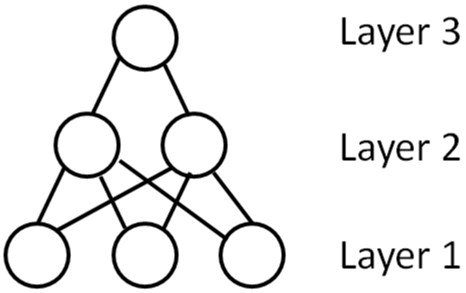
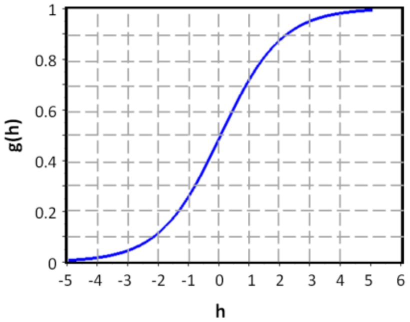
Final practice part 2

1. Consider the neural network below.





The initial weights are:

Layer 1:	$w^1_{1,1} = -10$	$w^1_{1,2} = 0$	$w^1_{1,3} = -5$	$w^1_{1,4} = 10$	$b^3_1 = 4$	Unit 1
	$w^1_{2,1} = 20$	$w^1_{2,2} = 0$	$w^1_{2,3} = 10$	$w^1_{2,4} = -5$	$b^3_1 = 4$	Unit 2
	$w^1_{3,1} = 0$	$w^1_{3,2} = -10$	$w^1_{3,3} = 0$	$w^1_{3,4} = 20$	$b^3_1 = 4$	Unit 3
Layer 2:	$w^2_{1,1} = 5$	$w^2_{1,2} = 10$	$w^2_{1,3} = 0$	$b^3_1 = -2$	Unit 1	
	$w^2_{2,1} = 0$	$w^2_{2,2} = -10$	$w^2_{2,3} = 15$	$b^3_1 = -2$	Unit 2	
Layer 3:	$w^3_{1,1} = 10$	$w^3_{1,2} = -20$	$b^3_1 = 5$			

Compute the output given the following inputs:

(a) Compute r^1_1, r^1_2, r^1_3 . Given the inputs: x_1= 5 x_2= -10 x_3= 10 x_4= 0

r^1_1: Sum inputs: $\sum_i x_i w_i$ = 5x-10 + -10x0 + 10x-5 + 0x10 + 4 = -50+0-50+4 = -96
Apply sigmoid function g( ) -> g(-96) : $\boldsymbol{r^1_1}$**=0**

r^1_2: Sum inputs: $\sum_i x_i w_i$ = 5x20 + -10x0 + 10x10 + 0x-5 + 4 = 100+0+100+4 = 204
Apply sigmoid function g( ) -> g(204) : $\boldsymbol{r^1_2}$**=1**

r^1_3: Sum inputs: $\sum_i x_i w_i$ = 5x0 + -10x-10 + 10x0 + 0x20 + 4 = 100+0+4 = 104
Apply sigmoid function g( ) -> g(104) : $\boldsymbol{r^1_3}$**=1**

(b) Compute r^3_1 . Given the lower-layer outputs: r^2_1=0.1 , r^2_2=0.6

(c) Compute r^2_2 . Given the lower-layer outputs: r^1_1=0.1 , r^1_2=0.3, r^1_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 $\Delta w^3_{1,2}$. Given the layer 2 rates: r^2_1=0.2 and r^2_2=0.8 ; layer 3 rates: r^3_1=0.1 ; the desired output from r^3_1 is 1.0

(e) Compute $\Delta w^1_{1,2}$. Given the features: x_1=10, x_2=-5, x_3=0, x_4=15; r^1_1=0.5, r^1_2=0.2, r^1_3=0.8; delta values: δ^2_1= -0.005, δ^2_2= 0.01

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

Take derivative and set to 0:
Derivative: f' : $\sum_i(-2h - 10x_i) = 0$

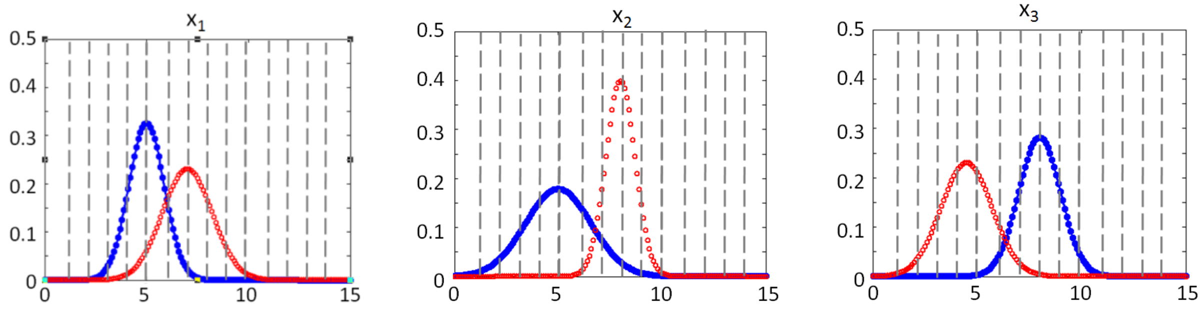
$-2Th - \sum_i 10x_i = 0$ where T is the number of x_i's summed together

$$\mathbf{h = -\frac{\sum_i 10x_i}{2T}}$$

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

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

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



i. We wish to multiply these likelihoods together to compute P($\mathbf{x}$|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

Answer: (d)

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

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

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

y=1 (blue): P(x_1=6 | y=1) = 0.15 P(x_2=7 | y=1) = 0.05 P(x_3=7 | y=1) = 0.15
-> .15x.05x.15=0.00113

y=0 (red): P(x_1=6 | y=0) = 0.15 P(x_2=7 | y=0) = 0.15 P(x_3=7 | y=0) = 0.01
-> .15x.15x.01 = .00023

Likelihood given y=1 (0.001) is greater than likelihood given y=0 (.0002)
Class y=1 is more probable

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($\mathbf{x}^1$, $\mathbf{x}^2$), for the specified input vectors.

K(**c**,**d**)=2$^{-(\boldsymbol{c}^T\boldsymbol{d}+2)}$

(a) $\boldsymbol{c} = \begin{bmatrix} 4 \\ 0 \\ -2 \end{bmatrix}$ $\boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \\ 1.5 \end{bmatrix}$

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

$\mathbf{c}^T\mathbf{d}$ = 3-1+2 = 4

$2^{-(4+2)} = 2^{-6} = \frac{1}{2^6} = \frac{1}{64} = \mathbf{0.016}$

K(**c**,**d**)=$(\boldsymbol{c}^T\boldsymbol{d} - \mathbf{4})^2 + 10\boldsymbol{c}^T\boldsymbol{d}$

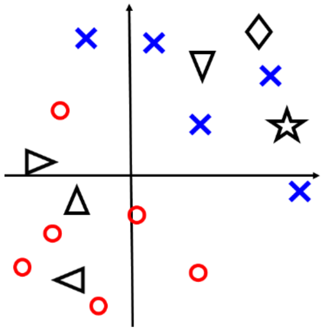
(c) $\boldsymbol{c} = \begin{bmatrix} 0 \\ 3 \\ -2 \end{bmatrix}$ $\boldsymbol{d} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$

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

$\mathbf{c}^T\mathbf{d}$ = -3+0+1 = -2

$(-2-4)^2 + 10 \times (-2) = (-6)^2 - 20 = 36-20 = \mathbf{16}$

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.

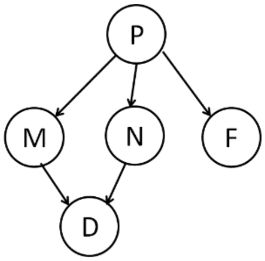
Class **P**erson-type: Teenager, YoungProfessional, Adult, SeniorCitizen
Features:
Daily-time-online: 1-2 hours, 3-4 hours, 5-8 hours
Number-of-online-friends: 0-10, 10-50, 50-200, 200-1000
Favored content: News, SocialPosts, Education, Entertainment
Money-spent-online: None, $1-$50, $50-$100, $100-$500,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

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

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

#Classes x (#Dfeats x #Nfeats x #Ffeats x #Mfeats - 1) = 4 x (3x4x4x5 - 1) = 4 x (240-1) = **956**

(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)?