CSE 512: Final Exam Dec. 8

Instructions:

- You must complete this exam on the specified date, from 8:30pm to 11:00 pm
- You may complete the exam virtually or in person, using the same format as the midterm. (Virtual = webcam on, showing widescreen pan of hands. In-person = in our usual lecture hall.)
- You are allowed 1 page (front and back) cheat sheet. If you take the exam virtually, the cheat sheet must be scanned (or photographed with high resolution) and submitted along with the exam solutions. If you are in person, you must submit all of your scratch paper + cheat sheet.
- You are also allowed a simple or graphing calculator. If you cannot get access to a calculator, you may use your phone, but it must be on calculator mode the whole time. You may not use your laptop.
- If you need extra sheets of paper, please label them carefully as to which question they are answering. Make your final answer clear.
- Your handwritten solutions must be clean and legible, and your scans must be clear. Anything we can't read, we will not grade.
- You may not discuss any problem with any other student while the exam submission portal is still open. You may not look for answers on the internet or in any notes outside of your cheatsheet.
- **The last question is extra credit.**

Name: ____________________________

Student ID: ____________________________

Scoring	
Q 1	______________ / 10
Q 2	______________ / 20
Q 3	______________ / 20
Q 4	______________ / 20
Total	______________ / 70
Q 5 (EC)	______________ / + 15

1. Quick responses. **(1 point each)**

   (a) Classify the following as a discriminative or generative model. (Circle.) **(1 point each)**

   i. Support vector machines **Ans.** Discriminative

   ii. Hidden Markov Model**Ans.** Generative

   iii. Gaussian mixture model**Ans.** Generative

   iv. Boosted decision trees**Ans.** Discriminative

   (b) True or false.

   i. **True** or **False**. It makes the most sense to use KNN on unlabeled data and K-means on labeled data. **Ans.** False. KNN requires labeled data, and K-means does not require labels.

   ii. **True** or **False**. Kernel SVMs are better than linear SVMs at modeling data that is inherently less complex , and tends to train more quickly. **Ans.** False. Kernel SVMs fit more complicated models, and require higher computational cost.

   iii. **True** or **False**. Dimensionality reduction approaches are powerful when they correctly identify the low-dimensional structure of the underlying data.
   Ans. True.

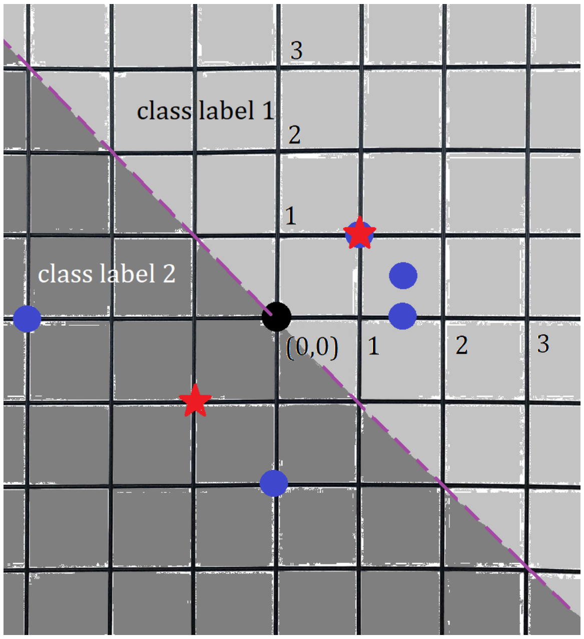
   iv. **True** or **False**. You can only use the One-Vs-All multiclass extension on SVMs; other classifiers are not compatible with this methodology.
   Ans. False. You can apply the One-vs-All extension on any type of binary classification scheme.

   v. **True** or **False**. Any deep neural network can be used as a dimensionality reduction tool. **Ans.** True.

   vi. **True** or **False**. PCA is a variation of gradient descent. **Ans.** False. Principle component analysis, or PCA, is a linear dimensionality reduction tool.

2. **Kmeans and Gaussian mixture models** In this problem you will use some clustering techniques on the following datapoints

	$x_{i,[1]}$	$x_{i,[2]}$
$i=1$	1.5	0
$i=2$	-3	0
$i=3$	1.5	0.5
$i=4$	1	1
$i=5$	0	-2



(a) **(1pt)** Plot the points from the table on the grid. Use large dots.

(b) **(1pt)** We will start with $K=2$ clusters and an initialization of centers $\mu_1=(1,1)$ and $\mu_2=(-1,-1)$. Draw these centers on the grid. Use stars $\star$ as the marker symbol.

(c) **(1pt)** Using these centers, roughly sketch the space on the graph corresponding to the two classification regions. (These are the Voronoi cells.)

(d) **(1pt)** Fill in the classification label (1 if closest to μ_1, 2 if closest to μ_2) in the table below.

(e) **(2pt)** Also, fill in the distance squared ($\|x_i-\mu_j\|_2^2$ from each point to its closest center, and report the average squared distance over all points.

Ans.

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
KNN classifications	1	2	1	1	2
KNN distances	1.25	5	0.5	0	2

It helps to first isolate these vectors:

$$x_1-\mu_1=\begin{bmatrix}0.5\\-1\end{bmatrix},\qquad x_2-\mu_2=\begin{bmatrix}-2\\1\end{bmatrix},\qquad x_3-\mu_1=\begin{bmatrix}0.5\\-0.5\end{bmatrix},\qquad x_4-\mu_1=\begin{bmatrix}0\\0\end{bmatrix},\qquad x_5-\mu_2=\begin{bmatrix}1\\-1\end{bmatrix}$$

which gives us

$$\begin{aligned}
\|x_1-\mu_1\|_2^2 &= (1.5-1)^2+(0-1)^2=1.25\\
\|x_2-\mu_2\|_2^2 &= (-3+1)^2+(0+1)^2=5\\
\|x_3-\mu_1\|_2^2 &= (1.5-1)^2+(0.5-1)^2=0.5\\
\|x_4-\mu_1\|_2^2 &= (1-1)^2+(1-1)^2=0\\
\|x_5-\mu_2\|_2^2 &= (0+1)^2+(-2+1)^2=2.
\end{aligned}$$

Average squared distance is 1.75.

(f) **(5pt)** Now assume that μ_1 and μ_2 represent the centers of a Gaussian mixture model. Return the best fit covariance matrices Σ_1 and Σ_2 and class weight priors α_1 and α_2. You may leave your answers as simple fractions, or as rounded to the nearest 0.01 decimal.

Ans. $\alpha_1 = 3/5$, $\alpha_2 = 2/5$.

$$\Sigma_1 = \frac{1}{3}\left(\begin{bmatrix}0.5\\-1\end{bmatrix}\begin{bmatrix}0.5 & -1\end{bmatrix} + \begin{bmatrix}0.5\\-0.5\end{bmatrix}\begin{bmatrix}0.5 & -0.5\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0 & 0\end{bmatrix}\right) = \frac{1}{12}\begin{bmatrix}2 & -3\\-3 & 5\end{bmatrix}$$

$$\Sigma_2 = \frac{1}{2}\left(\begin{bmatrix}-2 & 1\end{bmatrix}\begin{bmatrix}-2\\1\end{bmatrix} + \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}5 & -3\\-3 & 2\end{bmatrix}$$

(g) **(5pt)** Using your proposed GMM, compute probability density values for data samples $i = 4$ and $i = 5$, given *all* the model parameters (μ_i, σ_i, and α_i). Express your answers in scientific notation, with 3 significant digits, e.g. 2.34×10^{-4} or $3.15 \cdot 10^{0}$.

Hint: it may help to remember that the determinent and inverse of a 2x2 matrix:

$$\det(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = |ad - bc|, \qquad \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Hint: it may also help to remember the PDF of a *joint* Gaussian distribution:

$$p_X(\mu, \Sigma) = \frac{1}{\sqrt{2\pi \det(\Sigma)^n}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where n is the length of the random vector X.

Note the above is wrong, the PDF should be

$$p_X(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

Ans. It helps to first compute some determinants. Recall also that if a matrix Σ is 2x2, then $|c\Sigma| = c^2|\Sigma|$. So, we have

$$|\Sigma_1| = \frac{1}{12^2}(2 \cdot 5 - 3 \cdot 3) = \frac{1}{144} \approx 0.00694, \qquad |\Sigma_2| = \frac{1}{2^2}(5 \cdot 2 - 3 \cdot 3) = \frac{1}{4}.$$

and inverses

$$\Sigma_1^{-1} = 12 \begin{bmatrix} 5 & -3 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} 60 & -36 \\ -36 & 12 \end{bmatrix}, \qquad \Sigma_2^{-1} = 2 \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix} = \begin{bmatrix} 4 & -6 \\ -6 & 10 \end{bmatrix}$$

Then

$$\begin{aligned} p(x_4|\mu_1, \Sigma_1, \alpha_1) &= \frac{\alpha_1}{\sqrt{2\pi|\Sigma_1|^2}} \exp\left(-(x_4 - \mu_1)^T \Sigma_1^{-1} (x_4 - \mu_1)\right) \\ &= \frac{3}{5} \cdot \frac{144}{\sqrt{2\pi}} \exp \underbrace{\left(- \begin{bmatrix} 1 - 1 \\ 1 - 1 \end{bmatrix}^T \begin{bmatrix} 20 & -12 \\ -12 & 8 \end{bmatrix} \begin{bmatrix} 1 - 1 \\ 1 - 1 \end{bmatrix} \right)}_{0} \approx 3.45 \times 10^{1} \text{ or } 34.5 \\ p(x_5|\mu_2, \Sigma_2, \alpha_2) &= \frac{\alpha_2}{\sqrt{2\pi|\Sigma_2|^2}} \exp\left(-(x_5 - \mu_2)^T \Sigma_2^{-1} (x_5 - \mu_2)\right) \\ &= \frac{2}{5} \cdot \frac{4}{\sqrt{2\pi}} \exp \underbrace{\left(- \begin{bmatrix} 0 + 1 \\ -2 + 1 \end{bmatrix}^T \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} 0 + 1 \\ -2 + 1 \end{bmatrix} \right)}_{-13} \approx 1.44 \times 10^{-6} \end{aligned}$$

$$\begin{aligned} p(x_4|\mu_1, \Sigma_1, \alpha_1) &= \frac{\alpha_1}{2\pi\sqrt{|\Sigma_1|}} \exp\left(-(x_4 - \mu_1)^T \Sigma_1^{-1} (x_4 - \mu_1)\right) \\ &= \frac{3}{5} \cdot \frac{12}{\sqrt{2\pi}} \exp \underbrace{\left(- \begin{bmatrix} 1 - 1 \\ 1 - 1 \end{bmatrix}^T \begin{bmatrix} 20 & -12 \\ -12 & 8 \end{bmatrix} \begin{bmatrix} 1 - 1 \\ 1 - 1 \end{bmatrix} \right)}_{0} \approx 2.87 \\ p(x_5|\mu_2, \Sigma_2, \alpha_2) &= \frac{\alpha_2}{2\pi\sqrt{|\Sigma_2|}} \exp\left(-(x_5 - \mu_2)^T \Sigma_2^{-1} (x_5 - \mu_2)\right) \\ &= \frac{2}{5} \cdot \frac{2}{\sqrt{2\pi}} \exp \underbrace{\left(- \begin{bmatrix} 0 + 1 \\ -2 + 1 \end{bmatrix}^T \begin{bmatrix} 2 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} 0 + 1 \\ -2 + 1 \end{bmatrix} \right)}_{-13} \approx 7.21 \times 10^{-7} \end{aligned}$$

Either the black (as hinted) or red (technically correct) solution will be accepted.

(h) **(4pt)** Now increase $K = 3$. Propose new centers μ_1, μ_2, and μ_3 that fits the model best (lowest average distance). Return also the average squared Euclidean distance, either as a simple fraction or a decimal rounded to the nearest 0.01.

Ans. If we are allowed 3 clusters, then my best option is to pick a center that best clusters points $i = 1, 3, 4$, and then have one center each for $i = 2$ and $i = 5$. This corresponds to

$$\mu_1 = \begin{bmatrix} (1.5 + 1.5 + 1)/3 \\ (0 + 0.5 + 1)/3 \end{bmatrix} = \begin{bmatrix} 4/3 \\ 1/2 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}, \mu_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}.$$

Of course, which center goes with which cluster is up to you. Using this assignment, we now have

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
KNN labels	1	2	1	1	3

and the average distance is

$$\big(((1.5 - 4/3)^2 + (0 - 1/2)^2\big) + \big((1.5 - 4/3)^2 + (0.5 - 1/2)^2\big) + \big((1 - 4/3)^2 + (1 - 1/2)^2\big) + 0 + 0\big) \cdot \frac{1}{5} \approx 0.13$$

which is much smaller than that from our initial assignment (2 clusters)

3. It is the Intergalactic Summit on Climate Change, and aliens from around the universe have gathered. It is my job as host to greet them and give them all name tags according to their delegation. However, my memory is terrible, and I need to use machine learning to remember what species each alien is.

   I have some training data, given below, and I will use it to construct a decision tree, which I will use to classify future delegates. In this problem, round all values to the nearest 0.001, and use log base 2.

name	label	eyes	occupation
Admiral Ackbar	mon calamari	huge	military
Ardo Bardai	mon calamari	huge	politician
Lee Char	mon calamari	medium	politician
Chewbacca	wookie	medium	assistant
Tarfful	wookie	medium	military
Gungi	wookie	medium	military
R2-D2	droid	none	assistant
C-3PO	droid	medium	assistant
BB8	droid	none	military

   (a) What is the entropy of the labels over the entire dataset? **(3pts)**
   Ans. Defining

$$\mathbf{Pr}(Y = \text{mon calamari}) = \frac{3}{9}, \qquad \mathbf{Pr}(Y = \text{wookie}) = \frac{3}{9}, \qquad \mathbf{Pr}(Y = \text{droid}) = \frac{3}{9}$$

   the entropy of the collection is

$$\begin{aligned} H(X) &= -\mathbf{Pr}(Y = \text{mon calamari}) \log_2(\mathbf{Pr}(Y = \text{mon calamari})) - \mathbf{Pr}(Y = \text{wookie}) \log_2(\mathbf{Pr}(Y = \text{wookie})) \\ &\quad -\mathbf{Pr}(Y = \text{droid}) \log_2(\mathbf{Pr}(Y = \text{droid})) \\ &= -\frac{3}{9}\log_2\left(\frac{3}{9}\right) - \frac{3}{9}\log_2\left(\frac{3}{9}\right) - \frac{3}{9}\log_2\left(\frac{3}{9}\right) \\ &\approx 1.585 \end{aligned}$$

   (b) For each feature, compute the information gain if that feature was to be used in the first split. Assume that you are allowed to split more than just 2 ways, e.g. one child per feature value.
   Eyes (3pts)
   Ans.

$$\begin{aligned} H(X|\text{eye size}) &= -\mathbf{Pr}(Y = \text{mon calamari, eyes} = \text{huge}) \log_2(Y = \text{mon calamari}|\text{eyes} = \text{huge}) \\ &\quad -\mathbf{Pr}(Y = \text{mon calamari, eyes} = \text{medium}) \log_2(Y = \text{mon calamari}|\text{eyes} = \text{medium}) \\ &\quad -\mathbf{Pr}(Y = \text{wookie, eyes} = \text{medium}) \log_2(Y = \text{wookie}|\text{eyes} = \text{medium}) \\ &\quad -\mathbf{Pr}(Y = \text{droid, eyes} = \text{medium}) \log_2(Y = \text{droid}|\text{eyes} = \text{medium}) \\ &\quad -\mathbf{Pr}(Y = \text{droid, eyes} = \text{none}) \log_2(Y = \text{droid}|\text{eyes} = \text{none}) \\ &= -\frac{2}{9}\log_2(1) - \frac{1}{9}\log_2\left(\frac{1}{5}\right) - \frac{3}{9}\log_2\left(\frac{3}{5}\right) - \frac{1}{9}\log_2\left(\frac{1}{5}\right) - \frac{2}{9}\log_2(1) \\ &\approx 0.762 \end{aligned}$$

$$IG = H(X) - H(X|\text{eyes}) = 1.585 - 0.762 = 0.823$$

   Occupation (3pts)

Ans.

$$\begin{aligned}
H(X|\text{occupation}) &= -\mathbf{Pr}(Y = \text{mon calamari, occ. = military}) \log_2(Y = \text{mon calamari}|\text{occ. = military}) \\
&\quad -\mathbf{Pr}(Y = \text{mon calamari, occ. = politician}) \log_2(Y = \text{mon calamari}|\text{occ. = politician}) \\
&\quad -\mathbf{Pr}(Y = \text{wookie, occ. = assistant}) \log_2(Y = \text{wookie}|\text{occ. = assistant}) \\
&\quad -\mathbf{Pr}(Y = \text{wookie, occ. = military}) \log_2(Y = \text{wookie}|\text{occ. = military}) \\
&\quad -\mathbf{Pr}(Y = \text{droid, occ. = assistant}) \log_2(Y = \text{droid}|\text{occ.= assistant}) \\
&\quad -\mathbf{Pr}(Y = \text{droid, occ. = military}) \log_2(Y = \text{droid}|\text{occ.= military}) \\
&= -\frac{1}{9}\log_2\left(\frac{1}{4}\right) - \frac{2}{9}\log_2(1) - \frac{1}{9}\log_2\left(\frac{1}{3}\right) - \frac{2}{9}\log_2\left(\frac{2}{4}\right) - \frac{2}{9}\log_2\left(\frac{2}{3}\right) - \frac{1}{9}\log_2\left(\frac{1}{4}\right) \\
&\approx 0.973
\end{aligned}$$

$$IG = H(X) - H(X|\text{ear shape}) = 1.585 - 0.973 = 0.612$$

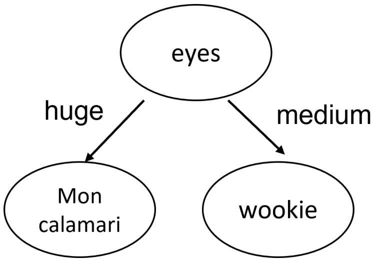
(c) **(1pt)** Report which feature has the largest information gain when revealed.
Ans. Eye size has the larger information gain.

(d) Now, suppose you decide to try an ensemble method. You take some minibatches, and from those form **binary** decision stumps (1-depth decision trees), which you will ensemble later.

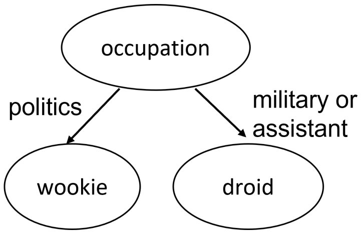
Draw the decision stumps for the minibatches specified below. Now each stump may only have one split (1 parent, 2 children), so you cannot group in 3 or 4 ways. (e.g. you cannot split in 4 children for red, blue, brown, and gold, but you can split into 2 children for (red and blue) and (white), etc.

Ans. From this point on, there was way too much ambiguity in the solutions. I tried to give as many points as possible with any solution I thought was "reasonable", and in places where it didn't exactly say to give a justification, I did give full credit for lucky guesses. If however you didn't guess lucky and you didn't have any justification, you didn't get any points.

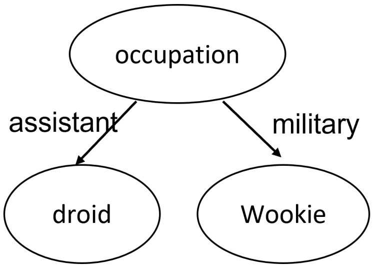
i. **(1pt)** Minibatch 1: {Admiral Ackbar, Ardo Bardai, Chewbacca, Gungi}
Ans.



ii. **(1pt)** Minibatch 2: {Ardo Bardai, Lee Chair,C-3PO, BB8} **Ans.**



iii. **(1pt)** Minibatch 3: {Tarful, Gungi, C-3PO, R2D2} **Ans.**



Ans. Other possible solutions include: one feature vs rest, or in the third minibatch eye size was also acceptable.

(e) **(2pts)** Give the classification formed by each decision stump over the *entire* dataset. (Note: if the decision stump doesn't provide a label, you may just label it "not specified" or NS.) (Give your answer by filling out the middle 3 columns of the table below.)

(f) **(2pts)** Now using a majority voting system (where not specified = no vote), classify each alien and report the final classification accuracy. (If there is no clear majority, then label it NS and count it as a misclassified training sample.) (Give your answer by filling out the last column of the table below.)

(It may help to use abbreviations: **MC** = mon calamari, **W** = wookie, **D** = droid.)

Ans. You must have consistent predictions based on your proposed stumps. Your majority vote must be consistent with your predictions.

name	true label	pred. by stump 1	pred. by stump 2	pred. by stump 3	Majority vote classification
Admiral Ackbar	mon calamari	MC	D	W	NS
Ardo Bardai	mon calamari	MC	MC	NS	MC
Lee Char	mon calamari	W	MC	NS	NS
Chewbacca	wookie	W	D	D	D
Tarfful	wookie	W	D	W	W
Gungi	wookie	W	D	W	W
R2-D2	droid	NS	D	D	D
C-3PO	droid	W	D	D	D
BB8	droid	NS	D	W	NS

classification accuracy is 5/9.

(g) **(3pts)** We will now construct a weighting scheme on each of the decision stumps such that we achieve maximum accuracy. That is, suppose that the predictions from each stump is represented by a 1-hot vector

$$\text{mon calamari} = \begin{bmatrix}1\\0\\0\end{bmatrix}, \qquad \text{wookie} = \begin{bmatrix}0\\1\\0\end{bmatrix}, \qquad \text{droid} = \begin{bmatrix}0\\0\\1\end{bmatrix}.$$

So, if the first stump predicts wookie, the second a droid, and the third mon calamari, then we will have

$$\hat{y}_1 = \begin{bmatrix}0\\1\\0\end{bmatrix}, \qquad \hat{y}_2 = \begin{bmatrix}0\\0\\1\end{bmatrix}, \qquad \hat{y}_3 = \begin{bmatrix}1\\0\\0\end{bmatrix}.$$

and we have the true label be identified by the index of the largest value in the aggregated vector

$$\hat{y} = \alpha_1\hat{y}_1 + \alpha_2\hat{y}_2 + \alpha_3\hat{y}_3.$$

Pick (circle) from the following the set of values for α_1,α_2, α_3 such that $\hat{y}$ achieves maximum accuracy over the training set.

(A) $\alpha_1 = 0.2$, $\alpha_2 = 0.2$, $\alpha_3 = 0.6$

(B) $\alpha_1 = 0.1$, $\alpha_2 = 0.6$, $\alpha_3 = 0.3$

(C) $\alpha_1 = 0.6$, $\alpha_2 = 0.1$, $\alpha_3 = 0.3$

Ans. If you had different stumps than me, or different predictions, I went off your answers. These are mine. Based on everything given, we know that $\hat{y}$ will look like the following, where each row corresponds to one row of the aliens in the training table. The bolded value also corresponds to the correct label.

$$\hat{y} = \begin{bmatrix}
\boldsymbol{\alpha_1} & \alpha_3 & \alpha_2 \\
\boldsymbol{\alpha_1 + \alpha_2} & 0 & 0 \\
\boldsymbol{\alpha_2} & \alpha_1 & 0 \\
0 & \boldsymbol{\alpha_1} & \alpha_2 + \alpha_3 \\
0 & \boldsymbol{\alpha_1 + \alpha_3} & \alpha_2 \\
0 & \boldsymbol{\alpha_1 + \alpha_3} & \alpha_2 \\
0 & 0 & \boldsymbol{\alpha_2 + \alpha_3} \\
0 & \alpha_1 & \boldsymbol{\alpha_2 + \alpha_3} \\
0 & \alpha_3 & \boldsymbol{\alpha_2}
\end{bmatrix}$$

Using each weighting scheme, we'd have

$$A : \hat{y} = \begin{bmatrix}
\mathbf{0.2} & 0.6 & 0.2 \\
\mathbf{0.4} & 0 & 0 \\
\mathbf{0.2} & 0.2 & 0 \\
0 & \mathbf{0.2} & 0.8 \\
0 & \mathbf{0.8} & 0.2 \\
0 & \mathbf{0.8} & 0.2 \\
0 & 0 & \mathbf{0.8} \\
0 & 0.2 & \mathbf{0.8} \\
0 & 0.6 & \mathbf{0.2}
\end{bmatrix}, \qquad B : \hat{y} = \begin{bmatrix}
\mathbf{0.1} & 0.3 & 0.6 \\
\mathbf{0.7} & 0 & 0 \\
\mathbf{0.6} & 0.1 & 0 \\
0 & \mathbf{0.1} & 0.9 \\
0 & \mathbf{0.4} & 0.6 \\
0 & \mathbf{0.4} & 0.6 \\
0 & 0 & \mathbf{0.9} \\
0 & 0.1 & \mathbf{0.9} \\
0 & 0.3 & \mathbf{0.6}
\end{bmatrix}, \qquad C : \hat{y} = \begin{bmatrix}
\mathbf{0.6} & 0.1 & 0.3 \\
\mathbf{0.9} & 0 & 0 \\
\mathbf{0.3} & 0.6 & 0 \\
0 & \mathbf{0.6} & 0.4 \\
0 & \mathbf{0.7} & 0.3 \\
0 & \mathbf{0.7} & 0.3 \\
0 & 0 & \mathbf{0.4} \\
0 & 0.6 & \mathbf{0.4} \\
0 & 0.1 & \mathbf{0.3}
\end{bmatrix}.$$

and from this we can see that A has 5 correct, B has 5 correct, and C has 7 correct. So C is the best weighting.

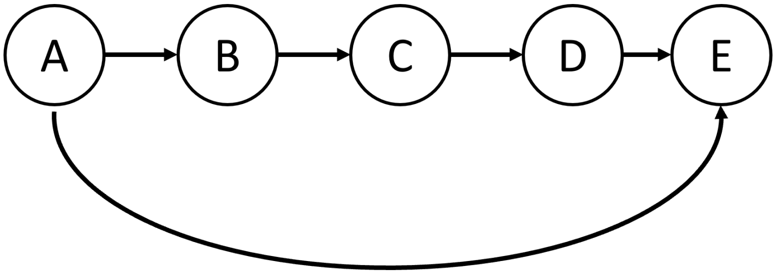
4. **The Nefarious British Baking Show.** It is time for the annual British Baking Competition, and the judges are super excited to taste delicious baked goods. This year, there are five contestants: Alice (A), Bob (B), Carlos (C), Dima (D), and Eve (E).

   (a) The first challenge is to bake a cake. The judges will taste each contestant's bread and give it either a pass or fail. On the evening after the bake is done but before the results are revealed, the contestants all get Mary Berry drunk, and she drops some hints as to who passed and who didn't.

   This is what drunk Mary tells the contestants.

   - If Alice passed, then Bob has a 50% chance of passing. Otherwise, Bob definitely failed.
   - If Bob passed, then Carlos definitely passed. If Bob failed, then Carlos definitely failed.
   - If Carlos passed, then Dima definitely failed. If Carlos failed, then Dima definitely passed.
   - If Dima and Alice both passed, then Eve has a 50% chance of passing. Otherwise, Eve definitely failed.

   i. **(1pts)** Draw the graphical model corresponding to this. Don't worry about labeling with probability values, just draw the nodes, label the nodes, and draw any arrows or edges needed. **Ans.**



   ii. **(3pts)** What is the probability that Eve passed, given that Alice has an 80% chance of passing? **Ans.** From the graphical model, we can decompose

$$\begin{aligned}
\mathbf{Pr}(E=1) &= \sum_{a,b,c,d} \mathbf{Pr}(E=1|A=a,B=b,C=c,D=d)\mathbf{Pr}(A=a,B=b,C=c,D=d) \\
&= \sum_{a,b,c,d} \mathbf{Pr}(E=1|A=a,D=d)\mathbf{Pr}(D=d|C=c)\mathbf{Pr}(C=c|B=b)\mathbf{Pr}(B=b|A=a)\mathbf{Pr}(A=a) \\
&= \mathbf{Pr}(E=1|A=1,D=1)\mathbf{Pr}(D=1|C=0)\mathbf{Pr}(C=0|B=0)\mathbf{Pr}(B=0|A=1)\mathbf{Pr}(A=1) \\
&= 50\% \cdot 100\% \cdot 100\% \cdot 50\% \cdot 80\% \\
&= 20\%
\end{aligned}$$

(b) It is now the second day of the challenge, and Dima and Eve have called in sick, so we are left only with Alice, Bob, and Carlos. The second challenge is now to bake a pie. Each contestant has a 75% chance of passing on their own merit. However, they're all a bit nervous, and sometimes fall prone to cheating. The contestants all bake in a row, so that

- Alice is in front, and cannot see anyone else's bake.
- Bob can see Alice's bake but no one else's
- Carlos can see Bob's bake but no one else's

The effect of this is some small chance of cheating; each contestant has a 25% chance of copying the bake of the person they can see, and a 75% chance that they just do what they know.

The judges make their decisions and tell each contestant if they passed or failed. The contestants then call their family members. We do not get to know what the judges told the contestants, but we get to observe the faces of the family members, who arrive later that day to take the contestants home.

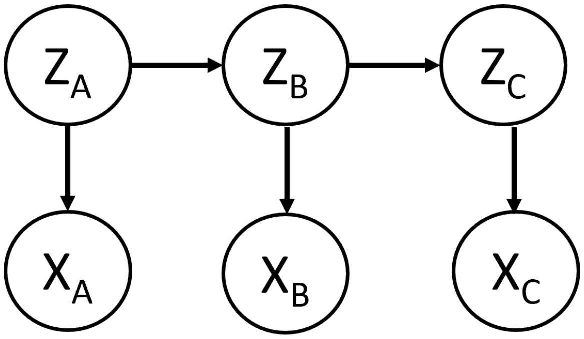
Here is what we know:

- Alice's husband loves to smile. If Alice passed, her husband has a 75% chance of smiling. If Alice failed, her husband will still smile 50% of the time.
- If Bob fails his bake, there's a 75% chance Bob's husband will cry. But even if Bob succeeds, there's a 50% chance his husband will cry, since the family dog died that day.
- Carlos' baby sister just smiles all the time.

We will denote the observations on the family members as $X_i = 1$ if the family of person i was smiling, and $X_i = 0$ if that person was frowning or crying, for $i \in \{A, B, C\}$. We denote whether this person passed or failed as $Z_i = 1$ (pass) and $Z_i = 0$ (failed).

Report all answers as either simple fractions or decimals rounded to the nearest 0.001.

i. **(1pts)** Draw a graphical model corresponding to this scenario, where we do not know the outcomes of the contestants' bakes, but we can see the reaction of their family members.
Ans.



ii. **(2pts)** Give the *emission probabilities*
Ans.

$$\begin{aligned}
\mathbf{Pr}(X_A = 1|Z_A = 1) &= 75\%, &\qquad \mathbf{Pr}(X_A = 1|Z_A = 0) &= 50\%,\\
\mathbf{Pr}(X_B = 1|Z_B = 1) &= 50\%, &\qquad \mathbf{Pr}(X_B = 1|Z_B = 0) &= 25\%,\\
\mathbf{Pr}(X_C = 1|Z_C = 1) &= 100\%, &\qquad \mathbf{Pr}(X_C = 1|Z_C = 0) &= 100\%,
\end{aligned}$$

iii. **(1pts)** Give the *transition probabilities*
Ans. Because of the symmetry of the problem,

$$\begin{aligned}
\mathbf{Pr}(Z_B = 1|Z_A = 1) &= \mathbf{Pr}(Z_C = 1|Z_B = 1) = 25\% \cdot 1 + 75\% \cdot 75\% = \frac{13}{16}\\
\mathbf{Pr}(Z_B = 1|Z_A = 0) &= \mathbf{Pr}(Z_C = 1|Z_B = 0) = 25\% \cdot 0 + 75\% \cdot 75\% = \frac{9}{16}
\end{aligned}$$

iv. **(6pts)** It turns out that every family member there is smiling. Compute the *forward messages*:

$$\alpha_i = \mathbf{Pr}(Z_i, X_i, X_{F(i)})$$

where $F(i)$ is the set of people strictly in front of person i. Give up to 3 decimal digits.
Ans.

$$\begin{aligned}
\alpha_A(a) &= \mathbf{Pr}(Z_A = a, X_A = 1) = \mathbf{Pr}(X_A = 1 | Z_A = a)\mathbf{Pr}(Z_A = a) = \begin{cases} 75\% \cdot 75\% = \frac{9}{16} & \text{if } a = 1 \\ 50\% \cdot 25\% = \frac{1}{8} & \text{if } a = 0 \end{cases} \\
\alpha_B(b) &= \mathbf{Pr}(Z_B = b, X_B = 1, X_A = 1) = \mathbf{Pr}(X_B = 1 | Z_B = b) \sum_{a \in \{0,1\}} \mathbf{Pr}(Z_B = b | Z_A = a)\mathbf{Pr}(Z_A = a, X_A = 1) \\
&= \begin{cases} 50\% \cdot (\frac{13}{16} \cdot \frac{9}{16} + \frac{9}{16} \cdot \frac{1}{8}) = \frac{135}{512} \approx 0.264 & \text{if } b = 1 \\ 25\% \cdot (\frac{3}{16} \cdot \frac{9}{16} + \frac{7}{16} \cdot \frac{1}{8}) = \frac{41}{512} \approx 0.080 & \text{if } b = 0 \end{cases} \\
\alpha_C(c) &= \mathbf{Pr}(Z_C, X_C = 1, X_B = 1, X_A = 1) \\
&= \mathbf{Pr}(X_C = 1 | Z_C = c) \sum_{b \in \{0,1\}} \mathbf{Pr}(Z_C = c | Z_B = b)\mathbf{Pr}(Z_B = b, X_B = 1, X_A = 1) \\
&= \begin{cases} 100\% \cdot (\frac{13}{16} \cdot \frac{135}{512} + \frac{9}{16} \cdot \frac{41}{512}) = \frac{2124}{8192} \approx 0.259 & \text{if } c = 1 \\ 100\% \cdot (\frac{3}{16} \cdot \frac{377}{512} + \frac{7}{16} \cdot \frac{471}{512}) = \frac{4428}{8192} \approx 0.541 & \text{if } c = 0 \end{cases}
\end{aligned}$$

v. **(4pts)** Compute the *backward messages*, e.g.

$$\beta_i = \mathbf{Pr}(X_{B(i)}|Z_i)$$

where $B(i)$ is the set of people strictly behind i. Give up to 3 decimal digits. Start with $\beta_C(c) = 1$ for $c = 0$ and $c = 1$.
Ans.

$$\begin{aligned}
\beta_B(b) &= \mathbf{Pr}(X_C = 1|Z_B = b) \\
&= \underbrace{\mathbf{Pr}(X_C = 1|Z_C = 0)}_{=100\%}\mathbf{Pr}(Z_C = 0|Z_B = b) \\
&\qquad + \underbrace{\mathbf{Pr}(X_C = 1|Z_C = 1)}_{=100\%}\mathbf{Pr}(Z_C = 1|Z_B = b) = 1 \\
\beta_A(a) &= \mathbf{Pr}(X_B = 1, X_C = 1|Z_A = a) \\
&= \sum_{b\in\{0,1\}} \mathbf{Pr}(X_C = 1|Z_B = b)\mathbf{Pr}(X_B = 1|Z_B = b)\mathbf{Pr}(Z_B = b|Z_A = a) \\
&= \begin{cases} 1 \cdot \frac{1}{2} \cdot \frac{13}{16} + 1 \cdot \frac{1}{4}\frac{3}{16} = \frac{29}{64} \approx 0.453 & \text{if } a = 1 \\ 1 \cdot \frac{1}{2} \cdot \frac{9}{16} + 1 \cdot \frac{1}{4}\frac{7}{16} = \frac{25}{64} \approx 0.391 & \text{if } a = 0 \end{cases}
\end{aligned}$$

vi. **(2pts)** Given the observations of the smilers, what are the chances that Alice passed her bake? Give your answer to the nearest 0.1 of a percent or as a simple fraction. **Ans.**

$$\begin{aligned}
\mathbf{Pr}(Z_A = 1|X) &= \rho\alpha_A(1)\beta_A(1) = \rho \cdot \frac{9}{16} \cdot \frac{29}{64} = \frac{261}{1024} \approx 0.255 \\
\mathbf{Pr}(Z_A = 0|X) &= \rho\alpha_A(0)\beta_A(0) = \rho \cdot \frac{1}{8} \cdot \frac{25}{64} = \frac{25}{512} \approx 0.0488
\end{aligned}$$

Normalizing, we get

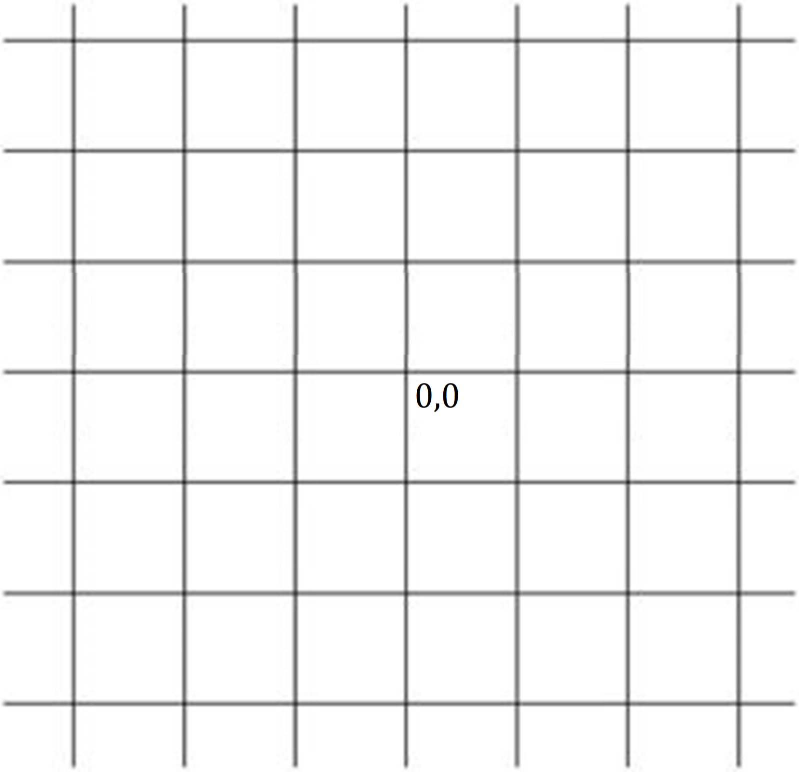
$$\mathbf{Pr}(Z_B = 1|X) = \frac{261}{261 + 50} \approx 83.9\%.$$

5. **(Extra credit) SVMs with duality.** Now consider the dataset

Dataset A

i	1	2	3	4
x_i	$(1, 0)$	$(0, 1)$	$(-1, 0)$	$(0, -1)$
y_i	1	-1	-1	1

You are not required to plot anything, but here is a grid that you may use to help your computations.



(a) **(3pts)** Recall that the primal hard-margin SVM is created by solving the following optimization problem.

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \frac{1}{2}\|\theta\|_2^2 \\ \text{subject to} & y_i x_i^T \theta \geq 1, \quad i = 1, ..., m \end{array} \tag{1}$$

Solve (1) for the given labeling, by providing the optimal θ in (1).
Hint: Solve for the direction and magnitude of θ separately.
Ans. Well, by eyeballing the picture, the best θ should be the one where

$$\theta^T \begin{bmatrix} a & a \end{bmatrix} = 0.$$

For example, a feasible $\theta = c \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Then

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \theta = c \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and we can see $c = 1$ is the largest feasible scaling. Therefore the optimal $\theta = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

(b) **(3pts)** Now recall that the dual hard-margin SVM is found by solving the optimization problem below.

$$\begin{array}{ll} \underset{u}{\text{maximize}} & -\frac{1}{2}u^T ZZ^T u + \sum_i u_i \\ \text{subject to} & 0 \leq u \end{array} \tag{2}$$

where the rows of Z are $y_i x_i^T$.
Optimize the dual solution by finding an optimal u.
Hint: there is some very special structure in ZZ^T which makes solving some linear system very easy by hand. At no point should you need a calculator.
Hint: If a solution is optimal for an unconstrained problem and feasible when constraints are added, it is still optimal for the constrained problem.
Ans. Actually, the key here is to see that

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad Z = \begin{bmatrix} I & I \\ I & I \end{bmatrix}.$$

Then one way to approach solving (2) is to first minimize the objective, and then check if that solution is feasible. We can do this by solving

$$ZZ^T u = \mathbf{1} \iff \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This is satisfied by $u = \frac{1}{2}\mathbf{1}$, which is feasible and therefore optimal. But, any other feasible solution satisfying the above equality is acceptable.

(c) **(3pts)** Based on your answer to the previous question, what are the support vectors? **Ans.** Whatever vector u is presented in the previous question, the support vectors should be the set $\{i : u_i > 0\}$. In my example, $\{1, 2, 3, 4\}$ would be the set of support vectors, but it is possible to also have $\{1, 2\}$, $\{1, 3\}$, $\{2, 4\}$, $\{3, 4\}$ as well, or any set of 3 points. All other combinations are not technically correct, but can be accepted if it corresponds with a (wrong) previously computed optimal u.

(d) **(3pts)** Now let's add an outlier to the mix:

Dataset B

i	1	2	3	4	5
x_i	$(1,0)$	$(0,1)$	$(-1,0)$	$(0,-1)$	$(0,-1)$
y_i	1	-1	-1	1	-1

Show that the dual solution (answer to problem (2)) is unbounded, e.g. show that $u = cv$ is dual feasible for some v and all positive scalars c, and the objective value D^* increases with c. [1]

Ans. Now our matrices are

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}, \qquad Z = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}.$$

and

$$\frac{1}{2}u^T Z^T Z u = \frac{1}{2}((u_1+u_3)^2 + (u_2+u_4-u_5)^2).$$

Many possible choices of $u = cv$ can be feasible for any $c > 0$. For example, we can pick

$$v = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \end{bmatrix}$$

and then

$$\frac{1}{2}u^T Z^T Z u - u^T \mathbf{1} = \frac{1}{2}((2c)^2) - 6c = 2c^2 - 6c.$$

Taking $c \to +\infty$ takes the dual objective to $+\infty$.

But in fact many answers here are possible.

(e) (Pretty hard) **(3pts)** Finally, we can consider the primal and dual pair of the soft-margin SVM problem

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \frac{1}{2}\|\theta\|_2^2 + \rho\sum_i s_i \\ \text{subject to} & y_i x_i^T\theta \geq 1 - s_i, \quad i = 1, ..., m \\ & s_i \geq 0, \quad i = 1, ..., m \end{array} \qquad \begin{array}{ll} \underset{u}{\text{maximize}} & -\frac{1}{2}u^T Z Z^T u + \sum_i u_i \\ \text{subject to} & 0 \leq u \leq \rho \end{array} \tag{3}$$

What is the smallest possible ρ such that the resulting θ, solved over Dataset B, is the same as that in Dataset A? Give a choice of optimal values of u for that chosen range of ρ.

You may make your argument by attacking either the primal or dual formulations.

If you pick a ρ using some systematic guess-and-check routine, you can get partial credit, but for full credit you should have some kind of justification.

Hint: Complementary slackness can help here.

Ans. While there may be multiple ways of answering this question, I think a more straightforward way to attack this is through the dual. Since

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \Rightarrow \theta = Z^T u = \begin{bmatrix} u_1 + u_3 \\ -u_2 - u_4 + u_5 \end{bmatrix}.$$

I want $\theta = (1, -1)$ so I can pick any combination of $u_1 + u_3 = 1$, and $u_5 - u_4 - u_2 = -1$.

[1] This is sometimes called finding the recession cone.

I also know from complementary slackness conditions, that I need either u_1 or u_3 to be positive, u_2 or u_4 to be positive, and since I would like u_5 to be classified as an outlier, $u_5 = \rho$.

This gets us to

$$u_1 + u_3 = 1, \qquad u_4 + u_2 = \rho + 1.$$

This suggests that at the very least, $\rho > 1$ is needed. Let's see if that works. In order for u to be dual-feasible, it must be that $\max_i u_i = \rho = u_5$, so this forces $u_4 = u_2 < \frac{\rho+1}{2} < 1$. However there are no new constraints on u_1 and u_3, so any choice

$$u = (u_1, \frac{\rho+1}{2}, u_3, \frac{\rho+1}{2}, \rho), \qquad u_1 + u_3 = 1$$

works, for $\rho > 1$. The smallest possible ρ is indeed $\rho = 1$.