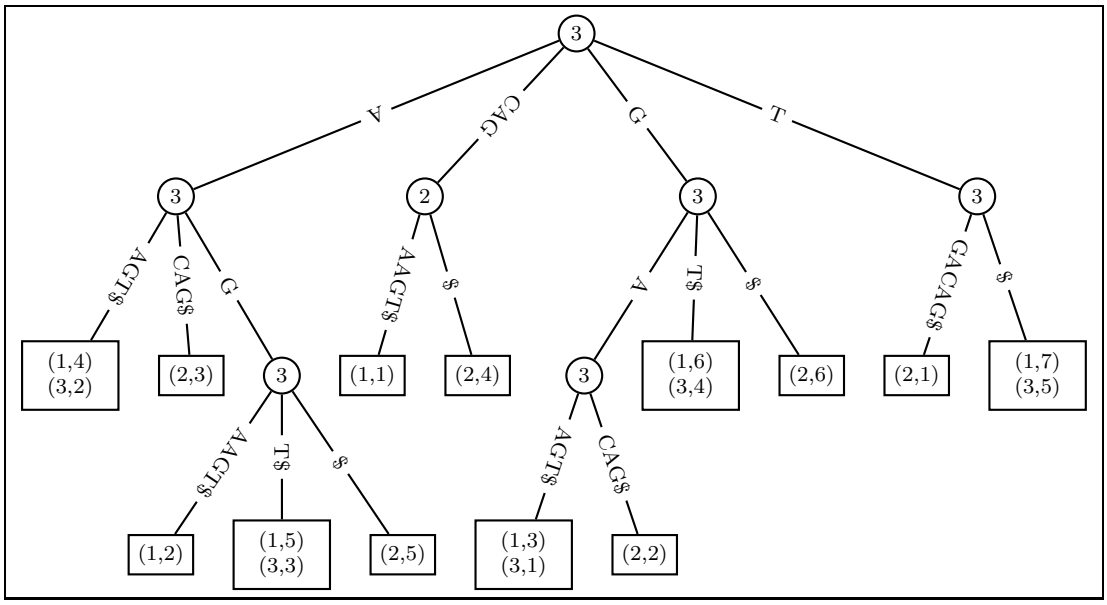
CSCI4390/6390 – Data Mining
Fall 2009, Exam II
Total Points: 100



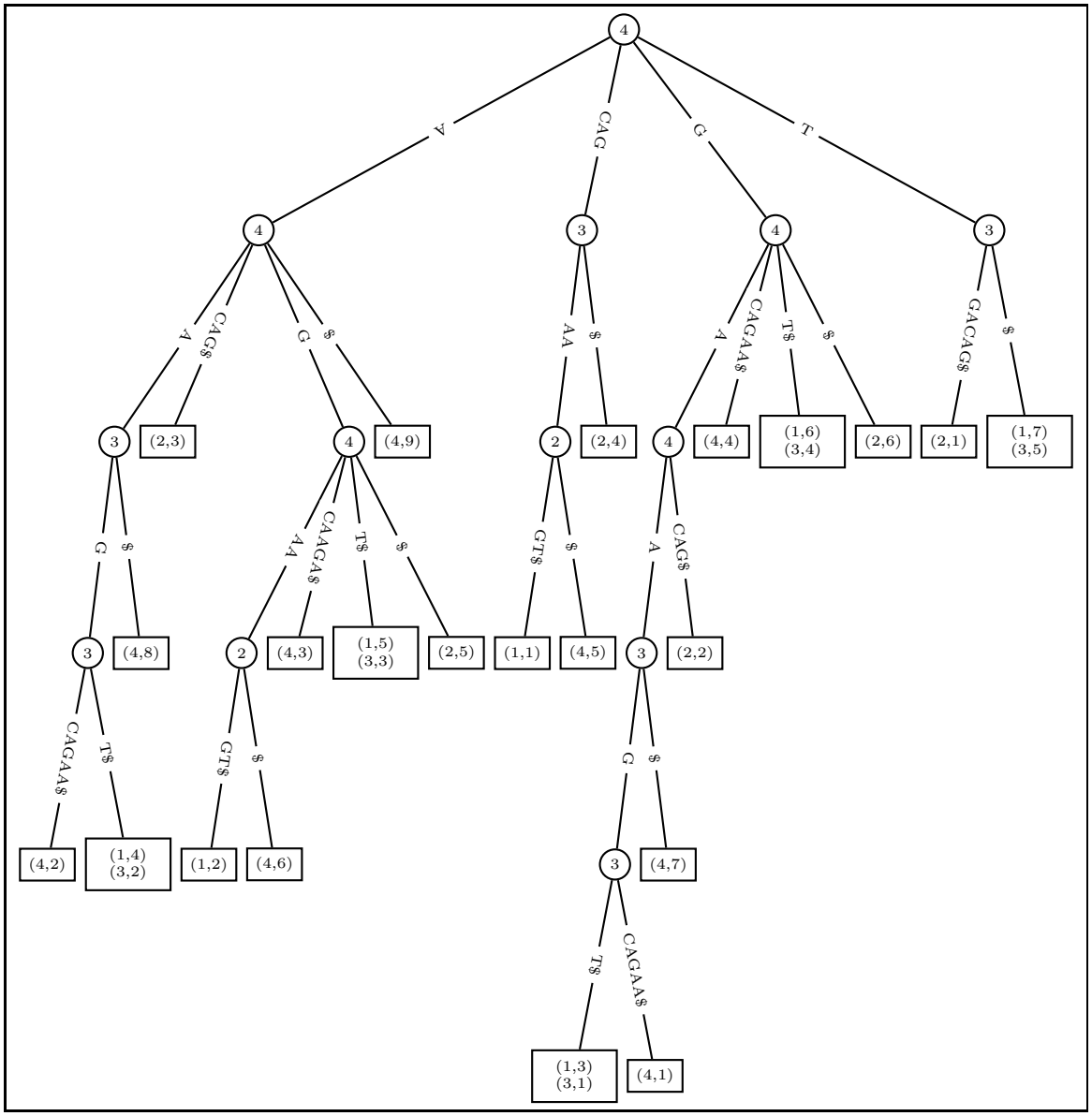
1. (25 points) The suffix tree shown above contains all suffixes of the three sequences $s_1 = CAGAAGT$, $s_2 = TGACAG$, $s_3 = GAAGT$. Note that a pair (i,j) in a leaf denotes the j-th suffix of sequence s_i; further, suffixes start at position 1, instead of 0.

   (a) (15 points) Add $s_4 = GAAGCAGAA$ to the existing suffix tree, using the Ukkonen algorithm. Show the last character position (e), along with the suffixes as they become explicit in the tree. Show the final suffix tree after all suffixes have become explicit.

   (b) (10 pts) Find all maximal frequent substrings with $minsup = 2$ using the final suffix tree.

Answer: (a) When adding s_4, we find that the following strings upto the current last character will all be found in the tree: G, GA, GAA, and $GAAG$. When looking at character 5, namely C, we find the first difference. At this point $e = 5$, and suffixes $l = 1, 2, 3, 4$ will become explicit. Suffix 5 does not become explicit, since C is already in the tree. In fact, all the remaining suffixes $CA, CAG, CAGA, CAGAA$ will be found in the tree and will remain implicit. Finally, when we consider the terminal character \$, all the suffixes will become explicit, i.e., when $e = 10$, $l = 5, 6, 7, 8, 9, 10$. The final suffix tree after adding sequence s_4 is shown below (without the last character \$).

(b) Now based on the tree above, the maximal frequent substrings, with $minsup = 2$ are:
CAGAA - 2
GAAGT - 2


4
A
CAG
G
T
4
3
4
3
A
CAG$
G
$
AA
$
A
CAGAA$
T$
$
GACAG$
$
3
(2,3)
4
(4,9)
2
(2,4)
4
(4,4)
(1,6)
(3,4)
(2,6)
(2,1)
(1,7)
(3,5)
G
$
A,A
CAAGA$
T$
$
GT$
$
A
CAG$
3
(4,8)
2
(4,3)
(1,5)
(3,3)
(2,5)
(1,1)
(4,5)
3
(2,2)
CAGAA$
T$
GT$
$
G
$
(4,2)
(1,4)
(3,2)
(1,2)
(4,6)
3
(4,7)
T$
CAGAA$
(1,3)
(3,1)
(4,1)

2. (15 points) Let $\mathbf{x}_1 = (4, 2.9)$ and $\mathbf{x}_2 = (3.5, 4)$ be two points belonging to class $+1$, and let $\mathbf{x}_3 = (2.5, 1)$ and $\mathbf{x}_4 = (2, 2.1)$ be two points from class -1. Find the best Linear Discriminant direction $\mathbf{w}$ to discriminate between the classes. Use the fact that the inverse of the matrix $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, is given as $\mathbf{A}^{-1} = \frac{1}{det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$. Show all intermediate steps for partial credit. No need to solve the eigen-decomposition to answer this question.

Answer:

$\boldsymbol{\mu}_{+1} = (3.75, 3.45)$, and $\boldsymbol{\mu}_{-1} = (2.25, 1.55)$.

$\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1} = (1.5, 1.9)$

$\mathbf{x}_1 - \boldsymbol{\mu}_{+1} = (0.25, -0.55)$ and $\mathbf{x}_2 - \boldsymbol{\mu}_{+1} = (-0.25, 0.55)$

$\mathbf{x}_3 - \boldsymbol{\mu}_{-1} = (0.25, -0.55)$ and $\mathbf{x}_4 - \boldsymbol{\mu}_{-1} = (-0.25, 0.55)$

$$\begin{aligned}
\mathbf{S}_{+1} &= \begin{pmatrix} 0.25 \\ -0.55 \end{pmatrix} (0.25, -0.55) + \begin{pmatrix} -0.25 \\ 0.55 \end{pmatrix} (-0.25, 0.55) \\
&= \begin{pmatrix} 0.0625 & -0.1375 \\ -0.1375 & 0.3025 \end{pmatrix} + \begin{pmatrix} 0.0625 & -0.1375 \\ -0.1375 & 0.3025 \end{pmatrix} \\
&= \begin{pmatrix} 0.125 & -0.275 \\ -0.275 & 0.605 \end{pmatrix} \\
&= \mathbf{S}_{-1}
\end{aligned}$$

$$\begin{aligned}
\mathbf{S}_w &= \mathbf{S}_{+1} + \mathbf{S}_{-1} \\
&= \begin{pmatrix} 0.25 & -0.55 \\ -0.55 & 1.21 \end{pmatrix}
\end{aligned}$$

$det(\mathbf{S}_w) = 1.21 \times 0.25 - (-0.55)^2 = 0.3025 - 0.3025 = 0$. Unfortunately, this means that $\mathbf{S}_w$ is **singular**, and does not have an inverse. For the purposes of expediency, I asked you to assume that $det(\mathbf{S}_w) = 1$. Taking that as such, we get $\mathbf{S}_w^{-1} = \begin{pmatrix} 1.21 & 0.55 \\ 0.55 & 0.25 \end{pmatrix}$.

The linear discriminant is then given as: $\mathbf{w} = \mathbf{S}_w^{-1} \times (\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}) = \begin{pmatrix} 1.21 & 0.55 \\ 0.55 & 0.25 \end{pmatrix} \begin{pmatrix} 1.5 \\ 1.9 \end{pmatrix} = \begin{pmatrix} 2.86 \\ 1.3 \end{pmatrix}$. The unit vector is then given as $\begin{pmatrix} 0.91 \\ 0.414 \end{pmatrix}$.

3. (20 points) Answer the following questions about the dataset below.

tid	itemset
1	ABD
2	DE
3	BCD
4	BCE
5	ABCE

(a) (10 points) Let Y be a closed itemset, we define an itemset X to be a *generator* for Y if $X \subseteq Y$ and $sup(X) = sup(Y)$. Furthermore, X is called a *minimal generator* if there is no subset of X that is a generator for Y. Find all minimal generators in the database above with $minsup = 2$.

(b) (10 points) Is BCE derivable or non-derivable? Show the upper and lower bound on its support (show all rules used to derive the bounds)

Answer: (a) We have the full set of frequent patterns as follows:
$\emptyset$(5), A(2), B(4), C(3), D(3), E(3)
AB(2), BC(3), BD(2), BE(2), CE(2)
BCE(2)

Thus the closed frequent itemsets are (omitting $\emptyset$): B(4), D(3), E(2), AB(2), BC(3), BD(2), BCE(2).

The minimal generators are as follows: For the single closed itemsets there is only one minimal generator, i.e., the set itself: B(4), D(3), E(2).

For AB(2) the minimal generator is clearly A(2), for BC(3) it is C(3), for BD(2) it is BD(2) itself, since no smaller subset has the same support. Finally for BCE(2) the minimal generators are BE(2) and CE(2).

(b) Looking at the subsets of BCE, we have: BC(3), BE(2), CE(2), B(4), C(3), E(3), $\emptyset$(5).

Trivially $sup(BCE) \in [0, 5]$.

Looking at immediate neighbors, namely BC, BE, CE, we have $sup(BCE) \leq 2$, so the range shrinks to: $sup(BCE) \in [0, 2]$.

Next lets consider the rules from B, C, and E, we get:
$sup(BCE) \geq sup(BC) + sup(BE) - sup(B) = 5 - 4 = 1$
$sup(BCE) \geq sup(BC) + sup(CE) - sup(C) = 5 - 2 = 2$
We do not have to do any more work, since the lower bound has now become 2, so that $sup(BCE) \in [2, 2]$. Thus BCE is derivable.

4. (10 points) Given the dataset below. Classify the point $(35, suv)$, using the naive bayes approach. Assume the domain of `Car` = {sports,suv,vintage}, and use Laplace correction when estimating probabilities.

Id	Age	Car	Class
1	20	sports	L
2	20	vintage	H
3	25	sports	L
4	45	suv	L
5	30	sports	H
6	25	suv	H

Answer: For `Age` the means for the two classes are: $\mu_H = 20 + 30 + 25 = 75/3 = 25$, and $\mu_L = 20 + 25 + 45 = 90/3 = 30$. The variance for the two classes are: $\sigma_H^2 = \frac{5^2+5^2+0}{3} = \frac{50}{3} = 16.67$ and $\sigma_L^2 = \frac{10^2+5^2+15^2}{3} = \frac{350}{3} = 116.67$.

For `Car` we find the following table of probabilities:

Class	Value	Frequency + Laplace	Probability
H	sports	1+1	1/3
H	suv	1+1	1/3
H	vintage	1+1	1/3
L	sports	2+1	1/2
L	suv	1+1	1/3
L	vintage	0+1	1/6

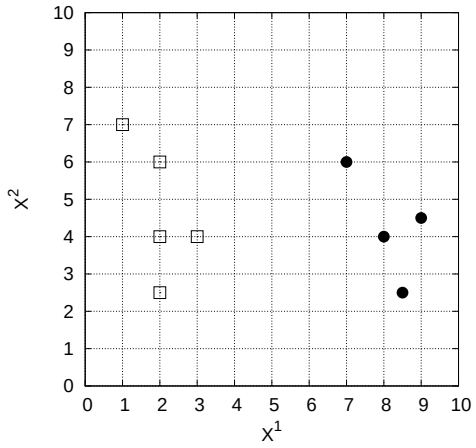
Since the probability of *suv* is 1/3 for both classes, the class is really determined by Age.

We have $P(35|H) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-(x-\mu)^2}{2\sigma^2}\} = \frac{1}{\sqrt{2\cdot\pi\cdot 16.67}} \exp\{\frac{-10^2}{2\cdot 16.67}\} = \frac{1}{10.234} e^{-3} = 0.00486.$

and $P(35|L) = \frac{1}{\sqrt{2\cdot\pi\cdot 116.67}} \exp\{\frac{-5^2}{2\cdot 116.67}\} = \frac{1}{27.075} e^{0.107} = 0.0332.$

Finally, since $P(H) = P(L) = 1/2$, it is clear that the predicted class is the same as the maximum likelihood class, i.e., $P(L|35, suv) = P(35|L) \times P(suv|L) \times P(L) > P(H|35, suv) = P(35|H) \times P(suv|H) \times P(H)$, since $P(H) = P(L) = 1/2$ and $P(suv|L) = P(suv|H) = 1/3$. Thus predicted class is L.

5. (30 points) Consider the dataset below, with two classes. The Lagrangian multipliers (α_i) for each point, obtained by solving the SVM dual objective, are also shown in the table below:



$\mathbf{x}$	x_1	x_2	y_i (class)	α_i
$\mathbf{x}_1$	2	6	1	0
$\mathbf{x}_2$	3	4	1	0.1003
$\mathbf{x}_3$	2	2.5	1	0
$\mathbf{x}_4$	2	4	1	0
$\mathbf{x}_5$	1	7	1	0
$\mathbf{x}_6$	7	6	-1	0.0933
$\mathbf{x}_7$	8	4	-1	0
$\mathbf{x}_8$	8.5	2.5	-1	0.0070
$\mathbf{x}_9$	9	4.5	-1	0

Answer the following questions:

(a) (5 points) List all the support vectors.

(b) (10 points) Compute the maximum margin hyperplane based on the Lagrangian multipliers. That is, compute the weight vector $\mathbf{w}$ and the bias b.

(c) (5 points) Plot the hyperplane and the margins on the figure.

(d) (5 points) What is the margin of this classifier?

(e) (5 points) Classify the point $(6, 3)$.

Answer: (a) The support vectors are $\boldsymbol{x}_2, \boldsymbol{x}_6, \boldsymbol{x}_8$.

(b) The weight vector is given as: $\boldsymbol{w} = \sum_{\alpha_i>0} \alpha_i y_i \boldsymbol{x}_i = 0.1003 \begin{pmatrix} 3 \\ 4 \end{pmatrix} - 0.0933 \begin{pmatrix} 7 \\ 6 \end{pmatrix} - 0.007 \begin{pmatrix} 8.5 \\ 2.5 \end{pmatrix} = \begin{pmatrix} -0.4117 \\ -0.1761 \end{pmatrix}$

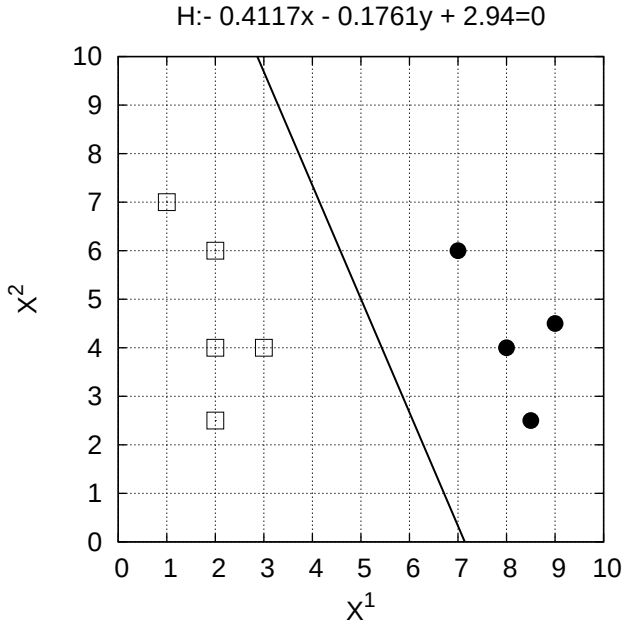
The bias is given the average of :

$$b_2 = y_2 - \boldsymbol{w}^T \boldsymbol{x}_2 = 1 - (-0.4117 \quad -0.1761) \begin{pmatrix} 3 \\ 4 \end{pmatrix} = 1 - (-1.94) = 2.94$$

$$b_6 = y_6 - \boldsymbol{w}^T \boldsymbol{x}_6 = -1 - (-0.4117 \quad -0.1761) \begin{pmatrix} 7 \\ 6 \end{pmatrix} = -1 - (-3.94) = 2.94$$

$$b_8 = y_8 - \boldsymbol{w}^T \boldsymbol{x}_8 = -1 - (-0.4117 \quad -0.1761) \begin{pmatrix} 8.5 \\ 2.5 \end{pmatrix} = -1 - (-3.94) = 2.94$$

The average $b = 2.94$, thus the equation of the hyperplane is $h(\boldsymbol{x}) \equiv (-0.4117 \quad -0.1761) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 2.94 = 0$, as plotted below (c).



(d) The margin of the classifier is $\frac{1}{\|\boldsymbol{w}\|} = \frac{1}{0.4478} = 2.23$.

(e) $h((6,3)) = (-0.4117 \quad -0.1761) \begin{pmatrix} 6 \\ 3 \end{pmatrix} + 2.94 = -2.9985 + 2.94 = -0.0585$, thus the class is -1.