methods for selecting the best k value in a KNN algorithm. Usually the best choice of k will vary depending on the data set. The difficulty is that smaller values of k can be noisy and will have higher influence of the result while larger values of k will have smoother decision boundaries which will mean lower variance but increased bias, in addition to being computationally expensive. You can a simple range of k values and use trial and error. In general data scientists chose k to be the square root of N, where N is the number of samples in your training dataset. Keeping the value of k odd is also a good practice to avoid confusion between two classes.

**QUESTION 2:**

**Run file q2.py for calculations: Charts formulated below –**

**Decision Tree on Education Level**

A close up of a map

Description automatically generated

**A close up of a map

Description automatically generatedDecision Tree on Career**

**A close up of a map

Description automatically generatedPRUNING ON CAREER:**

**QUESTION 3:**

**PolyKernal exponent 1:** A close up of a receipt

Description automatically generated

**PolyKernel exponent = 2: A picture containing text, receipt

Description automatically generated**

**PolyKernel exponent =A picture containing text, receipt

Description automatically generated**

**RBFKernel with gama = 0.01:**

**A picture containing receipt, text

Description automatically generated**

**RBFKernel with gama = 1.0**

**A picture containing receipt, text

Description automatically generated**

With the PolyKernel as the value of the exponent increases the flexibility of the decision boundaries become more flexible so when the kernel exponent was 1 there was very little flexibility and as a result there was greater error. The degree parameter is affecting the flexibility of the decision boundaries so when increased it may reach least error but as it becomes more and more flexible error increases again.

For the gama parameter RBFKernel impacts how far the influence of a single training example reaches, where lower values are ‘far’ and higher values are ‘close’. Since for the first example the gamma is very small the model is too constrained and cannot capture the complexity of the data model. In the second example the gama is larger and thus a more accurate result is produced as shown above.

**QUESTION 4:**

For a function K to be considered a kernel it must satisfy Mercer’s conditions. A kernel function is valid if and only if the kernel matrix for any particular set of data points has all non-negative eigenvalues.

Given x = (x1, x2), K(x, z) = x1\* z1+x1\*ez2+z1\*ex2+ex2+z2

= x1(z1+ ez2)+ ex2(z1+ ez2)

= (z1+ ez2)(x1+ ex2)

= K(z, x)

So K(x, z) = k(z, x) it is symmetric

x = (x1, x2) and x’ = (x1’, x2’)

Form a matrix in form:

[ K(x,x) K(x,x’) ]

[ K(x’,x) K(x’,x’) ]

And can no specify that:

K(x,x) = (x+ex)2

K(x’,x) = (x’ + ex’)(x + ex)

K(x,x’) = (x + ex’)(x’ + ex’)

K(x,’x’) = (x’ + ex’’)2

If we set n = x + ex and n’ = x’ + ex’

Can make the matrix by substituting n in

If we create a vector A = [ 1, 1] vertical and AT is [1,1]

Then A \* M \* AT would be greater or equal to zero satisfying the condition above that it is a positive definite.