CS390DM0 Spring 2017 Midterm 2 solutions, 11 April, 2017
Prof. Chris Clifton

Turn Off Your Cell Phone. Use of any electronic device during the test is prohibited. As previously noted, you are allowed notes: Up to two sheets of 8.5x11 or A4 paper, single-sided (or one sheet double-sided).

Time will be tight. If you spend more than the recommended time on any question, **go on to the next one**. If you can't answer it in the recommended time, you are either giving too much detail or the question is material you don't know well. You can skip one or two parts and still demonstrate what I believe to be an A-level understanding of the material.

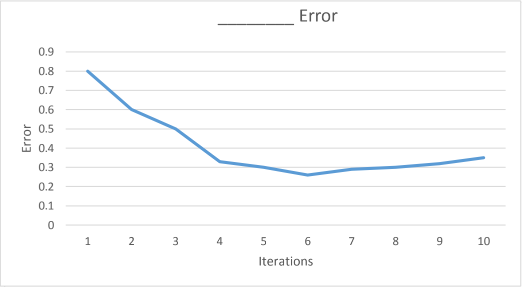
Note: It is okay to abbreviate in your answers, as long as the abbreviations are unambiguous and reasonably obvious.

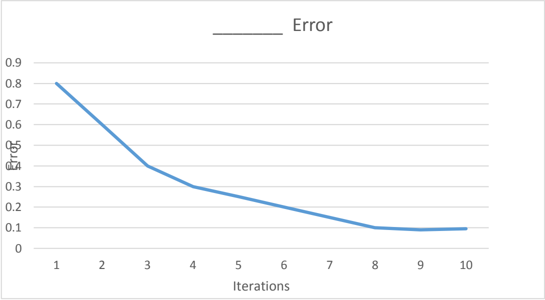
In all cases, it is important that you give some idea of how you derived the answer, not simply give an answer. Setting up the derivation correctly, even if you don't carry out the calculations to get the final answer, is good for nearly full credit.

This was a difficult exam (intentionally so). I would expect an A student to score at least 30, and a B student to score at least 22. I would expect a minimum "ready to go on" (C) to be in the neighborhood of 13.

1 Learning curve (4 minutes, 4 points)

The following graphs represent the learning and validation error curves for an iterative learning method that is guaranteed to converge to a local minimum.

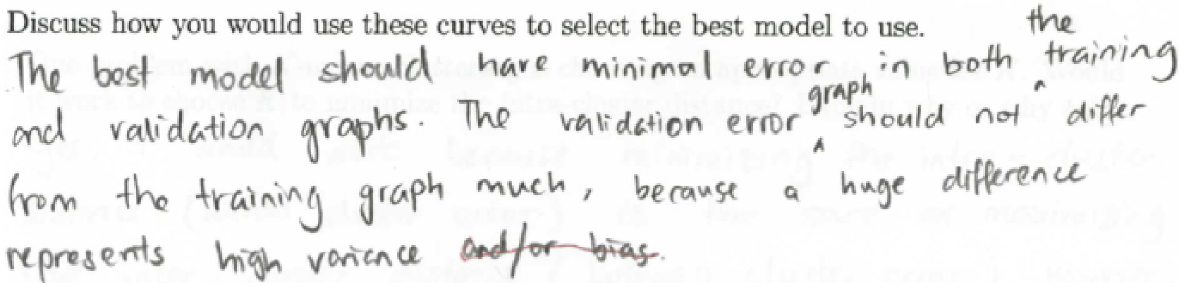




A. Correctly label the graphs as "training error" and "validation error" (i.e., which is which.)

   Left is validation, right is training. A method that "converges to a local minimum" wouldn't have the training error increase. *1 for each correct*

B. Discuss how you would use these curves to select the best model to use.



Discuss how you would use these curves to select the best model to use.

The best model should have minimal error in both the training and validation graphs. The validation error graph should not differ from the training graph much, because a huge difference represents high variance ~~and/or bias~~.

Courtesy of a student who prefers to be anonymous.

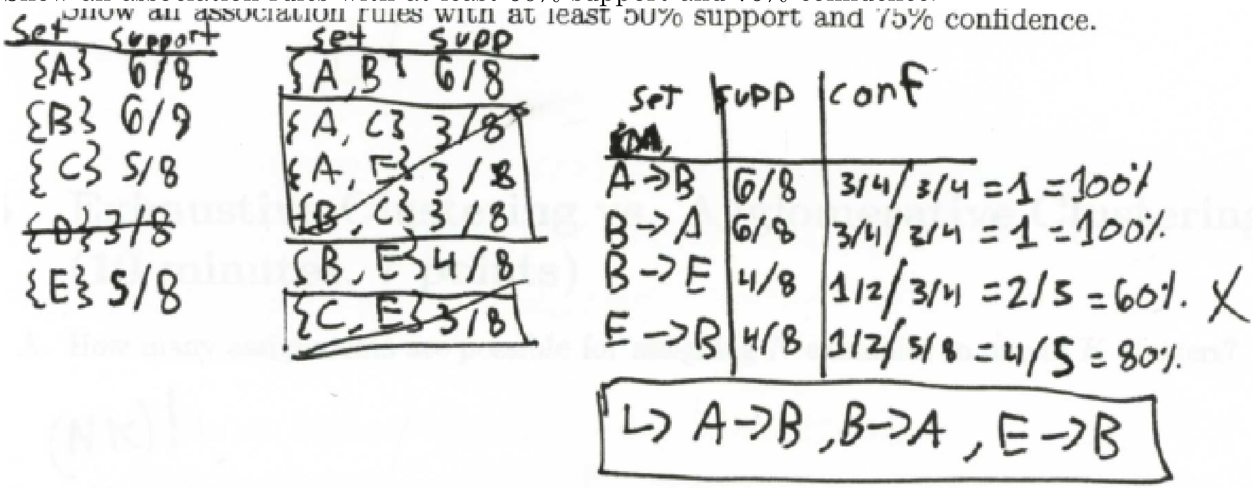
1 for discussing flattening of training error, 2 for discussing minimum validation error.

2 Association Rules (4 minutes, 4 points)

Given the following data:

Trans ID	*A*	*B*	*C*	*D*	*E*
1	+	+	0	+	+
2	+	+	+	0	+
3	+	+	+	0	0
4	+	0	+	0	0
5	0	+	+	+	+
6	0	0	+	+	+
7	+	+	0	0	+
8	+	+	0	0	0

Show all association rules with at least 50% support and 75% confidence.



Show all association rules with at least 50% support and 75% confidence.

Set	Support
{A}	6/8
{B}	6/9
{C}	5/8
~~{D}~~	~~3/8~~
{E}	5/8

Set	Supp
{A, B}	6/8
~~{A, C}~~	~~3/8~~
~~{A, E}~~	~~3/8~~
~~{B, C}~~	~~3/8~~
{B, E}	4/8
~~{C, E}~~	~~3/8~~

Set	Supp	conf
A->B	6/8	3/4/3/4 = 1 = 100%
B->A	6/8	3/4/3/4 = 1 = 100%
B->E	4/8	1/2/3/4 = 2/5 = 60%. X
E->B	4/8	1/2/5/8 = 4/5 = 80%

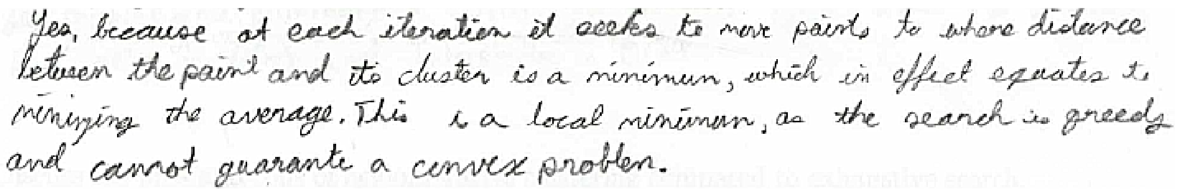
$\boxed{\hookrightarrow A \to B, B \to A, E \to B}$

Courtesy of Nicholas Semenza.

1 for getting minsupport items, 1 for all minsupport itemsets, 1 for some evidence of confidence, 1 for all correct

3 *K*-Means Clustering (6 minutes, 4 points)

A. Does K-means clustering minimize the average distance between a cluster center and points in the cluster? If so, explain why and note if this is a local or global minimum. If not, explain why not and what K-means does optimize.

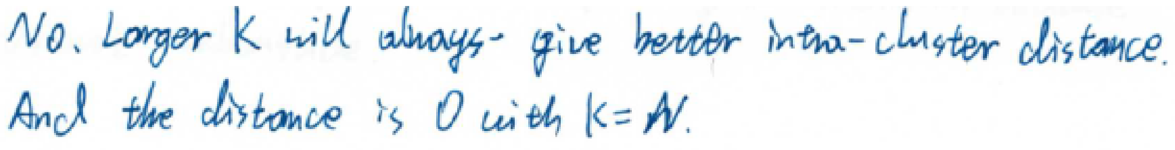


Yes, because at each iteration it seeks to move points to where distance between the point and its cluster is a minimum, which in effect equates to minimizing the average. This is a local minimum, as the search is greedy and cannot guarante a convex problen.

Courtesy of a student who prefers to be anonymous.

1 for local minimum, 1 for decent explanation

B. One problem with K-means clustering is choosing an appropriate value for K. Would it work to choose K to minimize the intra-cluster distance? Explain why or why not.



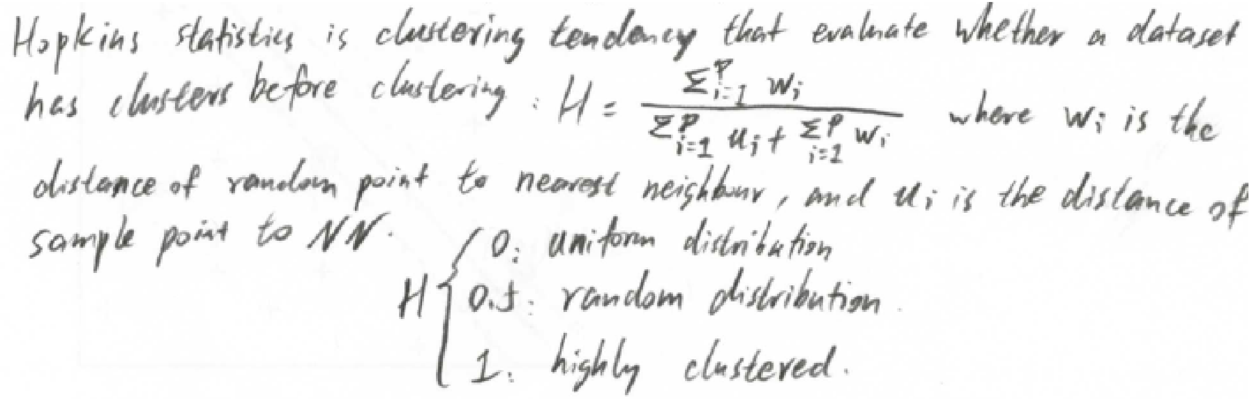
No. Larger K will always give better intra-cluster distance. And the distance is 0 with K = N.

Courtesy of Jingyang ZHANG.

1 for "no", 1 for explanation of always leads to K clusters

4 Clustering (4 minutes, 2 points)

Given a dataset, we would like to know if it makes sense to try to cluster it. Describe a metric that would be appropriate for measuring how likely we are to find good clusters (the metric need not be useful for actually finding clusters.)



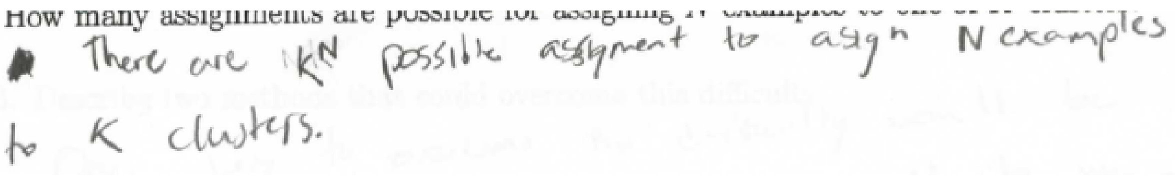
Hopkins statistics is clustering tendency that evaluate whether a dataset has clusters before clustering : $H = \frac{\Sigma_{i=1}^{P} w_i}{\Sigma_{i=1}^{P} u_i + \Sigma_{i=1}^{P} w_i}$ where w_i is the distance of random point to nearest neighbour, and u_i is the distance of sample point to NN.

$$H \begin{cases} 0: & \text{uniform distribution} \\ 0.5: & \text{random distribution.} \\ 1: & \text{highly clustered.} \end{cases}$$

Courtesy of a student who prefers to be anonymous.
1 for an appropriate metric, 1 for formula or description

5 Exhaustive Clustering vs. Agglomerative Clustering (10 minutes, 7 points)

A. How many assignments are possible for assigning N examples to one of K clusters?



How many assignments are possible for assigning

There are K^N possible assignment to asign N examples to K clusters.

Courtesy of Avnish Bablani.

2 for K^N, 1 for other exponential

B. What is the computational complexity of agglomerative clustering?

N^3 for brute force way, but by using union-find / priority queue, we can make this $N^2 \log N$, N is examples.

Courtesy of Tarang B Khanna.

1 for quadratic, 2 for cubic, 1 for discussion or better answer

C. Discuss the pros and cons of agglomerative clustering compared to exhaustive search.

Exhaustive search is guaranteed to give a global optima, however it is very computationally tasking.

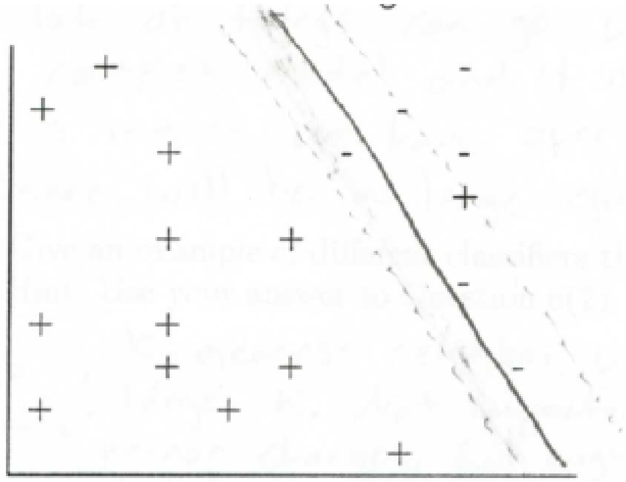
Agglomerative clustering gives a "good-enough" result w/o using up too much computational resources.

Courtesy of Derek Uche.

1 for noting substantial speed difference, 1 for may not find optimal

6 Support Vector Machine (10 minutes, 5 points)

Assume that you are trying to learn a Support Vector Machine classifier to separate the + and - classes of the following 2-dimensional dataset.

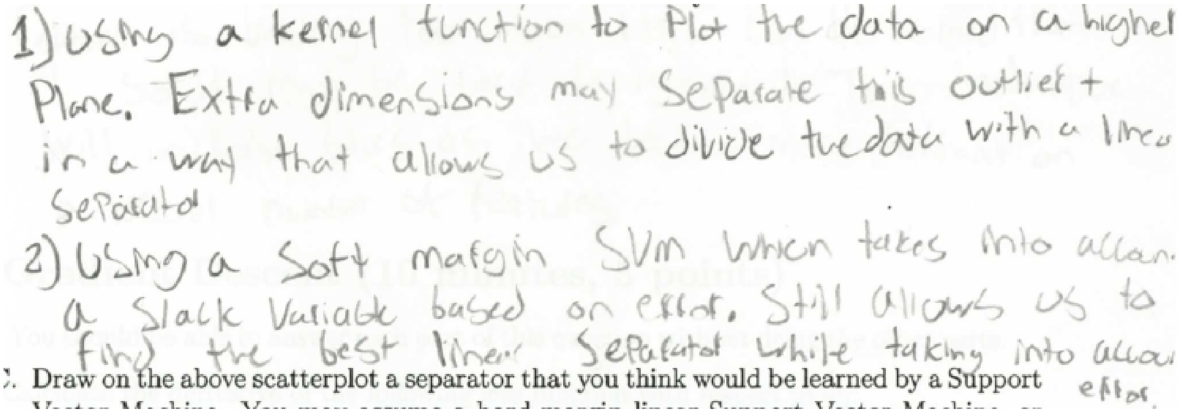


A. What would be the difficulty with learning a linear hard margin Support Vector Machine from this dataset?

The data isn't linearly separable, a hard margin SVM will be unable to find a linear separator.

1 for no linear separator

B. Describe two methods that could overcome this difficulty.



1) Using a kernel function to Plot the data on a higher Plane. Extra dimensions may Separate this outlier + in a way that allows us to divide the data with a linear Separator

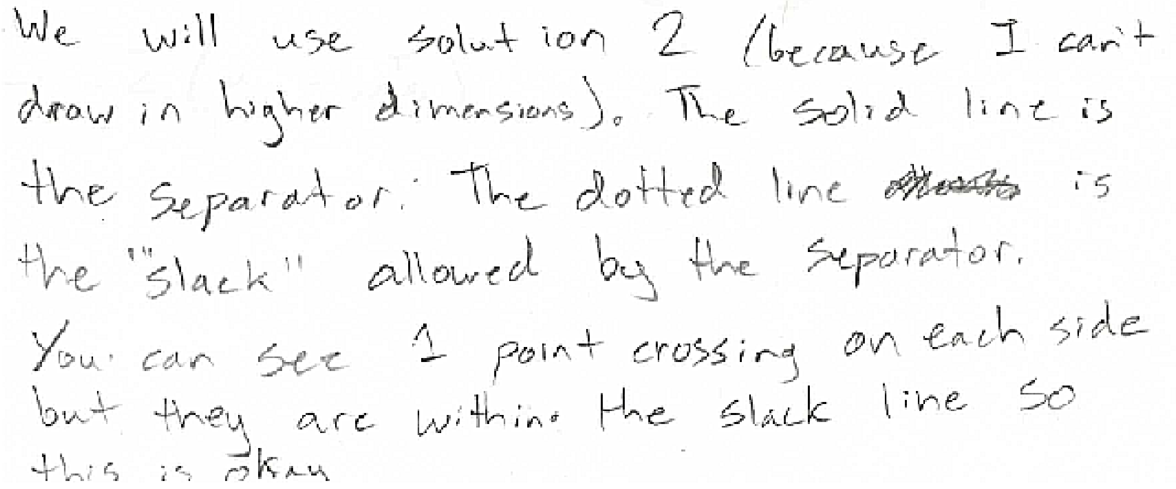
2) Using a Soft margin SVM which takes into accoun. a Slack Variable based on error. Still allows us to find the best linear Separator while taking into accou error.

. Draw on the above scatterplot a separator that you think would be learned by a Support

Courtesy of a student who prefers to be anonymous.

1 each, presumably soft margin and kernel method.

C. Draw on the above scatterplot a separator that you think would be learned by a Support Vector Machine. You may assume a hard-margin linear Support Vector Machine, or either of the two methods you describe in Part B, but specify which you assume.



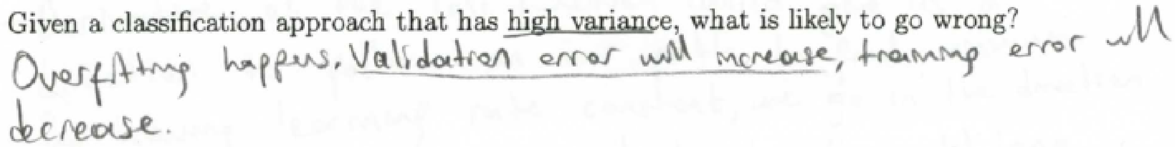
We will use solution 2 (because I can't draw in higher dimensions). The solid line is the separator. The dotted line ~~[illegible]~~ is the "slack" allowed by the separator. You can see 1 point crossing on each side but they are within the slack line so this is okay

Courtesy of a student who prefers to be anonymous.

1 for reasonable separator, 1 for matching their description

7 Bias/Variance (6 minutes, 5 points)

A. Given a classification approach that has high variance, what is likely to go wrong?



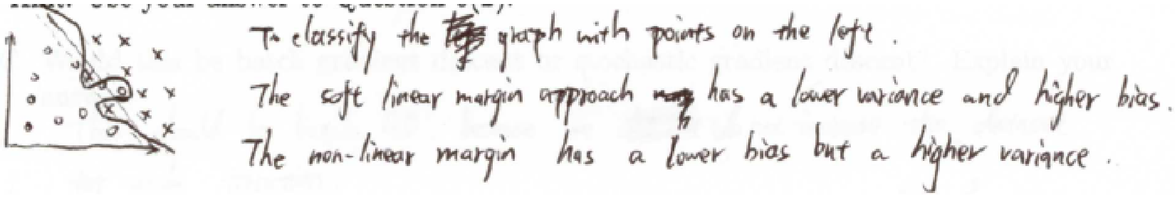
Given a classification approach that has high variance, what is likely to go wrong?

Overfitting happens. Validation error will increase, training error will decrease.

Courtesy of Muhammed Onus.

1 overfit

B. Give an example of different classifiers that have different levels of Bias and Variance. Hint: Use your answer to Question 6(B).



To classify the graph with points on the left.

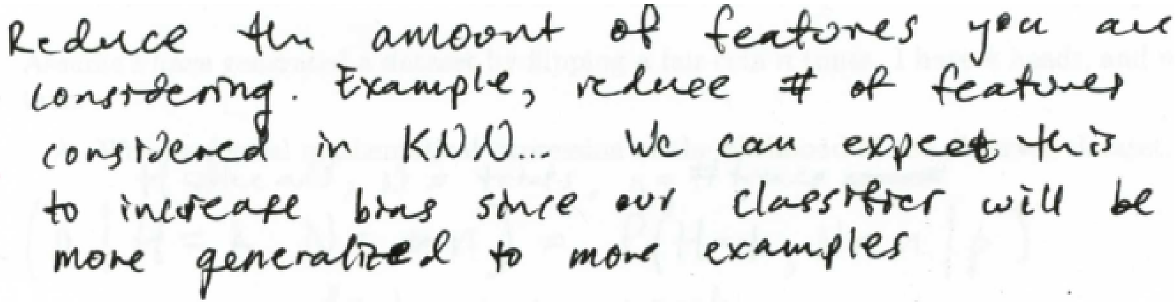
The soft linear margin approach ~~may~~ has a lower variance and higher bias.

The non-linear margin has a lower bias but a higher variance.

Courtesy of Yilang Fan.

1: More complex classifiers typically have lower bias but higher variance. 1: Example/explanation

C. Describe a simple approach to reducing variance that will work with any type of classifier. What would you expect this to do to bias?



Reduce the amount of features you are considering. Example, reduce # of features considered in KNN... We can expect this to increase bias since our classifier will be more generalized to more examples

Courtesy of Hans Allendorfer.

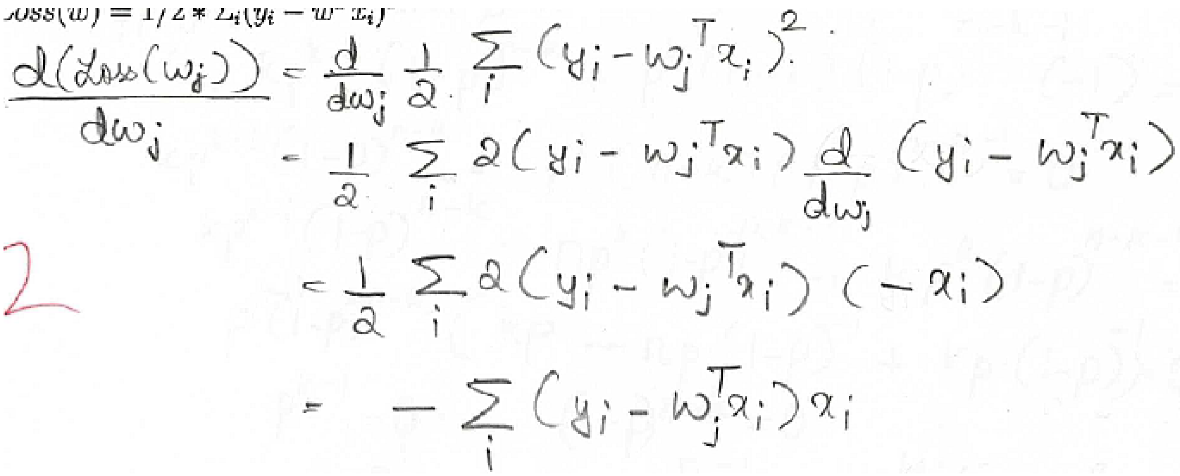
1 for method, 1 for correctly noting impact on bias

8 Gradient Descent (10 minutes, 5 points)

Note: You should be able to answer each part of this question without doing the other parts.

A. Calculate the derivative of the following loss function with respect to w_j:

$Loss(w) = 1/2 * \Sigma_i (y_i - w^T x_i)^2$



$Loss(w) = 1/2 * \Sigma_i(y_i - w^T x_i)$

$$\frac{d(Loss(w_j))}{dw_j} = \frac{d}{dw_j}\frac{1}{2}\sum_i (y_i - w_j^T x_i)^2$$

$$= \frac{1}{2}\sum_i 2(y_i - w_j^T x_i)\frac{d}{dw_j}(y_i - w_j^T x_i)$$

2

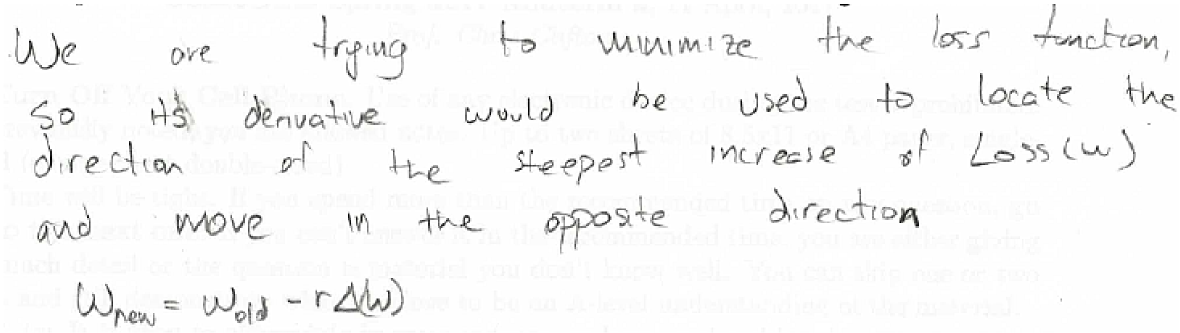
$$= \frac{1}{2}\sum_i 2(y_i - w_j^T x_i)(-x_i)$$

$$= -\sum_i (y_i - w_j^T x_i)x_i$$

Courtesy of Gouthami Kamalnath.

1 for showing some idea, 1 for correct.

B. Explain how the loss function and its derivative would be used in gradient descent.



We are trying to minimize the loss function, so its derivative would be used to locate the direction of the steepest increase of Loss(w) and move in the opposite direction

$W_{new} = W_{old} - r\Delta W$

Courtesy of Lev Zemlyanov.

1 for move in direction of greatest decrease, 1 for nothing that derivative gives us this direction.

C. Would this be batch gradient descent or stochastic gradient descent? Explain your answer.

Batch, because the loss function includes all items, so each update step is based on all items. *1 for correct answer and explanation*

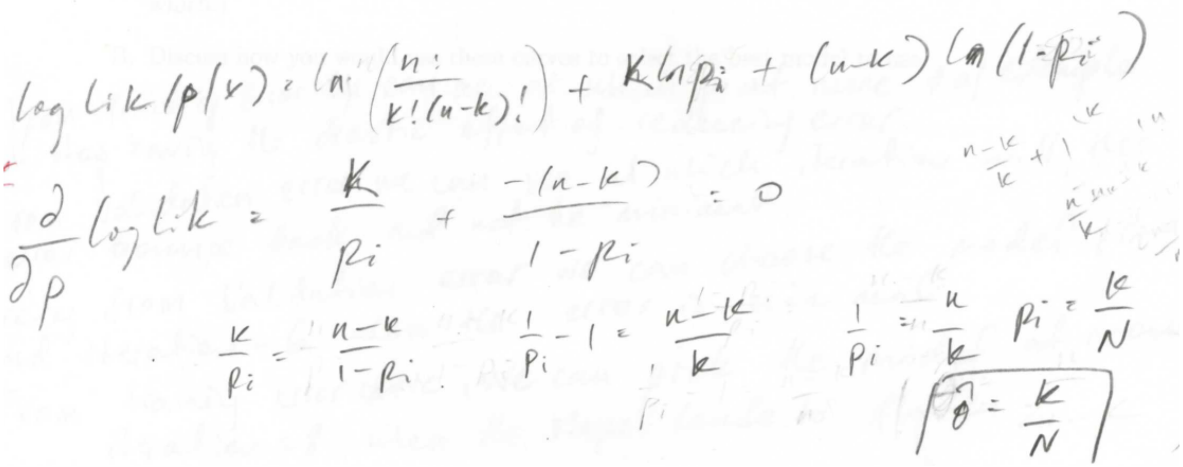
9 Maximum Likelihood Estimation (15 minutes, 4 points)

Assume I have generated a dataset by flipping a fair coin n times. I have k heads, and $n - k$ tails.

A. Write a formal mathematical expression of the likelihood of the observed dataset.

1 for some idea (e.g., formula for particular sequence), 1 for correct

B. Apply maximum likelihood estimation to estimate the model.



$$\log \mathrm{Lik}(p|x) = \ln\left(\frac{n!}{k!(n-k)!}\right) + k\ln p_i + (n-k)\ln(1-p_i)$$

$$\frac{\partial}{\partial p}\log \mathrm{Lik} = \frac{k}{p_i} + \frac{-(n-k)}{1-p_i} = 0$$

$$\frac{n-k}{k}+1$$

$$\frac{n}{k}$$

$$\frac{k}{p_i} = \frac{n-k}{1-p_i} \qquad \frac{1}{p_i} - 1 = \frac{n-k}{k} \qquad \frac{1}{p_i} = \frac{n}{k} \qquad p_i = \frac{k}{N}$$

$$\boxed{\hat{\theta} = \frac{k}{N}}$$

Courtesy of Jantsankhorloo Amgalan.

1 for idea, 1 for correct