1. Consider the neural network below.

The initial weights are:

Layer 1:
- \( w_{1,1} = -10 \), \( w_{1,2} = 0 \), \( w_{1,3} = -5 \), \( w_{1,4} = 10 \), \( b_{1} = 4 \)  
  - Unit 1
- \( w_{2,1} = 20 \), \( w_{2,2} = 0 \), \( w_{2,3} = 10 \), \( w_{2,4} = -5 \), \( b_{2} = 4 \)  
  - Unit 2
- \( w_{3,1} = 0 \), \( w_{3,2} = -10 \), \( w_{3,3} = 0 \), \( w_{3,4} = 20 \), \( b_{3} = 4 \)  
  - Unit 3

Layer 2:
- \( w_{1,1} = 5 \), \( w_{1,2} = 10 \), \( w_{1,3} = 0 \), \( b_{1} = -2 \)  
  - Unit 1
- \( w_{2,1} = 0 \), \( w_{2,2} = -10 \), \( w_{2,3} = 15 \), \( b_{2} = -2 \)  
  - Unit 2

Layer 3:
- \( w_{1,1} = 10 \), \( w_{1,2} = -20 \), \( b_{1} = 5 \)

Compute the output given the following inputs:

(a) Compute \( r_{1}, r_{2}, r_{3} \). Given the inputs: \( x_1 = 5 \), \( x_2 = -10 \), \( x_3 = 10 \), \( x_4 = 0 \)

(b) Compute \( r_{3} \). Given the lower-layer outputs: \( r_{1} = 0.1 \), \( r_{2} = 0.6 \)

Sum: \( \sum r_{i} w_{i} = 0.1 \times 10 + 0.6 \times -20 + 5 = 1 - 12 + 5 = -6 \)

Sigmoid: \( g(\cdot) \rightarrow g(-6) : r_{1} = 0 \)

(c) Compute \( r_{2} \). Given the lower-layer outputs: \( r_{1} = 0.1 \), \( r_{2} = 0.3 \), \( r_{3} = 0.6 \)

Compute the change in the specified weight based on the following input/outputs. In each case, presume the starting weight is as specified in the original list above. Assume \( \varepsilon = 1 \)

(d) Compute \( \Delta w_{1,2} \). Given the layer 2 rates: \( r_{1} = 0.2 \) and \( r_{2} = 0.8 \); layer 3 rates: \( r_{3} = 0.1 \); the desired output from \( r_{1} \) is 1.0
(e) Compute $\Delta w_{1,2}$. Given the features: $x_1=10$, $x_2=-5$, $x_3=0$, $x_4=15$; $r_1^1=0.5$, $r_2^1=0.2$, $r_3^1=0.8$; delta values: $\delta_1^2 = -0.005$, $\delta_2^2 = 0.01$

$\Delta w_{1,2} = \varepsilon (1-r_1^1) (\sum_n w_{n,1}^2 \delta_n^2) r_1^1 x_3 = 1(1-0.5)x(5x-0.005+0x.01)x.05x-5 = 1x.05x-0.025x.05x-5 = -0.0125x.05x-5 = 0.00625x-5 = \boxed{0.031}$

2. For each of the following functions $f(x; h)$, compute the value of $h$ that will maximize $f(x; h)$, assuming each function has a single maximum and no minimum.

(a) $f_1(x; h) = \sum_i (-h^2 - 10hx_i + 12x_i^2)$

(b) $f_2(x; h) = e^{-(h^2 + x^2)} = \exp(- (h^2 + x^2))$

Find derivative and set to 0. Equivalently, derivative of log and set to 0.

$log f_2 -> - (h^2 + x^2)$

Derivative of log $f_2 : -2h = 0$

Thus: $h=0$

(c) $f_3(x; h) = \prod_i 3h(x^i)$

3. Consider the following Gaussian likelihoods for features $x_1$, $x_2$, and $x_3$ given class $= 1$ (blue curves) or class $= 0$ (red curves).

i. We wish to multiply these likelihoods together to compute $P(x|y)$. Which type of classification is this:

(a) Naïve Bayes Max-Posterior classification
(b) Non-Naïve Bayes Max-Likelihood classification
(c) Naïve Bayes Max-Posterior classification
(d) Naïve Bayes Max-Likelihood classification
(e) Support Vector Machine classification
ii. For the feature values below, which class is more probable (based on $P(x|y)$ calculated from the plots above)?

(a) $x_1=5 \quad x_2=7 \quad x_3=6$

(b) $x_1=8 \quad x_2=8 \quad x_3=6$

Class $y=1$:  $P(x_1 | y=1) = 0 \quad P(x_2 | y=1) = 0.02 \quad P(x_3 | y=1) = 0.03 \quad -> \quad \text{total: 0}$
Class $y=0$:  $P(x_1 | y=0) = 0.15 \quad P(x_2 | y=0) = 0.4 \quad P(x_3 | y=0) = 0.1 \quad -> \quad \text{total: 0.006}$

Class 0 maximum probability

iii. Which class is more probable if we also incorporate the following prior:

$P(y=0) = 0.1 \quad P(y=1) = 0.9$

to compute $P(y|x)$?

(a) $x_1=4 \quad x_2=5 \quad x_3=9$

Multiply prior times likelihoods

Class $y=1$:  $P(x_1 | y=1) = 0.15 \quad P(x_2 | y=1) = 0.2 \quad P(x_3 | y=1) = 0.2 \quad P(y=1)=.9 \quad -> \quad \text{total: 0.005}$
Class $y=0$:  $P(x_1 | y=0) = 0.005P(x_2 | y=0) = 0 \quad P(x_3 | y=0) = 0 \quad P(y=0)=.1 \quad -> \quad \text{total: 0}$

Class $y=1$ is most probable

(b) $x_1=6 \quad x_2=7 \quad x_3=7$

iv. Provide a prior that would make class 1 more probable if the $x$ values are:

$x_1=6 \quad x_2=8 \quad x_3=6$

4. Using each of the following kernel functions, compute the result of $K(x^1, x^2)$, for the specified input vectors.

$K(c,d)=2^{-(c^Td+2)}$
(a) \( c = \begin{bmatrix} 4 \\ 0 \\ -2 \end{bmatrix} \quad d = \begin{bmatrix} 0 \\ 0 \\ 1.5 \end{bmatrix} \)

\[ c^T d = 0 + 0 - 3 = -3 \]

\[ 2^{-(3+2)} = 2^{-5} = 2^1 = 2 \]

(b) \( c = \begin{bmatrix} 1 \\ 0.5 \\ -2 \end{bmatrix} \quad d = \begin{bmatrix} 3 \\ -2 \\ -1 \end{bmatrix} \)

\[ K(c,d) = (c^T d - 4)^2 + 10c^T d \]

(c) \( c = \begin{bmatrix} 0 \\ 3 \\ -2 \end{bmatrix} \quad d = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \)

(d) \( c = \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix} \quad d = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} \)

5. Consider the following training data. Red circles are class 0, blue x’s are class 1, and all other shapes (triangles, stars, diamonds) are data points with known feature values but unknown labels.

Using the EM approach for learning, and assuming that we use a linear logistic classifier, how will the black triangles, diamonds, and star data points be used for learning? In the first round of EM, what y value do you expect each data point to be assigned, or no value at all?

6. Consider the classification problem with the following features and classes.

Class Person-type: Teenager, YoungProfessional, Adult, SeniorCitizen
Features:
Daily-time-online: 1-2 hours, 3-4 hours, 5-8 hours
Number-of-online-friends: 0-10, 10-50, 50-200, 200-1000
Favored content: News, SocialPosts, Education, Entertainment
Money-spent-online: None, $1-$50, $50-$100, $100-$500, $500+

(a) How many parameters given Naïve Bayes a posteriori classification?

$$\text{#Classes} \times ((\text{#Dfeats}-1) + (\text{#Nfeats}-1) + (\text{#Ffeats}-1) + (\text{#Mfeats}-1)) + (\text{#Classes}-1)$$

$$4 \times (2 + 3 + 3 + 4) + (4 - 1) = 4 \times (12) + 4 = 51$$

Consider a classifier hypothesis set of squares. A single hypothesis $h$ is a square with a fixed size and location. Four example hypotheses are shown.

And here is examples of $h$ that will help shatter a set of three data points.

For each data set:
- Pick four points and list a dichotomy that is not possible with four of the points provided.

Example 1:
Example 2:

A and D is 1, B and C is 0 (or the reverse)
Or A and F is 1, B and C is 0 (or the reverse)
Or C and E is 0, A and F is 1 (or the reverse)

Example 3:

What is the VC dimension for a cube classifier, where each h is a cube of some fixed size and fixed location \((x_1, x_2, x_3)\), where class 1 is assigned either to inside or outside the cube, and class 0 is assigned to the other region. We assume you can select points from anywhere in the 3D feature space.

You have data with three features and wish to use a decision tree to classify your data (should have been covered in your past data mining class). For example, a two layer decision tree may look like this:
if \( x_1 > 0.5 \) then: if \( x_2 < -1 \): y = 1

else: y = 0

else: if \( x_3 > 3 \): y = 0

else: y = 1

(First true-false measurement leads to second true-false, leading to a final class label.) We assume you can select points from anywhere in the 3D feature space.

**What is the VC dimension for a one-layer decision tree?**

\[ \text{VC} = 2 \]

**What is the VC dimension for a two-layer decision tree?**

Consider the following HMM. It uses a thermometer to attempt to predict the weather.

We begin with the following estimate for our HMM parameters:

\[ \Pi_{\text{snow}} = 0.2 \quad \Pi_{\text{rain}} = 0.3 \quad \Pi_{\text{sunny}} = 0.3 \quad \Pi_{\text{cloudy}} = 0.2 \]

\[ \phi_{o,i}: \]

<table>
<thead>
<tr>
<th></th>
<th>Cold</th>
<th>Mild</th>
<th>Hot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snow</td>
<td>0.8</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>Rain</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Sunny</td>
<td>0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>Cloudy</td>
<td>0.2</td>
<td>0.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>
We could actually learn a Gaussian function for the temperature for each state. Here, we’ll just do a discrete probability table.

We receive a new sequence of temperatures and wish to update our HMM parameters.

Sequence:
Cold Cold Hot Mild Hot

Correct alpha values are in black. Made-up alpha values are in color parentheses. You will have to find the real values below. You can use the made-up value in calculating $S_t$ values further below.

$\alpha_t(i)$

<table>
<thead>
<tr>
<th></th>
<th>t: 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snow</td>
<td>?? (.11)</td>
<td>.08</td>
<td>0</td>
<td>.00011</td>
<td>0</td>
</tr>
<tr>
<td>Rain</td>
<td>0.15</td>
<td>?? (.04)</td>
<td>.0082</td>
<td>.0017</td>
<td>.00049</td>
</tr>
<tr>
<td>Sunny</td>
<td>?? (.08)</td>
<td>0</td>
<td>.0056</td>
<td>?? (.0033)</td>
<td>.0020</td>
</tr>
<tr>
<td>Cloudy</td>
<td>0.04</td>
<td>.027</td>
<td>?? (.0044)</td>
<td>.0053</td>
<td>.00030</td>
</tr>
</tbody>
</table>

Correct beta values are in black. Made-up beta values are in color parentheses. You will have to find the real values below. You can use the made-up value in calculating $S_t$ values further below.

$\beta_t(i)$

<table>
<thead>
<tr>
<th></th>
<th>t: 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snow</td>
<td>.0067</td>
<td>.0062</td>
<td>.13</td>
<td>.05</td>
</tr>
<tr>
<td>Rain</td>
<td>.0097</td>
<td>?? (.011)</td>
<td>.13</td>
<td>?? (.08)</td>
</tr>
<tr>
<td>Sunny</td>
<td>.0028</td>
<td>.087</td>
<td>?? (.11)</td>
<td>.52</td>
</tr>
<tr>
<td>Cloudy</td>
<td>.0062</td>
<td>.047</td>
<td>.121</td>
<td>?? (.11)</td>
</tr>
</tbody>
</table>

Find the missing values in the tables above.

$\alpha_{t=2}(\text{Rain}): 0.05$
\( \alpha_{t=4}(Sunny) \): 0.3 \times (0 + 0 + .0056 \times .7 + .0044 \times .3) = .3 \times (.00392 + .00132) = 0.0016 
From: Snow, Rain, Sunny, Cloudy

\( \beta_{t=3}(Sunny) \): 0 + 0 + .52 \times .3 \times .7 + .11 \times .7 \times .3 = .109 + .0231 = .13 \quad 0.17 
To: Snow, Rain, Sunny, Cloudy

What are the values:

\( S_2(\text{cloudy}) \)

\( S_3(\text{snow,sunny}) = \)

\[
\begin{align*}
S_1(\text{rain}): & \quad \frac{.0097 \times .15}{.001455} = \frac{.097 \times .15 + .067 \times .11 + .028 \times .08 + .0062 \times .04}{.001455 + .000737 + .00024 + .000248} = .0543 \\
& \quad \frac{.001455}{.00268} = 0.543
\end{align*}
\]

Now let us presume the following S values (these are made-up values):

\( S_t(i) \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snow</td>
<td>0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Rain</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Sunny</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Cloudy</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\( S_t(i,j) \)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain, Cloudy</td>
<td>0.1</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Sunny, Rain</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
What are the resulting estimate of the following parameters?

\[ \Pi_{\text{rain}} \]
\[ \Pi_{\text{cloudy}} = S_1(\text{cloudy}) = 0.1 \]

\[
A_{\text{rain,cloudy}} = \frac{\sum_t S_t(\text{rain,cloudy})}{\sum_t S_t(\text{cloudy})} = \frac{.1+.4+.3+.2}{0.1+0.2+0.3+0.4} = \frac{1}{1} = 1
\]

*NOTE* THE S() values were made up, so we got an unrealistic probability for A

CORRECTED Apr 29 evening

\[ A_{\text{sunny,rain}} \]

\[ \phi_{\text{hot,rain}} \quad \phi_{\text{mild,sunny}} \]