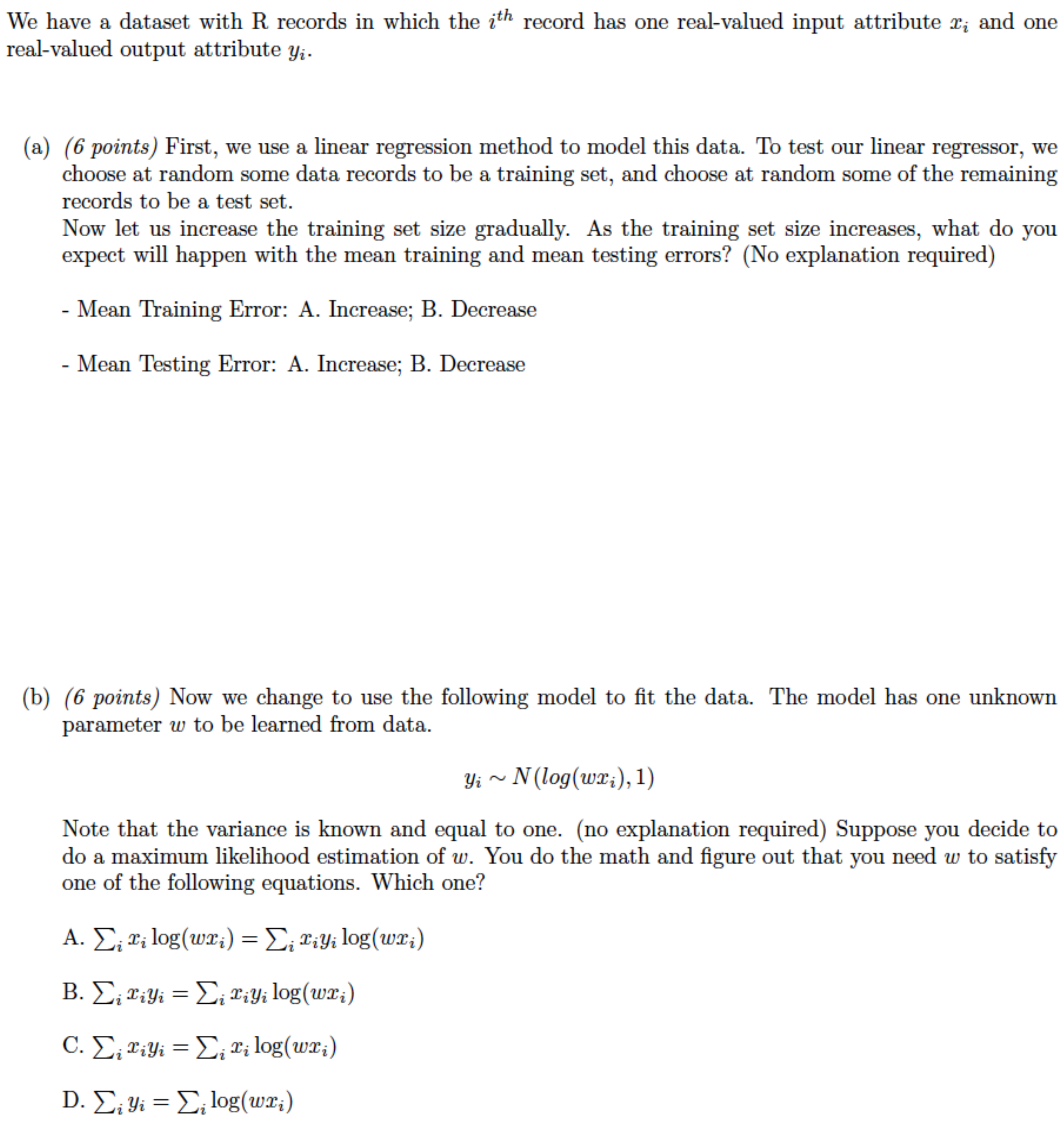
Q1-MLE, training error etc



We have a dataset with R records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i.

(a) *(6 points)* First, we use a linear regression method to model this data. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set.
Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors? (No explanation required)

- Mean Training Error: A. Increase; B. Decrease

- Mean Testing Error: A. Increase; B. Decrease

(b) *(6 points)* Now we change to use the following model to fit the data. The model has one unknown parameter w to be learned from data.

$$y_i \sim N(log(wx_i), 1)$$

Note that the variance is known and equal to one. (no explanation required) Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\sum_i x_i \log(wx_i) = \sum_i x_i y_i \log(wx_i)$

B. $\sum_i x_i y_i = \sum_i x_i y_i \log(wx_i)$

C. $\sum_i x_i y_i = \sum_i x_i \log(wx_i)$

D. $\sum_i y_i = \sum_i \log(wx_i)$

Q1-MLE, training error etc

We have a dataset with R records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i.

(a) *(6 points)* First, we use a linear regression method to model this data. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set.
Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors? (No explanation required)

- Mean Training Error: A. Increase; B. Decrease

- Mean Testing Error: A. Increase; B. Decrease

Answer:

The training error tends to increase. As more examples have to be fitted, it becomes harder to 'hit', or even come close, to all of them.

The test error tends to decrease. As we take into account more examples when training, we have more information, and can come up with a model that better resembles the true behavior. More training examples lead to better generalization.

(b) *(6 points)* Now we change to use the following model to fit the data. The model has one unknown parameter w to be learned from data.

$$y_i \sim N(log(wx_i), 1)$$

Note that the variance is known and equal to one. (no explanation required) Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?

A. $\sum_i x_i \log(wx_i) = \sum_i x_i y_i \log(wx_i)$

B. $\sum_i x_i y_i = \sum_i x_i y_i \log(wx_i)$

C. $\sum_i x_i y_i = \sum_i x_i \log(wx_i)$

D. $\sum_i y_i = \sum_i \log(wx_i)$

Answer: D.

$$y_i \sim N(log(wx_i), 1)$$

We could write the log likelihood as:

$$LL = log(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(y_i - log(wx_i))^2}{2\sigma^2})) = \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}} exp(-\frac{(y_i - log(wx_i))^2}{2}))$$

$$\frac{\partial LL}{\partial w} = 0 \Rightarrow \frac{\partial \sum_{i=1}^{n} (y_i - log(wx_i))^2}{\partial w} = 0$$

Q2. Regression

Suppose we want to learn a quadratic model:

$$\begin{array}{lllllllllllll}
y = & w_0 & + & w_1x_1 & + & w_2x_2 & + & w_3x_3 & + & & \dots & w_kx_k & + \\
 & & & w_{11}x_1^2 & + & w_{12}x_1x_2 & + & w_{13}x_1x_3 & + & & \dots & w_{1k}x_1x_k & + \\
 & & & & & w_{22}x_2^2 & + & w_{23}x_2x_3 & + & & \dots & w_{2k}x_2x_k & + \\
 & & & \vdots & & & & \vdots & & & & \vdots & \\
 & & & & & & & & & w_{k-1,k-1}x_{k-1}^2 & + & w_{k-1,k}x_{k-1}x_k & + \\
 & & & & & & & & & & + & w_{k,k}x_k^2 &
\end{array}$$

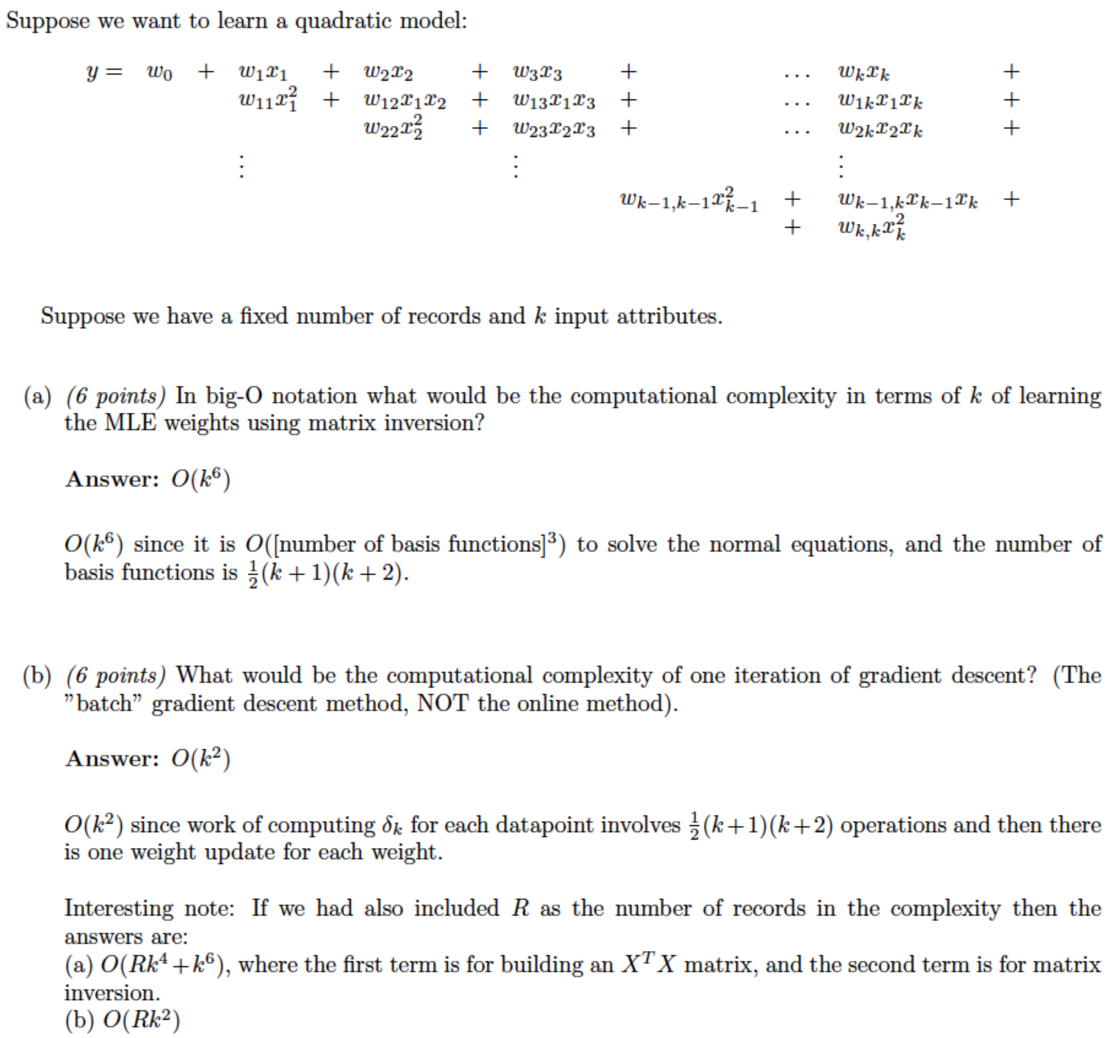
Suppose we have a fixed number of records and k input attributes.

(a) *(6 points)* In big-O notation what would be the computational complexity in terms of k of learning the MLE weights using matrix inversion?

$O(k^6)$ since it is $O([\text{number of basis functions}]^3)$ to solve the normal equations, and the number of basis functions is $\frac{1}{2}(k+1)(k+2)$.

(b) *(6 points)* What would be the computational complexity of one iteration of gradient descent? (The "batch" gradient descent method, NOT the online method).

Q2. Regression



Suppose we want to learn a quadratic model:

$$\begin{array}{cccccccccccccc}
y = & w_0 & + & w_1x_1 & + & w_2x_2 & + & w_3x_3 & + & & & \dots & w_kx_k & + \\
 & & & w_{11}x_1^2 & + & w_{12}x_1x_2 & + & w_{13}x_1x_3 & + & & & \dots & w_{1k}x_1x_k & + \\
 & & & & & w_{22}x_2^2 & + & w_{23}x_2x_3 & + & & & \dots & w_{2k}x_2x_k & + \\
 & & & \vdots & & & & \vdots & & & & & \vdots & \\
 & & & & & & & & & w_{k-1,k-1}x_{k-1}^2 & + & & w_{k-1,k}x_{k-1}x_k & + \\
 & & & & & & & & & & + & & w_{k,k}x_k^2 &
\end{array}$$

Suppose we have a fixed number of records and k input attributes.

(a) *(6 points)* In big-O notation what would be the computational complexity in terms of k of learning the MLE weights using matrix inversion?

Answer: $O(k^6)$

$O(k^6)$ since it is $O([\text{number of basis functions}]^3)$ to solve the normal equations, and the number of basis functions is $\frac{1}{2}(k+1)(k+2)$.

(b) *(6 points)* What would be the computational complexity of one iteration of gradient descent? (The "batch" gradient descent method, NOT the online method).

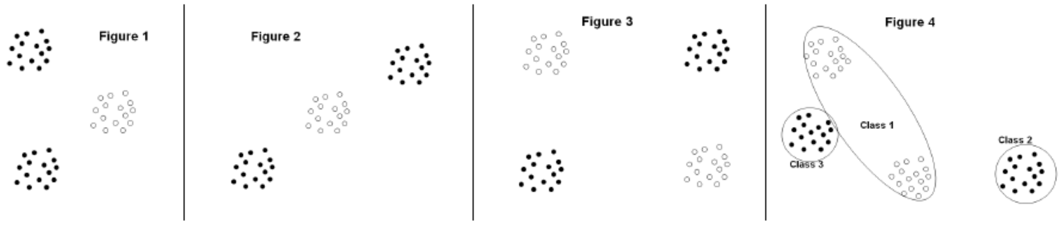
Answer: $O(k^2)$

$O(k^2)$ since work of computing δ_k for each datapoint involves $\frac{1}{2}(k+1)(k+2)$ operations and then there is one weight update for each weight.

Interesting note: If we had also included R as the number of records in the complexity then the answers are:
(a) $O(Rk^4+k^6)$, where the first term is for building an X^TX matrix, and the second term is for matrix inversion.
(b) $O(Rk^2)$

Q3-GNB

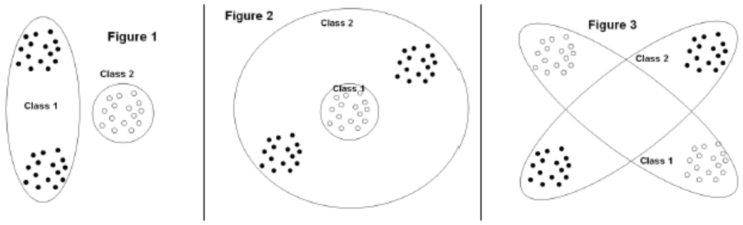
Figures 1, 2 and 3 present points from two different clusters: A (solid points) and B (hollow points). We would like to learn a classifier that achieves zero training error on this data. To do that we allow each classifier to divide the data into more than two classes, however, for each classifier there must be a subset of the classes that perfectly match class A and the complementary set of classes must match cluster B. For example, in Figure 4 classes 2 and 3 contain all of A's points and class 1 contains all of B's points and so this classification is a legitimate solution to this problem.



(a) *(6 points)* For a Gaussian Bayes classifier and for each of the three figures state the **minimum** number of classes required to achieve the above goal. For all figures you can assume equal class priors, that is P(A) = P(B).

	minimum number of classes
Figure 1	
Figure 2	
Figure 3	
Figure 4	3

Answer: The number of classes is 2 for all of the cases.



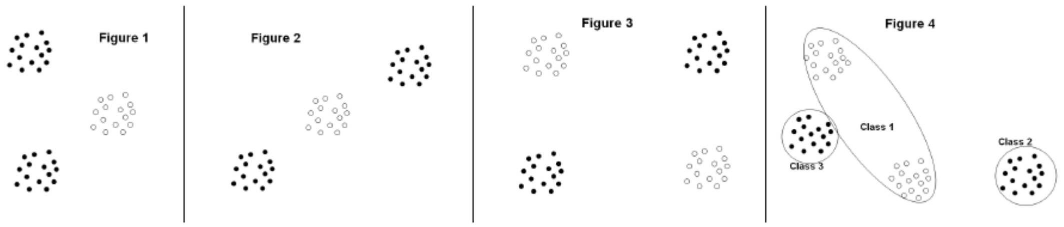
(b) *(6 points)* For the following figures, do we need a full covariance matrix for the classification or would a diagonal covariance matrix be enough

Figure 2? **Answer: Diagonal is enough. Note that the variance of the two clusters is different. A has a large variance for both the x and the y axis while B's variance is low in both direction. Thus, even though both have the same mean, the variance terms are enough to separate them.**

Figuer 3? Answer: Full is needed both mean and marginal variance are the same only the covariance terms are used to discriminate.

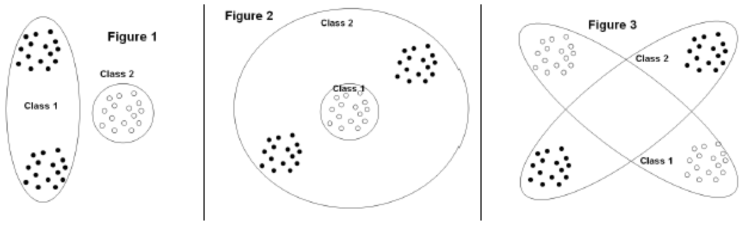
Q3-GNB

Figures 1, 2 and 3 present points from two different clusters: A (solid points) and B (hollow points). We would like to learn a classifier that achieves zero training error on this data. To do that we allow each classifier to divide the data into more than two classes, however, for each classifier there must be a subset of the classes that perfectly match class A and the complementary set of classes must match cluster B. For example, in Figure 4 classes 2 and 3 contain all of A's points and class 1 contains all of B's points and so this classification is a legitimate solution to this problem.



(a) *(6 points)* For a Gaussian Bayes classifier and for each of the three figures state the **minimum** number of classes required to achieve the above goal. For all figures you can assume equal class priors, that is P(A) = P(B).

	minimum number of classes
Figure 1	
Figure 2	
Figure 3	
Figure 4	3



(b) *(6 points)* For the following figures, do we need a full covariance matrix for the classification or would a diagonal covariance matrix be enough

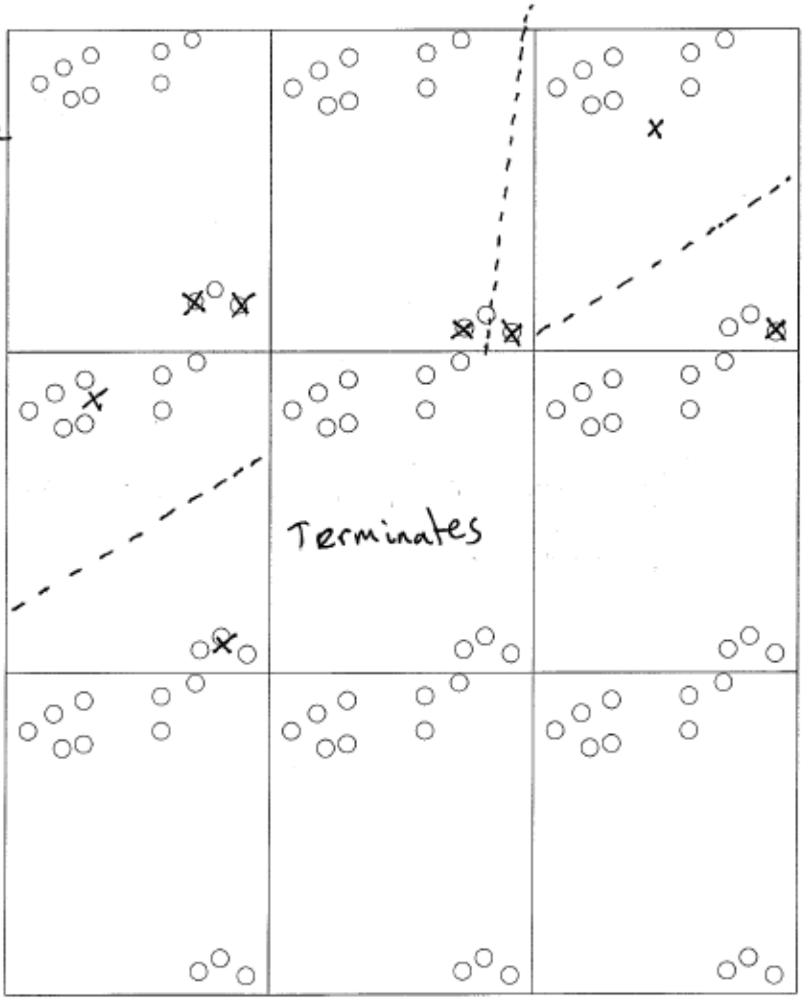
Q4-Kmeans/gmm

(a) What is the effect on the means found by k-means (as opposed to the true means) of overlapping clusters?

They are pushed futher apart than the true means would be.

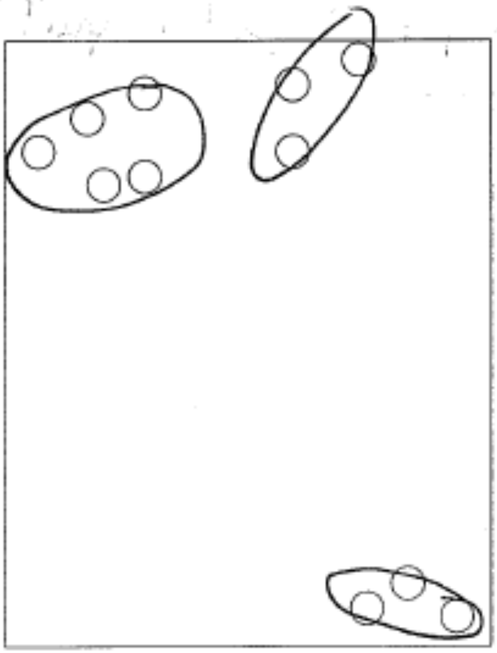
(b) Run k-means manually for the following dataset. Circles are data points and squares are the initial cluster centers. Draw the cluster centers and the decision boundaries that define each cluster. Use as many pictures as you need until convergence.
Note: Execute the algorithm such that if a mean has no points assigned to it, it stays where it is for that iteration.

I don't see
any squares
so I'll
tart somewhere
bitrary



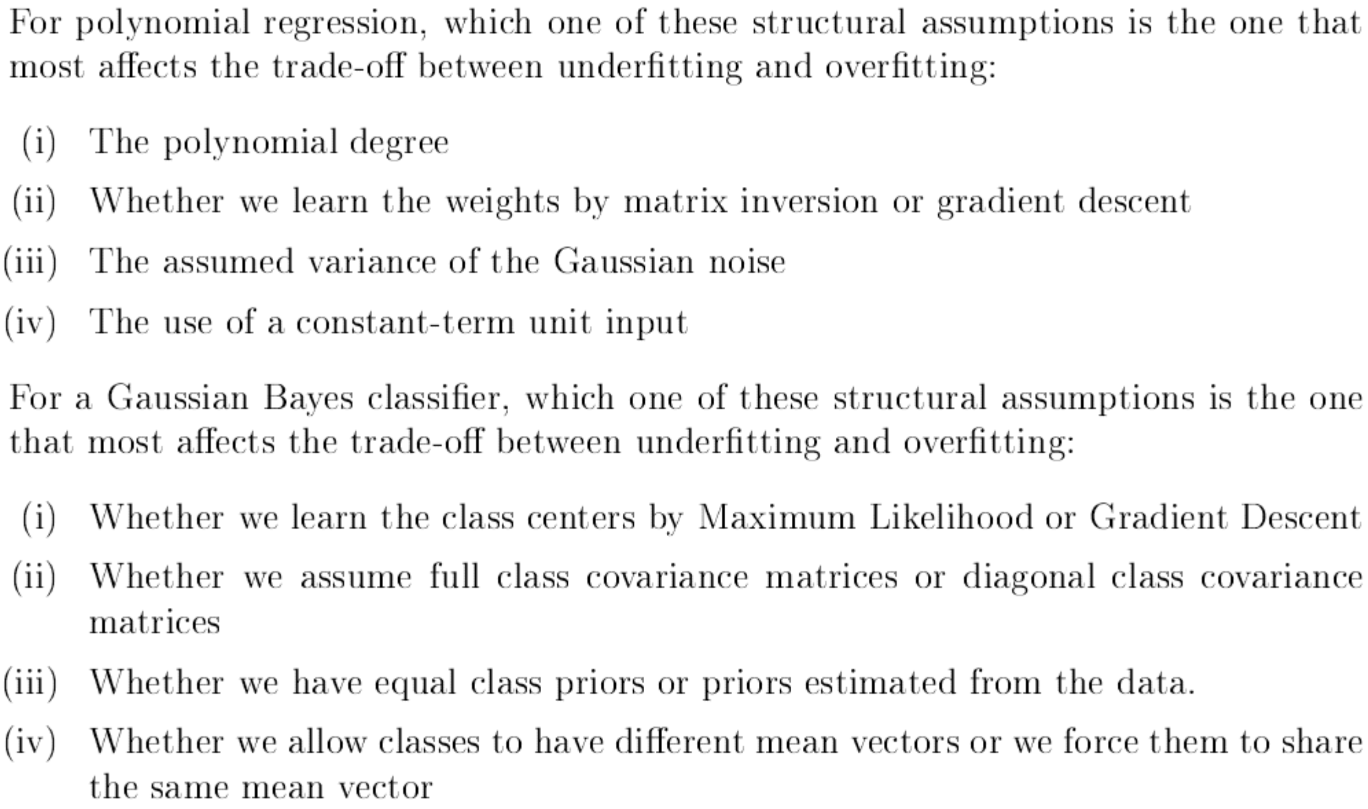
Q4-Kmeans/gmm continued.

(c) Now draw (approximately) what a Gaussian mixture model of three gaussians with the same initial centers as for the k-means problem would converge to. Assume that the model puts no restrictions on the form of the covariance matrices and that EM updates both the means and covariance matrices.



This is the result you'd get if no local optima

Q5-true/false



For polynomial regression, which one of these structural assumptions is the one that most affects the trade-off between underfitting and overfitting:

(i) The polynomial degree

(ii) Whether we learn the weights by matrix inversion or gradient descent

(iii) The assumed variance of the Gaussian noise

(iv) The use of a constant-term unit input

For a Gaussian Bayes classifier, which one of these structural assumptions is the one that most affects the trade-off between underfitting and overfitting:

(i) Whether we learn the class centers by Maximum Likelihood or Gradient Descent

(ii) Whether we assume full class covariance matrices or diagonal class covariance matrices

(iii) Whether we have equal class priors or priors estimated from the data.

(iv) Whether we allow classes to have different mean vectors or we force them to share the same mean vector

Q6-Bayes rule

(a) *(4 points)* I give you the following fact:

$$P(A|B) = 2/3$$

Do you have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

(b) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

(c) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \\ P(B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

(d) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \\ P(B) &= 1/3 \\ P(A) &= 4/9 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Q6-Bayes rule

(a) *(4 points)* I give you the following fact:

$$P(A|B) = 2/3$$

Do you have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Not Enough Info

(b) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Not enough Info

(c) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \\ P(B) &= 1/3 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|{\sim}B)P({\sim}B)} = \frac{2/3 \times 1/3}{2/3 \times \frac{1}{3} + \frac{1}{3} \times 2/3} = \frac{1}{2}$$

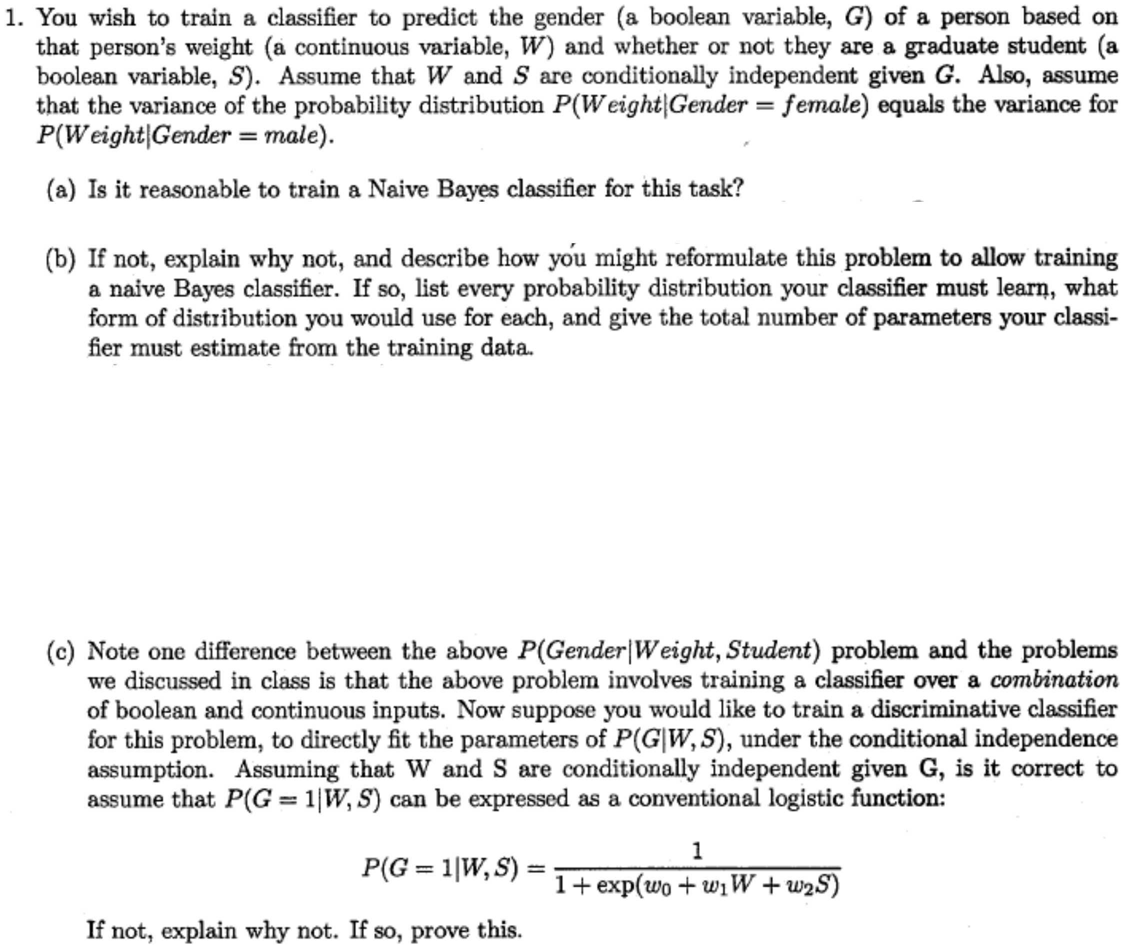
(d) *(5 points)* Instead, I give you the following facts:

$$\begin{aligned} P(A|B) &= 2/3 \\ P(A|{\sim}B) &= 1/3 \\ P(B) &= 1/3 \\ P(A) &= 4/9 \end{aligned}$$

Do you now have enough information to compute $P(B|A)$? If not, write "not enough info". If so, compute the value of $P(B|A)$.

Still $\frac{1}{2}$ (of course)

Q7-discriminative vs generative



1. You wish to train a classifier to predict the gender (a boolean variable, G) of a person based on that person's weight (a continuous variable, W) and whether or not they are a graduate student (a boolean variable, S). Assume that W and S are conditionally independent given G. Also, assume that the variance of the probability distribution $P(Weight|Gender = female)$ equals the variance for $P(Weight|Gender = male)$.

   (a) Is it reasonable to train a Naive Bayes classifier for this task?

   (b) If not, explain why not, and describe how you might reformulate this problem to allow training a naive Bayes classifier. If so, list every probability distribution your classifier must learn, what form of distribution you would use for each, and give the total number of parameters your classifier must estimate from the training data.

   (c) Note one difference between the above $P(Gender|Weight, Student)$ problem and the problems we discussed in class is that the above problem involves training a classifier over a *combination* of boolean and continuous inputs. Now suppose you would like to train a discriminative classifier for this problem, to directly fit the parameters of $P(G|W, S)$, under the conditional independence assumption. Assuming that W and S are conditionally independent given G, is it correct to assume that $P(G = 1|W, S)$ can be expressed as a conventional logistic function:

   $$P(G = 1|W, S) = \frac{1}{1 + \exp(w_0 + w_1 W + w_2 S)}$$

   If not, explain why not. If so, prove this.

Q7-discriminative vs generative

1. You wish to train a classifier to predict the gender (a boolean variable, G) of a person based on that person's weight (a continuous variable, W) and whether or not they are a graduate student (a boolean variable, S). Assume that W and S are conditionally independent given G. Also, assume that the variance of the probability distribution $P(Weight|Gender = female)$ equals the variance for $P(Weight|Gender = male)$.

   (a) Is it reasonable to train a Naive Bayes classifier for this task?

   Yes. W and S are conditionally independent given G.

   (b) If not, explain why not, and describe how you might reformulate this problem to allow training a naive Bayes classifier. If so, list every probability distribution your classifier must learn, what form of distribution you would use for each, and give the total number of parameters your classifier must estimate from the training data.

   We must estimate 6 parameters:

$P(G)$ Bernoulli; → $P(G=1) \equiv \pi$ (note $P(G=0)$ need not be estimated separately. It is $1 - P(G=1)$))

$P(S|G)$ Bernoulli → $P(S=1|G=1) \equiv \theta_1$

Bernoulli → $P(S=1|G=0) \equiv \theta_0$

$P(W|G)$ Normal → σ_w - variance for the Normal distributions governig W

$\mu_{w|G=1}$ - mean for $P(W|G=1)$

$\mu_{w|G=0}$ - mean for $P(W|G=0)$

   (c) Note one difference between the above $P(Gender|Weight, Student)$ problem and the problems we discussed in class is that the above problem involves training a classifier over a *combination* of boolean and continuous inputs. Now suppose you would like to train a discriminative classifier for this problem, to directly fit the parameters of $P(G|W, S)$, under the conditional independence assumption. Assuming that W and S are conditionally independent given G, is it correct to assume that $P(G = 1|W, S)$ can be expressed as a conventional logistic function:

$$P(G = 1|W, S) = \frac{1}{1 + \exp(w_0 + w_1 W + w_2 S)}$$

   If not, explain why not. If so, prove this.

   Yes. This can be shown by combinig the derivation in Tom's Naive Bayes chapter draft (which covers the case of Normal variables) with the solution to a question from homework 2 (which covers Boolean variables).

from eq 19 in Tom's handout, using our variables G, W, + S, we have:

$$P(G=1|WS) = \frac{1}{1 + \exp\left(\ln \frac{1-\pi}{\pi} + \ln \frac{P(W|G=0)}{P(W|G=1)} + \ln \frac{P(S|G=0)}{P(S|G=1)}\right)}$$

from Tom's handout this equals

$$\boxed{W\left(\frac{\mu_0 - \mu_1}{\sigma^2}\right) + \left(\frac{\mu_1^2 - \mu_0^2}{2\sigma^2}\right)}$$

from HW2, this is

$$\boxed{S \ln \frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)} + \ln \frac{1-\theta_0}{1-\theta_1}}$$

therefore:

$$w_0 = \ln \frac{1-\pi}{\pi} + \ln \frac{1-\theta_0}{1-\theta_1} + \frac{\mu_1^2 - \mu_0^2}{2\sigma^2}$$

$$w_1 = \frac{\mu_0 - \mu_1}{\sigma^2}$$

$$w_2 = \ln \frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}$$

Q1 Probability and MLE [20 pts]

1. (a) Suppose we wish to calculate $P(H|E_1, E_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation ?

   i. $P(E_1, E_2), P(H), P(E_1|H), P(E_2|H)$

   [✓ ii.] $P(E_1, E_2), P(H), P(E_1, E_2|H)$

   iii. $P(H), P(E_1|H), P(E_2|H)$

   Bayes' Rule : $P(H|E_1 E_2) = \frac{P(E_1 E_2|H) P(H}{P(E_1 E_2)}$

   (b) Suppose we know that $P(E_1|H, E_2) = P(E_1|H)$ for all values of H, E_1, E_2. Now which of the above three sets are sufficient ?

   (i) because $P(E_1 E_2|H) = P(E_1|H) P(E_2|H)$

   (ii) it just ignores the given independence relations.

2. Which of the following statements are true ? If none of them are true, write NONE.

   (a) If X and Y are independent then $E[2XY] = 2E[X]E[Y]$ and $Var[X + 2Y] = Var[X] + Var[Y]$.

   ~~$Var[X+2Y] = Var[X]+4Var[Y]$~~

   (b) If X and Y are independent and $X > 1$ then $Var[X+2Y^2] = Var[X]+4Var[Y^2]$ and $E[X^2-X] \geq Var[X]$.

   (c) If X are Y are not independent then $Var[X + Y] = Var[X] + Var[Y]$.

   (d) If X and Y are independent then $E[XY^2] = E[X]E[Y]^2$ and $Var[X + Y] = Var[X] + Var[Y]$.

   (e) If X and Y are not independent and $f(X) = X^2$ then $E[f(X)Y] = E[f(X)]E[Y]$ and $Var[X + 2Y] = Var[X] + 4Var[Y]$

   (b)

   OVER FOR REASONS →

3. You are playing a game with two coins. Coin 1 has a θ probability of heads. Coin 2 has a 2θ probability of heads. You flip these coins several times and record your results:

Coin	Result
1	Head
2	Tail
2	Tail
2	Tail
2	Head

   (a) What is the log-likelihood of the data given θ ?

   $L(\theta) = P(\text{data}|\theta) = P(\text{coin1} = \text{Head}) [P(\text{coin 2} = \text{Tail})]^3 P(\text{coin 2} = \text{Head})$

   $= \theta(1-2\theta)^3 2\theta = 2\theta^2(1-2\theta)3$

   $\ell(\theta) = \log L(\theta) = \log 2 + 2\log\theta + 3\log(1-2\theta)$

   (b) What is the maximum likelihood estimate for θ ?

   $0 = \frac{\partial\ell(\theta)}{\partial\theta} = \frac{2}{\theta} + \frac{3(-2)}{(1-2\theta)} \Rightarrow 2(1-2\theta) - 6\theta = 0 \Rightarrow \boxed{\hat{\theta}_{MLE} = 1/5}$

   ↑

u : $\hat{\theta}_{MLE} \overset{\Delta}{=} \underset{\theta}{\text{argmax}}\, L(\theta) = \underset{\theta}{\text{argmax}}\, \ell(\theta)$ b/c $\log(\cdot)$ is monotone + $\left[\begin{array}{l}\text{maximizing } L(\theta) \\ \text{directly is hard}\end{array}\right]$