CS 422 Exam 1 Name Solution

Short Answer (10 Points)

1. In the following optimization problem, what values of λ would result in overfitting? What about underfitting? Briefly explain.

$$\min L(w, b) = e^{y\hat{y}} + \lambda||w||^2$$

λ = large underfitting because we only care about making the weights small

λ = small overfitting because we ignore the regularization term focusing only on minimizing errors.

2. [✓ True]/False: False negatives are only important for calculating Recall. Briefly explain.

recall wants to remember everything it can

$$\frac{TP}{TP+FN}$$

a FN is something that wasn't recalled.

precision does not use this

$$\frac{TP}{TP+FP}$$

Short Answer (10 Points)

3. True/[✓ False]: There is no value of K for which K-NN will achieve 100% accuracy on the training data on every dataset. Briefly explain.

For K=1 you will always get 100% accuracy on the training data.

4. True/[✓ False]: The perceptron will always converge to the same w and b. Briefly explain.

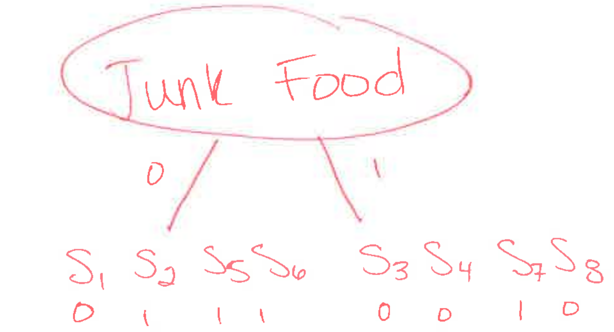
The ordering of the data affects the outcome of the algorithm.

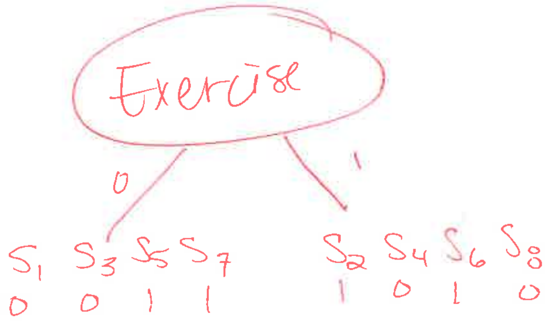
Decision Trees (~~10~~ 20 Points)

5. Assume you have the following training data. Using the Information Gain algorithm from class, build the best depth-1 decision tree for this data.

Sample	Junk Food	Exercise	Label
s_1	0	0	0
s_2	0	1	1
s_3	1	0	0
s_4	1	1	0
s_5	0	0	1
s_6	0	1	1
s_7	1	0	1
s_8	1	1	0

$H = 1$ (4 0s ; 4 1s)





$H = 0.5+.311$
$H = 0.811$

$H = 0.5+0.311$
$H = 0.811$

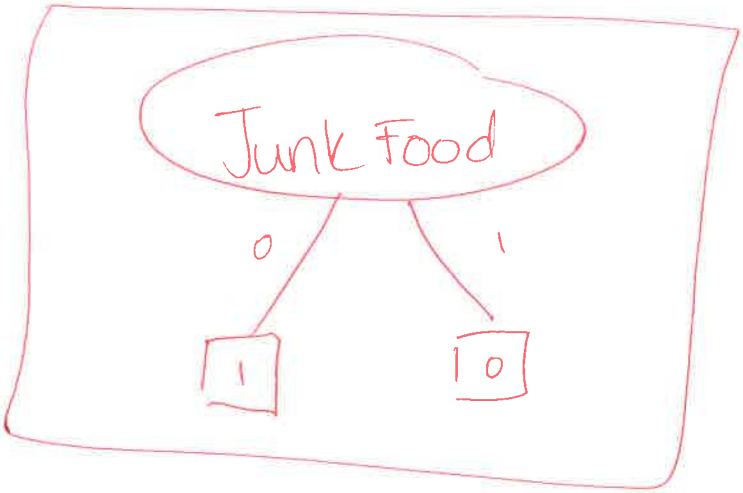
$H = 1$

$H = 1$

$IG = 1 - \frac{1}{2}(1) - \frac{1}{2}(1) = 0$

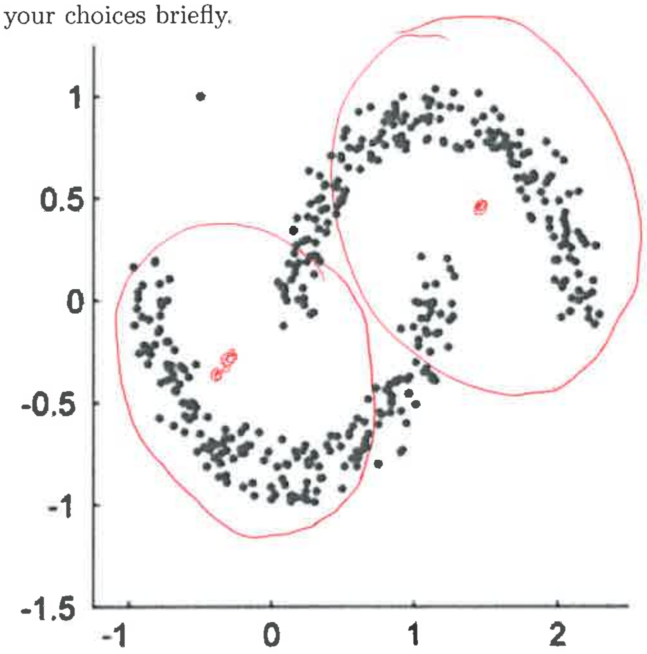
$IG = 1 - \frac{1}{2}(0.811) - \frac{1}{2}(0.811)$

$IG = 0.189$



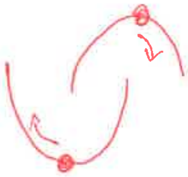
K-Means (10 Points)

6. How might K-means cluster the following data? Indicate the cluster centers and the rough clusters on the graph. Explain your choices briefly.



you can't get the half moon shapes with K-means because it's distance-based.

So you may start with cluster centers at the min/max of the half moons, but through the iterations they will shift away from eachother



Linear Classifiers (20 Points)

7. Give the gradient descent update rules for the following regularized loss function. Show your work for partial credit!

$$L(w, b) = \sum_n (y_n - (wx_n + b))^2 + \lambda ||w||^2$$

$$\frac{\partial L}{\partial b} = \sum_n -2(y_n - (wx_n + b))$$

$$b = b - \eta \sum_n -2(y_n - (wx_n + b))$$

$$\nabla_w L = \sum_n -2x_n(y_n - (wx_n + b)) + 2\lambda w$$

$$w = w - \eta\left(\sum_n -2x_n(y_n - (wx_n + b)) + 2\lambda w\right)$$

Perceptron (~~20~~ 10 Points)

8. Run the perceptron algorithm on the following data in the order provided for two epochs. Give the final w and b produced by the algorithm at the end of the ~~second~~ first epoch.

Sample	x_1	x_2	y
s_1	0	1	1
s_2	1	0	1
s_3	1	1	1
s_4	2	2	-1
s_5	2	1	-1
s_6	1	2	-1

$W=(0,0)$ $b=0$

Epoch 1

$a_1 = 0\cdot 0+0\cdot 1+0 = 0\cdot 1 \leq 0$ update! $w=(0,1)$ $b=1$

$a_2 = (0,1)\cdot(1,0)+1 = 1\cdot 1>0$ no update

$a_3 = (0,1)\cdot(1,1)+1 = 2\cdot 1>0$ no update

$a_4 = (0,1)\cdot(2,2)+1 = 3\cdot -1 \leq 0$ update! $\boxed{W=(-2,-1)\ b=0}$

$a_5 = (-2,-1)\cdot(2,1)+0 = -5\cdot -1 \geq 0$ no update

$a_6 = (-2,-1)\cdot(1,2)+0 = -4\cdot -1>0$ no update

Epoch 2

$a_1 = (-2,-1)\cdot(0,1)+0 = -1\cdot 1 \leq 0$ update! $w=(-2,0)$ $b=1$

$a_2 = (-2,0)(1,0)+1 = -1\cdot 1 \leq 0$ update! $w=(-1,0)$ $b=2$

$a_3 = (-$

Gradient Descent (10 Points)

9. Run gradient descent on the following function for two steps. Use the starting point $x_0 = 5$ and $\eta = 0.1$.

$$f(x) = x^2 + 3$$

$f'(x) = 2x$

$X = X - \eta f'(x)$

$X = X - 0.1 \cdot 2x = X - 0.2x$

iter 1

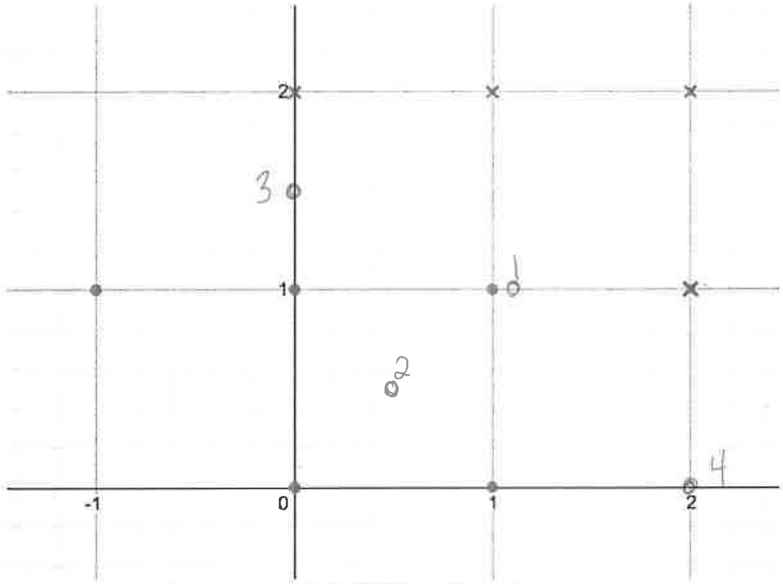
$X = 5 - 1 = 4$

iter 2

$X = 4 - 0.8 = \boxed{3.2}$

KNN (10 Points)

10. Using K=1, 3 and 5, classify the following test data using the plotted training data (x = negative, filled circle = positive). Use the table below to record your final answers. Indicate ties with the answer +/−.



	Sample	$K = 1$	$K = 3$	$K = 5$
1	(1.1,1)	+	+/−	+
2	(0.5,0.5)	+	+	+
3	(0,1.5)	+/−	+/−	+
4	(2,0)	+/\	+	+

CS 622 Exam 1 Name Solution

Short Answer (10 Points)

1. Briefly explain what each of the two terms in the following expression represent in English.

$$\min L(w,b) = \underset{\textcircled{1}}{e^{y\hat{y}}} + \lambda||\underset{\textcircled{2}}{w}||^2$$

① Surrogate loss function for 0/1 loss measuring error on training data (solving the problem)

② regularization term that minimizes weights and makes the solution simpler.

2. True/[✓ False]: True positives are only important for calculating Recall. Briefly explain.

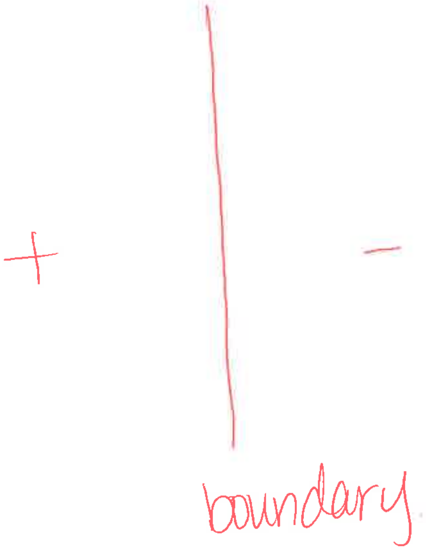
recall & precision need true positives.

$$\frac{TP}{TP+FN} \qquad \frac{TP}{TP+FP}$$

it's how they measure success.

Short Answer (10 Points)

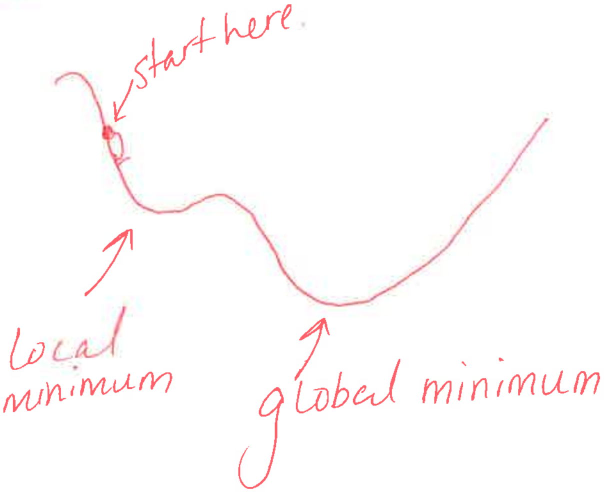
3. Give an example of a set of data for which K-NN and K-Means would produce the same decision boundary. Clearly indicate the values of K you chose for K-NN and for K-Means.



K-NN
K=1

K-means
K=2

4. True/[✓ False:] Gradient descent will always converge to the global minimum. Briefly explain.



depending on starting position and n you often times end up in a local minimum.

Decision Trees (~~10~~ 20 Points)

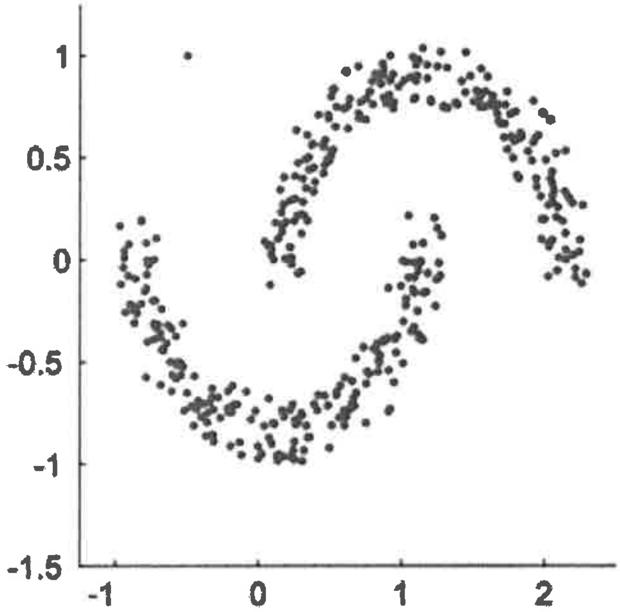
5. Assume you have the following training data. Using the Information Gain algorithm from class, build the best depth-1 decision tree for this data.

Sample	Junk Food	Exercise	Label
s_1	0	0	0
s_2	0	1	1
s_3	1	0	0
s_4	1	1	0
s_5	0	0	1
s_6	0	1	1
s_7	1	0	1
s_8	1	1	0

Same as 422 #5

K-Means (10 Points)

6. How might K-means cluster the following data? Indicate the cluster centers and the rough clusters on the graph. Explain your choices briefly.



Same as 422 #6

Linear Classifiers (20 Points)

7. Give the gradient descent update rules for the following regularized loss function. Show your work for partial credit! Note that $|w|$ is the L_1 norm. $|w| = \sum_i w_i$.

$$L(w, b) = \sum_n (y_n - (wx_n + b))^2 + \lambda |w|$$

$$\frac{\partial L}{\partial b} = \sum_n -2(y_n - (wx_n + b))$$

$$\nabla_w |w| \Rightarrow \frac{\partial |w|}{\partial w_i} = 1 \Rightarrow \vec{1}$$

$$\nabla_w L = -2x_n(y_n - (wx_n + b)) + \lambda \vec{1}$$

$$b = b + \eta \sum_n 2(y_n - (wx_n + b))$$

$$w = w + \eta \sum_n 2x_n(y_n - (wx_n + b)) - \eta\lambda\vec{1}$$

Perceptron (~~20~~ 10 Points)

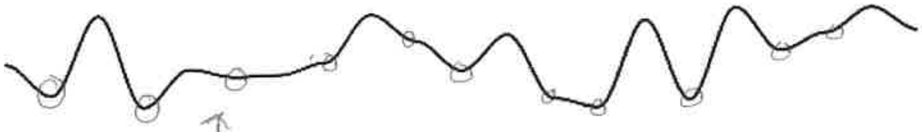
8. Run the perceptron algorithm on the following data in the order provided for two epochs. Give the final w and b produced by the algorithm at the end of the ~~second~~ first epoch.

Sample	x_1	x_2	y
s_1	0	1	1
s_2	1	0	1
s_3	1	1	1
s_4	2	2	-1
s_5	2	1	-1
s_6	1	2	-1

Same as 422 # 8

Gradient Descent (10 Points)

9. Indicate on the following function all the possible locations where we may end up after running gradient descent. Briefly explain your choice(s).

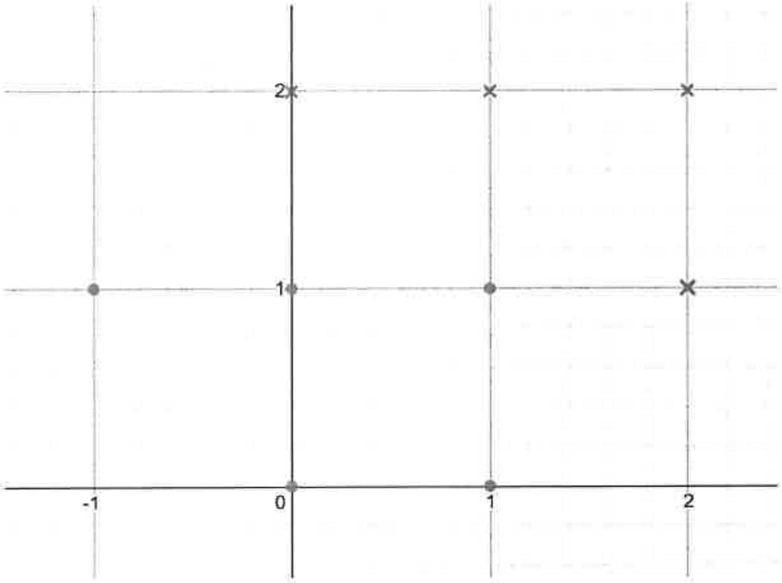


Any local minimum could be the result of running gradient descent.

Even these saddle points

KNN (10 Points)

10. Using K=1, 3 and 5, classify the following test data using the plotted training data (x = negative, filled circle = positive). Use the table below to record your final answers. Indicate ties with the answer +/−.



Sample	$K = 1$	$K = 3$	$K = 5$
(1.1,1)			
(0.5,0.5)			
(0,1.5)			
(2,0)			

Same as 422 #10.