<u>Question Bank</u>

Disclaimer: Please note that this question bank is a useful resource for practice, there is no guarantee that questions on the exam will be derived solely from this material.

Topics are- **SVM [Optimal Hyperplane, Primal and Dual Form, Kernel SVM]**
Unsupervised Learning: K Means
Convolutional Neural Networks
Application to Images, Fairness, Accountability and Transparency

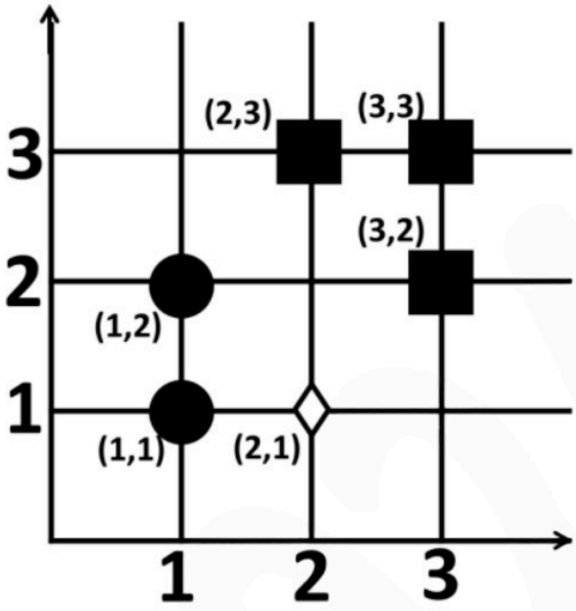
@@

Question - 1

Given the dataset: (1, 1), (3, 3), (4, 4), (5, 5), (6, 6), (9, 9), (0, 3), (3, 0) and assuming the initial centroids for (K = 3 – means clustering) to be $C1$ = (3, 3), C2 = (5, 5) and C3 = (6, 6). One iteration of the Expectation Maximization Algorithm for K-means clustering, will update $C3$ to (___, ___).

Question-2

Given the two-dimensional dataset consisting of 5 data points from two classes (circles and squares) and assume that the Euclidean distance is used to measure the distance between two points. The minimum odd value of k in k-nearest neighbor algorithm for which the diamond (◊) shaped data point is assigned the label square is _



Question - 3

We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0.$$

We now investigate a non-linear decision boundary.

(a) Sketch the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 \leq 4.$$

(c) Suppose that a classifier assigns an observation to the **blue** class if

$$(1+X_1)^2+(2-X_2)^2>4,$$

and to the **red** class otherwise. To what class is the observation $(0,0)$ classified? $(-1,1)$? $(2,2)$? $(3,8)$?

(d) Argue that while the decision boundary in part (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2, and X_2^2.

Question 4;

Explain how SVMs can be extended to handle multi-class classification problems, discussing the one-vs-one and one-vs-rest strategies.?

Question 5:

Explain the significance of the RELU Activation function in Convolution Neural Network?

Question 4

Answer the following questions about K-Means Clustering

(A) Calculate the percentage reduction in total variance after clustering.

(B) Analyze cluster stability with repeated K-means runs.

(C) Determine the effect of outliers on K-means clustering.

(D) Consider a dataset where points are distributed along two intersecting lines in a 2D plane. What challenges might K-means face in correctly clustering these points?

(E) Given a dataset with the following points: (10, 20), (20, 10), (10, 30), and assuming K-means is initialized with centroids at (10, 20) and (20, 10), calculate the centroids after the first iteration using Euclidean distance for clustering.

Q1. Prove that the convolution operation is linear and express a single-channel two-dimensional convolution operation as a matrix multiplication.

Q2. Discuss how zero-padding affects the Fourier Transform of the input signal and its implications on the learned features in a CNN.

Q3. Derive the formula for the computational complexity of a single convolutional layer in a CNN. Derive the computational complexity in terms of the number of multiplications required for this layer's forward pass.

Q4. Given an initial set of points: (2,3), (5,4), (9,6), and centroids (3,3), (8,5), calculate the new centroids after one iteration of K-means.

Q5. Analyze the convergence of K-means clustering. Assume you have a dataset and the K-means algorithm performs several iterations. How can you mathematically determine whether the algorithm has converged? What would you consider as a practical convergence criterion?

Q1. What is the role of max pooling in a CNN? How many parameters are there in a max-pooling layer?

Q2. A 28×28 input image is processed through a convolutional layer with 10 filters, each of size 5×5, with stride 1 and padding 0.

a) What is the dimension of the output feature map?

b) How many trainable parameters are in this layer (excluding biases)?

Q3. Write the mathematical expression minimized in the K-Means algorithm.

Q4.

A dataset consists of three points: $(2, 3), (6, 5), (8, 7)$. Perform **one iteration** of the K-Means algorithm with $k = 2$, starting with centroids $c_1 = (2, 3)$ and $c_2 = (8, 7)$. Assign clusters and update centroids.

Q5. Define the term "optimal hyperplane" in the context of SVMs.

Q6.

Given the final centroids of a K-Means clustering task, $c_1 = (2, 2)$, $c_2 = (7, 7)$, calculate the total within-cluster variance for the points: $(1, 1), (2, 3), (7, 6), (8, 8)$.

Q7. How does the kernel trick enable SVMs to handle non-linearly separable data?

Q8.

For the kernel $K(x, z) = (x^T z + 1)^2$, compute $K((1, 0), (0, 1))$ explicitly. Explain how this kernel maps data to a higher-dimensional space.

Q9. A facial recognition model predicts whether a person belongs to one of two groups (Group A or Group B) using image data. After testing, you observe the following confusion matrices:

- Group A: TP: 80, FP: 10, FN: 20, TN: 90
- Group B: TP: 40, FP: 20, FN: 60, TN: 80

Compute the precision and recall for each group and interpret any disparities.

Q1.

Explain why the K-Means algorithm converges to a local minimum of the objective function. Can it guarantee finding the global minimum? Justify your answer.

Q2.

The K-Means algorithm is sensitive to the initial cluster centers.

a. Demonstrate with an example how poor initialization can lead to suboptimal clustering.
b. Discuss strategies to mitigate this problem, such as K-Means++.

Q3.

K-Means is traditionally based on Euclidean distance. What challenges arise when using K-Means with non-Euclidean distance metrics, such as cosine similarity? How would you modify the algorithm to accommodate this?

Q4.

In high-dimensional data, the curse of dimensionality can affect K-Means performance.

a. Explain why this happens.
b. Propose and justify methods to address this issue before clustering.

Q5.

How does missing data affect the clustering results of K-Means?

a. Describe the marginalization approach for handling missing data in K-Means.

b. Derive the modified distance function when marginalizing over missing dimensions assuming standardized data with N(0,1) distribution.

Q1.

One of the most commonly used kernels in SVM is the Gaussian RBF kernel:$k(x\_i, x\_j) = \exp(-||x\_i - x\_j||^2 / 2\sigma)$.Suppose we have three points, z_1, z_2, and x. z_1 is geometrically very close to x, and z_2 is geometrically far away from x. What is the value of $k(z_1, x)$ and $k(z_2, x)$? Choose one of the following:

a. $k(z_1, x)$ will be close to 1 and $k(z_2, x)$ will be close to 0.

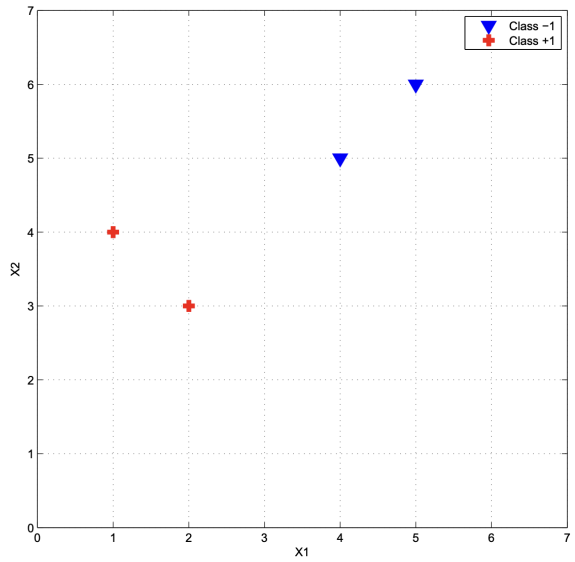
b. $k(z_1, x)$ will be close to 0 and $k(z_2, x)$ will be close to 1.

c. $k(z_1, x)$ will be close to c_1, $c_1 >> 1$ and $k(z_2, x)$ will be close to c_2, $c_2 << 0$, where $c_1, c_2 \in R$.

d. $k(z_1, x)$ will be close to c_1, $c_1 << 0$ and $k(z_2, x)$ will be close to c_2, $c_2 >> 1$, where $c_1, c_2 \in R$.

Q2.

Support vector machines learn a decision boundary leading to the largest margin from both classes.You are training SVM on a tiny dataset with 4 points shown in Figure 2. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label +1 (denoted with triangles). Find the weight vector w and bias b. What's the equation corresponding to the decision boundary, also circle the support vectors and draw the decision boundary?



Q3. You are solving the binary classification task of classifying images as cat vs. non-cat. You design a CNN with a single output neuron. Let the output of this neuron be z. The final output of your network, $\hat{y}$ is given by: $\hat{y} = \sigma(\text{ReLU}(z))$ You classify all inputs with a final value $\hat{y} \geq 0.5$ as cat images. What problem are you going to encounter?

Q4.

Prove that the kernel K(x1, x2) is symmetric, where xi and xj are the feature vectors for ith and jth examples.

Q5.

Now let us discuss an SVM classifier using a second-order polynomial kernel. The first polynomial kernel maps each input data x to $\Phi_1(x) = [x, x^2]^T$. The second polynomial kernel maps each input data x to $\Phi_2(x) = [2x, 2x^2]^T$.In general, is the margin we would attain using $\Phi_2(x)$:

A. Greater

B. Equal

C. Smaller

D. Any of the above...in comparison to the margin resulting from $\Phi_1(x)$?

Q1 - You come up with a CNN classifier. For each layer, calculate the number of weights, number of biases, and the size of the associated feature maps.

The notation follows the convention:

• CONV-K-N denotes a convolutional layer with N filters, each of them of size K×K, Padding and stride parameters are always 0 and 1 respectively.

• POOL-K indicates a K×K pooling layer with stride K and padding 0.

• FC-N stands for a fully-connected layer with N neurons.

Table:

Layer	Activation map dimensions	Number of weights	Number of biases
INPUT	128 × 128 × 3	0	0
CONV-9-32			
POOL-2			
CONV-5-64			
POOL-2			
CONV-5-64			
POOL-2			
FC-3			

Q2 -

Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $\mathbf{x} \in \mathbb{R}^d$ to a high dimensional feature space Q by giving the form of dot product in Q: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$.

Assume we use radial basis kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. Thus we assume that there's some implicit unknown function $\phi(\mathbf{x})$ such that

$$\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

Prove that for any two input instances $\mathbf{x}_i$ and $\mathbf{x}_j$, the squared Euclidean distance of their corresponding points in the feature space Q is less than 2, i.e. prove that $\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 < 2$.

Q3 - Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5),A3(8, 4),A4(5, 8),A5(7, 5),A6(6, 4),A7(1, 2),A8(4, 9)
Initial cluster centers are:

- A1(2, 10)
- A4(5, 8)
- A7(1, 2)

The distance function between two points $a = (x_1, y_1)$ and $b = (x_2, y_2)$ is defined as:

$d(a, b) = |x_2 - x_1| + |y_2 - y_1|$

Use the K-Means Algorithm to find the three cluster centers after the second iteration.

Q4 - Which of the following propositions are true about a CONV layer? (Check all that apply.)

(i) The number of weights depends on the depth of the input volume.

(ii) The number of biases is equal to the number of filters.

(iii) The total number of parameters depends on the stride.

(iv) The total number of parameters depends on the padding.

Q5- You have an AI-based loan approval system. During testing, you discover that the system has a gender bias. Which responsible AI principle does this violate?

A. Accountability

B. Reliability and safety

C. Transparency

D. Fairness

Q1. During a research work, you found seven observations as described with the data points below. You want to create three clusters from these observations using K-means algorithm. After first iteration, the clusters C1, C2, C3 has following observations:

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

If you want to run a second iteration then what will be the cluster centroids?

Q2. You are given a set of one-dimensional data points: {5, 10, 15, 20, 25, 30, 35}. Assume that k= 2 and first set of random centroid is selected as {15, 32} and then it is refined with {12, 30}.

1) Create two clusters with each set of centroid mentioned above following the K-means approach
2) Calculate the SSE (Sum of Squared Errors) for each set of centroid

Q3.

a) Data augmentation is often used to increase the amount of data you have. Should you apply data augmentation to the test set? Explain why.

b) Weight sharing allows CNNs to deal with image data without using too many parameters. Does weight sharing increase the bias or the variance of a model?

Q4. You are benchmarking runtimes for layers commonly encountered in CNNs. Which of the following would you expect to be the fastest (in terms of floating point operations)?

(i) Conv layer (convolution operation + bias addition)
(ii) Max pooling
(iii) Average pooling
(iv) Batch Normalization

Q5. You are building the next state-of-the-art CNN for a vision task, using a stack of modules that each look similar to the following:
(Layer Input) −→ (Conv Layer) −→ (Batch Norm) −→ (Activation) −→ (Next Layer Input)
You remember from your ML knowledge that each Conv layer has a set of learnable weights and biases. A colleague advises you to not learn biases (set them all to zero, forever) for these layers. Would performance be affected if you chose to follow their advice? Briefly explain (1-2 sentences) your answer.

Q1. Given a dataset with n points in a 2D space, the K-Means algorithm is applied to partition the data into k clusters.

(a) Explain the initialisation step and its impact on the algorithm's convergence.
(b) Propose a strategy to improve the initialisation process and justify its effectiveness in reducing sensitivity to local minima.

Q2. In a CNN, consider an input image of size 32×32×32x3. A convolutional layer has 16 filters of size 5×5 with a stride of 1 and no padding.

(a) Calculate the dimensions of the output after this layer.
(b) Discuss how increasing the stride to 2 affects the computational cost and the model's ability to capture spatial features.

Q3.Suppose you use a CNN-based model to classify medical images for disease diagnosis.

(a) Identify potential fairness issues that could arise in this context and suggest methods to mitigate bias in the model.
(b) How would you design a framework to ensure transparency and accountability when deploying this system in real-world healthcare?

Q4. Consider a linearly separable dataset in a 2D feature space. The equation of the hyperplane is given as xw^T+b=0. Derive the conditions for maximising the margin between the two classes and explain why these conditions lead to the optimal hyperplane.

Q5. Suppose you have a dataset that is not linearly separable in its original feature space.

(a) Explain the role of kernel functions in SVMs and derive the form of the decision boundary in the transformed space.
(b) For the Radial Basis Function (RBF) kernel, explain the significance of the hyperparameter γ\gammaγ and how it affects model performance.

Q1. Given a dataset with points A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), and A8(4, 9), and initial cluster centers:

1. A1(2, 10)
2. A4(5, 8)
3. A7(1, 2)

Using the Manhattan distance as the distance metric, perform two iterations of the K-Means algorithm and calculate the cluster centroids after the second iteration.

Q2. Explain how the ReLU activation function contributes to the performance of a Convolutional Neural Network (CNN). Discuss its advantages over other activation functions such as sigmoid or tanh.

Q3. In a Support Vector Machine (SVM), derive the conditions necessary for finding the optimal hyperplane in a linearly separable dataset. Explain the role of margin maximization in this context

Q4. A convolutional layer in a CNN has 32 filters of size 3×3 applied to an input of size 64×64×3. Assuming stride 1 and no padding: Calculate the dimensions of the output feature map. Determine the total number of trainable parameters in this layer.

Q5. K-Means clustering is sensitive to the choice of initial cluster centers.

(a) Provide an example illustrating how poor initialization can lead to suboptimal clustering.

(b) Propose a strategy such as K-Means++ to address this issue and explain its benefits.

Q1. Explain the kernel trick in Support Vector Machines (SVMs). How does it enable SVMs to classify data that is not linearly separable in the original feature space? Provide an example using the Radial Basis Function (RBF) kernel.

Q2. Consider a CNN with the following architecture:

- Input: 128×128×3
- CONV-3-16 (3×3 filters, stride 1, no padding)
- POOL-2 (2×2 pooling, stride 2)
- CONV-3-32 (3×3 filters, stride 1, no padding)

Calculate the size of the output feature map after each layer.

Q3. Discuss the potential fairness issues in applying a CNN model for facial recognition. Suggest methods to mitigate bias and ensure fairness in the model's predictions.

Q4. Suppose you have the following confusion matrices for two classes in a binary classification problem:

- Class A: TP = 90, FP = 20, FN = 10, TN = 80
- Class B: TP = 50, FP = 30, FN = 40, TN = 70

Compute the precision and recall for each class and interpret the results in the context of imbalanced datasets.

Q5. A K-Means clustering algorithm is applied to a dataset where points are distributed along two intersecting lines in a 2D space. Discuss the challenges this poses for K-Means and suggest alternative clustering approaches better suited for such data distributions.

Questions on GMM

Numerical type:

Problem:

Suppose we have a dataset with 6 one-dimensional data points:

$$x = [1.1, 1.9, 3.2, 4.5, 5.0, 5.8].$$

We want to fit a Gaussian Mixture Model (GMM) with $K = 2$ components using the Expectation-Maximization (EM) algorithm. Assume the following initial parameters:

1. Means: $\mu_1 = 2.0, \mu_2 = 5.0$.
2. Variances: $\sigma_1^2 = 1.0, \sigma_2^2 = 1.0$.
3. Mixing coefficients: $\pi_1 = 0.5, \pi_2 = 0.5$.

Perform one iteration of the EM algorithm to update the parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1, \pi_2$. Assume that the Gaussians are independent and have identical priors.

Theoretical:

Q1)

Consider a model where:

1. The marginal distribution $p(z)$ for the latent variable z is given as $p(z = k) = \pi_k$, where $\sum_{k=1}^{K} \pi_k = 1$, and $\pi_k > 0$ for all k.
2. The conditional distribution $p(x|z = k)$ for the observed variable x, given $z = k$, is a Gaussian distribution:

$$p(x|z = k) = \mathcal{N}(x|\mu_k, \Sigma_k),$$

where μ_k and Σ_k are the mean and covariance of the k-th Gaussian component.

Show that the marginal distribution $p(x)$, obtained by summing $p(z)p(x|z)$ over all possible values of z, is itself a mixture of Gaussians of the form:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k).$$

Q2)

Consider a probabilistic model with a mixture of distributions, where the covariance matrices Σ_k of the components are constrained to share a common value Σ. Derive the Expectation-Maximization (EM) algorithm equations for maximizing the likelihood function under this constraint.

Q3)

Consider a mixture model where data points are assigned responsibilities incrementally during the Expectation-Maximization (EM) algorithm. Specifically, this involves updating the parameters of the model iteratively as new responsibilities are calculated. Derive the explicit formulae for the Maximization (M-step) updates of:

1. The covariance matrices for the mixture components.
2. The mixing coefficients of the model.

Your derivation should be analogous to the incremental update formula for the means, where the parameters are updated progressively using the current and previous responsibilities.

Q4)

Prove that maximizing the expected complete-data log-likelihood function for a mixture of discrete distributions (in this case, Bernoulli distributions), where each distribution is parameterized by μ_k, yields the update equation for μ_k in the Maximization (M-step) of the Expectation-Maximization (EM) algorithm. Specifically, derive the M-step update equation for μ_k under this scenario.

Reference:
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

---

TOPIC:- SVM

Question:- We learned a two class linear SVM for a linearly separable input data. Let W and b be the parameters we obtained for the primal SVM formulation.

In the standard SVM formulation (SVM1) we use the following constraints for all x in class 1:
$W^T x + b \geq 1$

and for all x in class 0:
$W^T x + b \leq -1$

Assume that we learned a new SVM model (SVM2) using the following constraints instead, for all x in class 1:

$W^T x + b \geq 0$

and for all x in class 0: $W^T x + b < 0$

1. [2 pts] If we compare the margin of SVM2 to that of SVM1 we can say that:

a) The margin increased
b) The margin decreased

c) The margin stayed the same
d) Impossible to tell

Decrease (2). If we set the threshold at 0 then there will be no margin and since this is a linearly separable dataset the margin will decrease

2. Assume that we are using a new SVM, SVM3 which uses W/ 2 and b / 2 where W and b are the parameters learned for SVM1. With these new parameters

(a) [2 pts] Are we guaranteed that SVM3 would not make any mistakes on the training data? (recall that an SVM classifier determines the class based on the sign of WT x + b where x is the input).

a) Yes
b) No

Yes. This is a linearly separable problem and everything that was higher than 0 before remains higher now and similarly for lower than 0.

(b) [2 pts] How would the margin for SVM3 compare to the margin of SVM1?

a) The margin would increase
b) The margin would decrease
c) The margin would stay the same
d) Impossible to tell

The Margin would increase. The margin is $\frac{2}{\sqrt{(W^T W)}}$ and since W is divided by 2 it would increase.

(c) [3 pts] The number of support vectors for SVM3 compared to SVM1 would (recall that support vectors are those inputs that are either exactly on the +1 or -1 planes or those points that are between these planes and the decision boundaries)

a) The number of support vectors would likely increase
b) The number of support vectors would likely decrease
c) The number of support vectors would likely stay the same
d) Impossible to tell

The number of support vectors would likely increase. All previous support vectors now lie between the margin and the decision line and there could only be new support vectors added.

3. Now assume that we are using a new SVM, SVM4 which uses 2W and 2b where W and b are the parameters learned for SVM1. With these new parameters.

(a) [2 pts] How would the margin for SVM4 compare to the margin of SVM1?

a) The margin would increase
b) The margin would decrease
c) The margin would stay the same
d) Impossible to tell

The margin would decrease.

(b) [2 pts] The number of support vectors for SVM4 compared to SVM1 would (recall that support vectors are those inputs that are either exactly on the +1 or -1 planes or those points that are between these planes and the decision boundaries)

a) The number of support vectors would likely increase
b) The number of support vectors would likely decrease
c) The number of support vectors would likely stay the same
d) Impossible to tell

The number of support vectors would likely decrease. In fact, there will be no support vectors. All previous ones will now have a value of 2 or -2 and nothing has a lower value.

4. [3 pts] Assume that the number of support vectors for SVM1 is k. Can you give the exact number of support vectors for SVM3 or SVM4? Please choose only one of them, either SVM3 or SVM4 and provide the number of support vectors for that classifier only (specify which one you chose). Its fine for the result to be a function of k (but not big O notation, so 2k is possible answer while O(k 2) is not).

Question:- [2 points] Consider a 2 class classification problem with a dataset of inputs $\{x1 = (-1, -1)$, $x2 = (-1, +1)$, $x3 = (+1, -1)$, $x4 = (+1, +1)\}$ and a corresponding set of targets $\{t1, t2, t3, t4\}$ where $ti \in \{+1, -1\}$. Using this feature space (no kernel trick), can we build a SVM to perfectly classify this dataset regardless of values of ti 's?

No, since the decision boundary is linear. For examples, we cannot classify in the case $t1 = +1$, $t2 = t3 = -1$, $t4 = +1$

Question:- 3. [1 point] True/False After training a SVM, we can discard all examples which are not support vectors and can still classify new examples.

True.

Question:- [1 point] (True/False) Assume we are using the primal non linearly separable version of the SVM optimization target function. What do we need to do to guarantee that the resulting model is linearly separable?

Set $C = \infty$.
