1) If X and Y are independent events, and $p(Y)>0$, $p(X)=0.4$. What is the value of $p(X|Y)$?

x ,y independent iff p(xy) = p(x)p(y)
p(xly) = p(xy) / p(y) = p(x)p(y) / p(y) = p(x) = 0.4

0.4

x,y independent if p(xy) = p(x)p(y)

2) If X and Y are disjoint events and P(Y)>0. What is the value of P(X|Y)?

x, y disjoint iff p(xy) = 0
p(xly) = p(xy) / p(y) = 0 / p(y) = 0

0

x,y disjoint if p(xy) = 0

3)

1	2	3
0	1	4
5	6	0

Find the value of its inverse and its trace

tr 2

cofactor
-24,20,-5
18,-15,4
5,-4,1

adj
-24,18,5
20,-15, -4
-5,4,1

gauss-jordan elim or

get cofactor matrix:
c_ij = (-1)^(i+j) det(minor_ij)

get adjacency matrix:
transpose(cofactor)

get determinant of matrix

inv := (1/determinant) * adjacency matrix k classes, d features taking v values

NON-NAIVE:
k-1 prior probabilities
(v^d)-1 conditional probs per class
==> total: k*((v^d)-1) + k - 1

NAIVE:
k-1 prior probabilities
d features for conditional prob given per class
==> total: k*d + k - 1

4) Dataset for Naïve Bayes Classifier

Input Features x1, x2, x3			Label y
Temperature	Wind	Water	Picnic
Hot	High	warm	N
Cold	Low	warm	Y
Hot	Low	warm	N
Cold	High	cool	Y
Hot	High	cool	N
Cold	Low	warm	Y
Hot	HIgh	warm	N

1) How many independent parameters are present in the classifier? List Them
2) Give the estimations of these parameters?

k = 2 classes (Y/N)
d = 3 features (temp, wind, water) — all binary (hot/cold), (high/low), (warm/cool)

prior probs = k-1 = 1
cond probs = d features per class = d*k = 3*2 = 6
total params = 7

p(y=Y) ~ 3/7
p(temp=hot l y=Y) ~ 0
p(temp=hot l y=N) ~ 1
p(wind=high l y=Y) ~ 1/3
p(wind=high l y=N) ~ 3/4
p(water=warm l y=Y) ~ 2/3
p(water= warm l y=N) ~3/4

5) If A={2,5,4,7,8,9,12,15} and B = {15,4,5,7,9,8,12,2}
Check which of the following are true

1) $B \subset A$
2) $A \subset B$
3) B – A =Φ
4) A ∩ B =A

1 true, 2 true, 3 true, 4 true

6) If X is a uniformly distributed random variable that takes values from 2 to 9. What is the value of PMF(X=1)?

a) 1/8
b) 1/9
c) 0
d) 2/3

c

Midterm Practice Questions 2

QUESTION 1:

Suppose you are given eight independent samples that are drawn randomly from a normal distribution: $\{-2, 0, 1, 2, 4, 5, 8, 10\}$

a. What will be the maximum likelihood estimate for the mean (μ)?
b. What will be the maximum likelihood estimate for the variance (σ^2)?

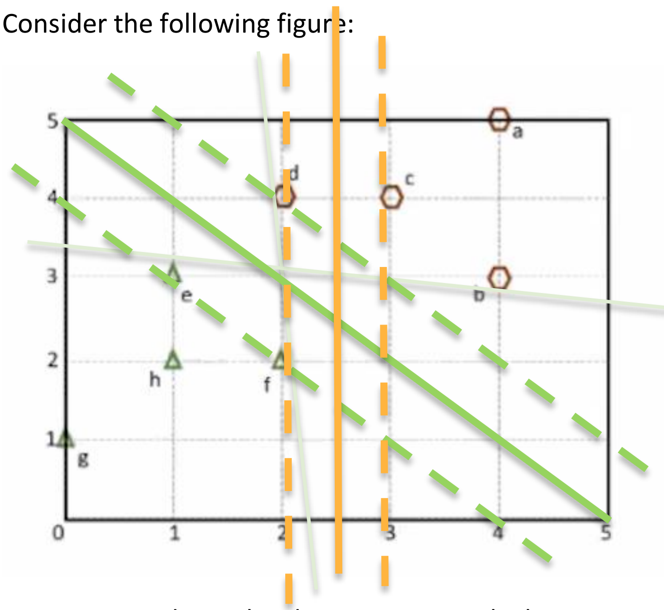
n = num samples

MLE(mu) = sum(x_i)/ n = 28 / 8 = 3.5

MLE(var) = sum((x_i - mu)^2)/ n = 116 / 8 = 14.5

QUESTION 2:

Consider the following figure:



support vectors —> take 2 closest points from each side then draw a like that reasonably fits

support vectors = e, f, d

decision boundary: y = -x + 5
margin: distance between e & d
sqrt((x_e - x_d)^2 + (y_e - y_d)^2) = sqrt(2)

margins become: x=2, x=3
decision boundary: x= 2.5

a. Considering hard-margin SVM, which points are the support vectors?
b. Find the equation of the decision boundary and calculate the margin.
c. If the point **'D'** is changed from a hexagon (⬡) to a triangle (△), will the decision boundary change? Write down the margin equations and calculate the decision boundary in this case.

kernel maps the non-linear separable data-set into a higher dimensional space where we can find a hyperplane that can separate the samples.
Kernel function defines the inner product in the transformed space

QUESTION 3:

a. What is the "kernel trick" and how is it useful?
b. What is the role of *C* in SVM? How does it affect the bias/variance trade-off?

In the given Soft Margin Formulation of SVM, C is a hyperparameter.
C hyperparameter adds a penalty for each misclassified data point.

Large C value implies a small margin, there is a tendency to overfit the training model. Such a model will have low bias and high variance.
Small Value of parameter C implies a large margin which might lead to underfitting of the model. Such a model will have high bias and low variance.

QUESTION 4:

Consider a biased coin that has a probability for heads as $p * (i+1)^2$ for i^{th} trial. After flipping the coin for three times, what is the maximum likelihood estimation of p for observing the pattern (*heads*, *heads*, *tails*)? $(0 < p < 1)$

d/dx ln(f(x)) = 1/f(x) * d/dx f(x)

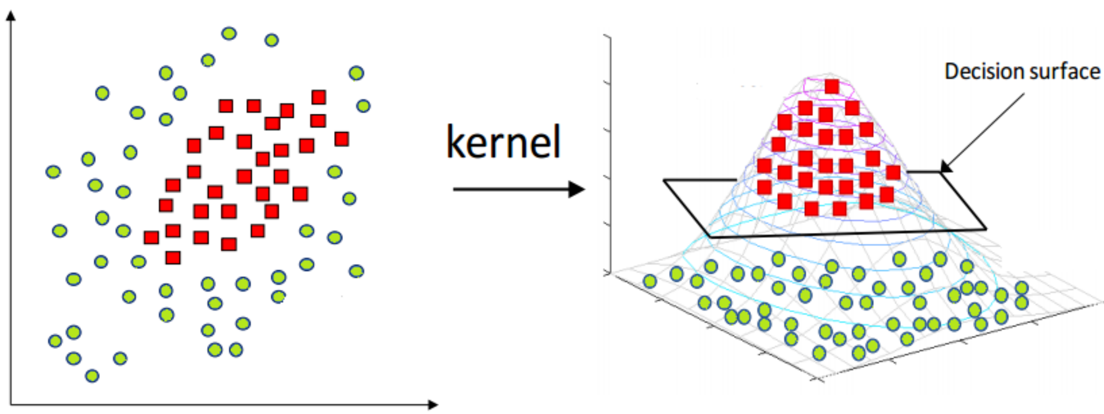
P(t=1, y=H) = 4p
P(t=2. y=H) = 9p
P(t=3, y=T) = 1-16p

P(HHT) = 4p * 9p * (1-16p)
apply log
log P(HHT) = log(4p) + log(9p) + log(1-16p)
for max LL: differentiate and set to 0

d/dx LL = 1/4p * (4) + 1/9p * (9) + 1/(1-16p) * (-16) = 0
2/p - 16/(1-16p) = 0
2/p = 16/(1-16p)
16p = 2 - 32p
16+32p = 2
48p = 2
p = 24

QUESTION 3 SOLUTION:

a. Earlier we have discussed applying SVM on linearly separable data but it is very rare to get such data. Here, kernel trick plays a huge role. The idea is to map the non-linear separable data-set into a higher dimensional space where we can find a hyperplane that can separate the samples.



It reduces the complexity of finding the mapping function. **So, Kernel function defines the inner product in the transformed space**. Application of the kernel trick is not limited to the SVM algorithm. Any computations involving the dot products (x, y) can utilize the kernel trick.

b. $\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{m}\xi_i^k$

In the given Soft Margin Formulation of SVM, *C* is a hyperparameter. *C* hyperparameter adds a penalty for each misclassified data point.

Large Value of parameter *C* implies a small margin, there is a tendency to overfit the training model. Such a model will have low bias and high variance.

Small Value of parameter *C* implies a large margin which might lead to underfitting of the model. Such a model will have high bias and low variance.