Manoj Kumar

GATE AIR - 13

M.Tech in Data Science From IIT Guwahati

Expertise in Machine Learning, Deep Learning, Artificial Intelligence, Probability and Statistics

Telegram: @Manoj_Gate_DSAI

Unacademy

CLICK

Some True/False

1 Using a model with less bias always results in a lower test error compared to a model with more bias.

2 **If we have two data points x_1 and x_2, then any function $k(x_1, x_2)$ that is written in terms of x_1 and x_2 is a valid kernel function.**

3 Decision trees are more suitable than logistic regression for predicting the probability of a patient having cancer based on cellular images.

4 When learning a decision tree, a feature x_j which is not correlated with the label y (i.e., $\text{corr}(x_j, y) = 0$) will never be split on and hence will never be used in the tree.

5 The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

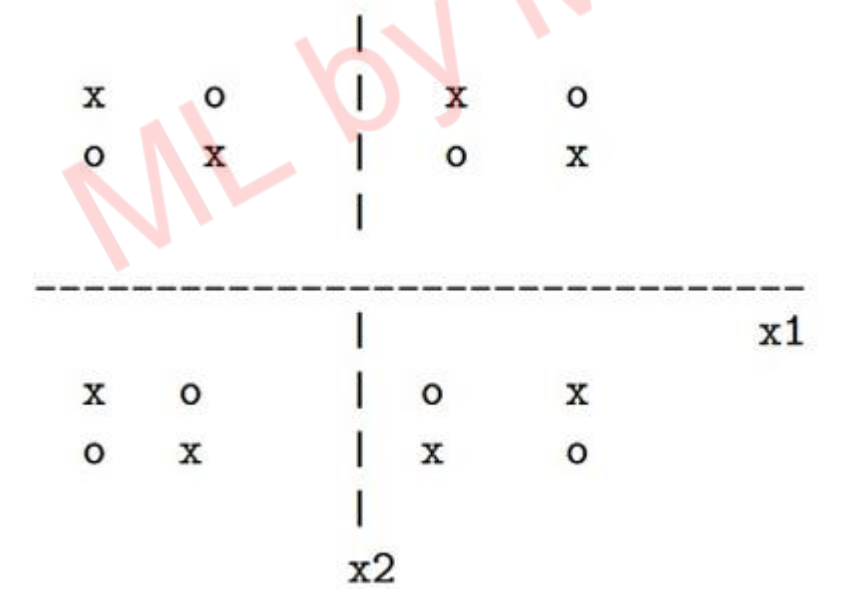
Some True/False

6 In k-fold cross-validation, a higher number of folds (larger k) generally leads to a higher estimated error.

7 The hinge loss used by SVMs always gives a little bit of weight to points that are correctly classified and distant from the decision boundary, similar to the log-loss used by logistic regression.

8 The kernel trick involves computing the dot product of input vectors in a higher-dimensional space without explicitly transforming them.

9 Radial basis functions can be used to make the following dataset linearly separable.



Some True/False

10 Leave-one-out cross-validation (LOOCV) generally gives less accurate estimates of true test error than 10-fold cross-validation.

11 In a soft margin SVM, only the points that are closest to the decision boundary on both sides are responsible for changes in the orientation and position of the decision boundary.

12 In a hard margin SVM, only the points that are closest to the decision boundary on both sides are responsible for changes in the orientation and position of the decision boundary.

13 When moving from a linear kernel to higher degree polynomial kernels in SVM, the support vectors always remain the same.

Question 1,2

What does the term 'support vectors' refer to in the context of SVM?

A) Parameters that directly influence the positioning and orientation of the decision boundary.

B) The maximum number of data points that can be classified correctly.

C) Data points that directly influence the positioning and orientation of the decision boundary.

D) The coefficients of the linear regression model used in SVM.

Consider an SVM using a Radial Basis Function (RBF) kernel, $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$. If you notice that the model is underfitting the training data, which parameter adjustment might improve the model's performance?

A) Increase the gamma (γ) parameter.

B) Decrease the gamma (γ) parameter.

C) Simplify the model by reducing the training iterations.

D) Remove the kernel trick and use a linear kernel.

Question 3[MSQ]

Consider two data points, $x_1 = (1, -1)$ and $x_2 = (2, 2)$, in a binary classification task using an SVM with a custom kernel function $K(x, y)$. The kernel function is applied to these points resulting in the following matrix, referred to as matrix A:

$$\begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) \\ K(x_2, x_1) & K(x_2, x_2) \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 6 \end{bmatrix}$$

Which of the following statements is correct regarding matrix A and the kernel function $K(x, y)$?

A) $K(x, y)$ is a valid kernel.

B) $K(x, y)$ is not a valid kernel.

C) Matrix A is positive semi-definite.

D) Matrix A is not positive semi-definite.

Question 4

When choosing the best feature to split on at a node in a Decision Tree, which of the following criteria should be maximized? (Here, H() means entropy, and P() means probability)

(a) $P(Y|X_j)$

(b) $P(Y) - P(Y|X_j)$

(c) $H(Y) - H(Y|X_j)$

(d) $H(Y|X_j)$

(e) $H(Y) - P(Y)$

Question 5,6

Consider the one-dimensional dataset with four positive data points {0, 1, 2, 3} and three negative data points {-3, -2, -1}. We want to find the optimal decision boundary (a point in 1D) using a soft-margin linear SVM.

Soft-Margin Linear SVM Formulation

The soft-margin linear SVM solves the following optimization problem:

```
Minimize: (1/2) w^2 + C * (Σ ξi)
```

```
Subject to: yi(w*xi + b) ≥ 1 - ξi   for all i (i = 1, 2, ..., 7)
            ξi ≥ 0                  for all i (i = 1, 2, ..., 7)
```

Question 5,6

In a Support Vector Machine (SVM), if the regularization parameter C = 0, how many support vectors do we have?

a) 2
b) 3
c) 5
d) 7

In a Support Vector Machine (SVM), if the regularization parameter C approaches infinity ($C \rightarrow \infty$), how many support vectors do we have?

a) 2
b) 3
c) 5
d) 7

Question 7,8

[2 points] Let $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be two symmetric, positive definite kernel functions. Which of the following *cannot* be a valid kernel function?

(a) $K(x, x') = 5 \cdot K_1(x, x')$

(b) $K(x, x') = K_1(x, x') + K_2(x, x')$

(c) $K(x, x') = K_1(x, x') + \dfrac{1}{K_2(x, x')}$

(d) All three are valid kernels.

If an SVM model is overfitting the training data, which of the following strategies could help in reducing the overfit?

A) Increase the polynomial degree of the kernel.

B) Decrease the penalty parameter C.

C) Increase the parameter C.

D) Use a more complex kernel.

Question 9,10

Given a kernel function K_1: R^n x R^n -> R and its corresponding feature map ϕ_1: R^n -> R^d, which feature map ϕ: R^n -> R^d would correctly produce the kernel $cK_1(x, z)$, where c is a positive constant?

a) $\phi(x) = c\phi_1(x)$

b) $\phi(x) = \sqrt{c}\,\phi_1(x)$

c) $\phi(x) = c^2\phi_1(x)$

d) No such feature map exists.

Given feature mappings ϕ_1: R^n -> R^d and ϕ_2: R^n -> R^d, and kernels $K_1(x, z)$ and $K_2(x, z)$ respectively, which feature mapping $\phi(x)$ would produce the kernel $K(x, z) = K_1(x, z) * K_2(x, z)$?

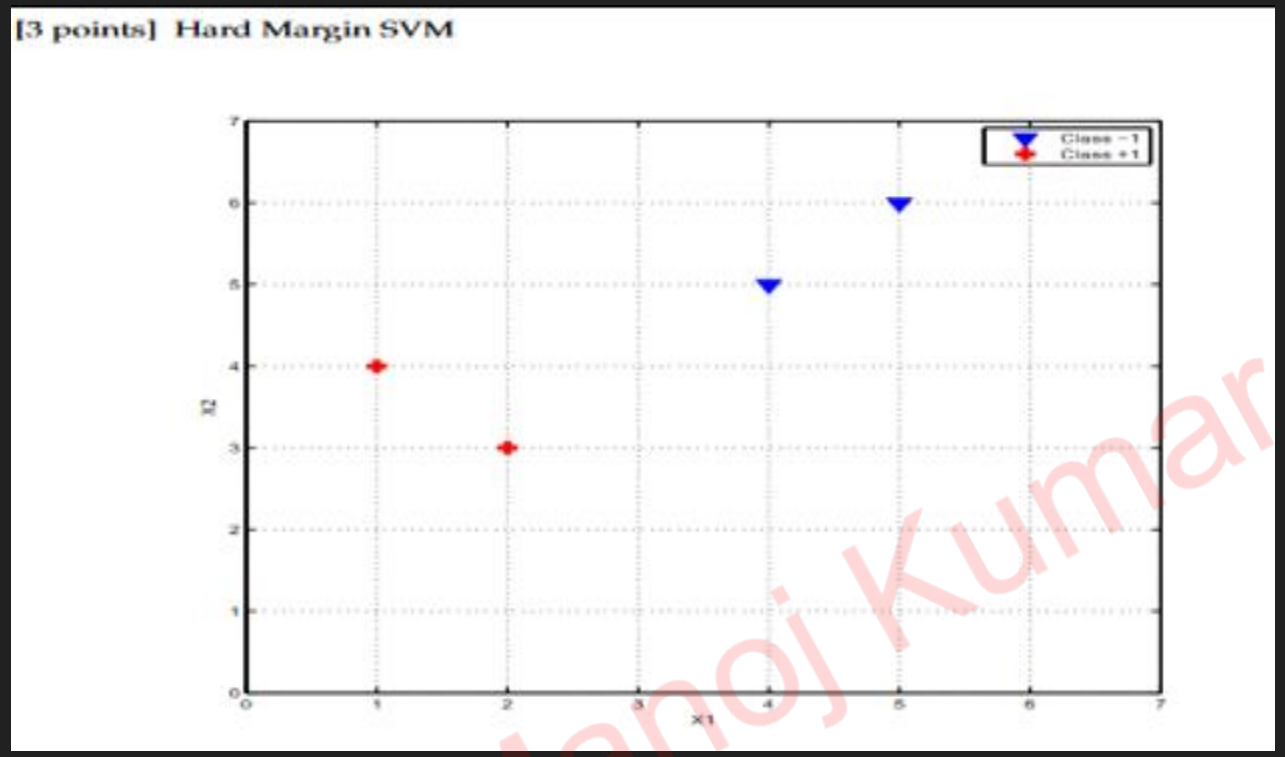
a. $\phi(x) = [(\phi_1(x) + \phi_2(x))^T , (\phi_2(x))^T]^T$

b. $\phi(x) = [\phi_1(x)^T , \phi_2(x)^T]^T$

c. $\phi(x) = \phi_1(x) + \phi_2(x)$

d. No such feature mapping exists in general.

Question 11



You are training an SVM on a tiny dataset with 4 points. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label +1 (denoted with triangles).

The points are:

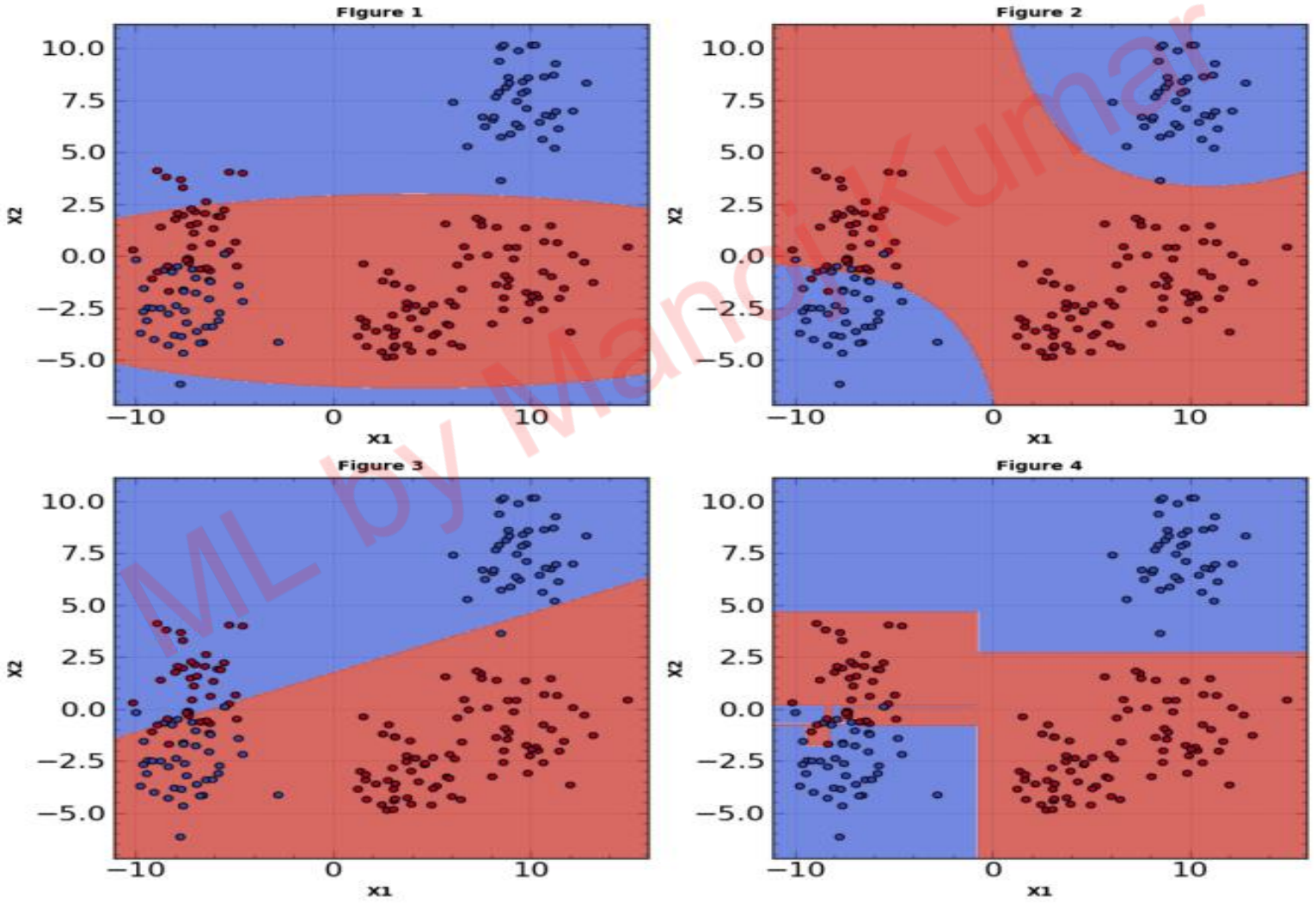
- Class -1: (1, 4), (2, 3)
- Class +1: (4, 5), (5, 6)

What is the equation corresponding to the decision boundary?

Question 12

Question 6. Match the following machine learning algorithms with their corresponding decision boundaries shown in the image below. Write the correct Figure Number (1, 2, 3, 4) in the space provided next to each algorithm.

- Logistic Regression [_____]
- Naive Bayes [_____]
- Support Vector Machine (SVM) [_____]
- Decision Tree [_____]



Question 13,14

Consider a dataset with 500 samples and 50 features. You perform 5-fold cross-validation. In this process, how many times is the error computed (N_1), what is the size of the data used for building each model (N_2), and what is the size of the data used for testing each model (N_3)?

a) $N_1 = 5$, $N_2 = 400$, $N_3 = 100$
b) $N_1 = 5$, $N_2 = 100$, $N_3 = 400$
c) $N_1 = 50$, $N_2 = 450$, $N_3 = 50$
d) $N_1 = 500$, $N_2 = 499$, $N_3 = 1$

Which of the following statements is true regarding LOOCV and k-fold cross-validation?

a) LOOCV has a lower computational cost than k-fold cross-validation.
b) LOOCV tends to have higher variance in its error estimates compared to k-fold cross-validation.
c) k-fold cross-validation always provides a better model than LOOCV.
d) LOOCV and k-fold cross-validation always produce the same test error.

Question 15,16

When performing k-fold cross-validation, what fraction of the dataset is used for training in each fold?

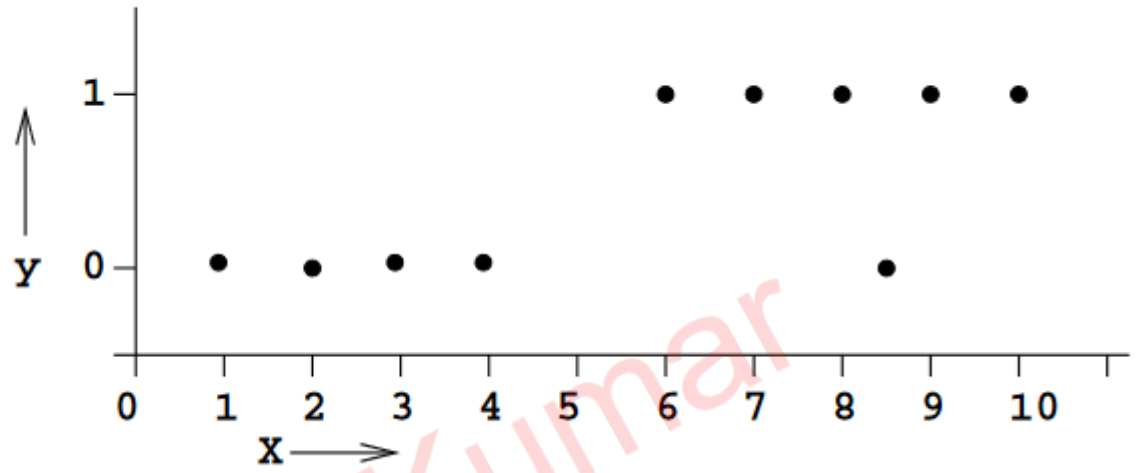
a) 1/k
b) (k-1)/k
c) k/(k-1)
d) 1 - (1/k)

In k-fold cross-validation, if the number of folds (k) is chosen to be very large (close to the number of samples), what is the likely impact on bias and variance?

a) Bias increases and variance decreases.
b) Bias decreases and variance increases.
c) Both bias and variance increase.
d) Both bias and variance decrease.

Question 17,18,19,20

X	Y
1	0
2	0
3	0
4	0
6	1
7	1
8	1
8.5	0
9	1
10	1



Suppose we are learning a classifier with binary output values Y=0 and Y=1. There is one real-valued input X.

Assume we will learn a decision tree on this data. Assume that when the decision tree splits on the real-valued attribute X, it puts the split threshold halfway between the attributes that surround the split. For example, using information gain as the splitting criterion, the decision tree would initially choose to split at X=5, which is halfway between the X=4 and X=6 data points.

Let Algorithm DT2 be the method of learning a decision tree with only two leaf nodes (i.e., only one split).

Let Algorithm DT* be the method of learning a decision tree fully with no pruning.

Question 17,18,19,20

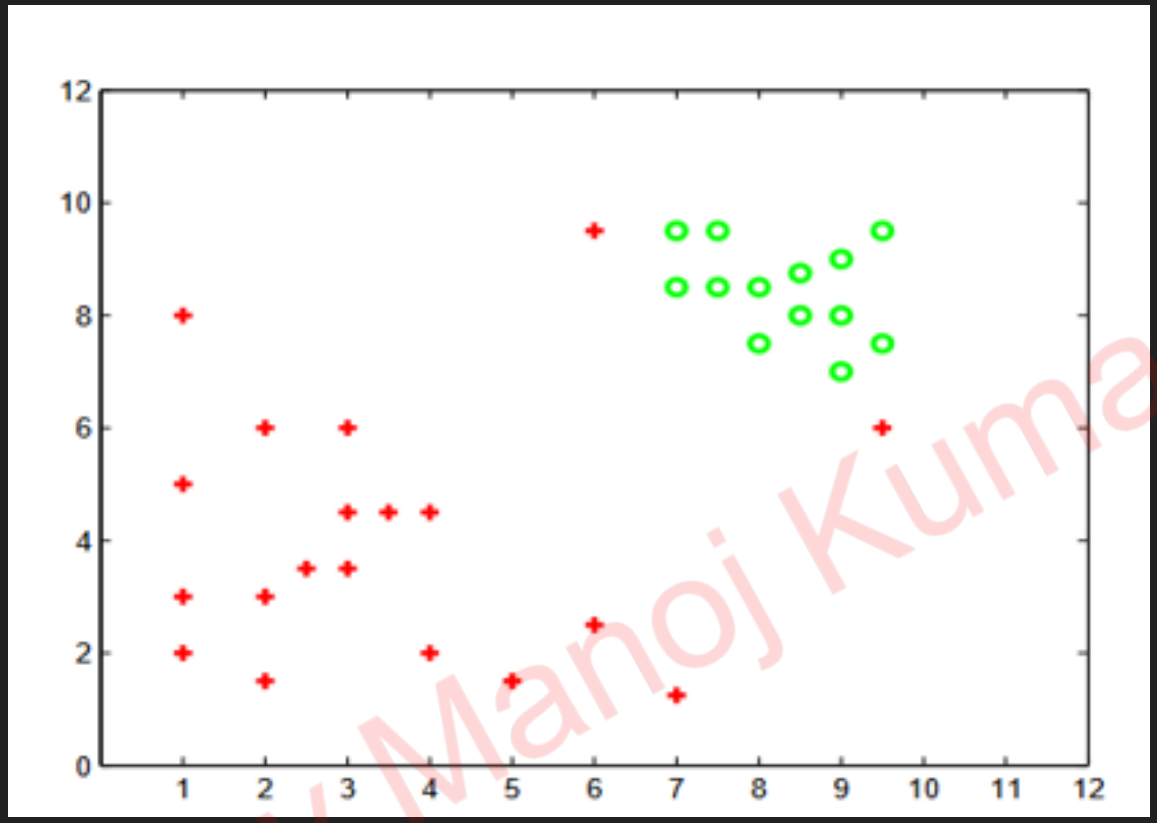
What will be the training set error of DT2 on our data? In this part, and all future parts, you can express your answer as the number of misclassifications out of 10.

What will be the leave-one-out-cross-validation error of DT2 on our data?

*What will be the training set error of DT on our data?**

*What will be the leave-one-out-cross-validation error of DT on our data?**

Question 21[MSQ]



Given a training data set shown in Figure and using an SVM with a quadratic kernel (polynomial kernel of degree 2), consider the following statements about the scenario when C tends to infinity:

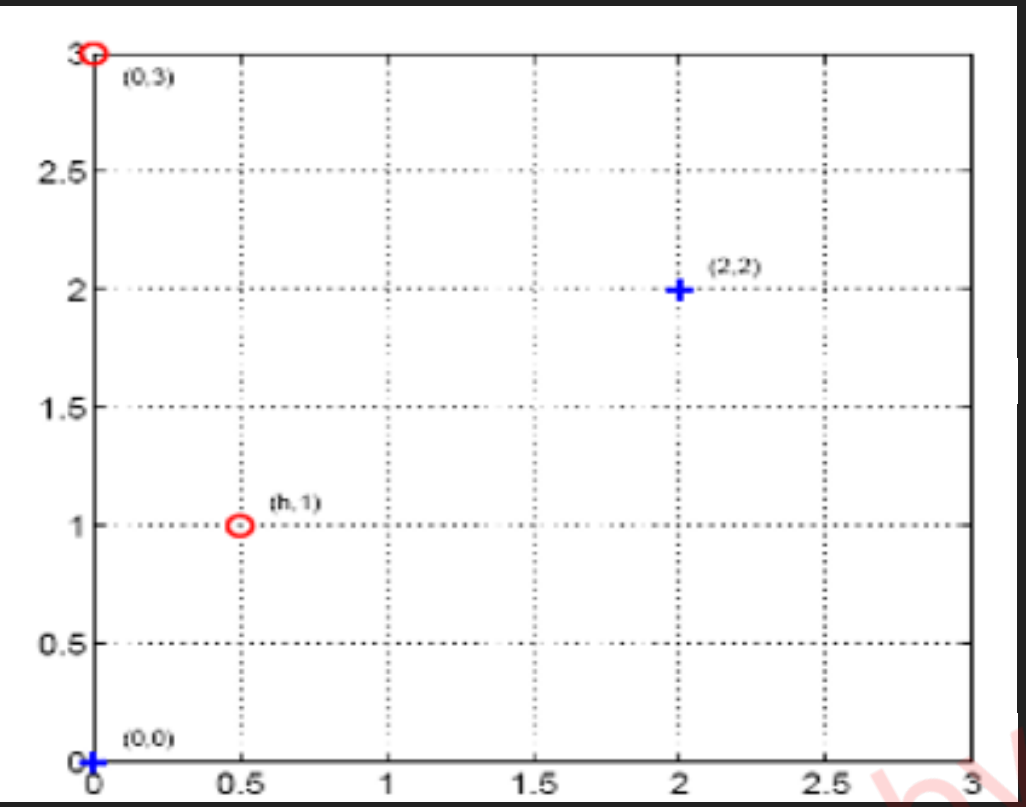
a) Adding a circle point (green point) at (9,8) will change the decision boundary

b) Adding a circle point (green point) at (6,6) will change the decision boundary

c) The decision boundary will prioritize minimizing misclassifications over maximizing the margin

d) Adding a point anywhere in this plane will change the decision boundary

Question 22,23,24



Given the scenario with four training examples in two dimensions as shown in the figure:

- Positive examples at $x_1 = [0, 0]$ and $x_2 = [2, 2]$
- Negative examples at $x_3 = [h, 1]$ and $x_4 = [0, 3]$
- We treat $0 \leq h \leq 3$ as a parameter.

How large can h be so that the training points are still linearly separable?

Does the orientation of the maximum margin decision boundary change as a function of h when the points are separable? (Yes/No)

Assume that we can only observe the second component of the input vectors. Without the other component, the labeled training points reduce to (0, y = 1), (2, y = 1), (1, y = -1), and (3, y = -1). What is the lowest order p of the polynomial kernel that would allow us to correctly classify these points?

Question 25

Consider a supervised learning problem in which the training examples are points in a 2-dimensional space. The positive examples are $(1, 1)$ and $(-1, -1)$. The negative examples are $(1, -1)$ and $(-1, 1)$.

1. Are the positive examples linearly separable from the negative examples in the original space?

   A. Yes

   B. No

2. Consider the feature transformation $\phi(x) = [1, x_1, x_2, x_1 x_2]$, where x_1 and x_2 are, respectively, the first and second coordinates of a generic example x. The prediction function is $y(x) = w^T \phi(x)$ in this feature space. Given the positive examples $(1, 1)$ and $(-1, -1)$ and negative examples $(1, -1)$ and $(-1, 1)$, which of the following weight vectors w defines a maximum-margin decision surface separating the positive examples from the negative examples?

   A. $w = [1, 0, 0, 0]$

   B. $w = [0, 0, 1, 0]$

   C. $w = [0, 0, 0, 1]$

   D. $w = [0, 1, 0, 0]$

Question 26

5 SVM - 12 points

Recall that the soft-margin primal SVM problem is

$$\begin{aligned} \min \quad & \tfrac{1}{2} w^T w + C \textstyle\sum_{i=1}^{n} \xi_i \\ s.t. \quad & \xi_i \geq 0, \forall i \in \{1, \cdots, n\} \\ & (w^T x_i + b) y_i \geq 1 - \xi_i, \forall i \in \{1, \cdots, n\} \end{aligned} \tag{1}$$

We can get the kernel SVM by taking the dual of the primal problem and then replace the product of $x_i^T x_j$ by $k(x_i, x_j)$ where $k(\cdot, \cdot)$ is the kernel function.

Figure 1 plots SVM decision boundaries resulting from using different kernels and/or different slack penalties. In Figure 1, there are two classes of training data, with labels $y_i \in \{-1, 1\}$, represented by circles and squares respectively. The SOLID circles and squares represent the support vectors. Label each plot in Figure 1 with the letter of the optimization problem below. You are NOT required to explain the reasons.

a) A soft-margin linear SVM with $C = 0.1$.

b) A soft-margin linear SVM with $C = 10$.

c) A hard-margin kernel SVM with $K(u, v) = u^T v + (u^T v)^2$.

d) A hard-margin kernel SVM with $K(u, v) = \exp(-\frac{1}{4}\|u - v\|_2^2)$.

e) A hard-margin kernel SVM with $K(u, v) = \exp(-4\|u - v\|_2^2)$.

f) None of the above.

Question 26

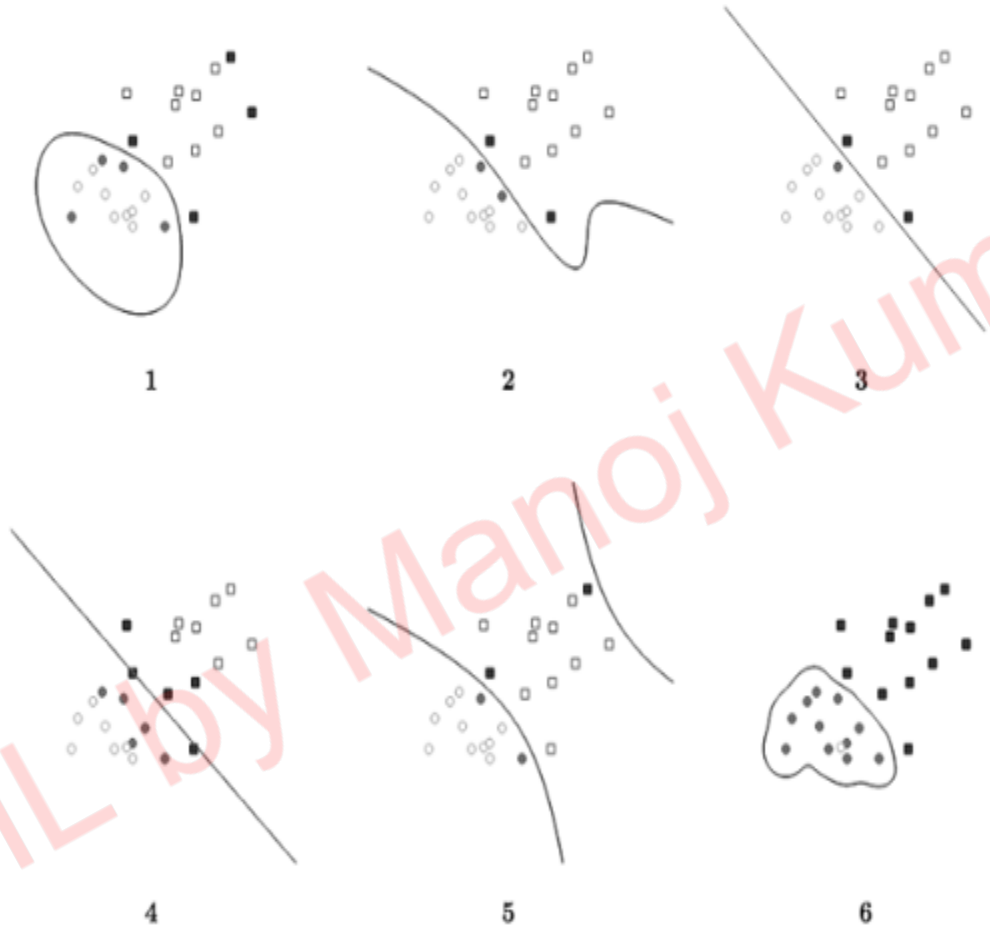


Figure 1: Induced Decision Boundaries

Question 27

Consider the 1-dimensional data set shown above, based on the single real-valued attribute x. Notice there are two classes (values of Y), and five data points.

We will learn Decision Trees from this data using the ID3 algorithm. Given real-valued attributes, ID3 considers splitting the possible values of the attribute into two sets based on a threshold (i.e., given a real valued attribute such as x, ID3 considers tests of the form $x > t$ where t is some threshold value). It considers alternative thresholds (data splits), and it selects among these using the same information gain heuristic that it uses for other discrete-valued attributes.

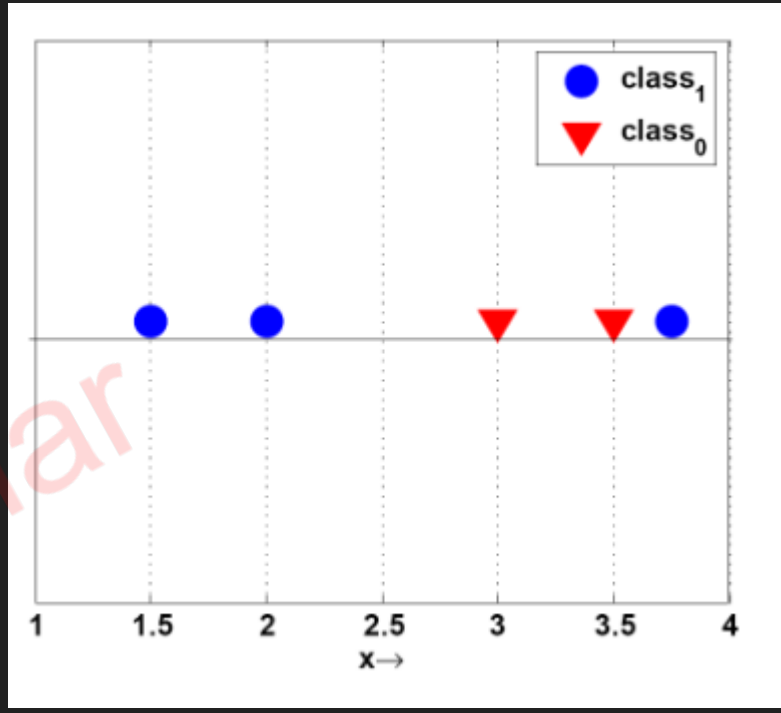
Given n training examples, ID3 considers exactly $n + 1$ possible values for the threshold. In particular, for each pair of adjacent training examples of x, it considers the threshold value midway between the two examples. In addition, it considers the threshold just to the right of the largest training example, and just to the left of the lowest training value.

Important: Assume that if your Decision Tree learner finds that two data splits $x < t_1$ and $x < t_2$ have the same information gain, then it breaks ties by picking the leftmost threshold. (e.g., if $t_1 < t_2$ then it picks the split $x < t_1$).

Question 27

Let algorithm $DT1$ be the algorithm that learns a decision tree with only one boolean split (a depth 1 tree). Let DT^* be the algorithm that learns a decision tree with as many boolean splits as necessary to perfectly classify the training data (using a different threshold for each tree node).

Note: The training error is equal to the number of misclassifications on the training data. The leave-one-out cross-validation (LOOCV) error is the total number of misclassifications at each step.



A. What is the training set error of $DT1$ on the data? What is its leave-one-out cross-validation error?

B. What is the training set error of DT^* on the data? What is its leave-one-out cross-validation error?

Question 28

Suppose you have 100 datapoints $\{(x^{(k)}, y^{(k)})\}_{k=1}^{100}$. Your dataset has one input and one output. The kth datapoint is generated as follows:

$$x^{(k)} = \frac{k}{100}$$

$$y^{(k)} \sim \text{Bernoulli}(p)$$

(A random variable with a Bernoulli distribution with parameter p equals 1 with probability p and 0 with probability $1 - p$.) Note that all of the $y^{(k)}$'s are just noise, drawn independently of all other $y^{(k)}$'s. You will consider an algorithm that always predicts zero.

What is the mean squared leave-one-out cross-validation (LOOCV) error for this algorithm?

Question 29,30,31

The following dataset will be used to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, color, and odor.

Shape	Color	Odor	Edible
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
C	W	3	No
D	W	3	No

What is the entropy $H(\text{Edible} \mid \text{Odor} = 1 \text{ or } \text{Odor} = 3)$?

Which attribute would the ID3 algorithm choose to use for the root of the tree (no pruning)?

Question 29,30,31

Suppose we have a validation set as follows:

Shape	Color	Odor	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes

What will be the training set error and validation set error of the tree? Express your answer as the number of examples that would be misclassified.

Question 32

Suppose we have three binary attributes A, B, and C and four training examples. We are interested in finding a minimum-depth decision tree consistent with the training data. The target concept is $A \text{ XOR } B$. Given the training data below, identify the target concept and the tree learned by ID3 (no pruning).

Training Data:

A	B	C	Class
1	1	0	0
1	0	1	1
0	1	1	1
0	0	1	0

Question 32

A. The target concept $A \text{ XOR } B$ results in the minimum-depth tree with 4 leaf nodes.

B. The ID3 algorithm correctly identifies the minimum-depth tree for the given training data.

C. The ID3 algorithm produces a decision tree with C as the root node, which is not the minimum-depth tree.

D. The minimum-depth tree and the tree learned by ID3 have the same structure.

Question 33

The following dataset will be used to learn a decision tree for predicting whether a person is happy (H) or sad (S) based on the color of their shoes, whether they wear a wig and the number of ears they have.

Color	Wig	Num. Ears	(Output) Emotion
G	Y	2	S
G	N	2	S
G	N	2	S
B	N	2	S
B	N	2	H
R	N	2	H
R	N	2	H
R	N	2	H
R	Y	3	H

a) What is the entropy H(Emotion | Wig = Y) in the dataset?

b) What is the entropy H(Emotion | Num. Ears = 3) in the dataset?

c) Which attribute is the decision-tree-building algorithm most likely to choose for the root of the tree, assuming no pruning?

Question 33

(d) Draw the full decision tree that would be learned for this data (assume no pruning).

(e) What would be the training set error for this dataset? Express your answer as the percentage of records that would be misclassified.

(f) Assuming that the output attribute can take two values (i.e., has arity 2), what is the maximum training set error (expressed as a percentage) that any dataset could possibly have?

Question 34,35,36,37

Consider a logistic regression model with two-dimensional input $\mathbf{x} = (x_1, x_2)$ and L2 regularization on the weights w_1 and w_2, where the regularization parameters λ_1 and λ_2 can be different. The objective function is given by:

$$F(\mathbf{w}, w_0) = \sum_i L(\mathbf{x}^i, y^i, \mathbf{w}, w_0) + \lambda_1 w_1^2 + \lambda_2 w_2^2$$

where $L(\mathbf{x}^i, y^i, \mathbf{w}, w_0)$ is the logistic loss function for example $(\mathbf{x}^i, y^i)$

Now suppose λ_1 and λ_2 are both 0. In the provided dataset graph, draw the decision boundary learned by logistic regression. (Note: for all these problems, your solution need not be exact. We are just looking for the correct points to be separated.)

Now suppose λ_1 is set to 0, but λ_2 is a very, very large value. Draw the resulting decision boundary.

Similarly, suppose λ_2 is set to 0, but now λ_1 is a very, very large value. Draw the resulting decision boundary.

Question 34,35,36,37

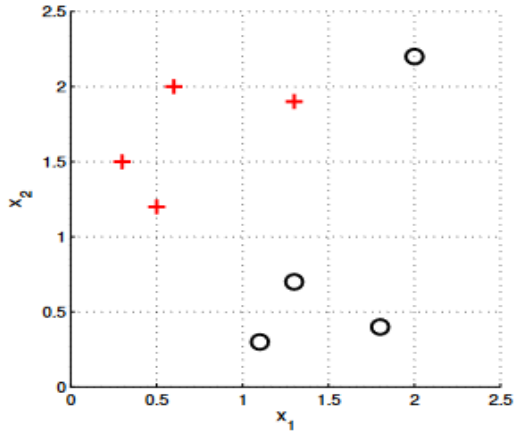
Suppose that we are given additional prior knowledge about the weights $\mathbf{w}$. In addition to being small, we also believe $\mathbf{w}$ should be close to some given parameters $\tilde{\mathbf{w}}$ and $\tilde{w}_0$. In this case, these weights are

$$\tilde{w}_0 = 0, \tilde{w}_1 = -1, \tilde{w}_2 = 1$$

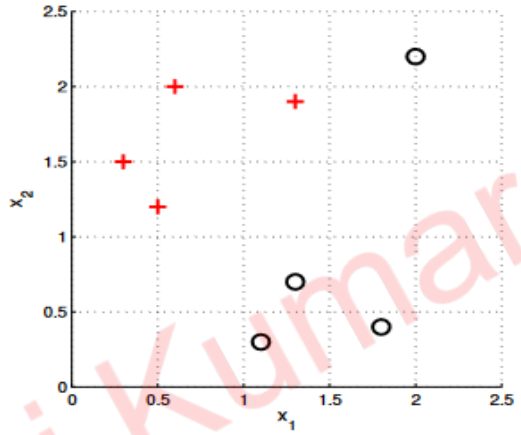
To ensure our estimate stays close to $\tilde{\mathbf{w}}$, we minimize the modified objective function $\tilde{F}(\mathbf{w}, w_0)$ with an added regularization term:

$$\tilde{F}(\mathbf{w}, w_0) = \sum_i L(x^i, y^i, \mathbf{w}, w_0) + \lambda_1 w_1^2 + \lambda_2 w_2^2 + \tilde{\lambda}(\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + (w_0 - \tilde{w}_0)^2)$$

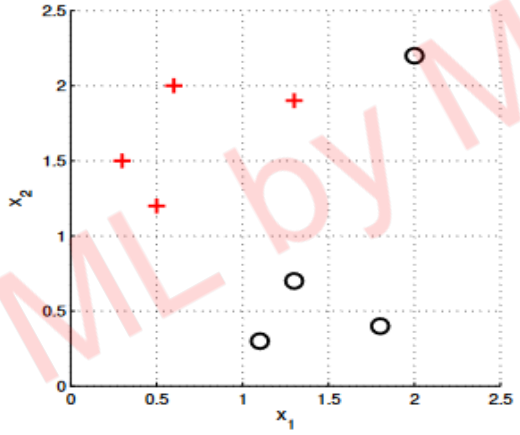
Assume the regularization parameters are chosen such that λ_1 and λ_2 are both quite small, but $\tilde{\lambda}$ is very, very large. Draw the resulting decision boundary.



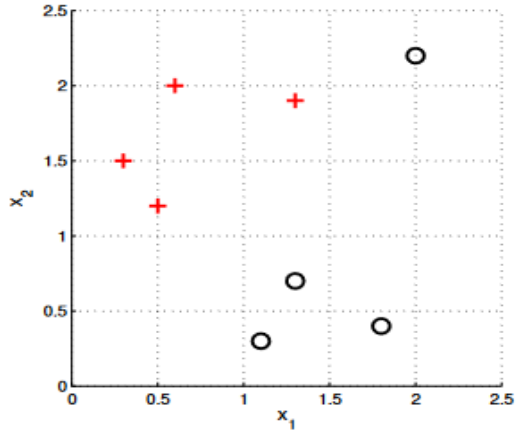
(a) Decision boundary for small λ_1 and λ_2



(b) Decision boundary for $\lambda_1 = 0$ and large λ_2



(c) Decision boundary for large λ_1 and $\lambda_2 = 0$



(d) Decision boundary for small λ_1 and λ_2 but large $\tilde{\lambda}$