PRACTICE EXAM 2

Exam will be Tuesday, November 2, 2021

- This will be a CLOSED BOOK exam.
- You will be permitted one sheet of handwritten notes, 8.5x11.
- Calculators and computers are not permitted.
- If you’re taking the exam online, you will need to have your webcam turned on. Your exam will appear on Gradescope at exactly 9:30am; you will need to photograph and upload your answers by exactly 11:00am.
- There will be a total of 100 points in the exam. Each problem specifies its point total. Plan your work accordingly.
- You must SHOW YOUR WORK to get full credit.
1. (16 points) Suppose you have an \( M \times D \) matrix, \( X = [\vec{x}_0, \ldots, \vec{x}_{M-1}]^T \), where \( \sum_{m=0}^{M-1} \vec{x}_m = 0 \). The eigenvalues of \( X^T X \) are \( \lambda_0 \) through \( \lambda_{D-1} \), its eigenvectors are \( \vec{v}_0 \) through \( \vec{v}_{D-1} \), and its principal components are \( Y = X V \).

(a) Write \( Y^T Y \) in terms of the eigenvalues, \( \lambda_0 \) through \( \lambda_{D-1} \).

(b) Write \( \sum_{m=0}^{M-1} \| \vec{x}_m \|_2^2 \) in terms of the eigenvalues, \( \lambda_0 \) through \( \lambda_{D-1} \).

(c) Write \( \vec{v}_i^T X^T X \vec{v}_j \) in terms of the eigenvalues, \( \lambda_0 \) through \( \lambda_{D-1} \), for \( 0 \leq i \leq j \leq D - 1 \).
2. (16 points) A 2-dimensional Gaussian random vector has mean \( \vec{\mu} \) and covariance \( \Sigma \) given by

\[
\vec{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}
\]

Draw a curve of some kind, on a two-dimensional Cartesian plane, showing the set of points \( \{ \vec{x} : p_X(\vec{x}) = \frac{1}{8\pi} e^{-\frac{1}{2}} \} \).
3. (16 points) A particular HMM-based speech recognizer only knows two words: word \( w_0 \), and word \( w_1 \). Word \( w_0 \) has a higher \textit{a priori} probability: \( p_Y(w_0) = 0.7 \), while \( p_Y(w_1) = 0.3 \). Each of the two words is modeled by a four-state Gaussian HMM \((N = 4)\) with three-dimensional observations \((D = 3)\). All states, in both HMMs, have identity covariance \((\Sigma = I)\). Both HMMs have \textit{exactly} the same transition probabilities and state-dependent means, given by:

\[
\text{Both Words: } A = \begin{bmatrix}
0.25 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.25 & 0.25 \\
\end{bmatrix}, \quad \vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{\mu}_3 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \vec{\mu}_4 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}
\]

But the initial residence probabilities are different:

\[
\text{Word 0: } \pi_i = \begin{cases} 1 & i = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{Word 1: } \pi_i = \begin{cases} 1 & i = 4 \\ 0 & \text{otherwise} \end{cases}
\]

Suppose that you have a two-frame observation, \( X = [\vec{x}_1, \vec{x}_2] \), where \( \vec{x}_t = [x_{1t}, x_{2t}, x_{3t}] \). The MAP decision rule, in this case, can be written as a linear classifier,

\[
\hat{y} = \begin{cases} w_1 \vec{w}_1^T \vec{x}_1 + \vec{w}_2^T \vec{x}_2 + b > 0 \\ w_0 \end{cases}
\]

Find \( \vec{w}_1 \), \( \vec{w}_2 \), and \( b \).
4. (16 points) In terms of $\alpha_t(i)$, $\beta_t(i)$, $a_{ij}$, $\pi_i$ and $b_i(\vec{x}_t)$, find

$$p(q_6 = i, q_7 = j | \vec{x}_1, \ldots, \vec{x}_{20})$$

5. (5 points) Suppose you have a dataset including the vectors

$$\vec{x} = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}, \quad \vec{z} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

Find a diagonal matrix $\Sigma$ such that $d_2^2(\vec{x}, \vec{y}) > d_2^2(\vec{x}, \vec{z})$. 

6. (10 points) Define \( \Phi(z) \) as follows:

\[
\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du
\]

Suppose \( \vec{X} = [X_1, X_2]^T \) is a Gaussian random vector with mean and covariance given by

\[
\vec{\mu} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix}
\]

(a) Sketch the set of points such that \( f_{\vec{X}}(\vec{x}) = \frac{1}{12\pi} e^{-\frac{1}{2}} \), where \( f_{\vec{X}}(\vec{x}) \) is the pdf of \( \vec{X} \).

(b) In terms of \( \Phi(z) \), find the probability \( \Pr \{ -1 < X_1 < 1, -1 < X_2 < 1 \} \).
7. (10 points) Suppose that a particular covariance matrix $\Sigma$ has the following eigenvector matrix, $U$, and eigenvalue matrix, $\Lambda$:

$$U = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

Let $\bar{y}(\bar{x}) = \begin{bmatrix} y_1(\bar{x}) \\ y_2(\bar{x}) \end{bmatrix} = U^T \bar{x}$ be the principal components of a vector space $\bar{x}$.

(a) Plot the set of vectors $\bar{x}$ such that $y_1(\bar{x}) = 3$.

(b) Find the squared Mahalanobis distance, $d^2_\Sigma(\bar{x}, \bar{\mu})$, between the vectors $\bar{x}$ and $\bar{\mu}$ where

$$\bar{x} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \quad \bar{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
8. (10 points) Suppose that, for a particular classification problem, the correct label of every data point is as follows:

\[ y^*(\vec{x}) = \begin{cases} 
1 & \|\vec{x}\|_2 < 1.5 \\
0 & \|\vec{x}\|_2 > 1.5 
\end{cases} \tag{1} \]

Unfortunately, you aren’t allowed to use the correct labeling function. Instead, you have to try to learn a nearest-neighbor or Bayesian classifier.

(a) Your nearest-neighbor classifier is trained using 25 training samples, taken at integer coordinates for \(-2 \leq x_1, x_2 \leq 2\). Fortunately, your training data are correctly labeled, using the labeling function shown in Eq. (1). Thus the complete training dataset is

\[ X = \begin{bmatrix} -2 & -2 & \ldots & 0 & 0 & \ldots & 2 \\ -2 & -1 & \ldots & 0 & 1 & \ldots & 2 \end{bmatrix}, \quad Y = [0, 0, \ldots, 1, 1, 0, \ldots, 0] \]

Using these 25 training examples, you construct a nearest-neighbor classifier. Draw the decision boundary of the resulting nearest-neighbor classifier.

(b) Suppose now that \( f_{\vec{X}|Y}(\vec{x}|0) \) and \( f_{\vec{X}|Y}(\vec{x}|1) \) are both zero-mean Gaussian pdfs, with the covariance matrices \( \Sigma_0 \) and \( \Sigma_1 \) respectfully, where

\[ \Sigma_0 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

Define \( \eta \) to be the odds ratio, \( \eta = p_Y(0)/p_Y(1) \). Find a value of \( \eta \) such that a Bayesian classifier gives exactly the decision boundary shown in Eq. (1) on the previous page.
9. (10 points) Suppose that you have $M$ different $D$-dimensional vectorized face images, $\vec{\Gamma}_m = [\gamma_{1m}, \ldots, \gamma_{Dm}]^T$, whose mean is $\vec{\Psi} = [\psi_1, \ldots, \psi_D]^T$. Define the data matrix to be $A = [\vec{\Gamma}_1 - \vec{\Psi}, \ldots, \vec{\Gamma}_M - \vec{\Psi}]$, and suppose that the eigenvectors and eigenvalues of $A^T A$ are given by $U = [\vec{u}_1, \ldots, \vec{u}_M]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_M)$.

(a) Find the numerical value of the vector $U^T \vec{u}_3$.

(b) Your goal is to find a $(D \times M)$ matrix $V = [\vec{v}_1, \ldots, \vec{v}_M]$ so that $\vec{\Omega}_m = V^T (\vec{\Gamma}_m - \vec{\Psi})$ is a vector containing the first $M$ principal components of the image $\vec{\Gamma}_m$. Write an equation showing how $V$ can be computed from $\vec{\Psi}$, $A$, $U$, and/or $\Lambda$. 
10. (10 points) Suppose that you have $M$ different $D$-dimensional vectorized face images, $\vec{\Gamma}_m = [\gamma_{1m}, \ldots, \gamma_{Dm}]^T$, whose mean is $\vec{\Psi} = [\psi_1, \ldots, \psi_D]^T$. Define the scatter matrix to be

$$S = \sum_{m=1}^{M} (\vec{\Gamma}_m - \vec{\Psi})(\vec{\Gamma}_m - \vec{\Psi})^T$$

Suppose that the eigenvectors and eigenvalues of $S$ are $V = [\vec{v}_1, \ldots, \vec{v}_D]$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_D)$. You want to find a value of $K$ such that the $K$-dimensional PCA projection $\vec{\Omega}_m = [\vec{v}_1, \ldots, \vec{v}_K]^T (\vec{\Gamma}_m - \vec{\Psi})$ has the following property:

$$\sum_{m=1}^{M} \left| \vec{\Omega}_m \right|^2 = (0.95) \sum_{m=1}^{M} \left| \vec{\Gamma}_m - \vec{\Psi} \right|^2$$

(2)

Specify an equation that, if satisfied, will guarantee the truth of Eq. 2. Your equation should only include the scalars $M$, $D$, $K$, and/or the eigenvalues $\lambda_d$ ($1 \leq d \leq D$); your equation should not include $\vec{\Gamma}_m$ or $\vec{\Psi}$. 
11. (20 points) A particular dataset has three data,

\[
\vec{x}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad \vec{x}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{x}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}
\]

Define \( X = [\vec{x}_1, \vec{x}_2, \vec{x}_3] \) and \( R = X^T X \). The matrix \( R \) is given by \( R = V \Lambda V^T \) where

\[
V = \begin{bmatrix} -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}
\]

Find a matrix \( W \) such that \( \vec{y}_i = W^T \vec{x}_i \), \( \vec{y}_i \) is two-dimensional, and the elements of \( \vec{y}_i \) are uncorrelated.
12. (16 points) A particular dataset has the scatter matrix $S = \sum_{k=1}^{n}(\vec{x}_k - \vec{m})(\vec{x}_k - \vec{m})^T$, whose first two eigenvectors are $\vec{v}_1$ and $\vec{v}_2$, characterized by eigenvalues $\lambda_1 = 450$ and $\lambda_2 = 150$. Define the transform $\vec{y}_k = [\vec{v}_1, \vec{v}_2]^T (\vec{x}_k - \vec{m})$. Define the $2 \times 2$ matrix $Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \sum_{k=1}^{n} \vec{y}_k \vec{y}_k^T$.

Find the numerical values of the elements $q_{11}, q_{12}, q_{21},$ and $q_{22}$ of matrix $Q$. 


13. (16 points) A particular dataset has six data vectors, given by

\[ \{ \vec{x}_1, \ldots, \vec{x}_6 \} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \right\} \]

By calling \texttt{np.random.rand}, you generate a 3 \times 2 random projection matrix \( V \), given by

\[ V = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \\ v_{31} & v_{32} \end{bmatrix} \]

Using this random projection matrix, you compute the transformed feature vectors \( \vec{y}_k = V^T \vec{x}_k \). The total energy of the transformed dataset can be written as

\[ E = \sum_{k=1}^{6} \vec{y}_k^T \vec{y}_k \]

Find the value of \( E \) in terms of the random projection matrix elements \( v_{ij} \).
14. (16 points) You want to classify zoo animals. Your zoo only has two species: elephants and giraffes. There are more elephants than giraffes: if $Y$ is the species,

$$p_Y(\text{elephant}) = \frac{e}{e+1}$$
$$p_Y(\text{giraffe}) = \frac{1}{e+1}$$

where $e = 2.718\ldots$ is the base of the natural logarithm. The height of giraffes is Gaussian, with mean $\mu_G = 5$ meters and variance $\sigma_G^2 = 1$. The height of elephants is also Gaussian, with mean $\mu_E = 3$ and variance $\sigma_E^2 = 1$. Under these circumstances, the minimum probability of error classifier is 

$$\hat{y}(x) = \begin{cases} 
\text{giraffe} & x > \theta \\
\text{elephant} & x < \theta 
\end{cases}$$

Find the value of $\theta$ that minimizes the probability of error.
15. (16 points) Random vector \( X \) is distributed as

\[
p_X(\vec{x}) = \sum_{k=1}^{2} c_k \mathcal{N}(\vec{x}|\vec{\mu}, \Sigma)
\]

where \( c_1 = c_2 = 0.5 \), and

\[
\vec{\mu}_1 = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}
\]

Draw a contour plot showing \( p_X(\vec{x}) \) as a function of \( \vec{x} \). Mark the modes of the distribution, and draw contour lines at levels of \( e^{-1/2} \) and \( e^{-2} \) times the height of the modes.
16. (16 points) A particular hidden Markov model is parameterized by $\lambda = \{\pi_i, a_{ij}, b_j(\vec{x})\}$ where $\pi_i$ is uniform ($\pi_i = \frac{1}{N}$). Devise an algorithm to compute $p(q_1 = k | \vec{x}_1, \ldots, \vec{x}_T, \lambda)$. Your algorithm should be similar to the forward algorithm, but with a different initialization.
17. (16 points) The scaled forward algorithm is provided for you on the formula page at the beginning of this exam. In terms of the quantities $\pi_i, a_{ij}, b_j(\vec{x}), \hat{\alpha}_t(j), g_t$, and/or $\tilde{\alpha}_t(j)$, find a formula for the quantity $p(q_{t-1} = i, q_t = j, \vec{x}_{t-1}, \vec{x}_t, \vec{x}_1, ..., \vec{x}_{t-2}, \lambda)$.

18. (16 points) The stock market alternates between long bull markets (state 1) and short bear markets (state 2). This HMM has the following parameters:

$$\vec{\pi} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0.999 & 0.001 \\ 0.005 & 0.995 \end{bmatrix}, \quad \mu_1 = 0.1, \quad \mu_2 = -0.3, \quad \sigma_1^2 = \sigma_2^2 = 1,$$

where $\pi_i = p(q_1 = i), a_{ij} = p(q_{t+1} = j|q_t = i)$, and $p(x_t|q_t = j) = \mathcal{N}(x_t; \mu_j, \sigma_j^2)$.

You observe $x_2$ on day 2.

For what values of $x_2$ does the forward algorithm yield probabilities $\alpha_t(i)$ such that $\alpha_2(2) > \alpha_2(1)$?

Asking exactly the same question in different words: for what values of $x_2$ would it be rational to conclude that a bear market has started?
19. (16 points) Consider two PDFs. Class $y = 0$ is Gaussian:

$$p(x|y = 0) = \mathcal{N}(x; \mu_0, \sigma_0^2)$$

Class $y = 1$ is mixture Gaussian, and for some reason, one of its mixture components is the Gaussian from class 0:

$$p(x|y = 1) = 0.9p(x|y = 0) + 0.1\mathcal{N}(x; \mu_1, \sigma_1^2)$$

where $\mu_0 = 0$, $\mu_1 = 3$, and $\sigma_0^2 = \sigma_1^2 = 1$.

For what values of $x$ is

$$\frac{p(x|y = 1)}{p(x|y = 0)} > 1?$$
20. (16 points) A pelican fishes by sweeping its beak through the water. Each sweep catches many fish. The total weight of fish caught in a single sweep is an instance of a random variable, $X$, that is well described by a Gaussian mixture model:

$$ p_X(x) = \sum_{k=1}^{2} c_k \mathcal{N}(x; \mu_k, \sigma_k^2) $$

Unfortunately, you don’t know what are the correct values of the parameters $c_k$, $\mu_k$, and $\sigma_k$.

(a) You have received the following suggestions for the parameters. For each candidate set of parameters, say whether or not $p_X(x)$ would be a valid probability density if computed using this set of parameters; if not, say why not.

i. Alice suggests $c_1 = 1, c_2 = 1, \mu_1 = 10, \mu_2 = 20, \sigma_1 = 10, \sigma_2 = 10$. Would $p_X(x)$ computed using this parameter set be a valid probability density? If not, why not?

ii. Barb suggests $c_1 = 0.1, c_2 = 0.9, \mu_1 = 0, \mu_2 = 20, \sigma_1 = 10, \sigma_2 = 10$. Would $p_X(x)$ computed using this parameter set be a valid probability density? If not, why not?

iii. Carol suggests $c_1 = 0.5, c_2 = 0.5, \mu_1 = 10, \mu_2 = 20, \sigma_1 = -10, \sigma_2 = 10$. Would $p_X(x)$ computed using this parameter set be a valid probability density? If not, why not?
(b) You follow a pelican named Pete, and measure the weight of fish he retrieves on four consecutive dips, resulting in the following training dataset:

\[ \{x_1, \ldots, x_4\} = \{5, 25, 15, 10\} \]

Using the parameter set \( c_1 = 0.5, c_2 = 0.5, \mu_1 = 10, \mu_2 = 20, \sigma_1 = 10, \sigma_2 = 10 \), compute \( \gamma_k(x_t) = \Pr\{k^{\text{th}} \text{Gaussian} | x_t\} \) for \( 1 \leq t \leq 4, 1 \leq k \leq 2 \). You might find the following table of Gaussian PDFs to be useful:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \frac{1}{\sqrt{2\pi}}e^{-x^2/2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.40</td>
</tr>
<tr>
<td>0.5</td>
<td>0.35</td>
</tr>
<tr>
<td>1</td>
<td>0.24</td>
</tr>
<tr>
<td>1.5</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>2.5</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
</tr>
</tbody>
</table>

(c) Recall that the training data are

\[ \{x_1, \ldots, x_4\} = \{5, 25, 15, 10\} \]

Suppose that, after a few iterations of EM, you wind up with the following gamma probabilities:

\[ \{\gamma_2(x_1), \gamma_2(x_2), \gamma_2(x_3), \gamma_2(x_4)\} = \{0.1, 0.8, 0.6, 0.6\} \]

Find the re-estimated values of \( c_2, \mu_2, \) and \( \sigma_2^2 \) resulting from this iteration of EM.
21. (16 points) You’re training an audiovisual bird classifier: based on measurements of the birdsong frequency \((f)\) and the bird color \((c)\), the bird is classified as a sparrow \((s = 1)\) if and only if
\[
\eta \ln p(c|s = 1) + (1 - \eta) \ln p(f|s = 1) > \eta \ln p(c|s = 0) + (1 - \eta) \ln p(f|s = 0)
\]
In truth, all sparrows have pitch \(f < 0.5\), and color \(c < 0.5\), while all other birds have pitch \(f > 0.5\) and color \(c > 0.5\). Unfortunately, your training algorithm is broken, so it learned these distributions:
\[
p(f|s = 0) = \begin{cases} 1 & 0 \leq f \leq 1 \\ 0 & \text{else} \end{cases}, \quad p(f|s = 1) = \begin{cases} 1 & 0 \leq f \leq 1 \\ 0 & \text{else} \end{cases}, \quad p(c|s = 0) = \begin{cases} 1 & 0 \leq c \leq 1 \\ 0 & \text{else} \end{cases}
\]
In fact, only one of the pdfs was learned to be non-uniform:
\[
p(c|s = 1) = \begin{cases} 2 - 2c & 0 \leq c \leq 1 \\ 0 & \text{else} \end{cases}
\]
Despite these horrible training results, it is still possible to choose a value of \(\eta\) so that your audiovisual fusion system has zero error. What value of \(\eta\) gives your classifier zero error?
22. (16 points) Good days and bad days follow each other with the following probabilities:

| $q_{t-1}$ | $p(q_t = G | q_{t-1} = \cdot)$ | $p(q_t = B | q_{t-1} = \cdot)$ |
|-----------|-------------------------------|-------------------------------|
| G         | 0.7                           | 0.3                           |
| B         | 0.4                           | 0.6                           |

In winter in Champaign, the temperature on a good day is Gaussian with mean $\mu_G = 50$, $\sigma_G = 20$. The temperature on a bad day is Gaussian with mean $\mu_B = 10$, $\sigma_G = 20$. A particular sequence of days has temperatures

$$\{x_1 = 10, x_2 = 20, x_3 = 30\}$$

What is the probability $p(X | q_1 = B)$, the probability of seeing this sequence of temperatures given that the first day was a bad day?
23. (25 points) Suppose that

\[
\begin{align*}
a_{ij} & = p(q_t = j | q_{t-1} = i) \\
b_j(\tau) & = p(x_\tau | q_t = j) \\
g_t & = p(x_\tau | x_1, \ldots, x_{\tau-1})
\end{align*}
\]

And define the scaled forward algorithm to compute

\[
\tilde{\alpha}_t(i) = p(q_t = i | x_1, \ldots, x_t) = \frac{p(x_\tau, q_t = i | x_1, \ldots, x_{\tau-1})}{g_t} \frac{p(x_1, \ldots, x_\tau | q_t = i)}{g_1 g_2 \cdots g_t}
\]

(a) Devise an algorithm to iteratively compute \(g_t\) and \(\tilde{\alpha}_t(i)\). Fill in the right-hand side of each equation, using only the terms \(a_{jk}, b_j(\tau), g_\tau,\) and \(\tilde{\alpha}_\tau(j)\) for \(1 \leq j \leq N, 1 \leq k \leq N, 1 \leq \tau \leq t\).

1. **INITIALIZE:** \(g_1 = \)
2. **INITIALIZE:** \(\tilde{\alpha}_1(i) = \)
3. **ITERATE:** \(g_t = \)
4. **ITERATE:** \(\tilde{\alpha}_t(i) = \)
5. **TERMINATE:** \(p(X) = \)

(b) Suppose \(\beta_t(i) = p(x_{t+1}, \ldots, x_T | q_t = i)\). Then

\[
\tilde{\alpha}_t(i) \beta_t(i) = p(f | g)
\]

for some list of variables \(f\), and some other list of variables \(g\). Specify what variables should be included in each of these two lists.
24. (20 points) The Maesters of the Citadel need to determine when winter starts. The temperature on
day $t$ is $x_t$. The state of day $t$ is either $q_t = 0$ (Autumn) or $q_t = 1$ (Winter). Nobody really knows
how cold this winter will be or how long it will last, but the Maesters have created an initial model
$\Lambda = \{a_{ij}, b_j(x)\}$ where $a_{ij} \equiv p(q_t = j|q_{t-1} = i)$ and $b_j(x) \equiv p(x_t = x|q_t = j)$.

(a) Suppose we have a particular three day sequence of measurements, $x_1$, $x_2$, and $x_3$. Given that
the preceding day was still autumn ($q_0 = 0$), we want to determine the joint probability that
it continued to be autumn for days 1, 2, and 3, and that the three observed temperatures were
measured. In other words, we want an estimate of

$$G_1 = p(q_1 = 0, x_1, q_2 = 0, x_2, q_3 = 0, x_3|q_0 = 0, \Lambda)$$

Find $G_1$ in terms of $a_{ij}$ and $b_j(x_t)$, for whatever particular values of $i$, $j$, and $t$ are most useful
to you.

(b) Suppose it is known that the preceding day was still autumn ($q_0 = 0$). Now, on day 1, the
Maesters have determined that the temperature is $x_1$. Find the conditional probability, given
this measurement, that it is still autumn, i.e., find

$$G_2 = p(q_1 = 0|x_1, q_0 = 0, \Lambda)$$

Find $G_2$ in terms of $a_{ij}$ and $b_j(x_t)$, for whatever particular values of $i$, $j$, and $t$ are most useful
to you.

(c) The Maesters have collected a long series of measurements, $\{x_1, \ldots, x_T\}$ for $T$ consecutive days.
From these measurements, the Maesters have applied the forward-backward algorithm in order
to calculate the following two quantities:

$$\alpha_t(i) \equiv p(x_1, \ldots, x_t, q_t = i|\Lambda), \quad \beta_t(i) \equiv p(x_{t+1}, \ldots, x_T|q_t = i, \Lambda)$$

Using these quantities, the Maesters wish to calculate the probability that Winter started on a
particular day, $t = w$. That is, they wish to find

$$G_3 = p(q_{w-1} = 0, q_w = 1|x_1, \ldots, x_T, \Lambda)$$

Find $G_3$ in terms of $\alpha_t(i)$, $\beta_t(i)$, $a_{ij}$ and $b_j(x_t)$, for whatever particular values of $i$, $j$, and $t$ are
most useful to you.
25. (20 points) A bimodal HMM uses a common state sequence, $Q = [q_1, \ldots, q_T]$, to explain two different observation sequences $X = [\vec{x}_1, \ldots, \vec{x}_T]$ and $Y = [\vec{y}_1, \ldots, \vec{y}_T]$. The HMM is parameterized by

\[
\begin{align*}
\pi_i &= p(q_1 = i) \\
\alpha_{ij} &= p(q_t = j|q_{t-1} = i) \\
b_j(\vec{x}_t) &= p_X(\vec{x}_t|q_t = j) \\
c_j(\vec{y}_t) &= p_Y(\vec{y}_t|q_t = j)
\end{align*}
\]

Define

\[
\begin{align*}
\alpha_t(i) &= p(\vec{x}_1, \vec{y}_1, \ldots, \vec{x}_t, \vec{y}_t, q_t = i) \\
\beta_t(i) &= p(\vec{x}_{t+1}, \vec{y}_{t+1}, \ldots, \vec{x}_T, \vec{y}_T|q_t = i)
\end{align*}
\]

(a) Specify initialization formulas for $\alpha_1(i)$ and $\beta_T(i)$ in terms of $\pi_i$, $a_{ij}$, $b_j(\vec{x}_t)$, and $c_j(\vec{x}_t)$.

(b) Specify iteration formulas for $\alpha_t(i)$ and $\beta_t(i)$ in terms of $\pi_i$, $a_{ij}$, $b_j(\vec{x}_t)$, $c_j(\vec{x}_t)$, $\alpha_{t-1}(j)$, and $\beta_{t+1}(j)$. 
26. (20 points) You are creating a recommender system that tries to recommend songs that will be considered to be similar to a given query. Each song is characterized by a two-dimensional vector \( \mathbf{x}_k = [b_k, v_k]^T \) where \( b_k \) is the number of beats per minute, and \( v_k \) is the fraction of air-time during which there is a human voice. Your customer considers the following four songs to be similar:

\[
[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4] = \begin{bmatrix}
120 & 140 & 140 & 120 \\
0.3 & 0.3 & 0.5 & 0.5
\end{bmatrix}
\]

You are given two more test data, \( \mathbf{x}_5 = [b_5, v_5]^T \) and \( \mathbf{x}_6 = [b_6, v_6]^T \), and you are asked whether or not \( \mathbf{x}_5 \) and \( \mathbf{x}_6 \) should be considered similar. Write formulas for the Mahalanobis distance between \( \mathbf{x}_5 \) and \( \mathbf{x}_6 \) under the following conditions:

(a) Estimate a diagonal data covariance matrix directly from the data, and use it to write the squared Mahalanobis distance \( d^2_{\Sigma}(\mathbf{x}_5, \mathbf{x}_6) \).

(b) Estimate a diagonal data covariance matrix from the data, then regularize it using regularization parameter \( \lambda = 0.01 \) before using the result to write the squared Mahalanobis distance \( d^2_{\Sigma}(\mathbf{x}_5, \mathbf{x}_6) \).