Midterm for CSC2515, Machine Learning Fall 2020 Thursday, Oct. 22 11:59am – Oct. 23 11:59am

READ ALL INSTRUCTIONS BEFORE STARTING.

RULES

- This is a open-book, open-note test.
- **You should not use the internet to search for solutions to the questions; this is a form of cheating, and you will be given a mark of zero.**
- You must **not discuss the contents of the midterm** with any other students until the grades have been released.

EXAM PERIOD

- Please answer ALL of the questions.
- During the exam, Piazza will be on exam mode. If you have any questions during the exam, you can make a private Piazza post, and we will try to answer within one hour.
- Any updates or clarifications to the exam will be communicated through Quercus and Piazza.

- The questions are NOT arranged in order of difficulty, so you should attempt every question.
- Questions that ask you to "briefly explain" something only require short (1-3 sentence) explanations. Don't write a full page of text. We're just looking for the main idea.
- None of the questions require long derivations. If you find yourself plugging through lots of equations, consider giving less detail or moving on to the next question.
- Many questions have more than one right answer.
- This examination contains 1 bonus mark. The bonus mark cannot raise your grade above 100%.

SUBMISSION

- Markus submissions will close exactly at Oct. 23 11:59am. We strongly suggest you to upload before Oct. 23 10:59am.
- Scan your exam (exactly 29 pages) as a single PDF named `midterm.pdf`, and upload to MarkUs for grading.
- The page that a question appears on the midterm (using the page number at the bottom of the page) **must correspond** to the page in which your solution appears in your submission. For example, Question 8b must appear in page 15 of your submission.
- We recommend either printing and scanning the midterm, or using an electronic writing device.
- You may also write your solutions from scratch or in LaTeX, but you may need to reproduce figures and add blank pages to your submission.

- Failure to submit a full, ordered, 29-page PDF may result in some questions going unmarked, and a loss of points.

OPTIONAL ADDENDUMS

- If you need extra room, you may submit continue your solutions in a separate pdf file (`supp.pdf`).
- To receive points, you must refer to this file in the original question with the exact page number. For example, "*solution continues on page 2 of supp.pdf*".

Q1:	_____	/	16
Q2:	_____	/	14
Q3:	_____	/	10
Q4:	_____	/	10
Q5:	_____	/	10
Q6:	_____	/	6
Q7:	_____	/	10
Q8:	_____	/	9
Q9:	_____	/	8
Q10:	_____	/	11
Q11:	_____	/	11
Q12:	_____	/	15
Q13:	_____	/	12
Q14:	_____	/	8
Bonus:	_____	/	+1

Final mark: ____ / 150

Start your scan at this page.

1. **True/False [16 points]** For each statement below, say whether it is true or false, and give a one or two sentence justification of your answer.

   (a) **[2 points]** As we increase the number of parameters in a model, both the bias and variance of our predictions should decrease since we can better fit the data.

   (b) **[2 points]** We can always improve the performance of clustering algorithms like k-means by removing features and reducing the dimensionality of our data.

   (c) **[2 points]** Bagging and boosting are both ways of reducing the variance of our models.

   (d) **[2 points]** Machine learning models work automatically without the need for the user to manually set hyperparameters.

(e) **[2 points]** Most machine learning algorithms, in general, can learn a model with a Bayes optimal error rate.

(f) **[2 points]** The k-means algorithm is guaranteed to converge to the global minimum

(g) **[2 points]** When optimizing a regression problem where we know that only some of our features are useful, we should use L_1 regularization.

(h) **[2 points]** For a given binary classification problem, a linear SVM model and a logistic regression model will converge to the same decision boundary.

2. **Probability and Bayes Rule [14 points]**

You have just purchased a two-sided die, which can come up either 1 or 2. You want to use your crazy die in some betting games with friends later this evening, but first you want to know the probability that it will roll a 1. You know know it came either from factory 0 or factory 1, but not which.

Factory 0 produces dice that roll a 1 with probability ϕ_0. Factory 1 produces dice that roll a 1 with probability ϕ_1. You believe initially that your die came from factory 1 with probability η_1.

(a) **[2 points]** Without seeing any rolls of this die, what would be your predicted probability that it would roll at 1?

(b) **[2 points]** If we roll the die and observe the outcome, what can we infer about where the die was manufactured?

(c) **[4 points]** More concretely, let's assume that:

- $\phi_0 = 1$: dice from factory 0 always roll a 1
- $\phi_1 = 0.5$: dice from factory 1 are fair (roll a 1 with probability 0.5)
- $\eta_1 = 0.7$: we think with probability 0.7 that this die came from factory 1

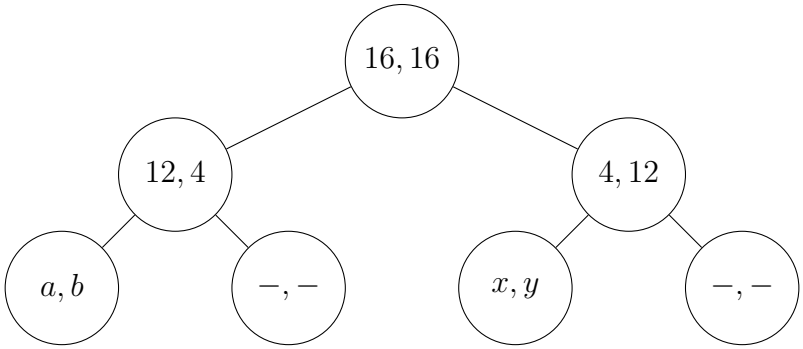
Now we roll it, and it comes up 1! What is your posterior distribution on which factory it came from? What is your predictive distribution on the value of the next roll?

(d) **[2 points]** You roll it again, and it comes up 1 again.
Now, what is your posterior distribution on which factory it came from? What is your predictive distribution on the value of the next roll?

(e) **[2 points]** Instead, what if it rolls a 2 on the second roll?

(f) **[2 points]** In the general case (not using the numerical values we have been using) prove that if you have two observations, and you use them to update your prior in two steps (first conditioning on one observation and then conditioning on the second), that no matter which order you do the updates in you will get the same result.

3. **Decision Trees & Information Gain [10 points]**



The figure above shows a decision tree. The numbers in each node represent the number of examples in each class in that node. For example, in the root node, the numbers $16, 16$ represent 16 examples in class 0 and 16 examples in class 1. To save space, we have not written the number of examples in two right leaf nodes as they can be deduced based on the values in the left leaf nodes.

(a) **[2 points]** What is the information gain of the split made at the root node?

(b) **2 points]** What values of a and b will give the **smallest** possible value of the information gain at this split, and what is the value?

(c) **[2 points]** What values of x and y will give the **largest** possible value of the information gain at this split, and what is the value?

(d) **[4 points]** Bayes rule for conditional entropy states that $\mathcal{H}(Y|X) = \mathcal{H}(X|Y) - \mathcal{H}(X) + \mathcal{H}(Y)$. Using the definition of conditional and joint entropy, show that this is the case.

Hint: First show that $\mathcal{H}(Y|X) = \mathcal{H}(X,Y) - \mathcal{H}(X)$

4. **Bayes Optimality [10 points]**

   (a) **[2 points]** Consider a regression task with the following data generating process:

   $$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$
   $$t|\mathbf{x} \sim \mathbf{w}^T\mathbf{x} + \alpha\mathbf{x}^T\mathbf{x} + \mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2)$$

   where $\mathbf{x}, \boldsymbol{\mu}_x \in \mathbb{R}^D, \boldsymbol{\Sigma}_x \in \mathbb{R}^{D\times D}$; $\mathbf{w} \in \mathbb{R}^D$ and $\alpha \in \mathbb{R}$ are parameters.

   At a particular query point $\mathbf{x}^* \in \mathbb{R}^D$, what is the output of the Bayes optimal regressor $y^* = f^*(\mathbf{x}^*)$?

   (b) **[2 points]** For the data generating process in part (a), at a particular query point $\mathbf{x}^* \in \mathbb{R}^D$, what is the expected squared error (defined as $(t - y^*)^2$) of the Bayes optimal regressor?

(c) **[2 points]** Consider a 3-class classification problem with the following data generating process:

$$x \sim \text{unif}(0, 1)$$
$$t|x \sim \begin{cases} 0, & p = 0.2 \\ 1, & p = 0.5x \\ 2, & p = 0.8 - 0.5x \end{cases}$$

where $x \in [0, 1]$ and $t \in \{0, 1, 2\}$.

What is the expression for the Bayes optimal classifier?

(d) **[4 points]** What is the misclassification rate of the Bayes optimal classifier in part (c)? Your answer should be a scalar.

5. **Linear/Logistic Regression [10 points]**

   (a) **[5 points]** In lecture, we considered using linear regression for binary classification on the targets $\{0, 1\}$. Here, we use a linear model

   $$y = \mathbf{w}^\top \mathbf{x} + b$$

   and squared error loss $\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$. We saw this has the problem that it penalizes confident correct predictions. Will it fix this problem if we instead use a modified hinge loss, shown below?

   $$\mathcal{L}(y, t) = \begin{cases} t = 0, & \max(0, y) \\ t = 1, & 1 - \min(1, y) \end{cases}$$

   Justify your answer mathematically.

(b) **[5 points]** Consider the logistic regression model $y = g(\mathbf{w}^T\mathbf{x})$, trained using the binary cross entropy loss function, where $g(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Your friend proposes a modified version of logistic regression, using

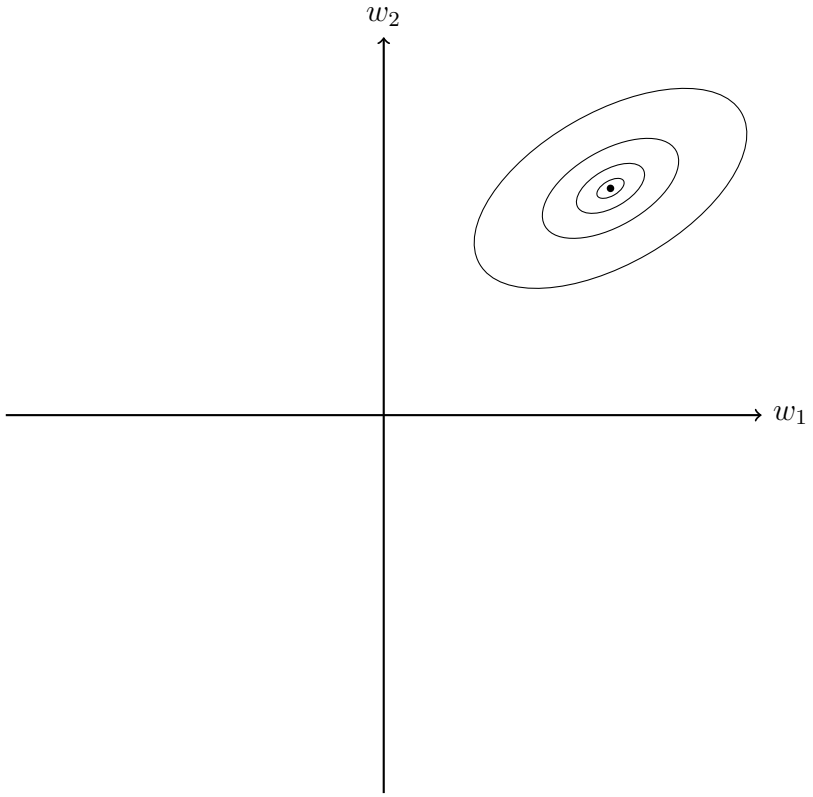
$$g(z) = \frac{e^{-z}}{1+e^{-z}}$$

The model would still be trained using the binary cross entropy loss. How would the learnt model parameters, as well as the model predictions, differ from conventional logistic regression? Justify your answer mathematically. A purely graphical explanation is not sufficient.

6. **L1/L2 Regularization [6 points]**

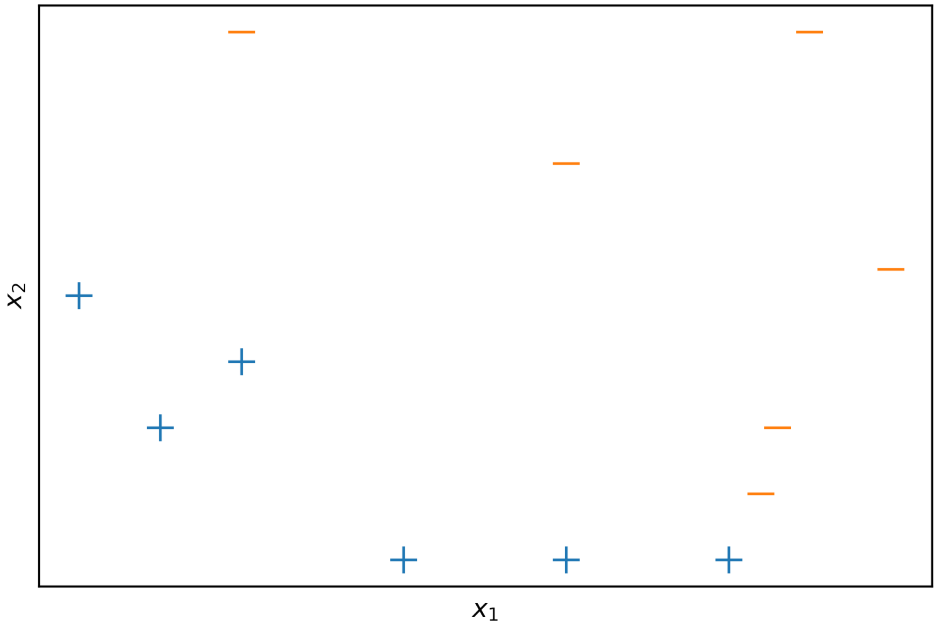
(a) **[2 points]** The L_∞ norm is defined on a vector x as $L_\infty|x| = max(\{x_i : i = 1, 2, \ldots, n\})$. As an example, if $x = [-6, 2, 4]$, then $L_\infty|x| = 6$ since the magnitude of the largest element in x is 6.

On the loss plot below, draw the L_∞ norm and draw a point at the set of weights the regularization term will select.



(b) **[4 points]** What kind of weights does the L_∞ norm encourage our models to learn? In which situations might this be useful?

7. **Classification Decision Boundaries [10 points]**



(a) **[1 point]** Assume this is the first step in the boosting algorithm. In the figure above, sketch the decision boundary for the decision stump, optimizing for weighted misclassification error. Label this boundary f_1.

(b) **[2 points]** Select an arbitrary point which f_1 incorrectly classifies. Draw a circle around the point you select. What is its weight in the next boosting iteration? Assume the weights initially sum to 1.

(c) **[1 point]** In the figure above, sketch the decision boundary for a 1-layer decision tree, optimizing for information gain. Label this boundary f_2. You do not have to show that the boundary is optimal.

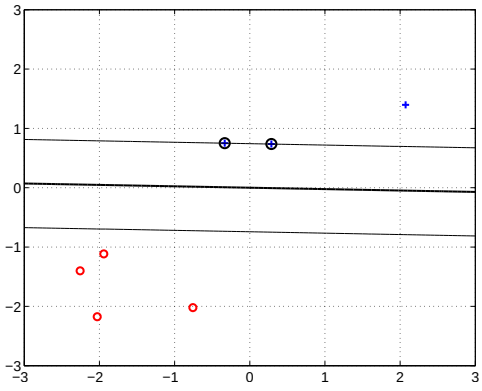
(d) **[3 points]** Compute the information gain of the split f_2.

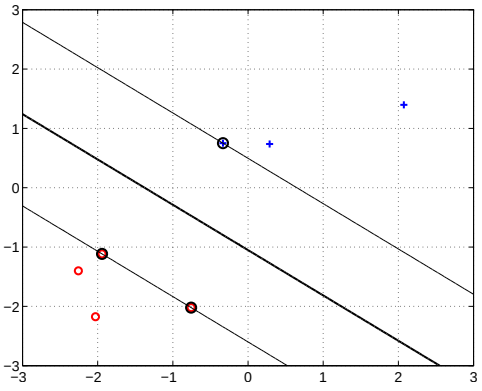
(e) **[1 point]** In the figure above, sketch the decision boundary for a decision tree with unlimited depth. Label this boundary f_3.

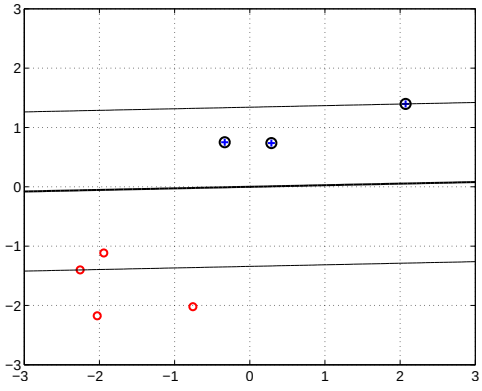
(f) **[2 points]** In the figure above, sketch the decision boundary for a hard-margin SVM. Label this boundary f_4. Is the boundary unique? You do not have to justify your answer.

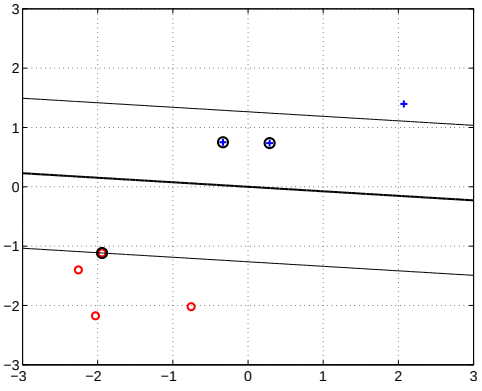
8. **SVM [9 points]**

Here are plots of $\mathbf{w}^T\mathbf{x} + b = 0$ for different training methods along with the support vectors. Points labeled $+1$ are in blue, points labeled -1 are in red. The line $\mathbf{w}^T\mathbf{x}+b = 0$ is shown in bold; in addition we show the lines $\mathbf{w}^T\mathbf{x} + b = -1$ and $\mathbf{w}^T\mathbf{x} + b = 1$ in non-bold. Support vectors have bold circles surrounding them.

1) 

2) 

3) 

4) 

For each method, list all figures that show a possible solution for that method and give a brief explanation. Note that some methods may have more than one possible figure.

(a) **[3 points]**

$$\min \ \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{t=1}^{n} \xi_t \ \text{ s.t. } \xi_t \geq 0, \ \ t^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \xi_t \ \ t = 1, \ldots, n$$

where $\gamma = \infty$.

(b) **[3 points]**

$$\min \ \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{t=1}^{n} \xi_t \ \text{ s.t. } \xi_t \geq 0, \ \ t^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)}) \geq 1 - \xi_t \ \ t = 1, \ldots, n$$

where $\gamma = \infty$.

(c) **[3 points]**

$$\min \ \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_{t=1}^{n} \xi_t \ \text{ s.t. } \xi_t \geq 0, \ \ t^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)}) \geq 1 - \xi_t \ \ t = 1, \ldots, n$$

where $\gamma = 1$.

9. **Gradient Descent [8 points]**

For a homework assignment, you have implemented a logistic regression model

$$y^{(i)} = \sigma(Wx^{(i)} + b)$$

You decide to train your model on a multi-class problem using gradient descent. However, before you can turn your assignment in, your friends stop by to give you some suggestions.

(a) **[2 points]** Pat sees your regressor implementation and says there's a much simpler way! Just leave out sigmoids, and let $f(x) = Wx + b$. The derivatives are a lot less hassle and it runs faster.

What's wrong with Pat's approach?

(b) **[2 points]** Chris comes in and says that your regressor is too complicated, but for a different reason. The sigmoids are confusing and basically the same answer would be computed if we used step functions instead.

What's wrong with Chris's approach?

(c) **[2 points]** Jody sees that you are handling a multi-class problem with 4 classes by using 4 output values, where each target $y^{(i)}$ is actually a length-4 vector with three 0 values and a single 1 value, indicating which class the associated $x^{(i)}$ belongs to.

Jody suggests that you just encode the target $y^{(i)}$ values using integers 1, 2, 3, and 4.

What's wrong with Jody's approach?

(d) **[2 points]** Evelyn overhears your conversation with Jody and suggests that you use as a target value $y^{(i)}$ a vector of two output values, each of which can be 0 or 1, to encode which one of four classes it belongs to.

What is good or bad about Evelyn's approach?

10. **Maximum Likelihood and Maximum A Posteriori [11 points]**

Your company has developed a test for COVID-19. The test has a false positive rate of α, and a false negative rate of β.

(a) **[2 points]** Assume that COVID-19 is evenly distributed through the population, and that the prevalence of the disease is γ. What is the accuracy of your test on the general population?

(b) **[2 points]** Your company conducts a clinical trial. They find n people, all of whom they know have COVID. What is the *likelihood* that your test makes n^+ correct predictions?

(c) **[5 points]** Derive the maximum likelihood estimate for β. You may assume all other parameters are fixed.

(d) **[4 points]** Derive the MAP estimate for β, assuming it has a prior $P(\beta) = \text{Beta}(a, b)$. You may assume all other parameters are fixed.

11. **Principal Component Analysis [11 points]**

Suppose that you are given a dataset of N samples, i.e., $\mathbf{x}^{(i)} \in \mathbb{R}^D$ for $i = 1, 2, ..., N$. We say that *the data is centered* if the mean of samples is equal to 0, i.e., $\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} = 0$.

For a given unit direction $\mathbf{u}$ such that $\|\mathbf{u}\|_2 = 1$, we denote by $\mathcal{P}_{\mathbf{u}}(\mathbf{x})$ the Euclidean projection of $\mathbf{x}$ on $\mathbf{u}$. Recall that projection is given by

$$\mathcal{P}_{\mathbf{u}}(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}\mathbf{u}$$

(a) **[3 points]** *Mean of data after projecting on* $\mathbf{u}$*:* Show that the projected data with samples $\mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)})$ in any unit direction $\mathbf{u}$ is still centered. That is show,

$$\frac{1}{N}\sum_{i=1}^{n} \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)}) = 0.$$

(b) **[4 points]** *Maximum variance:* Recall that the first principle component $\mathbf{u}_*$ is given by the largest eigenvector of the sample covariance (eigenvector associated to the largest eigenvalue). That is,

$$\mathbf{u}_* = \underset{\mathbf{u}\,:\,\|\mathbf{u}\|_2=1}{\operatorname{argmax}} \mathbf{u}^\top \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} [\mathbf{x}^{(i)}]^\top \mathbf{u}.$$

Show that the unit direction $\mathbf{u}$ that maximizes the variance of the projected data corresponds to the first principle component of the data. That is, show

$$\mathbf{u}_* = \underset{\mathbf{u}\,:\,\|\mathbf{u}\|_2=1}{\operatorname{argmax}} \sum_{i=1}^{N} \left\| \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)}) - \frac{1}{N} \sum_{j=1}^{N} \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(j)}) \right\|_2^2.$$

(c) **[4 points]** *Minimum error:* Show that the unit direction $\mathbf{u}$ that minimizes the mean squared error between projected data points $\mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)})$ and the original data points $\mathbf{x}^{(i)}$ corresponds to the first principal component $\mathbf{u}_*$. That is show,

$$\mathbf{u}_* = \underset{\mathbf{u}\,:\,\|\mathbf{u}\|_2=1}{\operatorname{argmin}} \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \mathcal{P}_{\mathbf{u}}(\mathbf{x}^{(i)})\|_2^2. \tag{1}$$

12. **Probabilistic Models [15 points]**

We would like to build a model that predicts a disease D in a patient. Assume we have two classes in our label set t: diseased and healthy. We have a set of binary patient features $\mathbf{x} = [x_1, ..., x_N]$ that describe the various comorbidities of our patient. We want to write down the joint distribution of our model.

(a) **[4 points]**W hat is the problem in calculating this joint distribution? How many parameters would be required to construct this model? How does the Naive Bayes model handle this problem?

(b) **[3 points]** Write down the joint distribution under the Naive Bayes assumption. How many parameters are specified under the model now?

(c) **[4 points]** Why does the Naive Bayes assumption allow for the model parameters to be learned efficiently? Write down the log-likelihood to explain why. Here, assume that you observed N data points $(\mathbf{x}^{(i)}, t^{(i)})$ for $i = 1, 2, ..., N$.

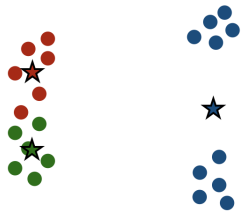
(d) **[1 points]** Show how you can use Bayes rule to predict a class given a data point.

(e) **[3 points]** Explain how placing a Beta prior on parameters is helpful for the Naive Bayes model.

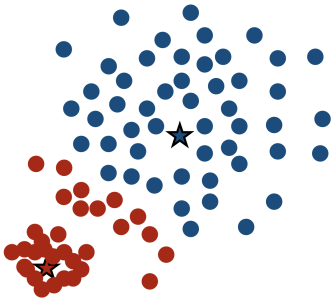
13. **K-means [12 points]**

In the following plots of k-means clusters, circles with different colors represent points in different clusters, and stars represent cluster centers. For each plot, describe the problem with the data or algorithm, and propose a solution to find a more appropriate clustering.

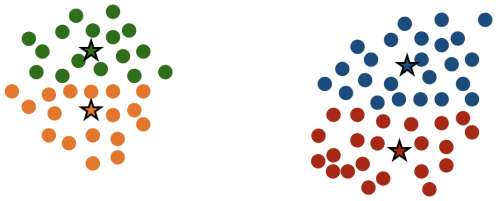
(a) **[2 points]**



(b) **[2 points]**



(c) **[2 points]**



(d) **[3 points]** The k-means algorithm is designed to reduce the mean squared error between the cluster centers and their assigned points. Will increasing the value of k ever increase this error term? Why or why not?

(e) **[3 points]** Suppose we achieve a very low error with a large value of k, and a higher error with a smaller value of k. Should we select the k value with the lowest error? Why or why not?

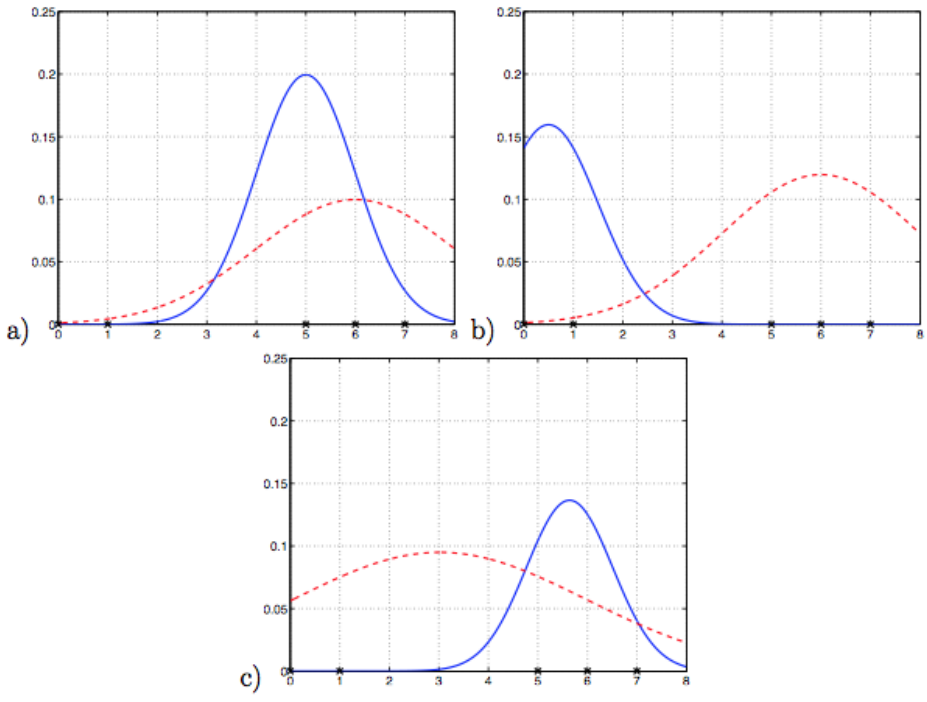
14. **Expectation-Maximization [8 points]**

Here we are estimating a mixture of two Gaussians via the EM algorithm. The mixture distribution over x is given by

$$P(x;\theta) = P(1)N(x;\mu_1,\sigma_1^2) + P(2)N(x;\mu_2,\sigma_2^2)$$

Any student in this class could solve this estimation problem easily. Well, one student, devious as they were, scrambled the order of figures illustrating EM updates. They may have also slipped in a figure that does not belong. Your task is to extract the figures of successive updates and explain why your ordering makes sense from the point of view of how the EM algorithm works. All the figures plot $P(1)N(x;\mu_1,\sigma_1^2)$ as a function of x with a solid line and $P(2)N(x;\mu_2,\sigma_2^2)$ with a dashed line.

(a) **[2 points]** (True/False) In the mixture model, we can identify the most likely T posterior assignment, i.e., j that maximizes $P(j \mid x)$, by comparing the values of $P(1)N(x;\mu_1,\sigma_1^2)$ and $P(1)N(x;\mu_2,\sigma_2^2)$



(b) **[2 points]** Assign two figures to the correct steps in the EM algorithm.
- Step 0: ( ) initial mixture distribution
- Step 1: ( ) after one EM-iteration

(c) **[4 points]** Briefly explain how the mixture you chose for "step 1" follows from the mixture you have in "step 0".

Bonus [+1 Point] Guess the percentage grade that you will receive on this examination out of 100% (excluding this bonus question and any formatting penalty). You will receive 1 point if your guess is within 5% of your actual mark.

Survey Questions

The following survey questions are **completely optional**, and will not impact your grade in any way. By answering these questions, you give permission for your data to be used in aggregate descriptive statistics. Results from these analyses will *only* be disseminated to students in the class.

A1. I have previously taken an "Introduction to machine learning" or equivalent course (at any university).

Yes No

A2. My home department for graduate studies is:

CS ECE Other: _____

A3. I would describe the overall difficulty of the course so far as:

Very easy Easy Average Hard Very hard