10-701: Machine Learning

Midterm Solutions

March 21, 2012

Name: ____

Andrew ID: ____

Instructions:

- Make sure that your exam has 15 pages and is not missing any sheets, then write your full name and **Andrew ID** on this page (and all others if you want to be safe).
- Write your answers in the space provided below the problem. If you make a mess, clearly indicate your final answer.
- The exam has 9 questions, with a maximum score of 100 points (+2 points extra credit). The problems are of varying difficulty. The point value of each problem is indicated.
- This exam is open book and open notes. You may use a calculator, but any other type of electronic or communications device is not allowed.

Question	Points	Score
Probability and Density Estimation	14	
K-Nearest Neighbors	6	
Linear Regression	10	
Logistic Regression	10	
Learning Theory	12	
Decision Trees	14	
SVMs	8	
Neural Networks	14	
Hierarchical Clustering	12 (+ 2)	
Total:	100 (+ 2)	

1 Probability and Density Estimation [14 points]

1. [**2 points**] If two binary random variables X and Y are independent, are $\bar{X}$ ($\bar{X}$ is the complement of X) and Y also independent? Prove your claim.

Yes they are. $P(\bar{X} \cap Y) = P(Y) - P(X \cap Y) = P(Y) - P(X)P(Y) = (1 - P(X))P(Y) = P(\bar{X})P(Y)$.

2. [**6 points**] There are two coins C1 and C2. C1 has a equal prior on a head (H=1) or tail (T=0) and the fate of C2 is dependent on C1. If C1 is a head, C2 will be a head with probability 0.7. If C1 is a tail, C2 will be a head with probability 0.5. C1 and C2 are tossed in sequence once, and the observed sum of the two coins, S = C1 + C2, is 1. What is the probability that C1=T and C2=H (*Hint: use Bayes theorem*)?

$P(C1, C2|S) = P(S|C1, C2)P(C1, C2)/Z = P(S|C1, C2)P(C2|C1)P(C1)/Z$ where Z is the normalization marginalized over all possible values (H or T) of C1 and C2. In this case, $P(C1 = T, C2 = H|S = 1) = P(S = 1|C1 = T, C2 = H)P(C2 = H|C1 = T)P(C1 = T)/Z$. Note there are two cases where $S = 1$, namely C1 is H and C2 is T or C1 is T and C2 is H. The catch is to work out this conditional probability in that sample space when $S = 1$, and work out the marginal Z by summing over all scenarios. This boils down to

$$P(C1 = 0, C2 = 1|S = 1) = \frac{P(S = 1|C1 = 0, C2 = 1)P(C2 = 1|C1 = 0)P(C1 = 0)}{\sum_{C1}\sum_{C2} P(S|C1, C2)P(C2|C1)P(C1)}$$

Since there are only two cases where $S = 1$, this further reduces to

$$P(C1 = 0, C2 = 1|S = 1) = \frac{1 \times .5 \times .5}{1 \times .5 \times .5 + 1 \times .3 \times .5} = \frac{5}{8}$$

3. [**2 points**] We estimate the head probability θ of a coin from the results of N flips. We use pseudo-counts to influence the "fairness" of the coin. This is equivalent to using which distribution as a prior for θ.

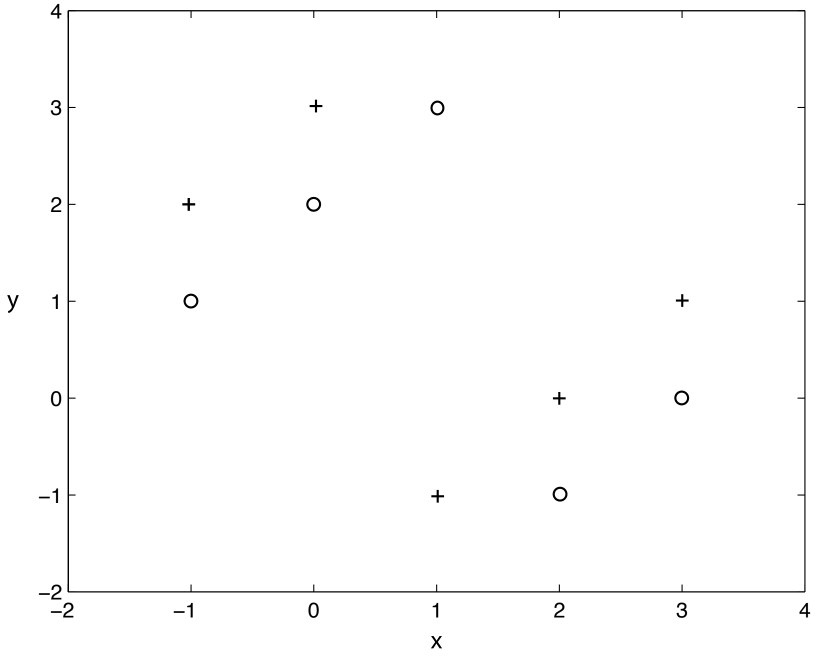
Beta distribution.

4. [**4 points**] Assume we computed the parameters for a Naive Bayes classifier. How can we use these parameters to compute the density P(X) of an observed vector $X = \{x_1, x_2, \ldots, x_n\}$?

Naive Bayes we compute: $P(X_i|\mathcal{C}_j)$ and $P(\mathcal{C}_j)$ for all feature i and class $\mathcal{C}_j$. For density estimation, we need to compute $P(X_1, \ldots, X_N) = \sum_{\mathcal{C}_j} \prod_i P(X_i|\mathcal{C}_j)P(\mathcal{C}_j)$.

2 K-Nearest Neighbors Classification [6 points]

Consider K-NN using Euclidean distance on the following data set (each point belongs to one of two classes: + and ∘).



1. [**2 points**] What is the leave one out cross validation error when using 1-NN?

Every point is misclassified. $10/10 = 1$.

2. [**4 points**] Which of the following values of k leads to the minimum number of validation errors: 3, 5 or 9? What is the error for that k?

All 3 values of k misclassify all of the points and have the same classification error as part 1: $10/10 = 1$.

3 Linear Regression [10 points]

1. [**6 points**] We would like to use the following regression function:

$$y = w^2 x + wx \tag{1}$$

where x is a single value variable. Given a set of training data points $\{(x_i, y_i)\}$ derive a solution for w. Simplify as much as you can.

$$(w^2 + w) = \sum_i \frac{x_i y_i}{x_i^2}$$

Note that this model is equivalent to:

$$y = (w^2 + w)x$$

And so the solution for $(w^2 + w)$ should be exactly the same as the solution for the original regression model discussed in class (model 2 below).

2. We would like to compare the regression model used in (1) to a the following regression model:

$$y = wx$$

(a) [**2 points**] Given limited training data, which model would fit the *training* data better:

i. Model 1
ii. Model 2
iii. Both will fit the data equally well
iv. Impossible to tell

3. As mentioned above, the solution to the optimization problem is the same and so the outcome will be the same as well (any value that can be expressed using w can be expressed using $(w^2 + w)$ and so the two models are exactly the same.

(b) [**2 points**] Given limited training data, which model would fit the *test* data better:

i. Model 1
ii. Model 2
iii. Both will fit the data equally well
iv. Impossible to tell

3. Same reason as above.

4 Logistic Regression [10 points]

1. **[2 points]** Logistic regression is named after the log-odds of success (the logit of the probability) defined as

$$\ln\left(\frac{\mathbb{P}[Y=1 \mid X=x]}{\mathbb{P}[Y=0 \mid X=x]}\right)$$

Show that log-odds of success is a linear function of x.

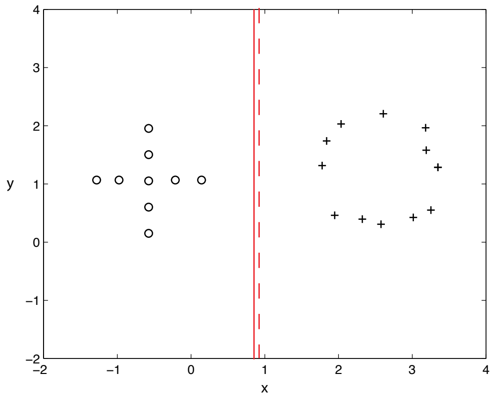
$$\ln\left(\frac{\mathbb{P}[Y=1 \mid X=x]}{\mathbb{P}[Y=0 \mid X=x]}\right) = \ln\left(\frac{\frac{\exp(w_0+w^\top x)}{1+\exp(w_0+w^\top x)}}{\frac{1}{1+\exp(w_0+w^\top x)}}\right) = \ln\big(\exp(w_0+w^\top x)\big)$$

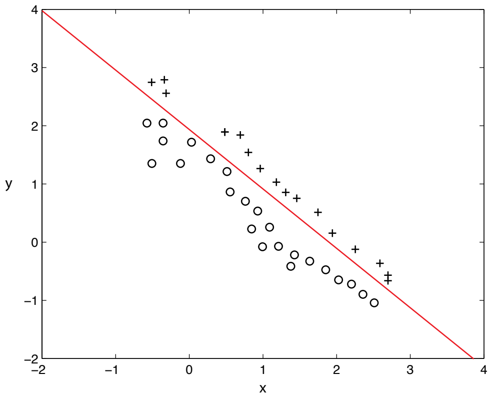
$$= w_0 + w^\top x$$

2. **[2 points]** How do the probabilities of outcomes y change as a function of x in the logistic regression model? How does this limit the classification ability of Logistic regression?

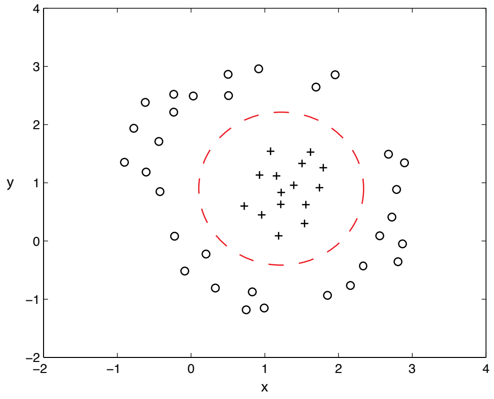
The logistic function means the probabilities are monotonically changing. This constrains the decision boundary to be close to linear. The next problem illustrates some of the limitations of Logistic regression.

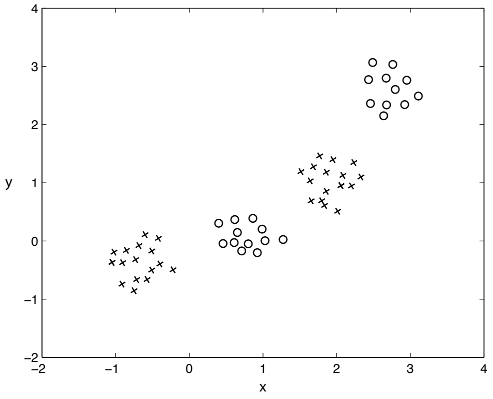
3. **[6 points]** Compare the decision boundaries of *Gaussian Naive Bayes* and *Logistic Regression* on the following data sets. Draw the Logistic Regression decision boundary with a solid line and the Gaussian Naive Bayes Boundary with a dashed line. If one (or both) of the methods cannot classify the data, indicate this fact and write **one sentence** explaining why not.





In the first figure, both methods classify the data correctly. In the second figure, only logistic regression can classify the data correctly. Naive Bayes assumes *uncorrelated features*, This assumption is violated.





In the first figure, there is no linear decision boundary, so logistic regression cannot classify. In the second figure, neither method can classify the data. There is no linear decision boundary and, if we run naive Bayes, the points at the center of each Gaussian would be misclassified.

5 Learning theory [12 points]

Assuming two classes and denoting the number of samples in each class as N and the dimension of the input space as D. Declare or compute the VC dimension of the following classifiers. State your assumptions and **briefly justify your answers**—use drawings if necessary.

1. [**3 points**] A K-nearest neighbor classifier with $K = 1$.

When $K = 1$ a 1NN can correctly classify all data points, hence the VC dimension is infinity.

2. [**3 points**] A single-layer perceptron classifier.

A perceptron is a linear classifier and hence the VC dimension is $D+1$.

3. [**6 points**] Assume $D = 2$. A square that assigns points within as one class and points outside as another class. Draw a scenario where this classifier shatters all points **for the VC dimension you have proposed**.

The VC dimesion is 3. Draw three points in 2D in a standard tripod structure and the square is able to shatter all labeling configurations. Note that a square can't cope with 4 points regardless of how they are placed. A rectangle can shatter 4 points if they are structured in a diamond-like shape.

6 Decision Trees [14 points]

1. We are trying to classify individuals as Males (M) or Females (F) based on weight and height information. The tables below summarize the information we collected for training a model:

Weight	>130	<=130
M	45	5
F	30	10

Weight	>150	<=150
M	30	20
F	10	30

Height	>5 feet	<=5
M	50	0
F	30	10

Weight	>6	<=6
M	15	35
F	0	40

H(3/5) = 0.97

H(1/3) = 0.92

H(3/4) = 0.81

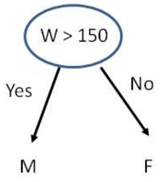
H(4/5)= 0.72

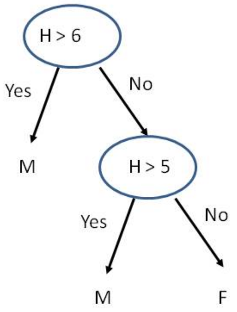
H(3/8) = 0.95

H(7/15) = 0.99

H(1) = 0

(a) **[6 points]** Using these tables, construct two decision tress (one only using weight and other only using height). You may find some of the H values listed above useful for constructing these trees.





(b) **[4 points]** What is the expected error of each of the trees you constructed?

W: 1/3
H: 1/3

(c) **[2 points]** Is there a way to combine the two trees to classify a new input sample?

A way to combine 2 two tress is to use both to classify a new input sample. If they agree, we are done. If they do not, we look at two leafs and count the *total* number of M and F training samples assigned to these two leafs. We output the label that is the most common among the samples.

(d) **[2 points]** Assume we have all the training data summarized in the tables above. We use this to construct a single tree that uses all the data (so basically we have, for each individual, 5 binary values: the label (M/F), whether their height is > 5, whether their height is > 6, whether their weight is > 130 and whether their weight is > 150). Select the most appropriate answer from below and briefly explain. When comparing this tree to the method you proposed in (c) using a test set the most likely outcome is:

i. The single tree would outperform the two trees approach.

ii. The two trees approach will outperform the single tree approach.

iii. Both methods would achieve similar results.

i. In a single tree we can utilize the dependency between the features (even when conditioned on the class label) whereas in combining the two trees we assume independence conditioned on the class label.

7 SVM [8 points]

1. **[1 point]** (True/False) When the data is not completely linearly separable, the linear SVM *without slack variables* returns $\mathbf{w} = 0$.

False, there is no solution.

2. **[1 point]** (True/False) Assume we are using the primal non linearly separable version of the SVM optimization target function. What do we need to do to guarantee that the resulting model is linearly separable?

Set $C = \infty$.

3. **[1 point]** True/False After training a SVM, we can discard all examples which are not support vectors and can still classify new examples.

True.

4. **[3 points]** Show that if $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ are kernels, then $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$ is also a kernel.

Assume ϕ_1 and ϕ_2 are two feature representation of the two kernels such that $k_1(\mathbf{x}, \mathbf{x}') = \phi_1(\mathbf{x})^T\phi_1(\mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}') = \phi_2(\mathbf{x})^T\phi_2(\mathbf{x}')$.

The feature representation of k is $\{\phi_1(x)_i\phi_2(x)_j\}$, for all i, j.

5. **[2 points]** Consider a 2 class classification problem with a dataset of inputs $\{\mathbf{x}_1 = (-1, -1)$, $\mathbf{x}_2 = (-1, +1)$, $\mathbf{x}_3 = (+1, -1), \mathbf{x}_4 = (+1, +1)\}$ and a corresponding set of targets $\{t_1, t_2, t_3, t_4\}$ where $t_i \in \{+1, -1\}$. Using this feature space (no kernel trick), can we build a SVM to perfectly classify this dataset regardless of values of t_i's?

No, since the decision boundary is linear. For examples, we cannot classify in the case $t_1 = +1, t_2 = t_3 = -1, t_4 = +1$.

8 Neural networks [14 points]

1. **[2 points]** (True/False) Increasing the number of layers always decrease the classification error of test data.

False

2. **[12 points]** We are trying to classify samples that only contain binary features. Can the following three classification algorithms be implemented using a feed-forward neural networks for such data, using units that are hard thresholds or linear activation functions? For each say yes / no. If yes, briefly explain how (no need to draw). If no, briefly explain why.

   (a) Naive Bayes with binary features

   (b) Decision Trees with binary features

   (c) 1-NN

(a) For NB: all parameters $P(C)$, $P(X|C)$ are used as parameters of the neural networks. For each class, one hidden unit computes $\log P(X, C) = \sum_i \log P(X_i|C) + \log P(C)$, this can be done because $X_i \in \{0, 1\}$, so $\log P(X_i|C) = \log P(X_i = 0|C)(1 - X_i) + \log P(X_i = 1|C)X_i$. The output unit uses a hard threshold function to output 1 if $\log P(X, C = 1) - \log P(X, C = 0) \geq 0$, and 0 otherwise. Note that the neural networks can only compute the log likelihoods because computing the likelihoods requires multiplying inputs, which is impossible using the given activation functions.

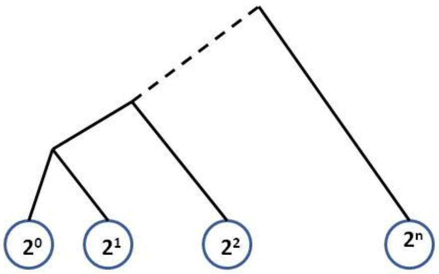
(b) For DT with only binary features. We can write the DT in terms of a logic sentence or furthermore in a conjunctive normal form sentence. And we already seen in the hw3, that neural networks can compute this logic sentence.

(c) Each internal node in the first layer is one input, and then we have a set of internal node to compute the distance between the input example with all examples in the data. Then we have a sets of internal nodes to find the smallest such distance and output the class of the closest example in the data.

9 Hierarchical Clustering [12 points + 2 points extra credit]

1. [**6 points**] Assume we are trying to cluster the points $2^0, 2^0, 2^1, 2^2, \ldots, 2^n$ (a total of n points where $n = 2^N$) using hierarchical clustering.
   If we are using Euclidian distance, draw a sketch of the hierarchical clustering tree we would obtain for each of the linkage methods (single, complete and average)

   Single link tree:



   Complete link tree:

   Same.

   Average link tree:

   Same.

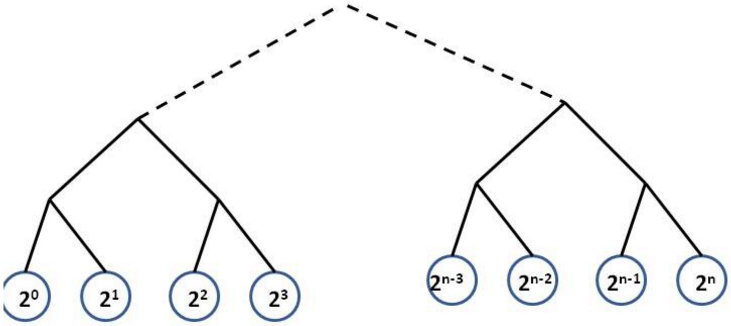
2. **[6 points]** Now assume we are using the following distance function: $d(A,B) = \max(A,B)/\min(A,B)$. Which of the linkage methods above will result in a *different* tree from the one obtained in (1) when using this distance function? If you think that one or more of these methods will result in a different tree, sketch the new tree as well.

   (a) Single link

   (b) Complete link

   (c) Average link

a. Single link does not change.
b. Complete link changes to the graph below.
c. Average link changes to the graph below.



3. Extra Credit: **[2 points]** We would like to use a hierarchical clustering tree as a decision tree. We select a linkage method and build the tree from the training data. For a test sample, at each node in the tree we compute the distance between the test sample and the *sub-cluster rooted at that node* using the linkage method we selected. This way each test sample is propagated down the tree. Specify a linkage method that, when using such method, will lead to a hierarchical clustering decision tree that is exactly the same as a classifier we discussed in class. What is this classifier? Be specific.

If we use single link, each sample will be routed to the nearest sample in our training set leading to a 1-NN classifier.