CS273a Final Exam

Introduction to Machine Learning: Fall 2013

Thursday December 12th, 2013

Your name: SOLUTIONS

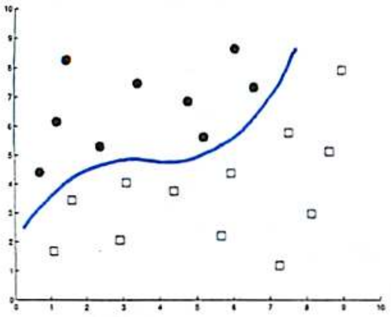
Your UCInetID (all caps):

Your Seat (row <u>and</u> number):

- Total time is 1:50. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Closed book; one page of (your own) notes
- Please **write clearly** and **show all your work.**
- If you need clarification on a problem, please raise your hand and wait for the instructor to come over.
- Turn in any scratch paper with your exam.

Problem 1: Separability

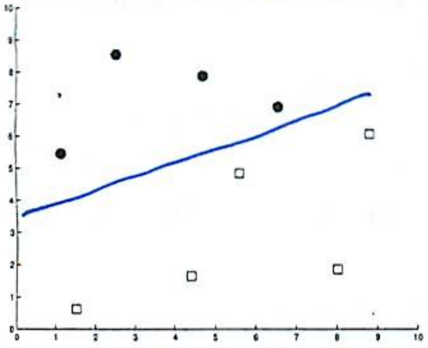
For each of the following examples of training data, sketch a classification boundary that separates the data. State whether or not the data are linearly separable, and if not, give a set of features that would allow the data to be separated.



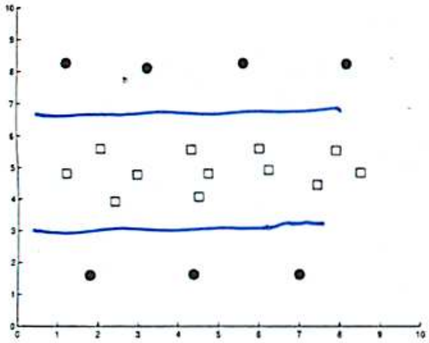
No, not linearly separable.

Based on the curve, $x_2 = a + bx_1 + cx_1^2 + dx_1^3$

the features $[1\ x_1\ x_1^2\ x_1^3\ x_2]$ should work.



Yes, linearly separable.



No, not linearly separable.

The decision bdy$_4$) can be obtained as

$ax_2^2 + bx_2 + c = \emptyset$

$\Rightarrow [1\ x_2\ x_2^2]$ is enough.

Problem 2:

Select the best choice to complete each statement.

Increasing the number of hidden nodes in a neural network will most likely
(increase [✓ decrease] not change) the bias.

Decreasing regularization on the weights in logistic regression will most likely
(increase decrease [✓ not change]) the VC dimension.

Increasing the amount of data will most likely
(increase [✓ decrease] not change) the variance.

Increasing the regularization on a perceptron will most likely
([✓ increase] decrease not change) the bias.

Decreasing the maximum depth of a decision tree will most likely
(increase [✓ decrease] not change) the VC dimension.

Increasing the depth of a decision tree will most likely
(increase [✓ decrease] not change) the bias.

Possible argument for no change - if depth is already very large compared to the amt of data

The predictions of a k-nearest neighbor classifier
([✓ will] will not) be affected by pre-processing to normalize the data.

Reducing the number of features using PCA will most likely
(increase [✓ decrease] not change) the variance.

Linear regression
([✓ can] cannot) be solved using either matrix algebra or gradient descent.

The predictions of a regression tree
(will [✓ will not]) be affected by pre-processing to normalize the features.

Problem 3: Regression

Suppose that we train a *non-linear* regression model on m data, where our prediction is

$$\hat{y}(x) = a + \exp(bx)$$

for two scalar parameters a, b.

(a) Write down the formula for the mean-squared error on the training data, and compute its gradient with respect to the parameters.

$$MSE = J(a,b) = \frac{1}{m} \sum_i \left(y^i - \hat{y}(x^i)\right) = \frac{1}{m} \sum_i \left(y^i - a - \exp(bx^i)\right)^2$$

$$\nabla J = \left[\frac{\partial J}{\partial a} \quad \frac{\partial J}{\partial b} \right]$$

$$\frac{\partial J}{\partial a} = \frac{1}{m} \sum_i \left(y - \hat{y}(x^i)\right)(-1). \qquad (1)$$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_i \left(y^i - \hat{y}(x^i)\right) \cdot (-1) \exp(b x^i) \cdot b. \qquad (2)$$

(b) Give pseudo-code for (batch) gradient descent on this problem. Be sure to specify initialization, the update itself (in enough detail to enable coding), and a stopping condition (again, in enough detail to enable coding).

```
Init a,b :   $a = \emptyset$.  $b = \emptyset$.     $a' = \inf$  $b' = \inf$.     $\alpha$ = stepsize.
while ($\neg$done)
    for $i = 1..m$,   $\hat{y}(x^i) = a + \exp(b x^i)$
    $a \leftarrow a - \alpha \frac{\partial J}{\partial a}$      as in (1)
                                              % Take a step (could update stepsize)
    $b \leftarrow b - \alpha \frac{\partial J}{\partial b}$      as in (2)
    done = $[\,(a-a')^2 + (b-b')^2 < \epsilon\,]$      % check for convergence
    $a = a'$;  $b = b'$;                            % save old values
```

(c) Give at least one advantage of batch gradient descent over stochastic gradient descent. In contrast, when would using stochastic gradient be more appropriate?

Batch - always a descent on J (for sufficiently small α)

$\Rightarrow$ easy to debug, easy to assess convergence, monotonic.

SGD - more appropriate than batch when m is very large (many data)

- in this case, batch updates will be very slow. (all data processed before each step).

Problem 4: Naïve Bayes

Consider the following table of measured data:

x_1	x_2	x_3	y
1	1	1	0
1	1	0	0
0	1	0	0
1	0	1	0
1	1	1	1
0	1	1	1
0	0	1	1

We will use the three observed features x_1, x_2, x_3 to predict class y. In the case of a tie, we will prefer to predict class $y = 1$.

(a) Write down the probabilities necessary for a naïve Bayes (NB) classifier:

$\Pr[y=1] = 3/7$

$\Pr[x_1=1 \mid y=0] = 3/4$ $\quad\Pr[x_2=1 \mid y=0] = 3/4$ $\quad\Pr[x_3=1 \mid y=0] = 1/2$

$\Pr[x_1=1 \mid y=1] = 1/3$ $\quad\Pr[x_2=1 \mid y=1] = 2/3$ $\quad\Pr[x_3=1 \mid y=1] = 1.$

(b) Using your NB model, what value of y is predicted given observation $(x_1, x_2, x_3) = (000)$.

Predict $y = \emptyset$. $\quad$ ($y=1$ has $\Pr[x_3=0 \mid y=1] = \emptyset$).

(c) Using your NB model, what is the probability $p(y = 1|x_1 = 1, x_2 = 1, x_3 = 1)$?

$$= \frac{3/7 \cdot 1/3 \cdot 2/3 \cdot 1}{(") + 4/7 \cdot 3/4 \cdot 3/4 \cdot 1/2} = \frac{6/9}{6/9 + 9/8} = \frac{48}{48+81} = \frac{48}{129}$$

(d) Using your NB model, what is the probability $p(y = 1|x_1 = 0)$?

$$= \frac{3/7 \cdot 2/3}{(") + 4/7 \cdot 1/4} = \frac{2/7}{2/7 + 1/7} = 2/3 .$$

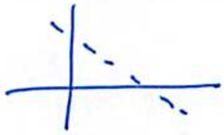
Problem 5: Perceptrons and VC Dimension

In this problem, consider the following perceptron model on two features:

$$\hat{y}(x) = \text{sign}(\,w_0 + w_1 x_1 + w_2 x_2\,)$$

and answer the following questions about the decision boundary and the VC dimension.

(a) Describe (in words, with diagrams if desired) the decision boundaries that can be realized by this classifier



Any hyperplane - eg, decision boundary is an arbitrary line.

(b) What is its VC dimension?

3 - VC dim of a perceptron is $d+1$, and $d=2$.

Now suppose that I also enforce an additional condition on the parameters of the model: that **at most two** of the weights w_i are non-zero (so, at least one weight is zero).

(c) Describe (in words, with diagrams if desired) the decision boundaries that can be realized by this classifier

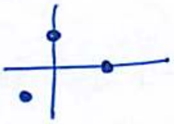
If $w_1 = 0$, horizontal decision

$w_2 = 0$ vertical decision

$w_0 = 0$ line through the origin

Decision boundary can be an arbitrary horizontal or vertical split, or a line through the origin.

(d) What is its VC dimension?

Still 3:  can still be shattered.

(2, can't be higher than part (b))

Finally, I enforce that **at most one** of the weights w_i is non-zero (so, at least two are zero).

(e) Describe (in words, with diagrams if desired) the decision boundaries that can be realized by this classifier

$w_0 \neq 0$ - entire plane same decision

$w_1 \neq 0$ - right vs left half-plane

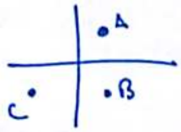
$w_2 \neq 0$ - upper vs lower half plane.

union of these functions.

(f) What is its VC dimension?

Now 2 - can't shatter above example; not possible to ~~get~~ arrange points to get all three ++- patterns...

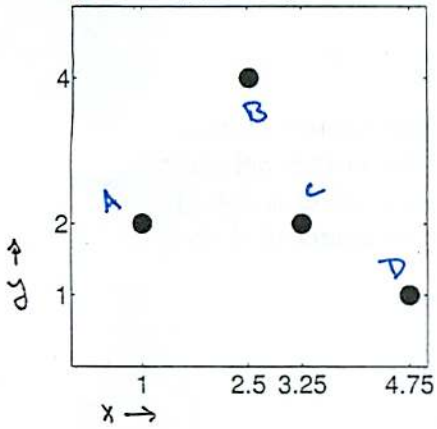
↳ (can't put points on the axes, though).



⇒ can't predict $A = C = +1$, $B = -1$.

Problem 6: Cross-validation

Consider a regression problem for predicting the following data points, using the k-nearest neighbor regression algorithm from class and the homework to minimize mean squared error (MSE). (Note: if you like, you may leave an arithmetic expression, e.g., leave values as "$(.6)^2$".)

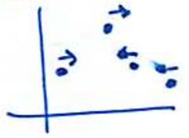


(a) For $k = 1$, compute the training error on the provided data.

ϕ.

(b) For $k = 1$, compute the leave-one-out cross-validation error on the data.

nearest nbr:



$\Rightarrow \frac{1}{4}\left[2^2 + 2^2 + 2^2 + 1^2\right] = 13/4$

(c) For $k = 3$, compute the training error on the provided data.

neighbors:	predns:
A: ABC	8/3
B: ABC	8/3
C: BCD	7/3
D: BCD	7/3

$\Rightarrow \frac{1}{4}\left[(2/3)^2 + (4/3)^2 + (1/3)^2 + (4/3)^2\right] = 37/\cancel{9\cdot 4}\ 36$

(d) For $k = 3$, compute the leave-one-out cross-validation error on the data.

A: BCD	7/3
B: ACD	5/3
C: ABD	7/3
D: ABC	8/3

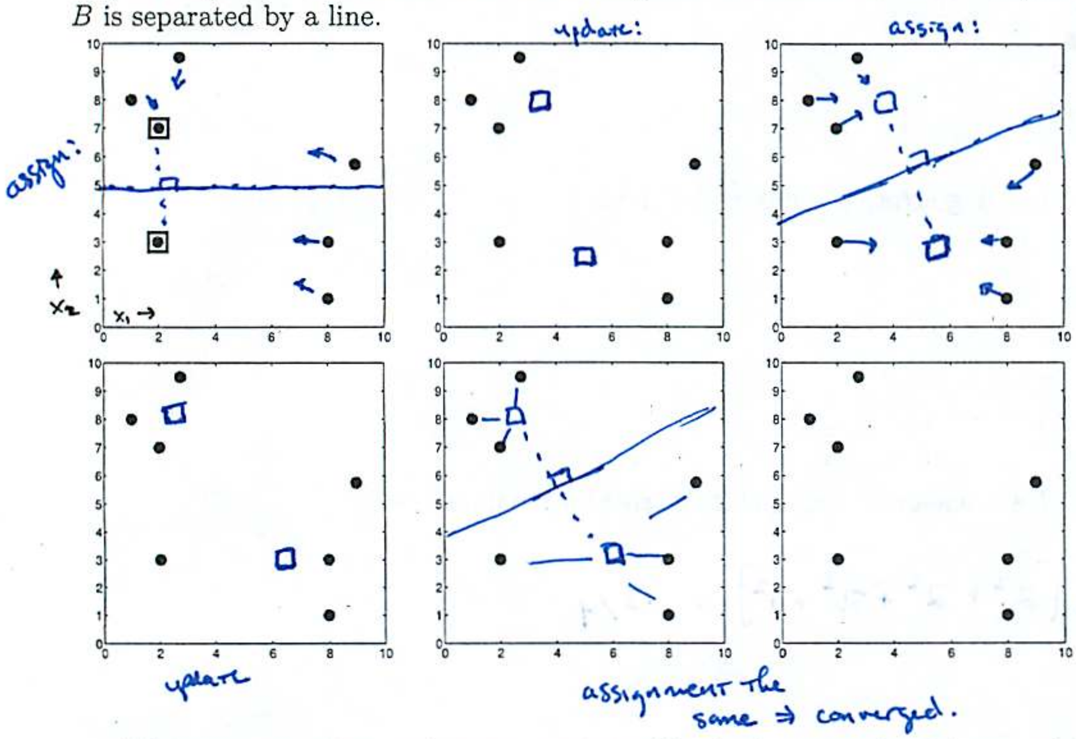
$\Rightarrow \frac{1}{4}\left[(1/3)^2 + (7/3)^2 + (1/3)^2 + (5/3)^2\right] = 76/36$

Problem 7: Clustering

Consider the two-dimensional data points plotted in each panel. In this problem, we will cluster these data using two different algorithms, where each panel is used to show an iteration or step of the algorithm.

k-means

(a) Starting from the two cluster centers indicated by squares, perform k-means clustering on the data points. In each panel, indicate (somehow) the data assignment, and in the next panel show the new cluster centers. Stop when converged, or after 6 steps (3 iterations), whichever is first. It may be helpful to recall from our nearest-neighbor classifier that the set of points nearer to A than B is separated by a line.



assignment the same ⇒ converged.

(b) Write down the cost function optimized by the k-means algorithm, explaining your notation.

Let μ_c be the center of cluster c.

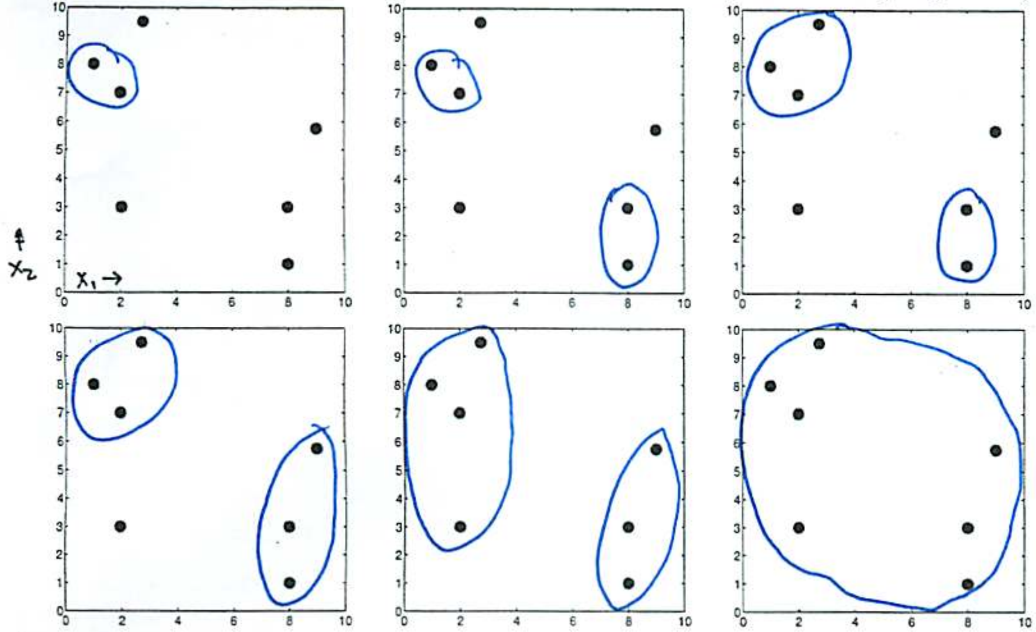
z_i be the cluster assignment of data point i.

$$J = \frac{1}{m} \sum_{i=1}^{m} \|x^{(i)} - \mu_{z_i}\|^2$$

where $\|\cdot\|^2$, is the Euclidean ~~distance~~ distance / length squared.

Linkage

(a) Now execute the hierarchical agglomerative clustering (linkage) algorithm on these data points, using "complete linkage" (maximum distance) for the cluster scores. Stop when the algorithm would terminate, or after 6 steps, whichever is first. Show each step separately in a panel.



Complete linkage ⇒ join nearest pair of clusters, where distance is given by farthest points.

(b) What is the algorithmic (computational) complexity of the hierarchical clustering algorithm? Briefly justify your answer.

$O(n^2)$.

Initial step calculates $\binom{n}{2} = O(n^2)$ pairwise distances

Each of n iterations (one join per step)

requires updating new cluster's distance to remaining clusters

$\Rightarrow (n-2) + (n-3) + (n-4) + \cdots (1) = O(n^2)$

and, $O(n^2) + O(n^2) = O(n^2)$.