Angela Krontiris
CISC 6930: Data Mining
Assignment 2
October 12, 2018

1a) Report test accuracies when k = 1,5,11,21,41,61,81,101,201,401 without normalizing the features.

k	**Test accuracies**
1	0.751847
5	0.754889
11	0.764885
21	0.746632
41	0.752282
61	0.737505
81	0.726641
101	0.728814
201	0.731421
401	0.719687

1b) Report test accuracies when k = 1,5,11,21,41,61,81,101,201,401 with z-score normalization applied to the features.

k	**Test accuracies**
1	0.856150
5	0.870056
11	0.878748
21	0.884398
41	0.885267
61	0.882660
81	0.877445
101	0.875272
201	0.860061
401	0.839635

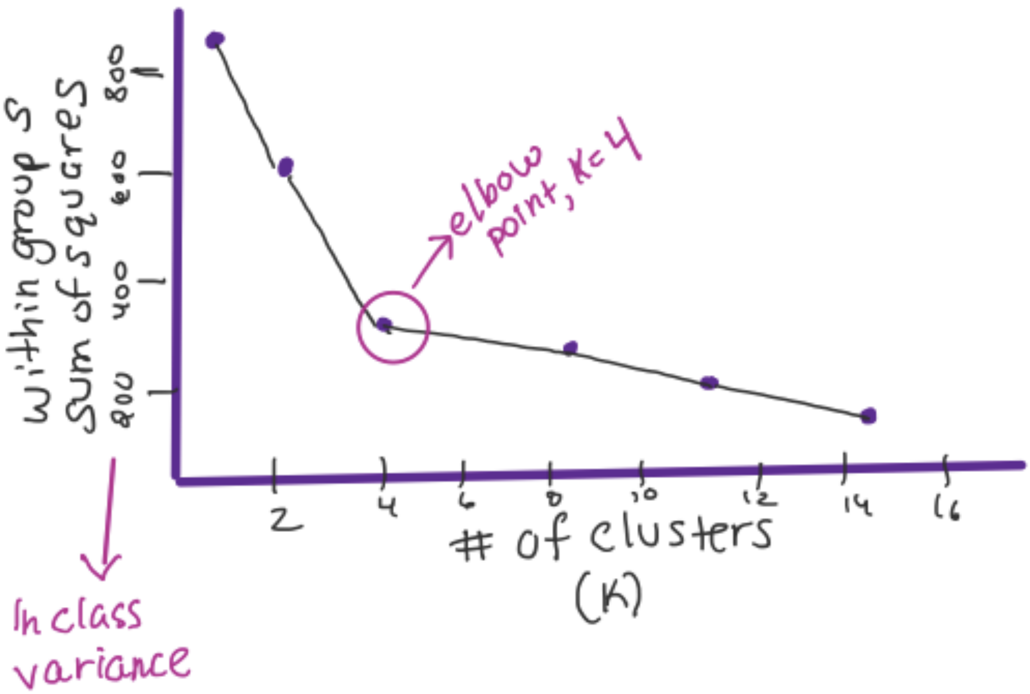
1c) Generate an output of KNN predicted labels for the first 50 instances (i.e. t1 – t50) when k = 1,5,11,21,41,61,81,101,201,401 (in this order)

	ID	1	5	11	21	41	61	81	101	201	401
0	t1	spam	spam	spam	spam	spam	spam	spam	spam	no	no
1	t2	spam	spam	spam	spam	spam	spam	spam	spam	no	no
2	t3	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
3	t4	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
4	t5	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
5	t6	spam	spam	no	spam	no	no	no	no	spam	spam
6	t7	spam	no	no	no	no	no	no	no	no	no
7	t8	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
8	t9	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
9	t10	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
10	t11	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
11	t12	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
12	t13	spam	spam	spam	spam	spam	spam	spam	no	no	no
13	t14	spam	spam	spam	spam	spam	spam	spam	spam	no	no
14	t15	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
15	t16	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
16	t17	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
17	t18	spam	spam	spam	spam	spam	spam	spam	spam	spam	no
18	t19	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
19	t20	no	spam	spam	spam	spam	spam	spam	spam	spam	spam
20	t21	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
21	t22	spam	spam	spam	spam	spam	spam	spam	no	no	no
22	t23	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
23	t24	no	spam	spam	spam	spam	spam	spam	spam	spam	spam
24	t25	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
25	t26	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
26	t27	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
27	t28	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
28	t29	spam	spam	spam	spam	spam	spam	spam	spam	no	no
29	t30	spam	spam	spam	spam	no	no	no	no	no	no
30	t31	spam	no	no	no	no	no	no	no	no	no
31	t32	spam	spam	spam	spam	spam	spam	spam	spam	no	no
32	t33	spam	spam	spam	spam	spam	no	no	no	no	no
33	t34	spam	spam	spam	spam	spam	no	no	no	no	no
34	t35	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
35	t36	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
36	t37	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
37	t38	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
38	t39	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
39	t40	no	no	no	no	no	no	no	no	no	no
40	t41	no	no	no	no	no	no	no	no	no	no
41	t42	spam	spam	spam	spam	spam	spam	spam	spam	no	no
42	t43	no	no	no	no	no	no	no	no	no	no
43	t44	no	no	no	no	no	no	no	no	no	no
44	t45	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
45	t46	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
46	t47	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
47	t48	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
48	t49	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
49	t50	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam

1d) Comparing the performances in part a and b, you can see that using z-score normalization on the features improves the test accuracy scores. In part a, k=11 gives the best the best test accuracy of 76.5% and in part b, k=41 gives the best test accuracy of 88.5%. Using z-score normalization, our test accuracy increased by 12%.

1e) Choosing Optimal K

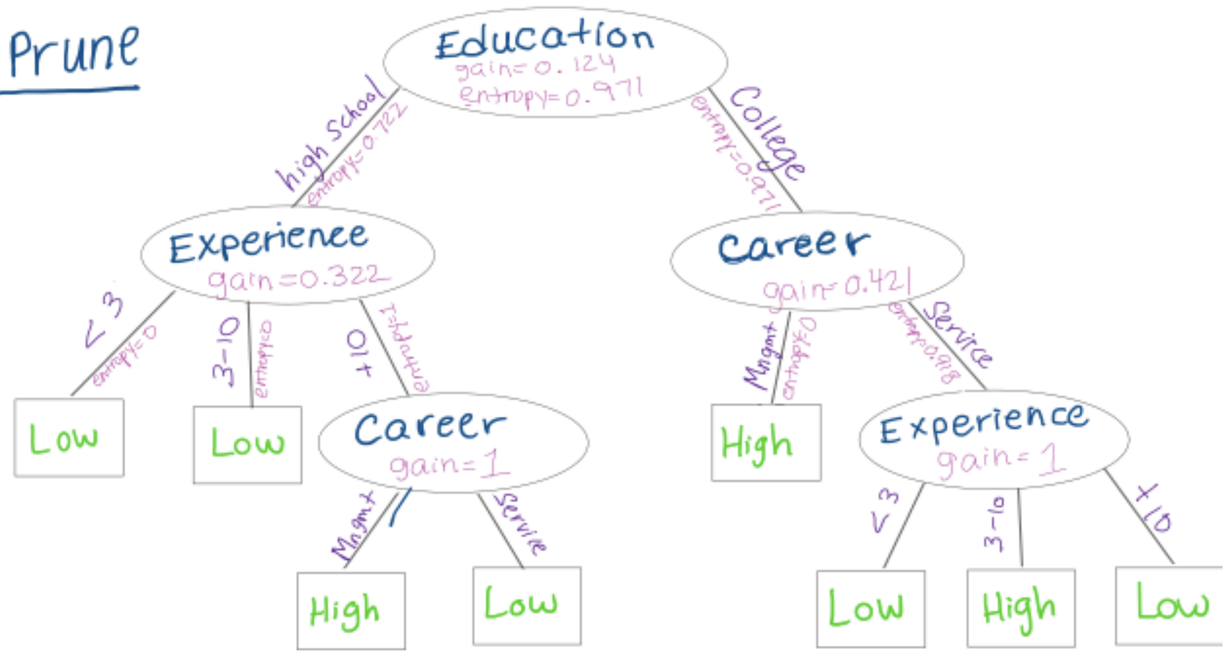
<u>"Knee" or "Elbow" Method</u>



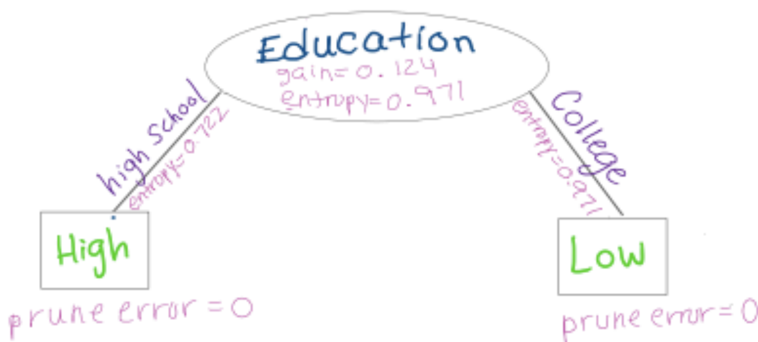
When K increases, the centroids (mean of points in a cluster) are closer to the clusters centroids. The improvements will decline, at some point rapidly, creating the elbow shape. That point is the optimal value for k. In the example above, k=4.

2. Create a decision tree including the number of low's and high's, entropy at each step and information gain for each feature at each node in the tree. Then, prune the tree you obtained using the validation data give in Table 2.

Pre-Prune



Post-Prune



Method used for post-pruning:
For every non-leaf node N:

- Test the accuracy of pruned tree on validation set. Checking to see if the pruned tree performs no worse than the original over the validation set.
- Remove the subtree that results in the greatest improvement in accuracy on validation set.

3) SVM using Weka

SMO Classifier (10-fold cross-validation**,** Classifier for classes: car, noncar)

First run

- Linear Kernel: K(x,y) = <x,y>
- Exponent = 1.0

Correctly Classified Instances	717	84.7518%
Incorrectly Classified Instances	129	15.2482%

Second run

- Poly Kernel: K(x,y) = <x,y>^2.0
- Exponent = 2.0

Correctly Classified Instances	810	95.7447%
Incorrectly Classified Instances	36	4.2553%

Third run

- Poly Kernel: K(x,y) = <x,y>^3.0
- Exponent = 3.0

Correctly Classified Instances	800	94.5626%
Incorrectly Classified Instances	46	5.4374%

Fourth run

- RBF Kernel: K(x,y) = exp(-0.01*(x-y)^2)
- Gamma = 0.01

Correctly Classified Instances	614	75.5768%
Incorrectly Classified Instances	232	27.4232%

Fifth run

- RBF Kernel: K(x,y) = exp(-1.0*(x-y)^2)
- Gamma = 1.0

Correctly Classified Instances	764	90.3073%
Incorrectly Classified Instances	82	9.6927%

The fourth run using RBF SVM with parameter gamma performed the lowest . When gamma is very small, the model is too constrained and cannot capture the complexity or “shape” of the data. You can see that as gamma increased in the fifth run using the same model, the number of correct classified instances improved by 15%. For the polynomial kernel, the exponent parameter controls the degree of the polynomial. The default is set to 1 for the linear kernel (i.e., no kernel at all, just a dot product). Setting the exponent to 2 for the quadratic kernel gave better results.

Question 4: Kernels

Wednesday, October 10, 2018 11:05 PM

Assume $x = (x_1, x_2)$ is a two dimensional vector and a function K defined as $K(x,z) = x_1 * z_1 + x_1 * e^{z_2} + z_1 * e^{x_2} + e^{x_2 + z_2}$. Prove that K is a kernel.

$$K(x,z) = \underbrace{\phi(x) \cdot \phi(z)}_{\text{dot product between two points}} = 1 + 2\sum_{i=1}^{d} x_i z_i + \sum_{i=1}^{d} x_i^2 z_i^2 + 2\sum_{i=1}^{d}\sum_{j=i+1}^{d} x_i x_j z_i z_j = (1 + x \cdot z)^2$$

Kernel Proof:

I. $K(x,z) = x_1 z_1 + x_1 e^{z_2} + z_1 e^{x_2} + e^{x_2} e^{z_2}$

$K(x,z) = x_1 (z_1 + e^{z_2}) + e^{x_2}(z_1 + e^{z_2})$

$K(x,z) = (x_1 + e^{x_2})(z_1 + e^{z_2})$

$K(z,x) = (z_1 + e^{z_2})(x_1 + e^{x_2})$

Commutative Property of multiplication $a \cdot b = b \cdot a$

According to Mercer's 1st condition, $K(x,z) = K(z,x)$ is symmetric.

II. Let $x = x_1 + e^{x_2}$

$z = z_1 + e^{z_2}$

Assume $z^T = [1 \ 1]$, $z = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$\text{Matrix} = \begin{bmatrix} x^2 & xz \\ zx & z^2 \end{bmatrix}$$

$$[1 \times N][N \times N][N \times 1] \geq 0\,[1 \times 1]$$

$$\underset{1\times 2}{\begin{bmatrix} 1 & 1 \end{bmatrix}} \underset{2\times 2}{\begin{bmatrix} x^2 & xz \\ zx & z^2 \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} 1 \\ 1 \end{bmatrix}} = 1 \times 1$$

$$\begin{bmatrix} (x^2 + zx) & (xz + z^2) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = x^2 + zx + xz + z^2$$

$$= x^2 + 2zx + z^2$$

$$= (x+z)^2 \rightarrow \text{this will always be positive}$$

According to Mercer's 2nd condition, the matrix is positive semi-definite.