CSCI4390/6390 – Data Mining
Fall 2011, Exam II
Total Points: 100 + 10 (bonus)

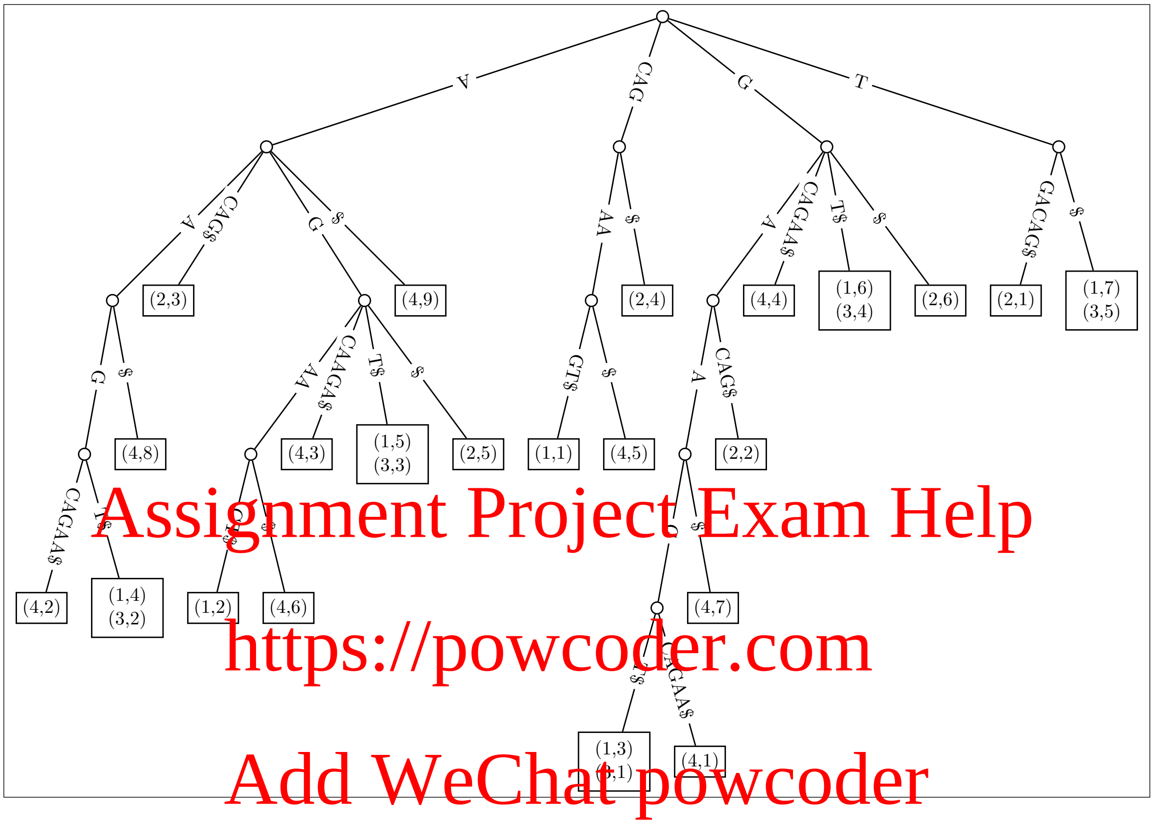


Figure 1: Suffix Tree

1. (35 points) Given the four sequences:
   $s_1 = CAGAAGT$
   $s_2 = TGACAG$
   $s_3 = GAAGT$
   $s_4 = GAAGCAGAA$

   (a) (10 points) The suffix tree for these four sequences is shown in Figure 1. Use this tree to find all the closed substrings with $minsup = 2$. Recall that a substring is a consecutive subsequence, and closed means that there is no superstring with the same support. In the tree, a leaf entry (i, j) means the j-th suffix of i-th sequence.

   Answer: From the suffix tree, in one traversal, we can record the following frequent substrings, namely those that appear in at least sequences: AAGT - 2, AA - 3, AGAA - 2, AGT - 2, AG - 4, A - 4, CAGAA - 2, CAG - 3, GAAGT - 2, GAAG - 3, GAA - 3, GA - 4, GT - 2, G - 4, T - 2.

   The closed frequent substrings are therefore: AG - 4, CAGAA - 2, CAG - 3, GAAGT - 2, GAAG - 3, GA - 4

(b) (10 points) Find all frequent maximal sequences with *minsup* = 3. You may use any method of your choice. Recall that maximal means that there is no frequent supersequence.

Answer: Let us apply a level-wise method to mine the frequent subsequences, and then we can extract the maximal ones.
level 1: A - 4, C - 3, G - 4, T - 3
level 2: AA - 4, AG - 4, GA - 4, GG - 4, CA - 3, CG - 3
level 3: AAG - 4, CAG - 3, GAA - 4, GAG - 4
level 4: GAAG - 4

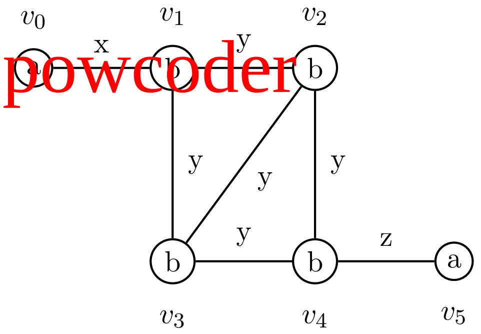
The maximal ones are: GAAG - 4, CAG - 3, T - 3

(c) (15 points) Find the *negative border* of the set of frequent sequences with *minsup* = 3. Negative border is defined as the set of minimal infrequent subsequences. That is, an infrequent sequence all of whose subsequences are frequent.

Answer: The minimal infrequent sequences can be found from those that were found to be infrequent in the question above. That is, once a subsequence is found to be infrequent, no supersequence of that can be minimal, thus, for each infrequent subsequence we have to check if any of its subsequences is infrequent. If so, it cannot be minimal. The set of minimal infrequent sequences checked in the question above are as follows:
AC, AT, AAA, AGA, AGG
CC, CT, CAA, CGA, CGG
GC, GT, GGA, GGG
TA, TC, TG, TT

	X_1	X_2	Class	α_i
$\boldsymbol{x}_1$	3.4	4	-1	1.70
$\boldsymbol{x}_2$	2	4	-1	0
$\boldsymbol{x}_3$	9.1	4.5	$+1$	0.39
$\boldsymbol{x}_4$	2	6	-1	0
$\boldsymbol{x}_5$	1.5	7	-1	0
$\boldsymbol{x}_6$	2.1	2.5	$+1$	1.31
$\boldsymbol{x}_7$	7	6.5	-1	0
$\boldsymbol{x}_8$	8	4	$+1$	0

(a) Classification Dataset



(b) Edge Labeled Graph

2. (30 points) Consider the classification dataset in Figure 2a. Let the Lagrangian multipliers be as shown in the last column. Answer the following questions

(a) (15 points) Compute the weight vector and the bias for the optimal hyperplane. You may assume that $C = 10$.

Answer: The weight vector is given as

$$\mathbf{w} = -1.7\begin{pmatrix}3.4\\4\end{pmatrix} + 0.39\begin{pmatrix}9.1\\4.5\end{pmatrix} + 1.31\begin{pmatrix}2.1\\2.5\end{pmatrix} = \begin{pmatrix}0.52\\-1.77\end{pmatrix}$$

The offset is computed from the three support vectors

$$
\begin{aligned}
b_1 &= -1 - \mathbf{w}^T \mathbf{x}_1 = 4.312 \\
b_3 &= 1 - \mathbf{w}^T \mathbf{x}_3 = 4.233 \\
b_6 &= 1 - \mathbf{w}^T \mathbf{x}_7 = 4.333
\end{aligned}
$$

Thus $b = (b_1 + b_3 + b_6)/3 = 4.293$

(b) (10 points) Write the equation for the optimal hyperplane. Plot the points and sketch the hyperplane by hand. The slope and y-offset must be accurate.

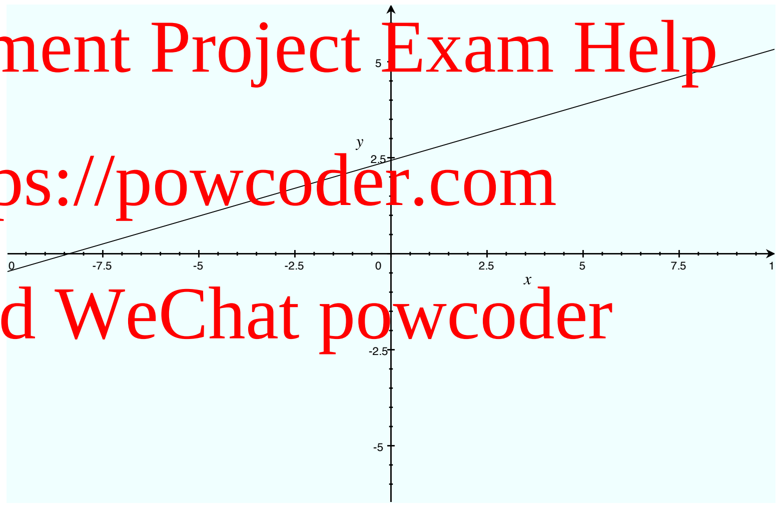
Answer: The hyperplane is given as

$$h(\mathbf{x}) : 0.52x - 1.77y + 4.293 = 0$$

The line is given as

$$h : y = 0.29x + 2.43$$

The line is given as:



(c) (5 points) Classify the point $(4.75, 3.5)$

Answer: The point $\mathbf{x} = (4.75, 3.5)$ is classified as follows:

$$y = sign(\mathbf{w}^T \mathbf{x} + b) = sign(-3.725 + 4.293) = sign(0.568) = +1$$

3. (20 points) Considering only the support vectors in Figure 2a, and assuming the homogeneous quadratic kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^2$, answer the following questions.

(a) (5 points) Compute the kernel matrix $\mathbf{K}$ for the support vectors

Answer: The support vectors are $\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_6$. The pair-wise kernels are given as:
$K(\mathbf{x}_1, \mathbf{x}_1) = (27.56)^2 = 759.6$
$K(\mathbf{x}_1, \mathbf{x}_3) = (48.94)^2 = 2395.1$
$K(\mathbf{x}_1, \mathbf{x}_6) = (17.14)^2 = 293.8$
$K(\mathbf{x}_3, \mathbf{x}_3) = (103.06)^2 = 10621.4$
$K(\mathbf{x}_3, \mathbf{x}_6) = (30.36)^2 = 921.7$

$K(\mathbf{x}_6, \mathbf{x}_6) = (10.66)^2 = 113.6$

The kernel matrix is therefore

$$\mathbf{K} = \begin{pmatrix} 759.6 & 2395.1 & 293.8 \\ 2395.1 & 10621.4 & 921.7 \\ 293.8 & 921.7 & 113.6 \end{pmatrix}$$

(b) (10 points) Compute the direction of most variance in feature space. For the eigenvector/eigenvalue computation, you need to carry out at most three iterations. There is no need to center the kernel matrix $\mathbf{K}$. Also, since the entires of $\mathbf{K}$ will be large, feel free to scale the matrix by dividing by some scalar, since that will not change the eigenvector (it only effects the eigenvalue). For example, if $\mathbf{K}' = \mathbf{K}/s$, for some scalar s (e.g., 100), and if λ' is the eigenvalue of $\mathbf{K}'$, then $\lambda = s\lambda'$ is the eigenvalue of $\mathbf{K}$. The eigenvector corresponding to λ' and λ is the same.

Answer: Since $\mathbf{K}$ have very large entries, we can scale it by any constant. For example, let $s = 100$, so that

$$\mathbf{K}' = \mathbf{K}/100 = \begin{pmatrix} 7.6 & 23.95 & 2.94 \\ 23.95 & 106.21 & 9.22 \\ 2.94 & 9.22 & 1.14 \end{pmatrix}$$

Let $\mathbf{x}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, we then find the eigenvector/eigenvalue as follows:

$$\mathbf{x}_1 = \begin{pmatrix} 34.49 \\ 139.38 \\ 13.3 \end{pmatrix} \rightarrow \begin{pmatrix} 0.247 \\ 1 \\ 0.0954 \end{pmatrix}$$

$$\mathbf{x}_2 = \begin{pmatrix} 26.1 \\ 113.03 \\ 10.05 \end{pmatrix} \rightarrow \begin{pmatrix} 0.231 \\ 1 \\ 0.089 \end{pmatrix}$$

$$\mathbf{x}_3 = \begin{pmatrix} 25.97 \\ 112.6 \\ 10 \end{pmatrix} \rightarrow \begin{pmatrix} 0.231 \\ 1 \\ 0.089 \end{pmatrix}$$

The eigenvalue is $\lambda' = 112.6$.

Thus $\mathbf{c} = \frac{\mathbf{x}_3}{\|\mathbf{x}_2\|} = \frac{1}{1.0302} \begin{pmatrix} 0.231 \\ 1 \\ 0.089 \end{pmatrix} = \begin{pmatrix} 0.224 \\ 0.971 \\ 0.086 \end{pmatrix}$.

The corresponding eigenvalue of $\mathbf{K}$ is $\lambda = 100 \times 112.6 = 11260$

We have to normalize $\mathbf{c}$ by $\frac{1}{\sqrt{\lambda}}$ to get a normalized direction $\mathbf{u}_1$ in feature space; we have:

$$\mathbf{c} = \frac{1}{\sqrt{11260}} \begin{pmatrix} 0.224 \\ 0.971 \\ 0.086 \end{pmatrix} = \frac{1}{106.11} \begin{pmatrix} 0.224 \\ 0.971 \\ 0.086 \end{pmatrix}$$

(c) (5 points) What is the projection of $\mathbf{x}_1$ onto the first kernel principal component.

Answer: Let $\mathbf{K}_1$ be the column of $\mathbf{K}$ corresponding to $K(\cdot, \mathbf{x}_1)$. The projection of $\mathbf{x}_1$ onto the principal component is given as:

$\mathbf{K}_1^T \mathbf{c} = \frac{2521.06}{106.11} = 23.76$

4. (15 points) Find the minimum DFS code for the graph in Figure 2b.

   Answer: The minDFS code is as follows:
   0,1,a,b,x
   1,2,b,b,y
   2,3,b,b,y
   3,1,b,b,y
   3,4,b,b,y
   4,2,b,b,y
   4,5,b,a,z

5. (**Bonus: 10 points**) For an edge labeled graph $G = (V, E)$, define its labeled adjacency matrix A as follows:

$$A(i,j) = \begin{cases} l(v_i) & \text{If } i = j \\ l(v_i, v_j) & \text{If } (v_i, v_j) \in E \\ 0 & \text{Otherwise} \end{cases}$$

   where $l(v_i)$ is the label for vertex v_i and $l(v_i, v_j)$ is the label for edge (v_i, v_j). In other words, the labeled adjacency matrix has the node labels on the main diagonal, and it has the label of an edge (v_i, v_j) in cell $A(i, j)$. Finally, a 0 in cell $A(i, j)$ means that there is no edge between v_i and v_j. Since the graph is assumed to be undirected, A is symmetric.

   Given a particular permutation of the vertices, the code for the corresponding labeled adjacency matrix is obtained by concatenating the lower triangular A row-by-row. For example, one possible matrix corresponding to the default vertex permutation $v_0v_1v_2v_3v_4v_5$ for the graph in Figure 2b is given as

a					
x	b				
0	y	b			
0	y	y	b		
0	0	y	y	b	
0	0	0	0	z	a

   The code for the matrix above is $axb0yb0yyb00yyb0000za$.

   Given the ordering on the labels/entires of the adjacency matrix $0 < a < b < x < y < z$, find the maximum matrix code for the graph in Figure 2b. That is, among all possible vertex permutations and the corresponding codes, you have to choose the lexicographically largest code.

   Answer: The maximum code is $bybyybyy0b00z0a000x0a$, which corresponds to the matrix:

b					
x	b				
y	y	b			
y	y	0	b		
0	0	z	0	a	
0	0	0	x	0	a

   corresponding to the permutation $v_3v_2v_4v_1v_5v_0$.