Solutions

10-601 Machine Learning **Name:** _____
Spring 2017
Mock Final Exam
05/04/2017
Time Limit: 60min Andrew ID _____

Disclaimer: This mock exam is designed to give you practice with problem-solving in a timed environment similar to the final exam. The material included in this mock exam is not comprehensive. You should expect topics and problem types to appear on the real final that are not included on the mock exam.

Instructions:

- Fill in your name and Andrew ID above. Be sure to write neatly, or you may not receive credit for your exam.
- This exam contains 16 pages (including this cover page) and 7 questions.
  The total number of points is 100.
- Clearly mark your answers in the allocated space **on the front of each page.** If needed, use the back of a page for scratch space, but you will not get credit for anything written on the back of a page. If you have made a mistake, cross out the invalid parts of your solution, and circle the ones which should be graded.
- Look over the exam first to make sure that none of the 16 pages are missing. The problems are of varying difficulty, so you may wish to pick off the easy ones first.
- You have 60 min to complete the exam. Good luck!

Question	Points
1	16
2	12
3	16
4	18
5	16
6	12
7	10
Total	100

1 K-Means [16 pts]

1. **[3 pts]** Which of the following statements is true about clustering using Lloyd's algorithm? **No justification necessary.**

   (a) Lloyd's algorithm always converges to a global optimum.

   (b) The choice of initialization greatly affects the performance of the algorithm.

   (c) With random initialization, increasing the selected number of clusters k makes it more likely that the initialization is better.

   (d) Lloyd's method is proven to converge only if initialized with k-means++.

   Solution: b.

2. **[3 pts]** Say you want to use k-means clustering to partition N data points, where $1 < k < N$. Which of the following is a sensible way to pick the value of k? **Select all that apply. No justification necessary.**

   (a) Choose k such that the k-means cost function is minimized on held out data.

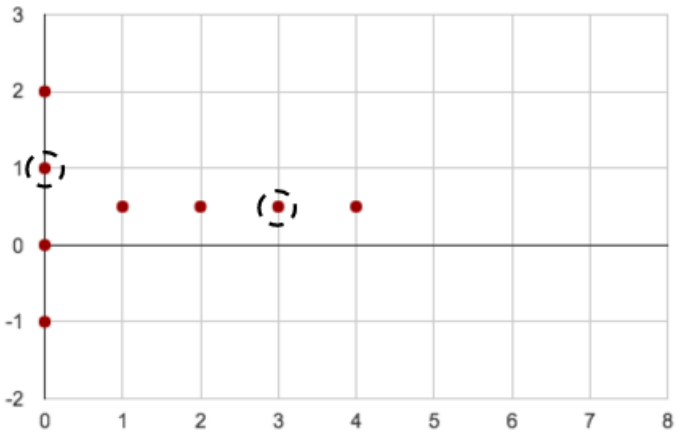
   (b) Choose k such that $k, k+1, \ldots, N$ have similar values of the cost function but that $1, \ldots, k-1$ have much higher costs, where all costs are evaluated from training data.

   (c) Set $k = \frac{n}{2}$.

   (d) Set $k = \log(\frac{n}{2})$.

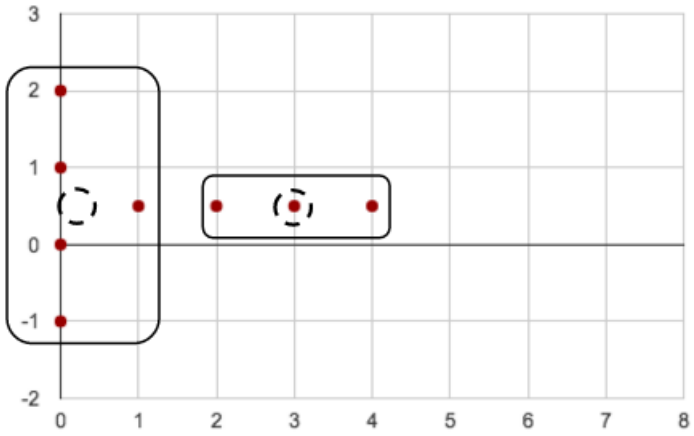
   Solution: a and b.

3. **[5 pts]** Consider the data given below.

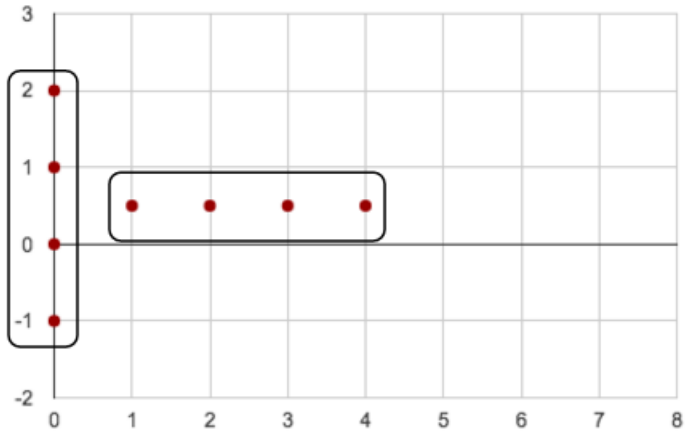


   If k-means clustering (k = 2) is initialized with the two points whose coordinates are (0, 1) and (3, 0.5), indicate the final clusters obtained along with the cluster centers (after the algorithm converges). Draw in the figure itself.

Solution -



4. **[5 pts]** Refer to the same figure as the previous question. Is there a specific cluster center initialization possible such that the 4 horizontal and 4 vertical points are assigned to different clusters after convergence as shown below?



(a) Yes (If you select this option, then mark one such initialization in the figure).

(b) No

Solution: No.
Notice that the assignment shown will change in one iteration of k-means.

2 Expectation Maximization [12 pts]

1. [**4 pts**] **True or False:** Iterating between the E-step and M-step will always converge to a local optimum of the parameter being estimated (which may or may not also be a global optimum)? **Briefly justify your answer**

   (a) True

   (b) False

   Solution: True, the lower bound increases on each iteration.

2. [**4 pts**] In the context of k-means clustering, characterize the E-step of Lloyd's algorithm.

   Solution: The E-step of K-means finds clusters by assigning each object o in the dataset to the cluster with the nearest centroid to o.

3. [**4 pts**]Suppose we clustered a set of N data points using two different clustering algorithms: k-means and a Gaussian Mixture Model (GMM). In both cases we obtained 5 clusters and in both cases the centers of the clusters are exactly the same. Can 3 points that are assigned to different clusters in the K-means solution be assigned to the same cluster in the Gaussian mixture solution? Explain your answer in 1 or 2 sentences or sketch an example.

   Solution: Yes, k-means assigns each data point to a unique cluster based on its distance to the cluster center. Gaussian mixture clustering gives soft (probabilistic) assignment to each data point. Therefore, even if cluster centers are identical in both methods, if Gaussian mixture components have large variances (components are spread around their center), points on the edges between clusters may be given different assignments in the Gaussian mixture solution.

3 Principal Component Analysis [16 pts]

1. **[3 pts]** For each of the following questions, circle **T** or **F**. **No justification required**.

   (a) **T or F** The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting an output variable. Solution: (F)

   (b) **T or F** The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation. Solution: (F)

   (c) **T or F** Subsequent principal components are always orthogonal to each other. Solution: (T)

2. **[4 pts]** Principal Component Analysis (PCA) is often applied to reduce the dimensionality of input data to avoid overfitting, particularly in case of small datasets. What are the two properties that the dataset must satisfy before PCA is applied?

   (a) The number of values in each bin should be the same if equal interval binning is done along any axis.

   (b) The data is centered i.e. the sample mean is zero.

   (c) The sample variance along each axis is 1.

   (d) Median values along each axis should be correlated.

   Explain in 1-2 sentences why this is necessary.

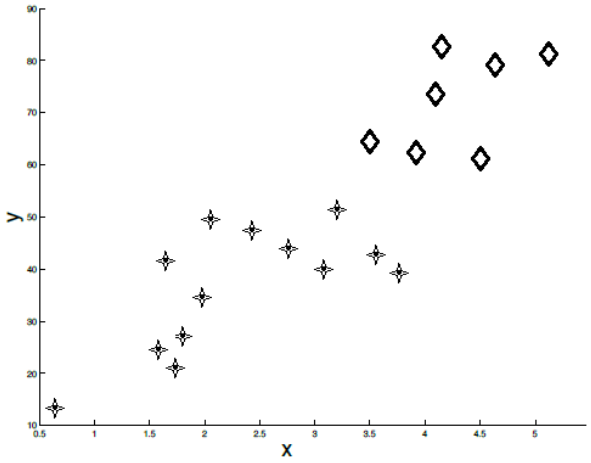
   Solution: b, c
   This normalization is important in PCA since it is a variance maximizing method.

3. In each of the following plots, a training set of data points X belonging to two classes on $\mathbb{R}^2$ are given, where the original features are the coordinates (x, y). For each, answer the following questions:

   (a) **[2 pts]** Draw and label all the principal components.

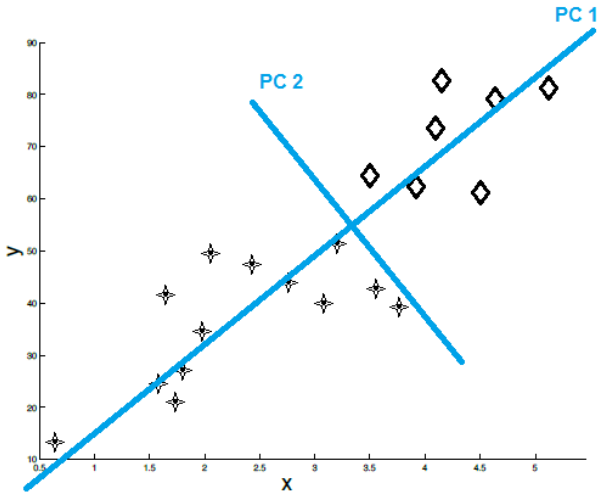
   (b) **[1 pts]** Can we correctly classify this dataset by using a threshold function after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain in 1–2 sentences why it is not possible.

Dataset 1:



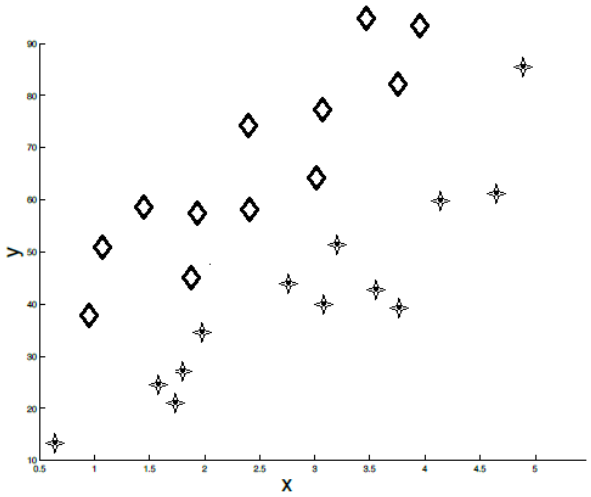
Response to part (b):

Solution:



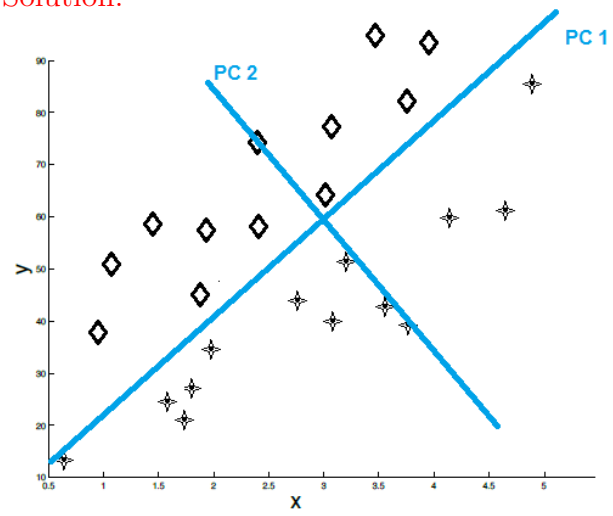
Response to part (b): Yes. On the projection onto the first principal component the two classes will be separable by a threshold c.

Dataset 2:



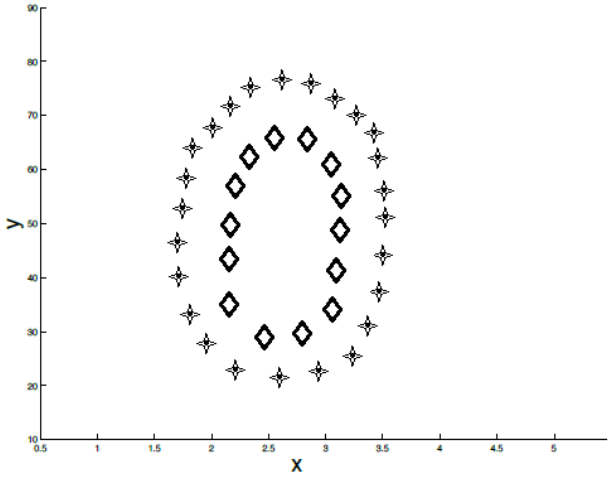
Response to part (b):

Solution:



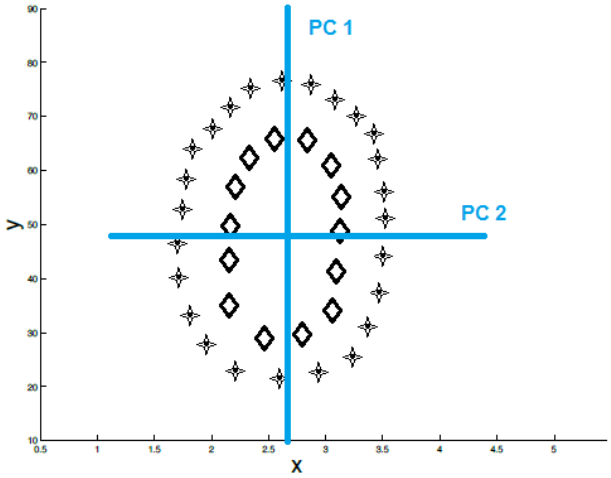
Response to part (b): Yes. On the projection onto the second principal component the two classes will be separable by a threshold c.

Dataset 3:



Response to part (b):

Solution:



Response to part (b): No. PCA fails to capture the non-linear structure of the data and as a result the two classes are indistinguishable using principal components.

4 Neural Networks [18 pts]

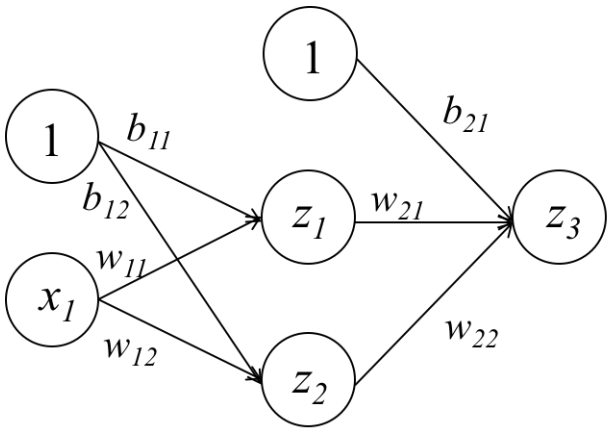
1. [**4 pts**] What is the purpose of the pooling layer in a convolutional neural network? Give at least two reasons for using a pooling layer.

   Solution: Prevents overfitting, reduces size of model representation, reduces total computation, provides a form of translational invariance.

2. [**4 pts**] If the size of the input to a max-pooling layer is 10×10, with a 3×3 filter and stride $s = 3$, what is the size of the output? The input is padded with 0s (i.e., $p = 1$ and the padded input has size $k + 2p \times k + 2p$).

   Solution: 4×4

3. [**4 pts**] Consider the neural network architecture shown below for a binary classification problem.



We define:

$a_1 = w_{11}x_1 + b_{11}$

$a_2 = w_{12}x_1 + b_{12}$

$a_3 = w_{21}z_1 + w_{22}z_2 + b_{21}$

$z_1 = \text{relu}(a_1)$

$z_2 = \text{relu}(a_2)$

$z_3 = \sigma(a_3)$, $\sigma(x) = \frac{1}{1+e^{-x}}$

Perform backpropagation on the bias b_{21} by deriving the expression for the gradient of the loss function $L(y, z_3)$ with respect to the bias term b_{21}, $\frac{\partial L}{\partial b_{21}}$, in terms of the partial derivatives $\frac{\partial \alpha}{\partial \beta}$, where α and β can be any of $L, z_i, a_i, b_{ij}, w_{ij}, x_1$ for all valid values of i, j. Your backpropagation algorithm should be as explicit as possible—that is, make sure each partial derivative $\frac{\partial \alpha}{\partial \beta}$ cannot be decomposed further into simpler partial derivatives. Do *not* evaluate the partial derivatives.

$\frac{\partial L}{\partial b_{21}} =$

Solution: $\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial z_3} \frac{\partial z_3}{\partial a_3} \frac{\partial a_3}{\partial b_{21}}$

4. **[6 pts]** Using the same neural network presented in the previous problem, perform backpropagation on the bias b_{12} by deriving the expression for the gradient of the loss function $L(y, z_3)$ with respect to the bias term b_{12}, $\frac{\partial L}{\partial b_{12}}$, in terms of the partial derivatives $\frac{\partial \alpha}{\partial \beta}$, where α and β can be any of $L, z_i, a_i, b_{ij}, w_{ij}, x_1$ for all valid values of i, j. Your backpropagation algorithm should be as explicit as possible—that is, make sure each partial derivative $\frac{\partial \alpha}{\partial \beta}$ cannot be decomposed further into simpler partial derivatives. Do *not* evaluate the partial derivatives.

$\frac{\partial L}{\partial b_{12}} =$

Solution: $\frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial z_3} \frac{\partial z_3}{\partial a_3} \frac{\partial a_3}{\partial z_2} \frac{\partial z_2}{\partial a_2} \frac{\partial a_2}{\partial b_{12}}$

5 Graphical Models [16 pts]

Consider the following Bayesian network.

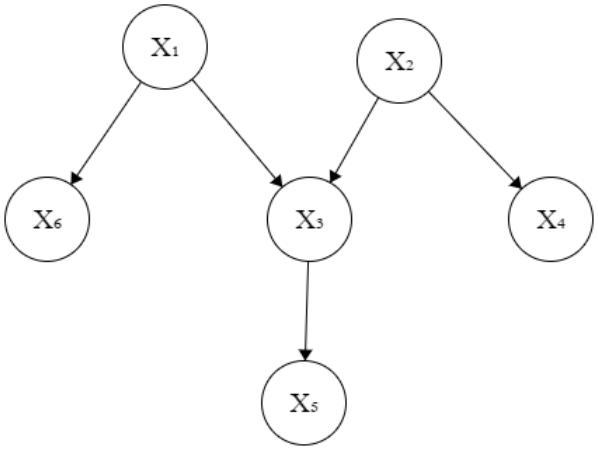


Figure 1: Bayesian network with six variables.

Answer whether following statements are implied by the model with brief justification.

1. **[2 pts]** $X_1 \perp X_4$:

   (a) True

   (b) False

   Justify your answer:
   Solution: TRUE / D-separated

2. **[2 pts]** $X_1 \perp X_4 | X_3$:

   (a) True

   (b) False

   Justify your answer:
   Solution: FALSE / Not D-separated

3. **[2 pts]** $X_6 \perp X_5$:

   (a) True

   (b) False

   Justify your answer:
   Solution: FALSE / Not D-separated

4. **[2 pts]** $X_6 \perp X_5 | X_1$:

   (a) True

(b) False

Justify your answer:
Solution: TRUE / D-separated.

5. **[4 pts]** Write out the joint probability for $P(X_1, X_2, X_3, X_4, X_5, X_6)$ in factorized form implied by Figure 1:

Solution:

$$P(X_1, X_2, X_3, X_4, X_5, X_6) = P(X_1)P(X_2)P(X_4|X_2)P(X_3|X_1, X_2)P(X_5|X_3)P(X_6|X_1)$$

6. **[2 pts]** Assume that each variable is binary with unknown distribution. What is the minimum number of parameters needed to model the joint distribution assuming the independence relationships given by Figure 1?

Solution: $1 + 1 + 2 + 4 + 2 + 2 = 12$

7. **[2 pts]** Again assume that each variable is binary with unknown distribution. How many parameters do we need to model the joint distribution if we are not given Figure 1?

Solution: $2^6 - 1 = 63$

6 Matrix Factorization [12 pts]

In the problem of matrix factorization, we define user vector $w_u \in R^r$ for each $u = 1, \cdots, n$ and item vector $h_i \in R^r$ for each $i = 1, \cdots, d$, where n is the number of users, d is the number of movies, index u represents the index of user and index i represents the index of movie. Then the rating for user u to movie i will be represented as $v_{ui} = w_u^T h_i$.

We want to solve the un-regularized problem:

$$\arg\min_{w,h} \sum_{(u,i)\in Z} \left(v_{ui} - w_u^T h_i\right)^2$$

where Z is the set of observed ratings, $Z = \{(u,i) : v_{ui} \neq 0\}$.

1. **[3 pts]** How many parameters do we need to estimate? Please write down the total number of parameters in terms of n, r, d and briefly justify your answer.

   Solution: $nr + dr$, we need to estimate user matrix $n \times r$ as well as the item matrix $d \times r$.

2. **[4 pts]** For a toy problem, we have

$$W = \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix}$$

$$H = \begin{pmatrix} 1 \\ 5 \\ 8 \end{pmatrix}$$

   where W is the user matrix, H is the item matrix. Please write down the rating matrix V.

   Solution:

$$V = \begin{pmatrix} 1 & 5 & 8 \\ 4 & 20 & 32 \\ 7 & 35 & 56 \end{pmatrix}$$

3. **[5 pts]** Please write down the updating rule for alternating least squares for the un-regularized matrix factorization problem.

Solution: Iteratively solving these two problems until converge:

$$\min_{w} \sum_{(u,i)\in Z} \left(v_{ui} - w_u^T h_i\right)^2$$

$$\min_{h} \sum_{(u,i)\in Z} \left(v_{ui} - w_u^T h_i\right)^2$$

7 Learning Theory [10 pts]

Let X be the feature space and there is a distribution D over X. We have training samples

$$S : (x_1, c^\star(x_1)), \cdots, ((x_n, c^\star(x_n))),$$

x_i i.i.d from D. We assume labels $c^\star(x_i) \in \{-1, 1\}$.

Let $\mathcal{H}$ be a hypothesis space and let $h \in \mathcal{H}$ be a hypothesis. We use

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{m} \mathbf{I}(h(x_i) \neq c^\star(x_i))$$

to denote the training error and

$$R(h) = \mathbf{Pr}_{x \sim D}(h(x) \neq c^\star(x))$$

to denote the true error. Recall from class that if $\mathcal{H}$ is finite, in the realizable case

$$n \geq \frac{1}{\epsilon}\left[\ln(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with probability at least $1-\delta$, all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$ or equivalently if $\hat{R}(h) = 0$, $R(h) \leq \epsilon$.

1. **[5 pts]** What is the full name of PAC learning? What is the correspondence between ϵ, δ and the full name?

   Solution: "Probably approximately correct." The hypotheses we find with m examples are *probably* (with probability $p \geq 1 - \delta$) *approximately* correct, with $R(h) \leq \epsilon$

2. **[5 pts]** Briefly explain what is the realizable case and what is agnostic case and why is the theorem given above not meaningful in the agnostic case?

   Solution:
   Realizable- the true classifier, $c*$, is in $\mathcal{H}$
   Agnostic- $c* \notin \mathcal{H}$, but "close"
   In the agnostic case, since $c* \notin \mathcal{H}$, we will not be able to find a hypothesis $h \epsilon \mathcal{H}$ for which $\hat{R}(h) = 0$