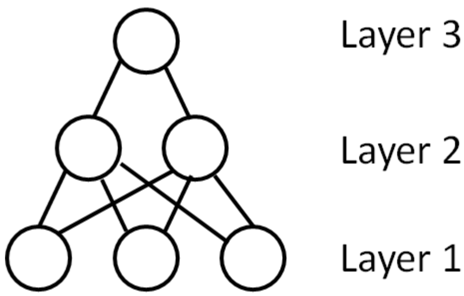
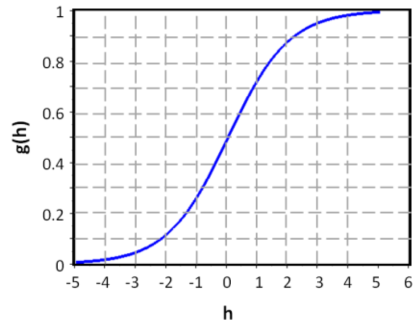
Final practice part 2

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

(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

Sum inputs: 0.1x0 + 0.3x-10 + 0.6x15 – 2 = 0-3+9-2 = 4
Use sigmoid g(4) : **r^2_2=0.99**

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

$\Delta w^3_{1,2} = \varepsilon(1 - r^3_1)(1 - r^3_1)r^3_1 r^2_2$ =1x(1-.1)x(1-.1)x.1x.8 = 1x.9x.9x.1x.8 = **0.06**

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

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

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

Set derivative equal to 0 … or set derivative of log equal to 0

log f_3 : $\sum_i(\log 3 + x^i \ \log h)$

~~Set equal to 0: $\sum_i 3x^i \ \log h = \log h \ \sum_i 3x^i = 0$~~

~~$\log h = 0$~~

~~**h=1**~~

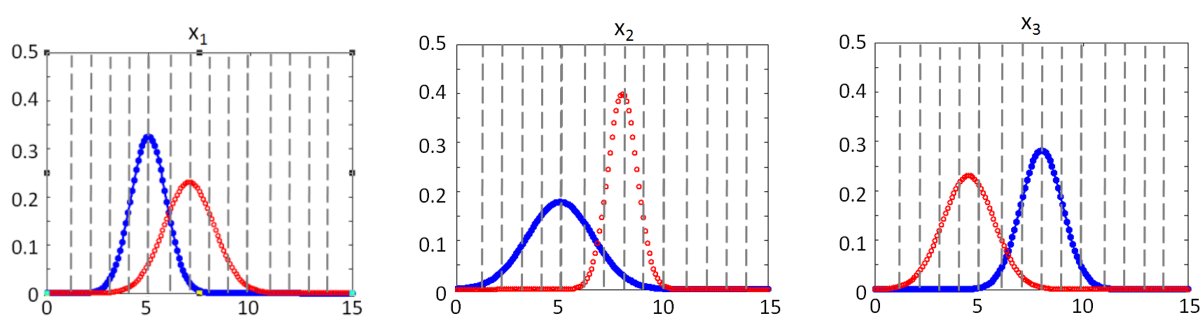
CORRECTION:

Set equal to 0: $\sum_i(\log 3 + x^i \ \log h) = \sum_i \log 3 \ + \log h \ \sum_i x^i = 0$

$$\log h = \frac{-\sum_i \log 3}{\sum_i x^i}$$

$$\mathbf{h=exp}\left(\frac{-\sum_i \log 3}{\sum_i x^i}\right) = \mathbf{exp}\left(\frac{-\#Data \times \log 3}{\sum_i x^i}\right)$$

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 x_2=7 x_3=6

Class y=1: P(x_1 | y=1) = 0.32 P(x_2 | y=1) = 0.1 P(x_3 | y=1) = 0.2 -> total: 0.006
Class y=0: P(x_1 | y=0) = 0.05 P(x_2 | y=0) = 0.2 P(x_3 | y=1) = 0.1 -> total: 0.001

Class y=1 most probable

(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

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

Class y=1: P(x_1 | y=1) = 0.2 P(x_2 | y=1) = 0.02 P(x_3 | y=1) = 0.05 -> total: 0.0002
Class y=0: P(x_1 | y=0) = 0.15 P(x_2 | y=0) = 0.4 P(x_3 | y=1) = 0.1 -> total: 0.006

Prior for y=1 (P(y=1)) needs to be at least 30x larger than prior for y=0:
e.g.,
P(y=1) = 0.98 **P(y=0) = 0.02**

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

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

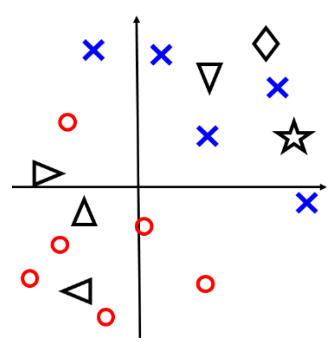
(c) $\boldsymbol{c} = \begin{bmatrix} 0 \\ 3 \\ -2 \end{bmatrix}$ $\boldsymbol{d} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$

$\mathbf{c}^T\mathbf{d}$ = 0+0-4 = -4

$(-4-4)^2$ + 10x-4 = $(-8)^2$ – 40 = 64-40 = **24**

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

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?

First, could assign random y values to the non-circle and non-x shapes, then learn classifier, then in next round use learned separator/classifier to assign better guesses of classes.

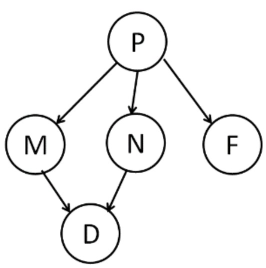
6. Consider the classification problem with the following features and classes.

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?

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



Probability computed using P(P), P(M|P), P(N|P), P(F|P) and P(D|M,N)
P(P) <- 4-1 = 3 parameters <- ACTUALLY NOT NEEDED FOR LIKELIHOOD COMPUTATION
P(M|P) = 4 x (5-1) = 4x4 = 16 (# of dependents variable values X # dependent var values)
P(N | P) = 4 x (4-1) = 4x3 = 12
P(F | P) = 4 x (4-1) = 4x3 = 12
P(D | M, N) = (5x4)x(3-1) = 20x2 = 40

In total: 16 + 12 + 12 + 40 = **80** **or**
Alternatively could be 4+16+12+12+40 = **84**

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

Want training data to be at least double the number of parameters to learn (preferably 10x more or even greater). Take your answer from (b) and multiply by 2 (or 10). So 160, or 800.