10-601: Machine Learning
Midterm Exam
November 3, 2010

Solutions

Instructions:

- Make sure that your exam has 16 pages (not including this cover sheet) and is not missing any sheets, then write your full name and **Andrew ID** on this page (and all the 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 8 questions, with a maximum score of 98 points. 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
Short Questions	16	
Naïve Bayes	12	
Linear Regression	10	
Learning Theory	13	
Decision Trees	10	
Neural Networks	12	
Support Vector Machines	13	
Clustering	12	
Total:	98	

Question 1: Short Questions (16 points)

(a) Which of the following are true? Prove or give a counter example (in terms of probability and conditional probability tables).

i. (2 points) If A and B are conditionally independent given C, are A and B independent?

Solution: False. For example, A = B = C.

ii. (2 points) If A and $\{B, C\}$ are conditionally independent given D (e.g $p(A, B, C|D) = p(A|D)p(B, C|D)$), are A and B conditionally independent given D? *Hint: you can use the fact* $P(X) = \sum_Y P(X, Y)$.

Solution: True.

$$\begin{aligned} P(A, B|D) &= \sum_C P(A|D)P(B, C|D) \\ &= P(A|D)P(B|D) \end{aligned}$$

(b) (2 points) Which of the following statements are true for k-NN classifiers (circle all answers that are correct).

1. The classification accuracy is better with larger values of k.
2. The decision boundary is smoother with smaller values of k.
3. k-NN is a type of instance-based learning.
4. k-NN does not require an explicit training step.
5. The decision boundary is linear.

Solution: 3 and 4

(c) (3 points) Is it possible for a 2-class 1-NN classifier to always classify all new examples as positive even though there are negative examples in the training data? If yes, show an example. If no, briefly explain.

Solution: No. Pick a negative example and its closest positive example. Any example on the line connecting them but closer to the negative example must be classified as negative.

(d) (2 points) Consider AdaBoost with decision stumps (decision tree of depth 1) as the weak classifier, Figure 1 illustrates the decision threshold (horizontal line in the middle) at iteration 1, where the points below the line are predicted to have class label +1 and those above -1. Note that the ensemble decision boundary coincides with the decision threshold at iteration 1. Assuming the true class label for the squares is +1 and that for the circles is -1. Where would you predict the decision threshold and the ensemble decision boundary would lie at the next iteration?

1. The new decision threshold lies below the threshold at iteration 1 and the ensemble boundary lies further below the new threshold.

2. The new decision threshold lies above the threshold at iteration 1 and the ensemble boundary lies between the two stump thresholds.
3. The new decision threshold lies below the threshold at iteration 1 and the ensemble boundary lies between the two stump thresholds.
4. Both the new decision threshold and the ensemble boundary coincide with the threshold at iteration 1.

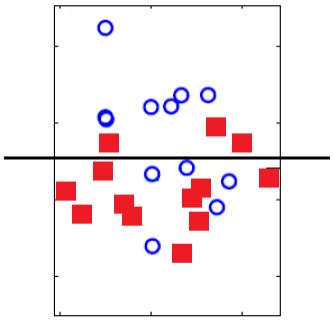

Figure 1: Threshold of decision stump at iteration 1 in AdaBoost.

Solution: 3

(e) (3 points) Imagine we would like to cluster houses around Pittsburgh without using their exact addresses. For each house, we map properties of the house to a numeric value. For instance, the house's location is mapped as Oakland = 0, Shadyside = 1, Squirrel Hill = 2, etc., the exterior material is brick = 0, aluminum = 1, wood = 2, etc., the kitchen color is white = 0, green = 1, tan = 2, etc. We have 50 such features so each house can be represented as a vector in $\mathbb{R}^{50}$. Which of the three clustering algorithms learned in class (hierarchical clustering, k-means and Gaussian mixture models) would be most appropriate for this task? Explain briefly for each algorithm.

Solution: Hierarchical clustering is most appropriate. Even though we converted our categorical data to a numeric format, the mean of these vectors is meaningless so we should not use k-means. Likewise, the data does not obey a Gaussian distribution so Gaussian mixture models are a poor choice. However, hierarchical clustering with a suitable distance function can reasonably cluster this data because hierarchical clustering can handle categorical features.

(f) (2 points) Which of the following adaptations of the Gaussian mixture models algorithms will make it *most* similar to k-means? No need to explain.

1. Restrict each Σ_i to have all off-diagonal entries be 0
2. Restrict each Σ_i to take the form $r_i I$ where r_i is a real-valued scalar and I is the identity matrix
3. Restrict each Σ_i to take the form rI where r is a single real-valued scalar shared by all clusters and I is the identity matrix

Solution: 3. If all of the Gaussian distributions are restricted in this manner, the points will be most likely to belong in the cluster whose mean they are closest to.

Question 2: Naïve Bayes (12 points)

(a) (3 points) $X = (X_1, X_2)$ is drawn from a two dimensional Gaussian distribution with a diagonal covariance matrix.

$$X = (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma)$$
$$\Sigma = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

where a and b are some real numbers.

Are X_1 and X_2 independent? Explain as succinctly as possible.

Solution: Yes, since the covariance between X_1 and X_2 is 0, $p(X)$ can be factored into $p(X_1)$ and $p(X_2)$.

(b) (3 points) We now assume that the two dimensional vector $X = (X_1, X_2)$ is generated by a Gaussian Mixture Model with 3 mixture components (k=3). All mixture components use a *diagonal* covariance matrix similar to the one used in part a. Are X_1 and X_2 (the two components of the input vector X) independent? Explain as succinctly as possible.

Solution: No, $p(X)$ cannot be factored into $p(X_1)$ and $p(X_2)$. An acceptable argument: knowing X_1 tells some information about the mixture component, which tells how likely the value of X_2 is.

(c) (6 points) We have a training set consisting of samples and their labels. All samples come from one of two classes, 0 and 1. Samples are two dimensional vectors. The input data is the form $\{X1, X2, Y\}$ where $X1$ and $X2$ are the two values for the input vector and Y is the label for this sample.

After learning the parameters of a Naïve Bayes classifier we arrived at the following table:

Table 1: Naïve Bayes conditional probabilities

	$Y=0$	$Y=1$
$X1$	$P(X1=1\|Y=0)=1/5$	$P(X1=1\|Y=1)=3/8$
$X2$	$P(X2=1\|Y=0)=1/3$	$P(X2=1\|Y=1)=3/4$

Denote by w_1 the probability of class 1 (that is $w_1 = P(Y=1)$). If we know that the likelihood of the following two samples: {1,0,1},{0,1,0} given our Naïve Bayes model is 1/180, what is the value of w1? You do not need to derive an explicit value for w_1. It is enough to write a (correct ...) *equation* that has w_1 as the only unknown and that when solved would provide the value of w_1. Simplify as best as you can.

Solution: The likelihood of the data given the model is:

$$P(X1=1|Y=1)*P(X2=0|Y=1)*P(Y=1)* \\ P(X1=0|Y=0)*P(X2=1|Y=0)*P(Y=0)$$

Using the values in the table, set $P(Y = 1)$ to w_1 and $P(Y = 0)$ to $(1 - w_1)$ which leads to:

$$3/8 * 1/4 * w_1 * 4/5 * 1/3 * (1 - w_1) = 1/180$$
$$1/40 * w_1 * (1 - w1) = 1/180$$
$$9w_1^2 - 9w_1 + 2 = 0$$

(d) (1 bonus point) Derive an explicit value for w_1.

Solution: Solving the quadratic equation above leads to $w_1 = 1/3$ or $2/3$

Question 3: Linear Regression (10 points)

(a) (6 points) We are given a set of two dimensional inputs and their corresponding output pair: $\{x_{i,1}, x_{i,2}, y_i\}$. We would like to use the following regression model to predict y:

$$y_i = w_1^2 x_{i,1} + w_2^2 x_{i,2}$$

Derive the optimal value for w_1 when using least squares as the target minimization function (w_2 may appear in your resulting equation). Note that there may be more than one possible value for w_1.

Solution: We first write the function we would like to minimize:

$$\sum_i \left(y_i - w_1^2 x_{i,1} - w_2^2 x_{i,2}\right)^2$$

We now take the derivative w.r.t. w_1:

$$\frac{\partial}{\partial w_1} \sum_i \left(y_i - w_1^2 x_{i,1} - w_2^2 x_{i,2}\right)^2 = \sum_i 2w_1 x_{i,1} \left(y_i - w_1^2 x_{i,1} - w_2^2 x_{i,2}\right)$$

We now equate to 0 to find the minimum and get:

$$\sum_i 2w_1 x_{i,1} \left(y_i - w_1^2 x_{i,1} - w_2^2 x_{i,2}\right) = 0$$

One possible answer is that $w_1 = 0$. To find the other one we divide by $2w_1$ which leads to:

$$\sum_i x_{i,1} \left(y_i - w_1^2 x_{i,1} - w_2^2 x_{i,2}\right) = 0$$
$$\sum_i x_{i,1} y_i - w_1^2 x_{i,1}^2 - w_2^2 x_{i,1} x_{i,2} = 0$$
$$\sum_i w_1^2 x_{i,1}^2 = \sum_i x_{i,1} y_i - w_2^2 x_{i,1} x_{i,2}$$
$$w_1^2 = \frac{\sum_i x_{i,1} y_i - w_2^2 x_{i,1} x_{i,2}}{\sum_i x_{i,1}^2}$$

And so w_1 is either 0 or if the sum on the right is ≥ 0 we can set:

$$w_1 = \sqrt{\frac{\sum_i x_{i,1} y_i - w_2^2 x_{i,1} x_{i,2}}{\sum_i x_{i,1}^2}}$$

(b) (2 points) Now assume we only observe a single input for each output (that is, a set of $\{x, y\}$ pairs). We would like to compare the following two models on our input dataset (for each one we split into training and testing set to evaluate the learned model). Assume we have an unlimited amount of data:

A: $y = w^2 x$

B: $y = wx$

Which of the following is correct (chose the answer that best describes the outcome):

1. There are datasets for which A would perform *better* than B
2. There are datasets for which B would perform *better* than A
3. Both 1 and 2 are correct.
4. They would perform equally well on all datasets.

Solution: 2. Model A can only account for positive relationships but cannot account for models in which the w parameters should be negative (for example, $y = -5x$). Model B can account for both positive and negative settings.

(c) (2 points) For the data above we are now comparing the following two models:

$$\text{A:} \quad y = w_1^2 x + w_2 x$$
$$\text{B:} \quad y = wx$$

Note that model A now uses two parameters (though both multiply the same input value, x). Again we assume unlimited data. Which of the following is correct (chose the answer that best describes the outcome):

1. There are datasets for which A would perform *better* than B
2. There are datasets for which B would perform *better* than A
3. Both 1 and 2 are correct.
4. They would perform equally well on all datasets.

Solution: 4. Since we have unlimited data, the parameters learned for model B would be equal to the some of the parameters learned for model A : $w = w_1^2 + w_2$.

Question 4: Learning Theory (13 points)

(a) (2 points) Which of the following procedures is sufficient and necessary and most efficient for proving that the VC dimension of a learner is N?

1. Show that the classifier can shatter all possible dichotomies with N points.
2. Show that the classifier can shatter a subset of all possible dichotomies with N points.
3. Show that the classifier can shatter all possible dichotomies with N points and that it cannot shatter any of the dichotomies with N+1 points.
4. Show that the classifier can shatter all possible dichotomies with N points and that it cannot shatter one of the dichotomies with N+1 points.
5. Show that the classifier can shatter a subset of all possible dichotomies with N points and that it cannot shatter one of the dichotomies with N+1 points.

Solution: 4

(b) (4 points) Figure 2 illustrates exclusive-OR with two inputs (x_1 and x_2), where the squares are labeled as class 1 and the circles are labeled as class 0. As the figure illustrates and as we have discussed in the class and problem set, the VC dimension of a linear classifier in 2D is 3. Assuming that we would like to correctly shatter any set of 4 points with linear decision boundaries, what would you do with the input points to allow successful classification?

Solution: Introduce a new dimension such as $|x_1 - x_2|$ or $(x_1 - x_2)^2]$

(c) (3 points) What is the VC dimension of the resulting classifier?

Solution: 4

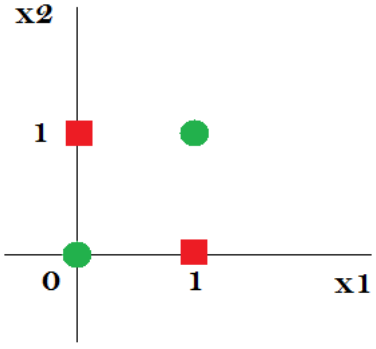

Figure 2: Visualization of XOR points in the input plane x_1-x_2.

(d) (4 points) Consider the following fixed balanced binary decision tree of depth 2 (Figure 3) with features f_1 and f_2 each of which takes the value 0 or 1.

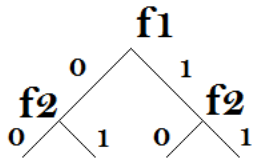

Figure 3: Balanced binary decision tree of depth 2.

Which of the following correctly depicts the size of the space of distinct hypotheses that this tree can represent?

1. 4
2. 7
3. 12
4. 14
5. 16
6. 18

Solution: 5

Question 5: Decision Trees (10 points)

(a) Our goal is to construct a decision tree classifier for predicting flight delays. We have collected data for a few months and a summary of the data is provided in Table 2.

Table 2: Decision tree data

Feature	Value = yes	Value = no
Rain	Delayed - 30, not Delayed - 10	Delayed - 10, not Delayed - 30
Wind	Delayed - 25, not Delayed - 15	Delayed - 15, not Delayed - 25
Summer	Delayed - 5, not Delayed - 35	Delayed - 35, not Delayed - 5
Winter	Delayed - 20, not Delayed - 10	Delayed - 20, not Delayed - 30
Day	Delayed - 20, not Delayed - 20	Delayed - 20, not Delayed - 20
Night	Delayed - 15, not Delayed - 10	Delayed - 25, not Delayed - 30

i. (4 points) Based on the table, which feature should be at the root of the decision tree (briefly explain, no need to provide exact values for information gain)?

Solution: The root would be Summer. It is easy to see that using Summer would lead to the fewest mistakes (10 overall) and so would lead to the highest information gain.

ii. (4 points) Based on the table, which feature should be on the second level (the level just beneath the root) of the decision tree (briefly explain, no need to provide exact values for information gain)?

Solution: It is impossible to tell. To determine the second level feature we need to know the breakdown of delays for the other features given the value of Summer. Since we only have summaries in this table it is not enough information for determining the feature that would lead to the highest information gain AFTER we used Summer.

(b) (2 points) Which of the following statements are true for BOTH decision trees and Naïve Bayes classifiers (you may chose more than one statement):

1. In both classifiers a pair of features are assumed to be independent
2. In both classifiers a pair of features are assumed to be dependent
3. In both classifiers a pair of features are assumed to be independent given the class label
4. In both classifiers a pair of features are assumed to be dependent given the class label

Solution: 2. In both classifiers features are not assumed to be independent. In Naïve Bayes they are assumed to be independent only when given the class label. For decision trees they are not assumed to be independent even if the class label is provided.

Question 6: Neural Networks (12 points)

Can a neural network be used to model the following machine learning algorithms? If so, state the neural network structure (how many hidden layers are required) and the activation function(s) used at the internal and output nodes. If not, describe why not in one or two sentences.

(a) (3 points) k-nearest neighbors

Solution: No. kNN is an instance-based learning algorithm and does not have any parameters to train.

(b) (3 points) Linear regression

Solution: Yes. There are no hidden layers and the activation function at the output layer is the identity function.

(c) (3 points) Logistic regression

Solution: Yes. There are no hidden layers and the activation function at the output layer is the sigmoid function.

(d) (3 points) L1-regularized logistic regression

Solution: No. While the neural network backpropagation algorithm could be extended to include a regularization term, standard backpropagation is unable to include a function of the weights in the objective.

Question 7: Support Vector Machines (13 points)

(a) For each of the following cases, state whether it would be best to use the primal or dual SVM formulation.

i. (2 points) We apply a feature transformation that maps the input data into a feature space with infinite dimension.

> **Solution:** Dual. The primal would have an infinite number of components in the weight vector w and be unsolvable.

ii. (2 points) We apply a feature transformation that doubles the dimension of the input data. The input data has billions of training examples and is linearly separable.

> **Solution:** Primal. The dual formulation would have billions of α variables, and if the data is linearly separable we do not need an ϵ parameter for each data point in the primal.

(b) (3 points) In the linearly separable case, how can we use the solution of the *primal* formulation to determine which points are the support vectors?

> **Solution:** The primal solution gives us w and b. For each x_i, we can compute $y_i\left(w^T x_i + b\right)$. x_i is a support vector if and only if this quantity is 1.

(c) (6 points) Recall that the primal form of the SVM is

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{n} \epsilon_i \\ \text{subject to} \quad & y_i\left(w^T x_i + b\right) \geq 1 - \epsilon_i, \quad \forall i \\ & \epsilon_i \geq 0, \quad \forall i \end{aligned}$$

In the following figures, $C = 0.1$, 1, 10, or 100. The SVM decision boundary has been drawn and all support vectors are circled. Below each figure, write the value of C.

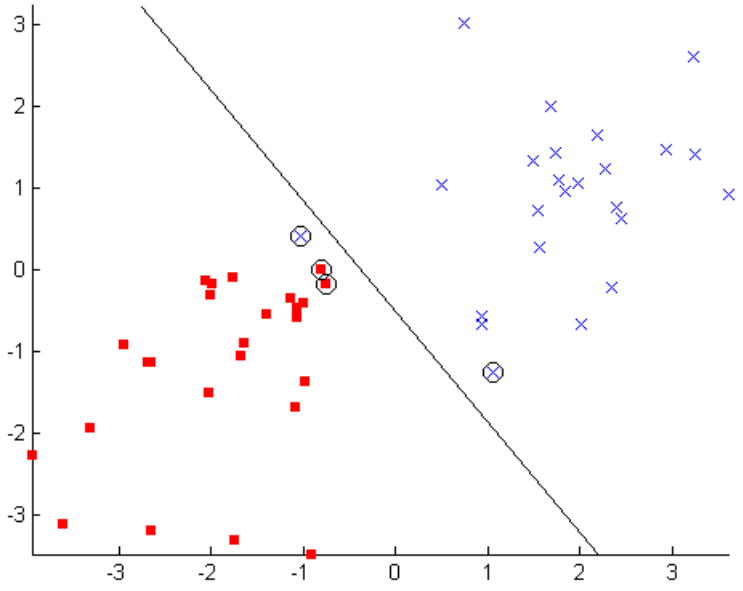

Figure 4: $C = 10$

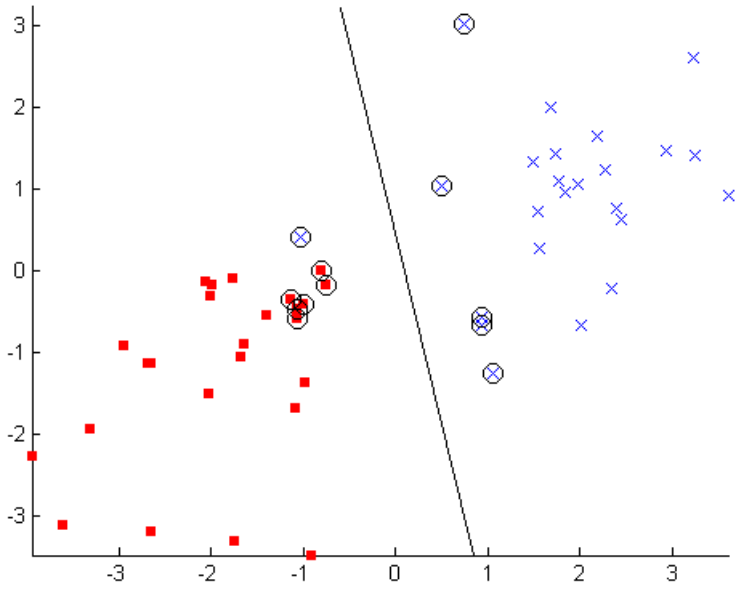

Figure 5: $C = 0.1$

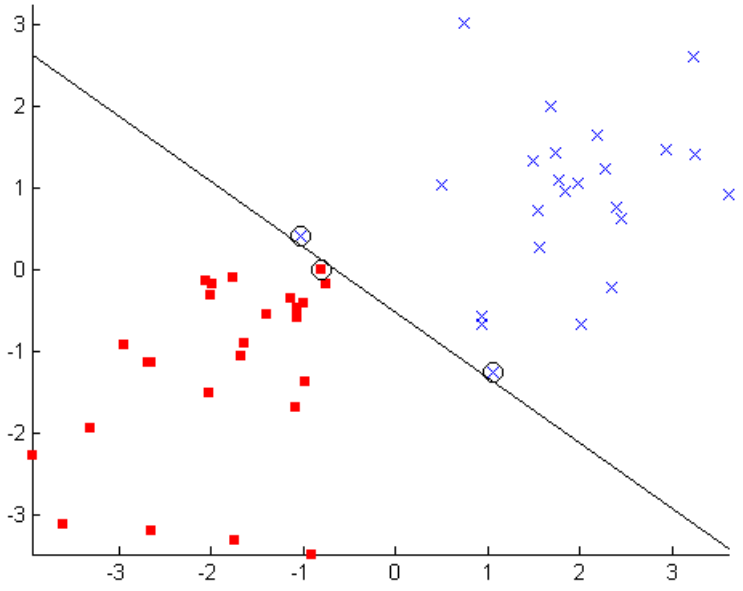

Figure 6: $C = 100$

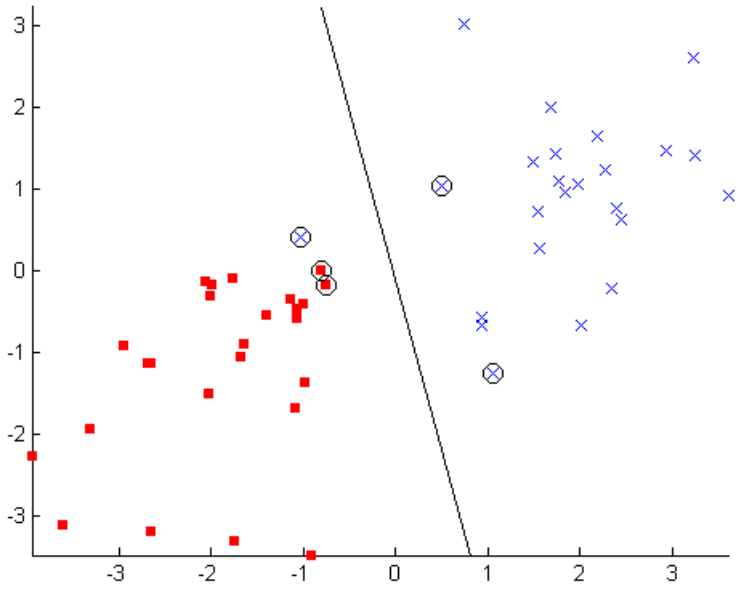

Figure 7: $C = 1$

Question 8: Clustering (12 points)

(a) (6 points) We would like to cluster the points in Figures 8 and 9 (which are the same) using k-means and GMM, respectively. In both cases we set $k = 2$. We perform several random restarts for each algorithm and chose the best one as discussed in class. For each method show the resulting *cluster centers* in the appropriate figure (k-means on Figure 8 and GMM on Figure 9).

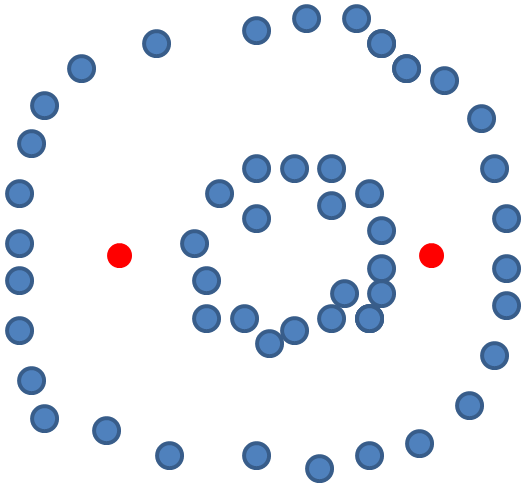

Figure 8: k-means

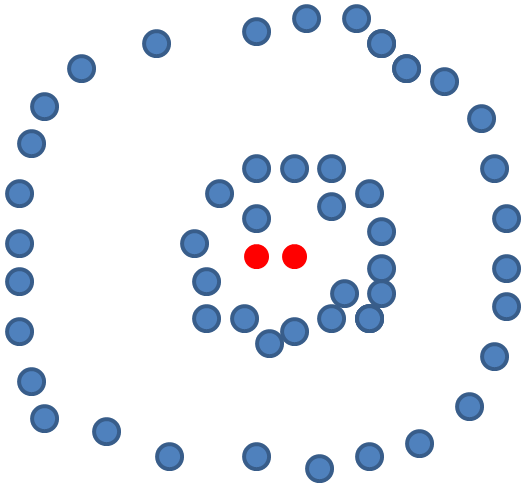

Figure 9: GMM

(b) (6 points) For the same figure (which is repeated in Figures 10 and 11) we would like to use hierarchical clustering. We will use the Euclidian distance as the distance function. In both cases we cut the tree at the second level to obtain two clusters. For two of the linkage models learned in class, single and average link, *circle the resulting groups* of points on each of the figures (Figure 10 - single link, Figure 11 - average link).

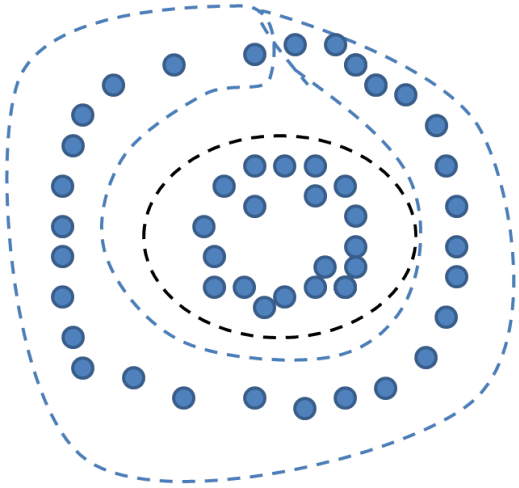

Figure 10: Single link

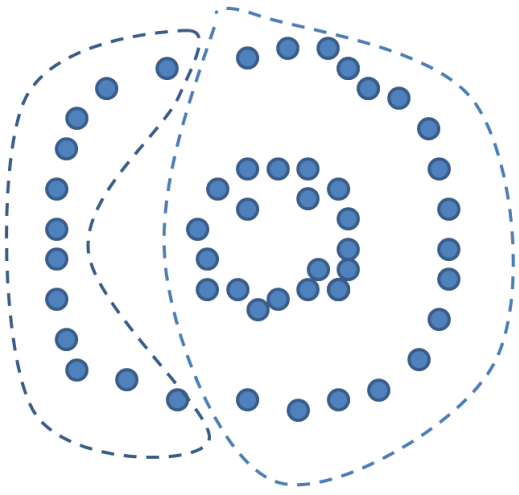

Figure 11: Average link