1. Consider the following four vectors:

(i) $\boldsymbol{x}^1 = \begin{bmatrix} -1 \\ 0.5 \\ 2 \\ 0 \end{bmatrix}$ (ii) $\boldsymbol{x}^2 = \begin{bmatrix} -0.5 \\ 0 \\ 1 \\ -2 \end{bmatrix}$ (iii) $\boldsymbol{x}^3 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0.5 \end{bmatrix}$

(a) What is the magnitude of each vector?

(b) What is the result of each dot product below?

$\mathbf{x}^{1\,\mathrm{T}}\,\mathbf{x}^2$ $\mathbf{x}^{3\,\mathrm{T}}\,\mathbf{x}^2$ $\mathbf{x}^{1\,\mathrm{T}}\,\mathbf{x}^3$

2. We wish to use a Bayesian classifier to distinguish between two classes of birds: y^i=D (for Duck) or y^i=G (for Goose). Each data point contains 5 features, measuring: motion speed, weight, size, number of daily hours-of-sleep, and typical depth-of-dive into water.

Presume we use a Gaussian Naïve Bayes classifier – we assume each $P(x^i_j \mid y^i)$ is Gaussian.

(a) How will we calculate the posterior probability: $P(y^i \mid x^i_{speed}, x^i_{weight}, x^i_{size}, x^i_{sleepHours}, x^i_{diveDepth})$?
(What other probabilities will we use for this calculation?)

(b) How many parameters will we learn under the Naïve Bayes assumption?

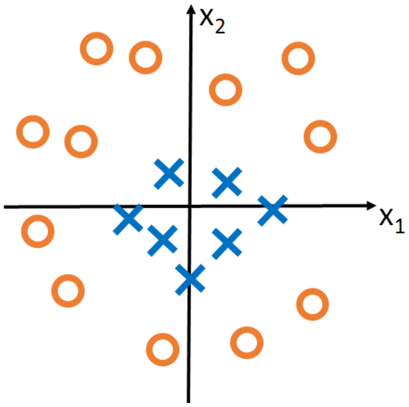
(c) Let us now assume we will use a logistic classifier instead on the same data set. How many parameters must we learn to determine the separating hyperplane?

3. For each example below, which of the following mapping functions will make these points linearly separable?

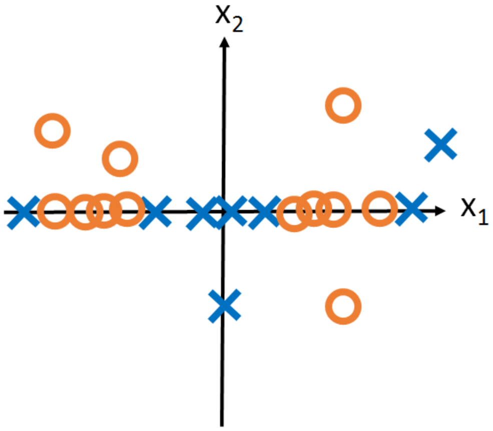
Possible mapping function: φ_1= ([x_1, x_2])->[(x_1+x_2)^2]

φ_2= ([x_1, x_2])->[cos(x_1),cos(2x_1),cos(3x_1)] φ_3= ([x_1, x_2])->[|x_1|, |x_2|]

(a)



(b)



4. Consider the following optimization covered in class:

$\min_{\mathbf{w},b}$ $\mathbf{w}^T\mathbf{w}$ + C$\sum_j \xi_j$

such that

$\mathbf{w}^T x^i + b \geq +1 - \xi_i$ if x^i is class +1

$\mathbf{w}^T x^i + b \leq -1 + \xi_i$ if x^i is class -1

(a) Which term(s) in the optimization is/are used to permit limited classification errors?

(b) Which term(s) in the optimization is/are used to maximize the margin?

(c) Which term(s) in the optimization is/are used to encourage proper classification?

5. Consider each set of support vector and find the resulting **w**

(a) $\boldsymbol{x^1} = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}, \ y^1 = +1, \ \alpha^1 = 0.5$ $\quad \boldsymbol{x^2} = \begin{bmatrix} -3 \\ 3 \\ 0 \end{bmatrix}, \ y^2 = -1, \ \alpha^2 = 1$

$\boldsymbol{x^3} = \begin{bmatrix} 1 \\ 4 \\ -4 \end{bmatrix}, \ y^3 = +1, \ \alpha^3 = 0.5$

(b) $\boldsymbol{x^1} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \ y^1 = +1, \ \alpha^1 = 1.0$ $\quad \boldsymbol{x^2} = \begin{bmatrix} 0 \\ -3 \\ -1 \end{bmatrix}, \ y^2 = -1, \ \alpha^2 = 0.7$

$\boldsymbol{x^3} = \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}, \ y^3 = +1, \ \alpha^3 = 0.5$ $\quad \boldsymbol{x^4} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \ y^4 = -1, \ \alpha^4 = 0.8$

(c) $\boldsymbol{x^1} = \begin{bmatrix} -3 \\ -1 \\ 4 \end{bmatrix}, \ y^1 = +1, \ \alpha^1 = 1.0$ $\quad \boldsymbol{x^2} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}, \ y^2 = -1, \ \alpha^2 = 0.7$

$\boldsymbol{x^3} = \begin{bmatrix} -4 \\ -2 \\ 1 \end{bmatrix}, \ y^3 = +1, \ \alpha^3 = 0.5$ $\quad \boldsymbol{x^4} = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}, \ y^4 = -1, \ \alpha^4 = 0.8$

6. List two useful applications for logarithms in Machine Learning (they can be either practical engineering uses or mathematical derivation uses).

7. For the following functions, set the derivative with respect to h to 0:

(a)

f(h) = $\frac{\mathrm{h}^2-3}{5\mathrm{x}^3}$ (Presume $\mathrm{x} \neq 0$)

(b)

f(h) = $\prod_i 5\mathrm{h}^{2\mathrm{i}}\mathrm{e}^{-3\mathrm{h}}$ (Assume $\mathrm{h} \neq 0$)

8. Compute AIC given the following log-likelihoods and number of parameters

(a) log(L)=-42 # params= 12

(b) log(L)=-44 # params= 9

(c) log(L)=-33 # params=10

9. Let us define a logistic regression classifier with initial weight w^0=[1.2 -2.3 0.5].

Let us begin by optimizing for “maximum likelihood”, i.e., assuming no prior constraints. Also, assume $\varepsilon = 0.01$

(a) What will be the update to w^0 if we see the data point:
x^1=[3 2 1], y^1=1

(b) What will be the update to w^0 if we see the data point:
x^1=[3 2 1], y^1=0

(c) What will be the update to w^0 if we see the data point:
x^1=[1 2 3], y^1=1

(d) What will be a potential effect of decreasing ε.

(e) How will we change the optimization process if we include L2 regularization?

10. We are trying to determine the probability that Microsoft stock will go up (M=yes) based on the following yes/no features:
Recent increase in laptop sales (L=yes)
Recent increase in silicon price (S=yes)
Recent decrease in rain in Seattle, Washington (R=yes)

For the rest of this section: Let us assume R is independent of L and S, given M. However, Assume L and S are NOT independent given M.

(a) Write the expression for the likelihood of L, S, and R, given M. Simplify the expression by including the fewest number of variables possible in each probability term.

Let us say we have the following data from past days. Y means yes, N means no and ? means "data not available." Whenever data is unavailable for a given feature, it is not used in estimating probabilities relating to that feature. For example, the first data point:

	L	S	R	M
Day 1:	Y	?	N	Y

can be used in the estimation of P(L=yes) and P(M=yes | R=no), but it cannot be used in the estimation of P(L=yes,S=yes|M=yes)

Data:

	L	S	R	M
Day 1:	Y	?	N	Y
Day 2:	Y	Y	N	N
Day 3:	?	Y	N	Y
Day 4:	?	?	Y	N
Day 5:	N	N	?	Y
Day 6:	N	?	?	Y
Day 7:	Y	?	Y	Y
Day 8:	?	N	Y	Y

(b) Which of the probabilities below have Maximum Likelihood Estimate of 0? What are the non-zero values?

(b-1) P(M=yes)

(b-2) P(L=yes,S=yes|M=no)
(b-3) P(L=no,S=no,R=yes|M=yes)
(b-4) P(L=no,S=yes|M=yes)

(c) To calculate the a posteriori probability of P(L=no,S=no|M=yes), how do we incorporate a prior belief that there is a 20% chance Microsoft stock will fall or stay the same?

11. Let us consider a binary classification problem. For several objects in an online store, we wish to train a classifier to predict the label: "Does this object make people happy?" (variable H={yes,no}) We will use five features:
How big is it?
How old is it?
How expensive is it?
How familiar is it?
How natural is it?

Each feature will take on an integer value from 1 to 10, inclusive.

(a) Assuming all features are independent, how many parameters must be learned for the classifier?

(b) How many parameters do we learn if we wish to categorize each object into three different "make people happy" classes (including scores 1 – really makes people sad, 2 – neutral in affecting happiness, and 3 – really makes people happy).

12. Compute triangle distribution
Let us assume a single-feature data set in which each data point comes from one of two distributions:

For class 0: a **uniform** distribution starting x and ending x =a, $f(x) = \begin{cases} \frac{1}{a} & 0 \le x \le a \\ 0 & otherwise \end{cases}$

Class 1: a **triangle** distribution, starting at x=0 and ending at x=b:

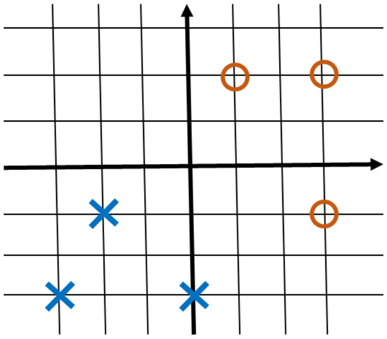
$$h(x) = \begin{cases} \frac{2}{b}\left(1 - \frac{x}{b}\right) & 0 \le x \le b \\ 0 & otherwise \end{cases}$$

(a) Given N data points (x^i,y^i) in a training set, what is the formula for the likelihood, given the parameters a and b? (You may express your answer in terms of f(x) and h(x).)

(b) Assuming all data points are positive, what is the maximum likelihood estimate for a and b? **Calculus will not help you here. You actually have to use your intuition!**

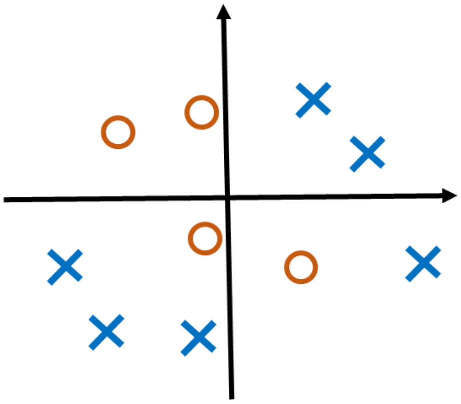
13. Consider points below and select four as likely support vectors for a linear separator.

(a) Draw your estimate of the separator and calculate the w value, assuming all support vectors have alpha=1.

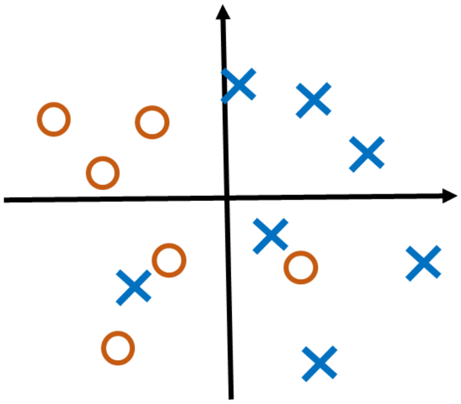


Suggest the two classification methods from the list below most fitting the following data:

(b)



(c)



Bayes classifier

Logistic classification
Linear SVM (no mapping function)
Kernel SVM (i.e., SVM with dimension mapping function)
SVM with slack variables

(d) What is overfitting? What are potential causes?

14. We wish to use a Maximum Likelihood Bayesian classifier to determine whether we are at a fruit stand (variable F) based on the presence of bananas (variable B), milk (variable M), and cash registers (variable C). We assume B, M, and C are all independent of one another given the value of F. Based on training data, we have found the following probabilities:

$P(B=\text{yes} \mid F=\text{yes}) = 0.8$	$P(M=\text{yes} \mid F=\text{yes})=0.1$	$P(C=\text{yes} \mid F=\text{yes})=0.7$
$P(B=\text{yes} \mid F=\text{no})=0.2$	$P(M=\text{yes} \mid N=\text{no})=0.5$	$P(C=\text{yes} \mid F=\text{no})=0.6$

(a) We observe cash registers, but no bananas or milk. Does the Maximum Likelihood classifier conclude we are in a fruit stand?

(b) We observe bananas, but no milk or cash registers. Does the Maximum Likelihood classifier conclude we are in a fruit stand?