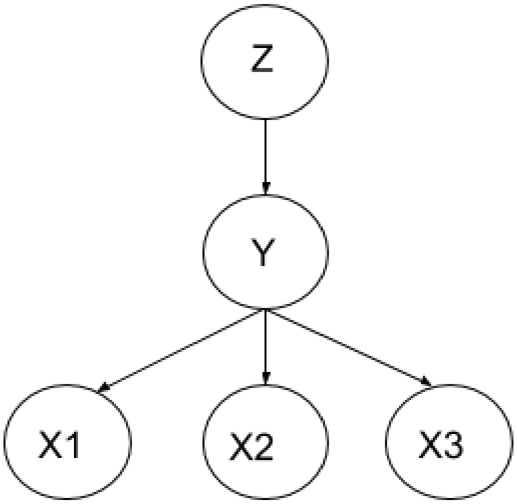
CSC591- HW1 (Group: ALL_12)

Anusha (amanur)
Kartik (kcshah)
Shailaja (smallic)

1. With the given scenario,
    a. The decision tree may vary depending on the interpretation of the feature variable, if specifically applied to nuclear plant scenario. The tree below is a generalised version as per the given question. Making Z as the parent of Y makes sense as if the sensor i.e., the core temperature (as given by Z) is abnormal (high), then Y value can be predicted to malfunction. The feature variables are accordingly the child nodes of Y.



    b. P(Y=1| Z= missing) = $\frac{P(Z=missing|\ Y=1)\ P(Y=1)}{P(Z=missing\ |\ Y=1)\ P(Y=1) + P(Z=missing|\ Y=0)\ P(Y=0)}$

    By substituting the given values,

    P(Y=1| Z= missing) = $\frac{0.15\theta_{Y=1}}{0.15\theta_{Y=1} + 0.03(1-\theta_{Y=1})}$

    c. The likelihood is given by:

    P($X_1, X_2, X_3 | \lambda$) = $(1/2)^{20}\ (\lambda)^{24}\ (1-3\lambda)^{4}$

    The log likelihood is given by:

    Log P($X_1, X_2, X_3 | \lambda$) = 20 log ½ + 24 log λ + 4 log ($1-3\lambda$)

    In order to calculate the value of λ:

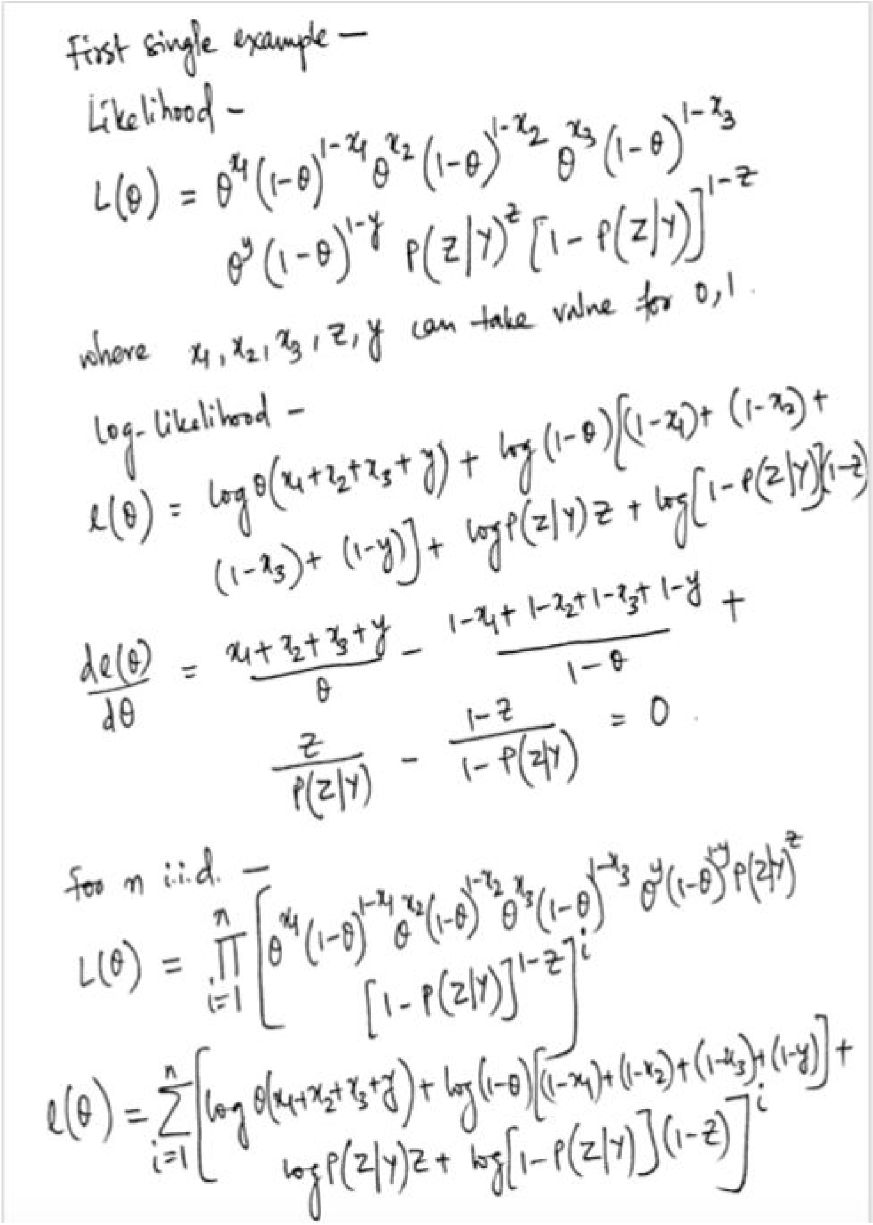
$\frac{d \log P(X_1,X_2,X_3|)}{d\lambda}$ = 24/ λ + 4/($1-3\lambda$)(-3) = 0

⇒ 24/ λ - 12/($1-3\lambda$) = 0

⇒ 24($1-3\lambda$) -12 λ = 0

⇒ $\lambda = 2/7$

d. For single example, as all the variables are Bernoulli, (for example, in coin toss, it is $p^k(1-p)^{(1-k)}$ where *k* can take value of either 0,1 for head or tail), similarly this is applied to our case as follows-



First single example –

Likelihood –

$$L(\theta) = \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\,\theta^{x_3}(1-\theta)^{1-x_3}$$
$$\theta^{y}(1-\theta)^{1-y}\;P(z|y)^{z}\left[1-P(z|y)\right]^{1-z}$$

where x_1, x_2, x_3, z, y can take value for 0,1.

log-likelihood –

$$\ell(\theta) = \log\theta(x_1+x_2+x_3+y) + \log(1-\theta)\left[(1-x_1)+(1-x_2)+\right.$$
$$\left.(1-x_3)+(1-y)\right] + \log P(z|y)z + \log\left[1-P(z|y)\right](1-z)$$

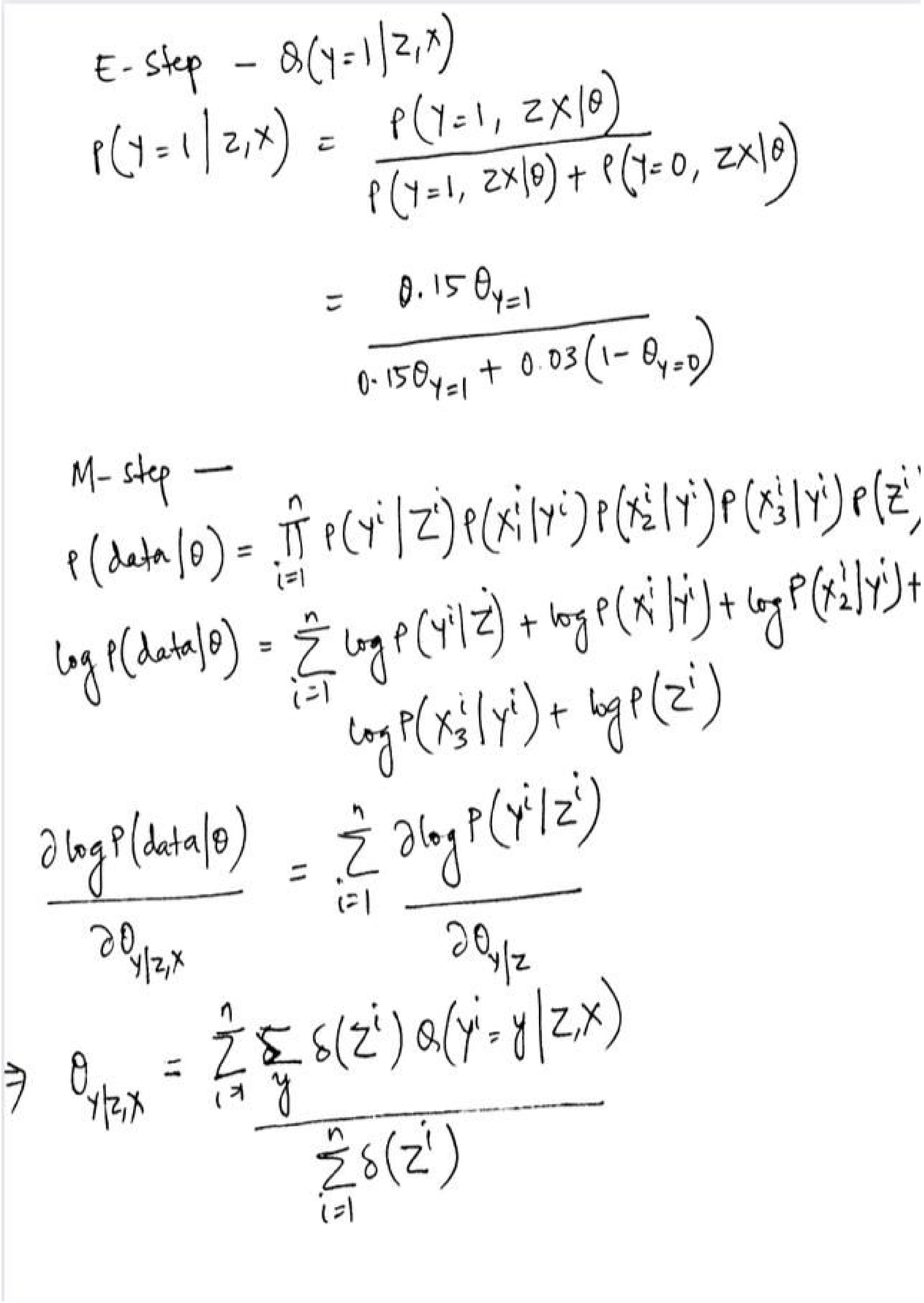
$$\frac{d\ell(\theta)}{d\theta} = \frac{x_1+x_2+x_3+y}{\theta} - \frac{1-x_1+1-x_2+1-x_3+1-y}{1-\theta} +$$
$$\frac{z}{P(z|y)} - \frac{1-z}{1-P(z|y)} = 0.$$

for n i.i.d. –

$$L(\theta) = \prod_{i=1}^{n}\left[\theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2}\theta^{x_3}(1-\theta)^{1-x_3}\,\theta^{y}(1-\theta)^{1-y}P(z|y)^{z}\right.$$
$$\left.\left[1-P(z|y)\right]^{1-z}\right]^{i}$$

$$\ell(\theta) = \sum_{i=1}^{n}\left[\log\theta(x_1+x_2+x_3+y) + \log(1-\theta)\left[(1-x_1)+(1-x_2)+(1-x_3)+(1-y)\right] +\right.$$
$$\left.\log P(z|y)z + \log\left[1-P(z|y)\right](1-z)\right]^{i}$$

e.



E-step — $Q(y=1|z,x)$

$$p(y=1|z,x) = \frac{p(y=1, zx|\theta)}{p(y=1, zx|\theta) + p(y=0, zx|\theta)}$$

$$= \frac{0.15\,\theta_{y=1}}{0.15\,\theta_{y=1} + 0.03(1-\theta_{y=0})}$$

M-step —

$$p(data|\theta) = \prod_{i=1}^{n} p(y^i|z^i)\,p(x_1^i|y^i)\,p(x_2^i|y^i)\,p(x_3^i|y^i)\,p(z^i$$

$$\log p(data|\theta) = \sum_{i=1}^{n} \log p(y^i|z) + \log p(x^i|y^i) + \log p(x_2^i|y^i) +$$

$$\log p(x_3^i|y^i) + \log p(z^i)$$

$$\frac{\partial \log p(data|\theta)}{\partial \theta_{y|z,x}} = \sum_{i=1}^{n} \frac{\partial \log p(y^i|z^i)}{\partial \theta_{y|z}}$$

$$\Rightarrow \theta_{y|z,x} = \frac{\sum_{i=1}^{n} \sum_{y} \delta(z^i)\,Q(y^i=y|z,x)}{\sum_{i=1}^{n} \delta(z^i)}$$

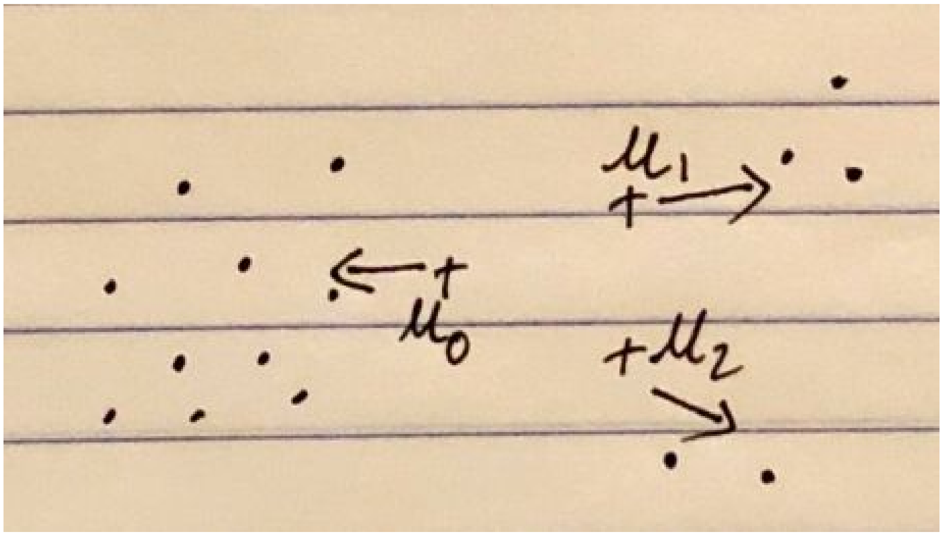
2. GMM
    a. No, the same cluster centers will not be produced as K-Means performs hard clustering, meaning, every datapoint is assigned to a single cluster. The K-Means cluster center is the means of the the data points belonging to that cluster.

Whereas GMM trained with EM performs soft clustering. In GMM, every data point is given a probability of the point belonging to every cluster. Each point is assigned to all clusters with some probability or weight. This affects the mean value of the cluster which is the center.

This difference in hard and soft clustering in the 2 methods leads to different cluster centers.

b. EM

   i.

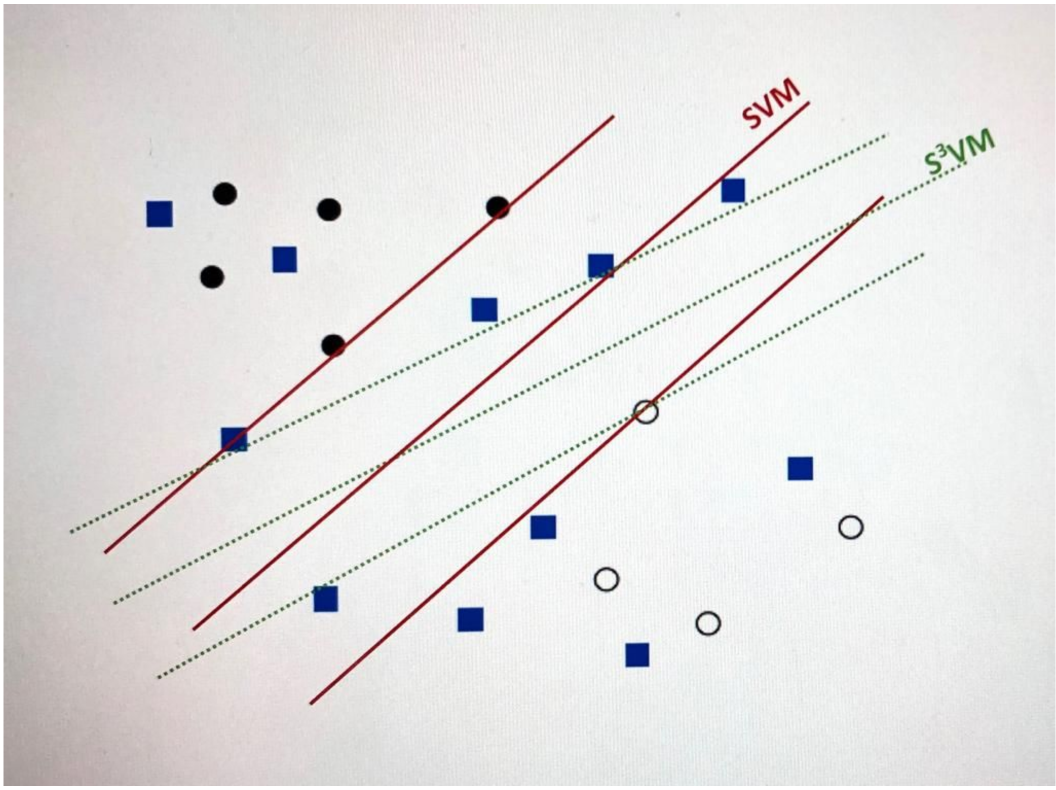


   ii. Marginal likelihood increases for every iteration of EM until it reaches local optimum.

   iii. The estimate of π_0 will increase on the next EM iteration because each point is assigned to all the clusters and π_0 is calculated considering the sum of all the probabilities of all the points in cluster 1. Since, μ_1 & μ_2 are close to μ_0 it wasn't getting a higher probability. So, in every EM iteration π_0 will increase.

3. With the given labeled and unlabeled data points,

   a. The marginal boundaries and separating hyper-plane for both *SVM* and S^3VM are shown below:



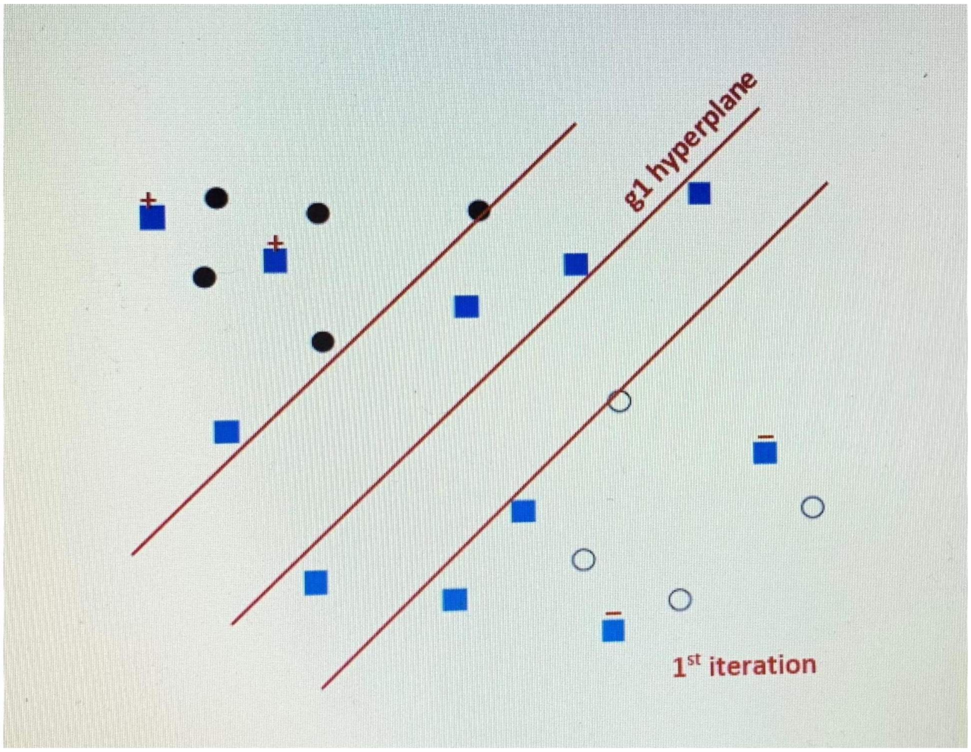
b.

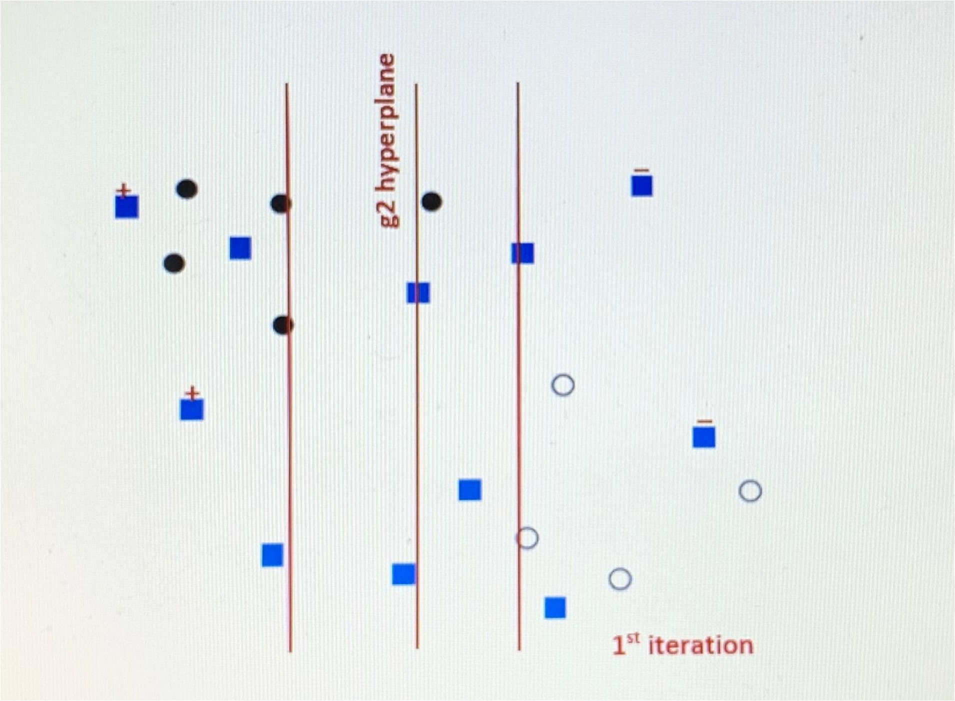
i. Datasets whose features naturally partition into two sets, and algorithms that use this division, fall into the co-training and the main underlying assumptions are (1) each set of features is sufficient for classification, and (2) the two feature sets of each instance are conditionally independent and not redundant.

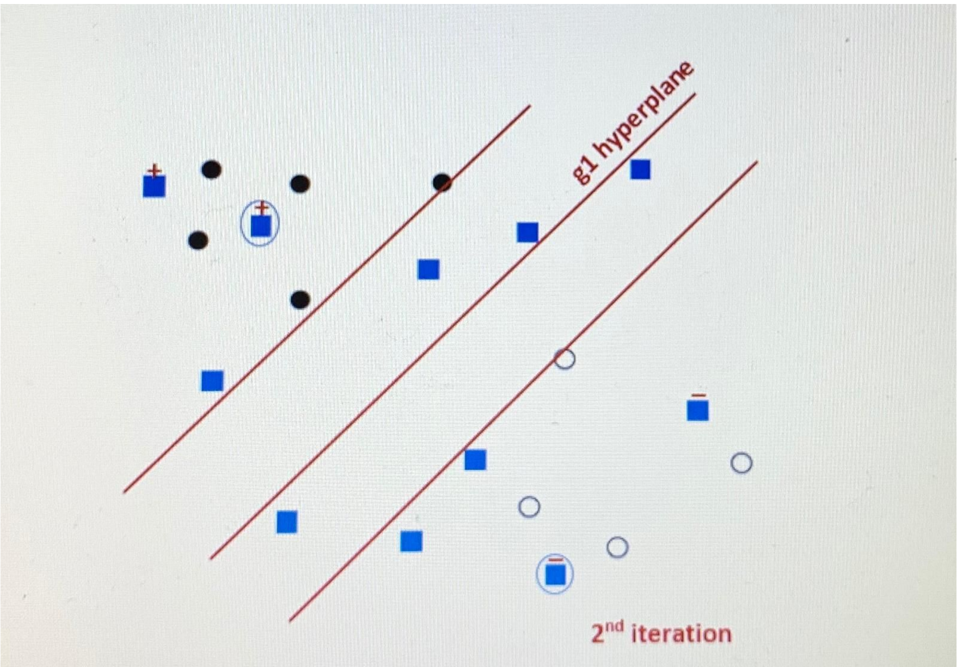
Idea is that if two algorithms use different representations of their hypothesis, they can learn two diverse models that can complement each other by labeling some unlabeled data & helping the training set of others. (Blum and Mitchel (1998)).

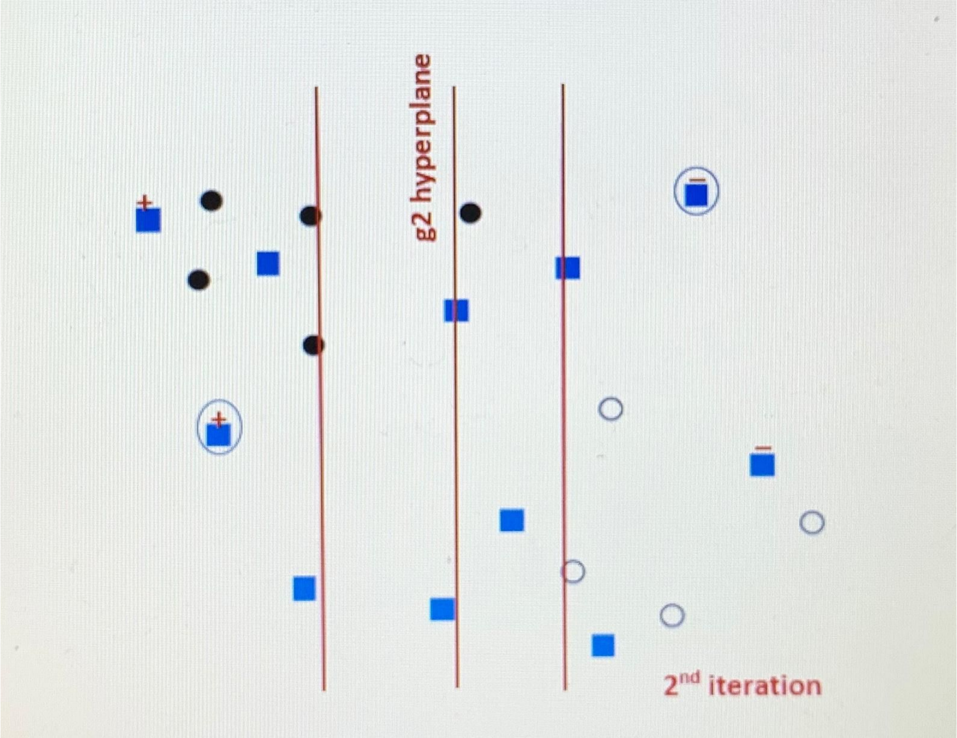
The two classifiers that can be used are SVM classifiers with different views i.e. different hyperplane separating the data points.

ii.


g1 hyperplane
+
+
−
−
1st iteration


g2 hyperplane
+
+
−
−
1st iteration





In the 2nd iteration, only the circled data points are labelled.

c. The hyper plane won't change as the co-training is performed based on the farthest distance from the hyper-plane, the nearer unlabeled points won't change and the hyper plane would still remain the same as S^3VM.

4. Data sequence: S= ({A,B} {B,C} {D,E} {A,D} {B,E,F})
   With $n_g = 0, x_g = 2, m_s = 5, w_s = 1,$
   a. No, it is not a subsequence of S as it violates the min_gap constraint for the event {A}{B}{C}. If we choose {A}{B} to be in the same element in the data sequence, e1, this leads to a min gap of 0 and violates the constraint. Similarly , for {B}{C}, If we choose {B}{C} to be in the same element in the data sequence, e2, this leads to a min gap of 0 and violates the constraint.
   b. Yes, it is a subsequence of S as it satisfies all the time constraints.
   c. Yes, it is a subsequence of S as it satisfies all the time constraints.
   d. No, it is not a subsequence of S as it violates the max_gap constraint for the event {C,D,E} {A,F} in w as x_g = 3.
   e. No, it is not a subsequence of S as it violates the max_gap constraint for the event {A,B,C,D} {E,F} in w as x_g = 4.

5. From the given frequent 3-sequence,
   a. Candidate 4-sequences are:
      i. <{1,2,3} {4}>
      ii. <{1,2,3} {5}>
      iii. <{1,3} {4,5}>
      iv. <{1,3} {4} {5}>
      v. <{1} {2,4} {4}>
      vi. <{2,4} {4,5}>
      vii. <{2,3} {4,5}>
      viii. <{2,3} {4} {5}>

   b. Candidate 4-sequences that are pruned are:
      i. <{1,2,3} {4}> as <{1,2} {4}> not there.
      ii. <{1,2,3} {5}> as <{1,2} {5}> not there.
      iii. <{1,3} {4,5}> as <{1,3} {5}> not there.
      iv. <{1,3} {4} {5}> as <{1,3} {5}> not there.
      v. <{1} {2,4} {4}> as <{1} {2} {4}> not there.
      vi. <{2,4} {4,5}> as <{2,4} {5}> not there.
      vii. <{2,3} {4} {5}> as <{2} {4} {5}> not there

      Only <{2,3} {4,5}> is not pruned

   c. With the given time constraint, the candidates that are pruned are:
      i. <{1,2,3} {4}> as {1,2} {4} contiguous subsequence is not frequent.
      ii. <{1,2,3} {5}> as {1,2} {5} contiguous subsequence is not frequent.

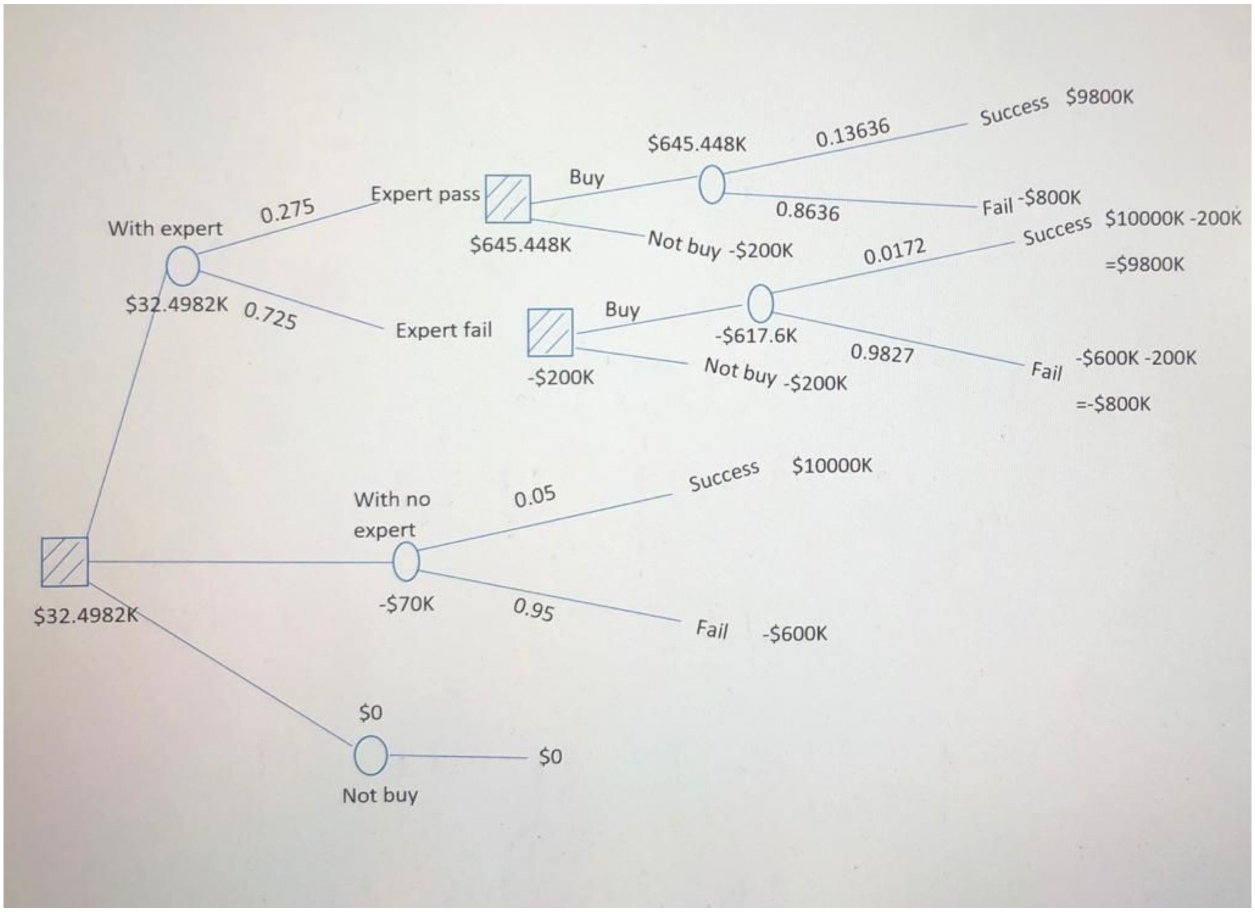
iii. <{1,3} {4,5}> as {1,3} {5} contiguous subsequence is not frequent.

iv. <{1} {2,4} {4}> as <{1} {2} {4}> contiguous subsequence is not frequent.

v. <{2,4} {4,5}> as <{2,4} {5}> contiguous subsequence is not frequent.

vi. <{2,3} {4} {5}> as <{2} {4} {5}> contiguous subsequence is not frequent.

6.



U(with no expert) => (0.05*10000K - 600 * 0.95) => -70K

Type	Expert Pass	Expert Fail
Company success	0.75*0.05	0.25*0.05
Company fail	0.25*0.95	0.75*0.95

P(Expert Pass) => 0.75*0.05 + 0.25*0.95 => 0.275
P(Company success|expert pass) => 0.75*0.05/0.275 => 0.13636
P(Company fail | exp pass) => 0.25*0.95/0.275 => 0.8636
U(Buy Company given exp pass) => 0.13636*9800K-800K*0.8636 => $645.448K

P(Expert fail) => 0.25*0.05+0.75*0.95 => 0.725
P(Company success | exp fail) => 0.25*0.05/0.725 => 0.0172
P(Company fail | exp fail) => 0.75*0.95/0.725 => 0.9827
U(Buy Company given exp fail) => 0.0172*9800K - 0.9827*800K => -$617.6K

U(with expert) => 0.275*645.448K - 0.725*200K => 32.4982K

Conclusion: We would only buy the company by hiring the experts and taking their suggestion since the chance of getting profit is more but, if they suggest no then we won’t buy. If we don’t take suggestions then it will be a higher loss.