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**
  - \( w_{1,1} = 20 \)
  - \( w_{1,2} = 0 \)
  - \( w_{1,3} = 10 \)
  - \( w_{1,4} = -5 \)
  - \( b_1 = 4 \)
- **Unit 2**
  - \( w_{1,1} = 0 \)
  - \( w_{1,2} = -10 \)
  - \( w_{1,3} = 0 \)
  - \( w_{1,4} = 20 \)
  - \( b_1 = 4 \)
- **Layer 2:**
  - \( w_{2,1} = 5 \)
  - \( w_{2,2} = 10 \)
  - \( w_{2,3} = 0 \)
  - \( b_1 = -2 \)
  - **Unit 1**
  - \( w_{2,1} = 0 \)
  - \( w_{2,2} = -10 \)
  - \( w_{2,3} = 15 \)
  - \( b_1 = -2 \)
- **Unit 2**
- **Layer 3:**
  - \( w_{3,1} = 10 \)
  - \( w_{3,2} = -20 \)
  - \( b_3 = 5 \)

Compute the output given the following inputs:

(a) Compute \( r_1, r_2, r_3 \). Given the inputs: \( x_1 = 5 \quad x_2 = -10 \quad x_3 = 10 \quad x_4 = 0 \)

(b) Compute \( r_3 \). Given the lower-layer outputs: \( r_1^2 = 0.1 \), \( r_2^2 = 0.6 \)

Sum: \( \sum r_i w_i = 0.1 \times 10 + 0.6 \times (-20) + 5 = 1 - 12 + 5 = -6 \)

Sigmoid: \( g( ) \rightarrow g( -6 ) : r_3^3 = 0 \)

(c) Compute \( r_2^2 \). Given the lower-layer outputs: \( r_1^2 = 0.1 \), \( r_2^1 = 0.3 \), \( r_3^1 = 0.6 \)

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: \( r_1^2 = 0.2 \) and \( r_2^2 = 0.8 \); layer 3 rates: \( r_1^3 = 0.1 \); the desired output from \( r_1^3 \) 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: $\delta^2_2=-0.005, \delta^2_3=0.01$

$\Delta w_{1,2} = \epsilon (1 - r_1^1 ) ( \sum_n w_{n,1}^2 / \sigma_n^2 ) r_1^1 x_3 = 1 \times (1 - 0.5) x (5 \times -0.005 + 0 \times 0.01) x 0.5 \times -5 = 1 \times 0.5 \times -0.025 \times 0.5 \times -5 = -0.0125 \times 0.5 \times -5 = 0.00625 \times -5 = -0.031$

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) $f_2(x; h) = e^{-(h^3 + x^2)} = \exp\left(- (h^2 + x^2)\right)$

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

$\log f_2 \rightarrow -(h^2+x^2)$

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

Thus: $h=0$

(c) $f_3(x; h) = \prod_l 3h(x^l)$

3. Consider the following Gaussian likelihoods for features $x_1, x_2, \text{ and } x_3$ 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) $x_1 = 5$  $x_2 = 7$  $x_3 = 6$

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

Class $y=1$:  $P(x_1 | y=1) = 0$  $P(x_2 | y=1) = 0.02$  $P(x_3 | y=1) = 0.03$  -> total: 0

Class $y=0$:  $P(x_1 | y=0) = 0.15$  $P(x_2 | y=0) = 0.4$  $P(x_3 | y=0) = 0.1$  -> total: 0.006

**Class 0 maximum probability**

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

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

to compute $P(y|x)$?

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

Multiply prior times likelihoods

Class $y=1$:  $P(x_1 | y=1) = 0.15$  $P(x_2 | y=1) = 0.2$  $P(x_3 | y=1) = 0.2$  $P(y=1)=.9$  -> total: 0.005

Class $y=0$:  $P(x_1 | y=0) = 0.005$  $P(x_2 | y=0) = 0$  $P(x_3 | y=0) = 0$  $P(y=0)=.1$  -> total: 0

**Class $y=1$ is most probable**

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

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

$x_1 = 6$  $x_2 = 8$  $x_3 = 6$

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) \( \mathbf{c} = \begin{bmatrix} 4 \\ 0 \\ -2 \end{bmatrix} \) \( \mathbf{d} = \begin{bmatrix} 0 \\ 0 \\ 1.5 \end{bmatrix} \)
\[ \mathbf{c}^T \mathbf{d} = 0 + 0 - 3 = -3 \]
\[ 2^{-3+2} = 2^{1} = 2 \]

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

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

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

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

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.

![Diagram](image)

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?

\[
\text{#Classes} \times ( (\text{#Dfeats}-1) + (\text{#Nfeats}-1) + (\text{#Ffeats}-1) + (\text{#Mfeats}-1) ) + (\text{#Classes}-1)
\]
\[
4 \times (2 + 3 + 3 + 4) + (4-1) = 4 \times 12 + 4 = 51
\]

(b) How many parameters without Naïve Bayes (nor any Bayes net) likelihood classification?

(c) How many with the following Bayes nets likelihood classification?

(d) What is the minimum number of training examples you would advise to use for the Bayes net from (c)?