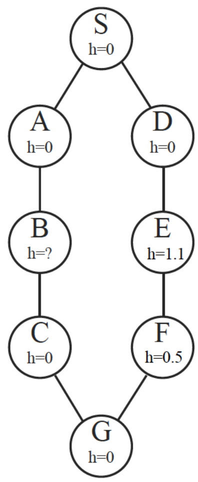
CSE 415 Winter 2018 Review Session Worksheet

Your Name: ______________________________

1 A* search

Consider the following graph, where S is the start and G is the goal state. Each edge has a cost of 1.

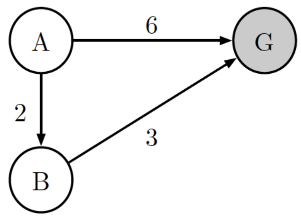


1. Say h(F) = 0.5 is only given (ignore the other h values for now). What are the ranges of h(D) for admissibility and consistency?
   Admissibility: $h(n) \leq h^*(n)$ for all nodes
   Consistency: $h(n) \leq c(n,m) + h(m)$ for all nodes n and their successors m
   $h(D) \in [0,3]$ for admissibility; since path cost from D to G is 1+1+1=3
   $h(D) \in [0,2.5]$ for consistency; since $h(D) \leq c(D\text{-}F) + h(F) = 2+0.5$

2. Now consider all the values and assume ties are broken alphabetically. For what range of values of h(B) will (a) B be expanded before E expanded before F, (b) E be expanded before B expanded before F?
   (a) $c(S\text{-}B) \leq c(S\text{-}E) \Rightarrow 2+h(B) \leq 2+h(E) \Rightarrow h(B) \in [0,1.1]$
   (b) $c(S\text{-}B) \geq c(S\text{-}E) \Rightarrow h(B) \geq 1.1$ and $c(S\text{-}B) \leq c(S\text{-}F) \Rightarrow 2+h(B) \leq 3+0.5$
   Hence, $h(B) \in [1.1,1.5]$

2 A* search

Consider the directed graph below with A: start, G: goal.



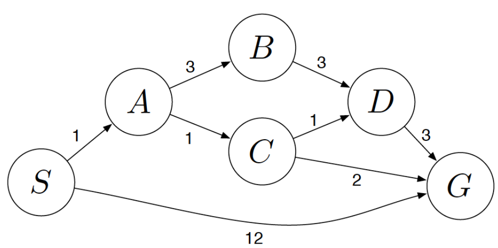
	h(A)	h(B)	h(G)
I	4	1	0
II	5	4	0
III	4	3	0
IV	5	2	0

1. Check for admissibility and consistency of each heuristic I, II, III and IV
   I: admissible, inconsistent (h(A)-h(B) $\geq$ c(A,B))
   II: inadmissible, inconsistent (h(B) $\geq$ 3)
   III: admissible, consistent
   IV: admissible, inconsistent (h(A)-h(B) $\geq$ c(A,B))

2. For one heuristic to dominate another, **all** of its values must be greater than or equal to the corresponding values of the other heuristic. If not, the two heuristics have no dominance relationship. What is the dominance relationship of the given heuristics?
   IV dominates I; III and IV have no dominance relationship etc.

3 A* search

Consider the directed graph below with S: start, G: goal. c(n,m) denotes the cost of the edge from node 'n' to node 'm'.



State	h_1	h_2
S	5	4
A	3	2
B	6	6
C	2	1
D	3	3
G	0	0

1. Find the optimal solution path for the graph given any consistent heuristic.
   S-A-C-G (considering h=0 i.e. uniform cost search)

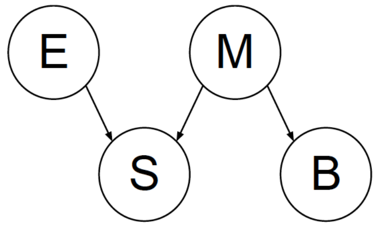
2. Check for admissibility and consistency of h1 and h2.
   h1: inadmissible, inconsistent (h(S) $\geq$ 4)
   h2: admissible, inconsistent (h(S) - h(A) $\geq$ c(S,A))

4 Joint Distribution and Naive Bayes

Consider the following tables. An arrow from A to B indicates B is dependent on A, but not the other way round. Also, P(A,B) denotes joint distribution of A and B and P(A|B) denotes conditional distribution of A conditioned on B.

$P(E)$	
$+e$	0.4
$-e$	0.6

$P(S\|E,M)$			
$+e$	$+m$	$+s$	1.0
$+e$	$+m$	$-s$	0.0
$+e$	$-m$	$+s$	0.8
$+e$	$-m$	$-s$	0.2
$-e$	$+m$	$+s$	0.3
$-e$	$+m$	$-s$	0.7
$-e$	$-m$	$+s$	0.1
$-e$	$-m$	$-s$	0.9



$P(M)$	
$+m$	0.1
$-m$	0.9

$P(B\|M)$		
$+m$	$+b$	1.0
$+m$	$-b$	0.0
$-m$	$+b$	0.1
$-m$	$-b$	0.9

1. Calculate P(e, s, m, b) and P$(+m| + s, +e, +b)$

   P(e, s, m, b) = P(e)P(m)P(s|e,m)P(b|m)

   P$(+m| + s, +e, +b) = \frac{P(+m,+s,+e,+b)}{P(+s,+e,+b)} = \frac{P(+m,+s,+e,+b)}{\sum_m P(m,+s,+e,+b)} = \frac{P(+m,+s,+e,+b)}{P(+m,+s,+e,+b)+P(-m,+s,+e,+b)}$

   Now using the first equation of joint distribution and putting values from the tables:
   P$(+m| + s, +e, +b) = \frac{0.4*0.1*1.0*1.0}{0.4*0.1*1.0*1.0+0.4*0.9*0.8*0.1} \simeq 0.581$

2. Calculate P$(+b)$, P$(+m| + b)$ and P$(+e| + m)$

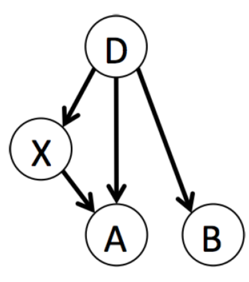
   P$(+b)$ = P(+b|+m)P(+m) + P(+b|-m)P(-m) = 1.0*0.1 + 0.1*0.9 = 0.19 [Bayes rule expansion]

   P$(+m| + b) = \frac{P(+b|+m)P(+m)}{P(+b)} = \frac{1.0*0.1}{0.19} = \frac{10}{19}$ [Naive Bayes]

   P$(+e| + m)$ = P$(+e)$ = 0.4 [conditional independence]

5 Bayes Net

$P(A\|D,X)$			
$+d$	$+x$	$+a$	0.9
$+d$	$+x$	$-a$	0.1
$+d$	$-x$	$+a$	0.8
$+d$	$-x$	$-a$	0.2
$-d$	$+x$	$+a$	0.6
$-d$	$+x$	$-a$	0.4
$-d$	$-x$	$+a$	0.1
$-d$	$-x$	$-a$	0.9



$P(D)$	
$+d$	0.1
$-d$	0.9

$P(X\|D)$		
$+d$	$+x$	0.7
$+d$	$-x$	0.3
$-d$	$+x$	0.8
$-d$	$-x$	0.2

$P(B\|D)$		
$+d$	$+b$	0.7
$+d$	$-b$	0.3
$-d$	$+b$	0.5
$-d$	$-b$	0.5

1. Find P$(-d, +a)$, P$(+d| + a)$ and P$(+d| + b)$. Is B independent of A?

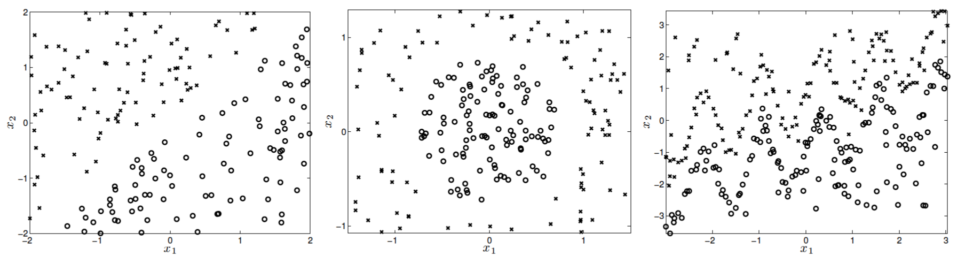
   P$(-d, +a) = \sum_X P(-d, X, +a) = \sum_X P(+a|X, -d)P(X| - d)P(-d)$
   = P$(+a| + x, -d)$P$(+x| - d)$P$(-d)$ + P$(+a| - x, -d)$P$(-x| - d)$P$(-d)$
   = 0.6*0.8*0.9 + 0.1*0.2*0.9 = 0.45

   P$(+d| + a) = \frac{P(+a,+d)}{P(+a)} = \frac{P(+a,+d)}{P(+a,+d)+P(+a,-d)} = \frac{0.087}{0.087+0.45} \simeq 0.162$ [using above derivation]

$P(+d| + b) = \frac{P(+b|+d)P(+d)}{P(+b)} = \frac{P(+b|+d)P(+d)}{P(+b|+d)P(+d)+P(+b|-d)P(-d)} = \frac{0.7*0.1}{0.7*0.1+0.5*0.9} \simeq 0.135$

No, B will be independent of A only if D is given.

6 Perceptrons with kernels



(A) (B) (C)

(i) Linear kernel: $K(x, z) = x^\top z = x \cdot z = x_1 z_1 + x_2 z_2$

(ii) Polynomial kernel of degree 2: $K(x, z) = (1 + x^\top z)^2 = (1 + x \cdot z)^2$

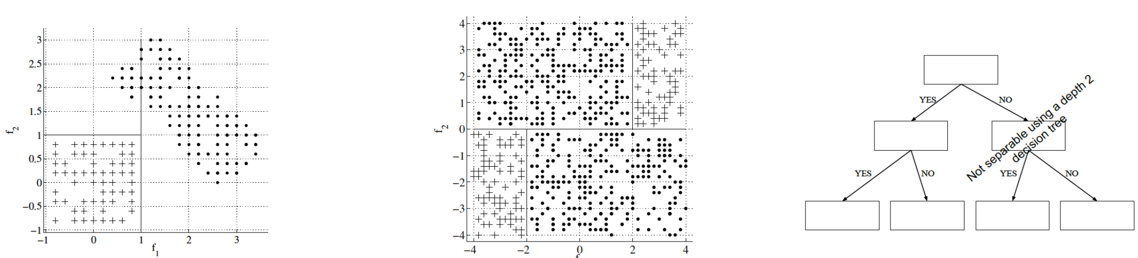
(iii) RBF (Gaussian) kernel: $K(x, z) = \exp\left(-\frac{1}{2\sigma^2}\|x - z\|^2\right) = \exp\left(-\frac{1}{2\sigma^2}(x - z)^\top (x - z)\right)$

1. Datasets are A, B, C and $x = [x_1, x_2]^T$. Which kernel(s) is suitable for which dataset? Do you need bias for the linear kernel? Name a classifier that is good for A and B but not for C.

   Kernels i, ii and iii for A, kernels ii and iii for B and kernel iii for C. Each kernel is superior compared to the previous, hence the reasoning. Intuitively, A can be divided into 2 parts by a straight line through origin, B can be divided by a circle through the empty portion, and C needs a complex zigzag divider. No bias is needed since line will pass through origin

   K-Nearest Neighbour; it may misclassify some points in C due to close proximity at some locations but not in A and B.

7 Decision Trees



(a) (b) (c)

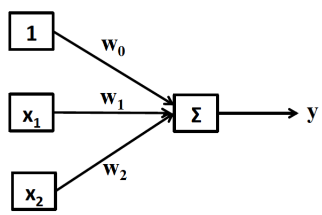
1. Construct the decision tree in (c) for each of (a) and (b). If construction is not possible, write not separable as shown in (c).
   For A:
   Box 1: $f_1 < 1$
   Box 1 YES (Box 2): $f_2 < 1$

Box 2 YES: *Plus*; Box 2 NO: *Dot*
Box 1 NO (Box 3): *Dot* - END

For B:
Box 1: $f_2 < 0$;
Box 1 YES (Box 2): $f_1 < -2$
Box 2 YES: *Plus*; Box 2 NO: *Dot*
Box 1 NO (Box 3): $f_1 > 2$
Box 3 YES: *Plus*; Box 3 NO: *Dot*

2. The conditional independence assumption of Naive Bayes model is that features are conditionally independent when given the class. Do the set of points in (a) and (b) satisfy the assumption?
   No for both set of points. For each class, there is overlap of feature space. You can't decide based on just f1 or just f2, you need both.

8 Perceptron weight update



$$y = \begin{cases} 1 & \text{if } (w_0 + w_1 \cdot x_1 + w_2 \cdot x_2) > 0 \\ 0 & \text{otherwise} \end{cases}$$

1. For what values of w_0 will the perceptron correctly classify $y = x_1 XOR x_2$ and $y = x_1 AND x_2$?
   Consider the tables for XOR and AND
   For XOR:
   x1 = 1, x2 = 1 gives y = 0 $\Rightarrow$ w0 + w1*1 + w2*1 ≤ 0
   x1 = 0, x2 = 1 gives y = 1 $\Rightarrow$ w0 + w1*0 + w2*1 > 0
   x1 = 1, x2 = 0 gives y = 1 $\Rightarrow$ w0 + w1*1 + w2*0 > 0
   x1 = 0, x2 = 0 gives y = 0 $\Rightarrow$ w0 + w1*0 + w2*0 ≤ 0
   which is contradictory. Hence, no solution is possible.
   For AND:
   x1 = 1, x2 = 1 gives y = 1 $\Rightarrow$ w0 + w1*1 + w2*1 > 0
   x1 = 0, x2 = 1 gives y = 0 $\Rightarrow$ w0 + w1*0 + w2*1 ≤ 0
   x1 = 1, x2 = 0 gives y = 0 $\Rightarrow$ w0 + w1*1 + w2*0 ≤ 0
   x1 = 0, x2 = 0 gives y = 0 $\Rightarrow$ w0 + w1*0 + w2*0 ≤ 0
   Combining, (-w1-w2) $<$ w0 $\leq$ min(0,-w1,-w2)

2. Suppose initial weights for a multi-class perceptron is $w_A = [1, 0, 0], w_B = [0, 1, 0]$ and $w_C = [0, 0, 1]$ and A is chosen over B over C in case of tie. What will be the weights after a feature vector (training data) $[1, -2, 3]$ with label A is added? In the next step, what will be the weights after a feature vector (training data) $[1, 1, -2]$ with label B is added? In the next step, what will be the weights after a feature vector (training data) $[1, -1, -4]$ with label B is added?

feature vector	label	w_A	w_B	w_C
		[1 0 0]	[0 1 0]	[0 0 1]
[1 -2 3]	A	[2 -2 3]	[0 1 0]	[-1 2 -2]
[1 1 -2]	B	[2 -2 3]	[1 2 -2]	[-2 1 0]
[1 -1 -4]	B	[2 -2 3]	[1 2 -2]	[-2 1 0]

Step 1: $w_A * f$ = 1*1+0*(-2)+0*(3) = 1; $w_B * f$ = 0*1+1*(-2)+0*(3) = -2; $w_C * f$ = 0*1+0*(-2)+1*(3) = 3; hence predicted label = C but actual label = A.
So, $w_A = w_A + f = [1,0,0] + [1,-2,3] = [2,-2,3]$ and $w_C = w_C - f = [-1,2,-2]$

Step 2: Use weights of step 1 and follow same procedure as step 1. Misclassified as C, so $w_B = w_B + f$ and $w_C = w_C - f$
Step 3: correctly classified as B, so no change

9 Naive Bayes

Given the table below where W stands for word and given W_1 = perfect and W_2 = note,

W	note	to	self	become	perfect
$P(W \mid Y = \text{spam})$	1/6	1/8	1/4	1/4	1/8
$P(W \mid Y = \text{ham})$	1/8	1/3	1/4	1/12	1/12

find the values of P(Y=spam) for which the classifier will produce label ”spam”.

$$P(Y = spam|W_1, W_2) > P(Y = ham|W_1, W_2)$$
$$\Rightarrow P(W_1, W_2|Y = spam)P(Y = spam) > P(W_1, W_2|Y = ham)P(Y = ham)$$
$$\Rightarrow P(perfect|Y = spam)P(note|Y = spam)P(Y = spam)$$
$$> P(perfect|Y = ham)P(note|Y = ham)[1 - P(Y = spam)]$$
$$\Rightarrow (1/8) * (1/6)P(Y = spam) > (1/12) * (1/8) * [1 - P(Y = spam)]$$
$$\Rightarrow P(Y = spam) > 1/3$$

10 Maximum likelihood and Laplacian smoothing

1. $P(X = k) = (1 - \theta)^{k-1}\theta$ for $k \in \{2, 4, 7, 9\}$. Write the log-likelihood expression as a function of θ and find the maximum likelihood estimate for θ.

   Expression = log[P(X=2)P(X=4)P(X=7)P(X=9)] = $\log[(1-\theta)^{18}\theta^4]$ = $18\log(1-\theta) + 4\log\theta$

   Derivative = $\frac{-18}{1-\theta} + \frac{4}{\theta} = 0 \Rightarrow \theta = \frac{4}{22}$
   Second derivative = $\frac{18}{(1-\theta)^2} + \frac{-4}{\theta^2}$. Putting $\theta = \frac{4}{22}$, the second derivative is negative. Hence ML estimate is $\frac{4}{22}$. [P.S. if value was positive, that θ would give minimum value of f(θ)]

2. Consider Bayes' net with two variables A, B. Find the ML and k = 2 Laplace estimates for each of the table entries based on the following data: $(+a, -b), (+a, +b), (+a, -b), (-a, -b), (-a, -b)$.

A	$P^{\text{ML}}(A)$	$P^{\text{Laplace, k=2}}(A)$
$+a$	$\frac{3}{5}$	$\frac{5}{9}$
$-a$	$\frac{2}{5}$	$\frac{4}{9}$

A	B	$P^{\text{ML}}(\mathbf{B} \mid \mathbf{A})$	$P^{\text{Laplace, k=2}}(\mathbf{B} \mid \mathbf{A})$
$+a$	$+b$	$\frac{1}{3}$	$\frac{3}{7}$
$+a$	$-b$	$\frac{2}{3}$	$\frac{4}{7}$
$-a$	$+b$	0	$\frac{2}{6}$
$-a$	$-b$	1	$\frac{4}{6}$

3 out of 5 given data have +a and 2 have -a, hence the values of $P^{ML}(A)$.
Add 2 dummy data for +a (since k=2) and 2 dummy data for -a. Total data: 9, out of which 3+2=5 have +a and 2+2 = 4 have -a, hence the values of $P^{Laplace}(A)$.

Given +a, 1 given data has +b and 2 have -b. Given -a, 0 given data has +b and 2 have -b. Hence the values of $P^{ML}(B|A)$.
Add 2 dummy data for each case. Now, given +a, 3 given data has +b and 4 have -b. Given -a, 2 given data has +b and 4 have -b. Hence the values of $P^{ML}(B|A)$.

11 Miscellaneous (FYI)

1. See questions file

2. If H1 and H2 are both admissible heuristics, min(H1,H2), max(H1, H2) and (H1+H2)/2 are also admissible. Any weighted average of H1 and H2 is also admissible. However, they are not necessarily consistent.
   P.S. What will happen is H1 is admissible and H2 is not?
   Only min(H1,H2) still remains admissible, rest don't
   P.P.S. Suppose a heuristic h's values are multiplied by 'n'. Will the solution of A* be still optimal and what will be the cost of that solution?
   The cost will be 'n' times the original cost but the path may not be optimal since the new heuristic may not be admissible.

3. See questions file

4. See questions file

5. How to determine and avoid overfitting and underfitting during classifier training?
   Overfitting: High training accuracy, low validation accuracy
   Actions: Add and remove features; Add more training data
   Increase Laplace smoothing parameter to make the classifier less sensitive to rare patterns.
   If classifier is Bayes Net, you can remove some edges.
   Underfitting: low training and validation accuracies
   Actions: Add more features; Add more training data