Classification

	0.17	0.35	0.5	0.67	0.83	1
TP	100	100	86	27	1	0
FP	100	76	1	0	0	0
TN	0	24	99	100	100	100
FN	0	0	14	73	99	100
TPR	1	1	86/100	27/100	1/100	0
FPR	1	76/100	1/100	0	0	0

1.5 points, -15% penalty for a wrong answer

The true positive rate (TPR) and false positive rate (FPR) are defined as:

$$TPR = \frac{TP}{TP+FN}$$

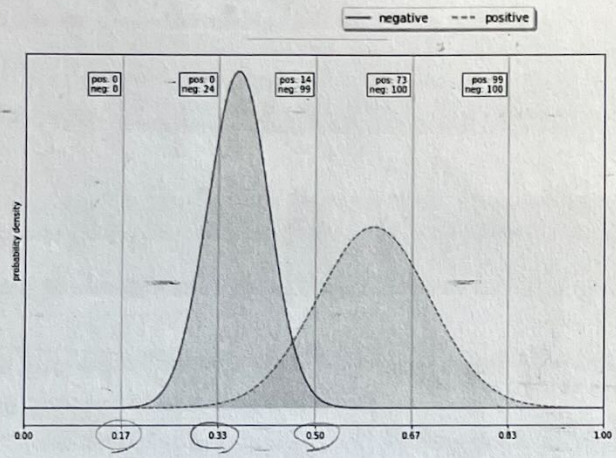
$$FPR = \frac{FP}{FP+TN}$$

Where TP = true positive, FP = false positive, TN = true negative, FN = false negative

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the TPR (y axis) against the FPR (x axis) at various threshold settings.

A binary classifier (output: positive, negative) produces a posterior probability x = P(class=positive| t) for any test instance t. The class label is obtained based on a discrimination threshold applied to *x* . Values above the threshold are predicted *positive* , values below the threshold are predicted *negative* .

The following plot shows how the posterior probabilities are distributed within a dataset comprised of 200 records (100 positive records, 100 negative records).



- (a) (1/2, 1/2), (100/176, 76/176)
- (b) (1, 0), (76/100, 0)
- (c) (100/101, 1), (100/127, 1)
- (d) (86/100, 1/100), (27/100, 0)
- (e) The information provided is not sufficient to answer the question
- (f) (1, 0), (0, 73/127)

TP = 100 - Pos
FP = 100 - Neg
TN = N
FN = P

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

In particular, the 5 boxes in the upper part of the figure show the cumulative number of records that belong to the positive (pos) and negative (neg) classes, from to the $-\infty$ threshold defined by the respective vertical line.

For example, there are 14 positive records and 99 negative records that have $x \leq 0.5$.

Consider the ROC curve for this classifier. Which of the following pairs of points belong to the curve? (each of the points is provided as a pair of (x, y) coordinates).

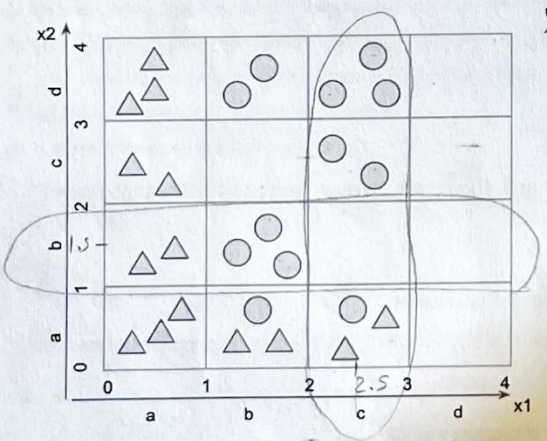
Let X be a feature vector, $X \in \mathbf{R}^N$, ie X = (x1, x2, ..., xn).

Naive Bayes classifiers compute the output class based on the following definition:

P(class|X)=P(x1|class)*P(x2|class)*...*P(xn|class)*P(class) It is

given the dataset represented in figure, where x1 and x2 are two continuous features, while triangle and circle are two class labels.

The two features are discretized in four bins [a,b,c,d], depicted in the figure with a grid.



1) $P(\Delta|X) = P(x_1|\Delta) \cdot P(x_2|\Delta) \cdot P(\Delta)$

$= \frac{2}{14} \cdot \frac{2}{14} \cdot \frac{14}{26} = \frac{1}{91}$

2) $P(O|X) = P(x_1|O)\,P(x_2|O)\,P(O)$

$= \frac{6}{12} \times \frac{3}{12} \times \frac{12}{26} = \frac{3}{52}$

3)

Suppose that a Naive Bayes classifier has been trained on the provided dataset, which contains 14 triangles and 12 circles.

Classify the test data sample X with features x1=2.5, x2=1.5

Write in the box the following values (**one answer per row**):

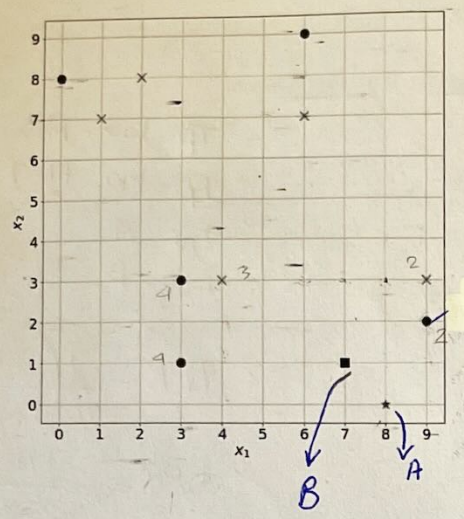
1. P(triangle|X) = ?
2. P(circle|X) = ?
3. The class assigned to P (circle or triangle?)

The KNN algorithm can assign a class probability based on the votes assigned to each class. If $vote_Y(x)$ represents the unnormalized vote assigned to class Y for sample x, then, for any class $c \in C$: $p(c|x) = \frac{vote_c(x)}{\sum_{i \in C} vote_i(x)}$

The Chebyshev distance (L_∞ norm) is the limit case for $r \to \infty$ of the Minkowski distance, as shown below for two points $x, y \in \mathbb{R}^n$:

$lim_{r\to\infty}(\sum_{i=1}^{n}|x_i - y_i|^r)^{\frac{1}{r}} = max_{i=1}^{n}|x_i - y_i|$

The figure below shows a 2-dimensional dataset. Points in this dataset belong to either of two classes: circles (in blue) or crosses (in red). You are given two test points to be labelled, A (green star) and B (black square).



The K-NN algorithm (with Chebyshev distance and K = 3) is used to label A and B.

Answer the following questions.

Question 1) What is the label assigned to A and B, when neighbors' votes are weighted uniformely?

Question 2) What are the class probabilities for A and B, if the vote cast by each neighbor is $w = \frac{1}{1+distance}$?

Use the following notation:

```
A1) A=class_for_A, B=class_for_B
A2)
p(x|A)=prob_cross_A, p(o|A)=prob_circle_A
p(x|B)=prob_cross_B, p(o|B)=prob_circle_B
```

For example

```
A1) A=circle, B=circle
A2)
p(x|A)=0.5, p(o|A)=0.5
p(x|B)=0.9, p(o|B)=0.1
```

A binary classifier is trained to separate between images of cats and dogs. The test set used to evaluate this model is balanced, with 10,000 images of dogs and 10,000 images of cats.

The classifier only predicts 50 images as being cats. All of those predictions are correct.

What can be said about such a classifier?

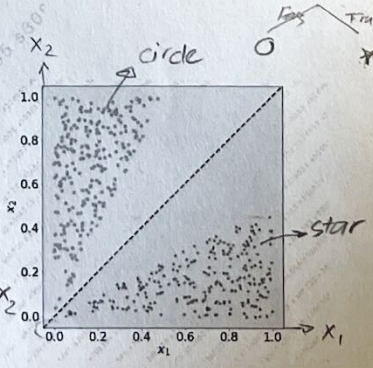
- ○ (a) None of the other answers is correct
- ○ (b) It has low precision for the class "cat"
- ○ (c) It has high recall for the class "dog"
- ○ (d) It has high F1 score for the class "dog"
- ○ (e) It has high recall for the class "cat"
- ○ (f) It has high accuracy
- ○ (g) It has high precision for the class "dog"
- ○ (h) It has high F1 score for the class "cat"

The following image represents a two-dimensional dataset with two numerical features, x1 and x2, and two classes, blue (star) and orange (circle).

A classifier is trained to distinguish between the blue (star) class and the orange (circle) one.

The dashed line represents the decision boundary identified by the classifier to separate between the two classes.

Which of the following statements is correct?



- ● (a) A decision tree can generate the boundary represented in the figure, when trained with the feature x1 / x2
- ○ (b) A decision tree can generate the boundary represented in the figure, when trained with the feature x1
- ○ (c) A decision tree can generate the boundary represented in the figure, when trained with the feature x1 * x2
- ○ (d) A decision tree can generate the boundary represented in the figure, when trained with the features x1 and x2
- ○ (e) A decision tree can generate the boundary represented in the figure, when trained with the feature x2
- ○ (f) A decision tree can generate the boundary represented in the figure, when trained with the feature x1

A classification model can be tuned on P hyperparameters. You identify a subset of p < P hyperparameters that require tuning. For each hyperparameter, you identify n candidate values.

For n = 10, p = 5, P = 20, which of the following statements is correct regarding the number of configurations to be assessed?

- (a) Introducing an additional value for each hyperparameter introduces fewer new configurations to be assessed than if introducing an additional hyperparameter (with n candidate values)
- (b) Introducing an additional hyperparameter (with n candidate values) does not increase the number of configurations to be assessed
- (c) Introducing an additional candidate value for each hyperparameter does not increase the number of configurations to be assessed
- (d) The number of candidate values n cannot exceed P
- (e) Introducing an additional hyperparameter (with n candidate values) introduces fewer new configurations to be assessed than if introducing an additional value for each hyperparameter
- (f) It is advisable to always choose p = P and the largest possible value for n

Consider the training process of a Random Forest classifier on a dataset X characterized by a set of features F.

What is used to train *each individual tree* of the forest?

- (a) All the records from X and all the features F. However, each tree is initialized to a different random state
- (b) None of the other answers is correct
- (c) All the records from X and a subset of features from F. However, only a subset of X is considered at each split
- (d) A sample of records from X and a sample of features from F. The samples of records and features can overlap among the trees
- (e) A sample of records from X and a sample of features from F. The samples of records and features do not overlap among the trees
- (f) A sample of records from X and a sample of features from F. The samples of records can overlap among the trees but the samples of features do not

A classifier is incremental if, when adding a single new point in the training set, the training algorithm analyzes the new data point and does not require the availability of the training set.

Inference time is the time required to classify a new test object.

Which of the following statements on the Naive Bayes classification algorithm is correct?

- (a) The inference time of Naive Bayes is proportional to the training set size
- (b) Naive Bayes requires storing the entire training set to make new inferences
- (c) Naive Bayes cannot be used with continuous attributes
- (d) The Naive hypothesis is used to compute the prior probability (i.e. p(C) for a class C)
- (e) Naive Bayes is not an incremental algorithm
- (f) None of the other answers is correct

Consider the training process of a Random Forest classifier on a dataset X characterized by a set of features F.

What is used to train *each individual tree* of the forest?

- (a) A sample of records from X and a sample of features from F. The samples of records and features do not overlap among the trees
- (b) All the records from X and a subset of features from F. However, only a subset of X is considered at each split
- (c) None of the other answers is correct
- (d) All the records from X and all the features F. However, each tree is initialized to a different random state
- (e) A sample of records from X and a sample of features from F. The samples of records and features can overlap among the trees
- (f) A sample of records from X and a sample of features from F. The samples of records can overlap among the trees but the samples of features do not

A classifier is incremental if, when adding a single new point in the training set, the training algorithm analyzes the new data point and does not require the availability of the training set.

Inference time is the time required to classify a new test object.

Which of the following statements on the Naive Bayes classification algorithm is correct?

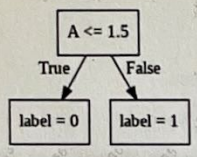
- ○ (a) Naive Bayes is not an incremental algorithm
- ○ (b) Naive Bayes requires storing the entire training set to make new inferences
- ○ (c) Naive Bayes cannot be used with continuous attributes
- ● (d) None of the other answers is correct
- ○ (e) The inference time of Naive Bayes is proportional to the training set size
- ○ (f) The Naive hypothesis is used to compute the prior probability (i.e. p(C) for a class C)

You are training a decision tree classifier and need to evaluate its performance. To this end, you use 3-fold cross-validation. The following are the decision trees obtained during the validation – the fold used for testing is shown with each tree.

Fold 1

A	B	label
0	0	9	1
1	6	2	1
2	0	2	0
3	2	5	0

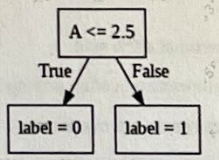
Decision tree 1



Fold 2

A	B	label
0	9	6	1
1	9	4	1
2	0	9	0

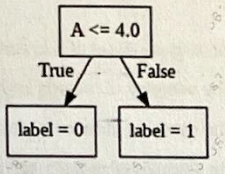
Decision tree 2



Fold 3

A	B	label
0	8	4	1
1	3	8	1
2	0	5	0

Decision tree 3



1. Compute the accuracy for each fold. Use the syntax:
   fold1: <accuracy>
   fold2: <accuracy>
   fold3: <accuracy>
2. Compute the overall accuracy. Use the syntax: *overall: <accuracy>*

Example:

fold1: 0.25
fold2: 0.3
fold3: 0.35
overall: 0.3

A random forest is trained on a problem with C classes (0, 1, ... C-1), using K decision trees.

For each input point x, each decision tree returns a class probability for each of the C classes.

The random forest assigns a class label to each point using soft voting.

Soft voting works as follows: given a point x, each of the K decision trees produces a class probability for each of the C classes. For each class, the respective K probabilities are then added together. Finally, the random forest assigns to x the label of the class with the highest sum.

You are given a test set X containing N points.

The following syntax will be used:

- $y_i \in \{0, 1, \ldots, C-1\}^N$ represents the vector of labels predicted for all N instances of X by the i-th decision tree

When applying this random forest to X, we notice that $y_1 = y_2 = y_3 = \ldots = y_k$.

Which of the following statements is true? (Multiple answers may be true)

Scegli una o più alternative:

- ☐ (a) There is no benefit in having an ensemble, since all predictions by the random forest will be the same obtained with a single decision tree
- ☐ (b) All K models have learned the same tree even though they have been trained on different subsets of the training set, hence the same predictions
- ☐ (c) The optimal number of features used at each split is of $\sqrt{C}$
- ■ (d) For a different test set Z, different decision trees in the random forest may yield different predictions
- ☐ (e) The optimal number of features used at each split is of $\sqrt{N}$
- ☐ (f) All K decision trees have been trained on the same subsets of data, hence the same predictions

You are building a K-NN classifier A on a data set D split into a training set $D_{A,tr}$ and a test set $D_{A,te}$, such that $D = D_{A,tr} \cup D_{A,te}$.

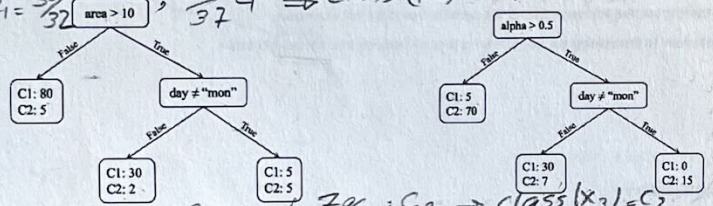
Suppose you build a second K-NN classifier B using the same hyper-parameters and the same dataset D, split into a training set $D_{B,tr}$ and a test set $D_{B,te}$, such that $D_{A,tr} \cup D_{A,te} = D_{B,tr} \cup D_{B,te}$.

Which of the following statements is true?

- ○ (a) None of the other statements is correct
- ○ (b) A and B will have the same performance on the respective test sets since they are trained using the same hyperparameters
- ○ (c) We can obtain a meaningful evaluation of A by testing it on $D_{B,te}$
- ● (d) If $D_{A,tr} = D_{B,tr}$, A and B will make the same predictions on the respective test sets
- ○ (e) A and B will have the same performance on the respective test sets since they are evaluated on the same test sets
- ○ (f) If we choose the same K for both classifier A and B, they will make the same predictions on the respective test sets

Given a point x to classify by means of a random forest, each tree of the ensemble assigns to point x a probability for each class. For each tree, the probability for a class C is given by the fraction of the training set points at the leaf reached by point x that belong to class C. The random forest assigns to point x the class with the highest average probability across all trees.

Consider the Random Forest shown in the figure below, which is composed of two trees. The represented trees show the split conditions and, for each leaf, the number of training samples that belong to each class (i.e., after applying Hunt's algorithm).



	Importance	alpha	area	day
x1	high	0.7	5	wed
x2	low	0.8	100	mon
x3	low	0.4	50	fri
x4	high	0.6	20	wed

Consider the training process of a Random Forest classifier on a dataset X characterized by a set of features F.

What is used to train *each individual tree* of the forest?

- ○ (a) A sample of records from X and a sample of features from F. The samples of records and features can overlap among the trees
- ○ (b) None of the other answers is correct
- ○ (c) All the records from X and all the features F. However, each tree is initialized to a different random state
- ○ (d) A sample of records from X and a sample of features from F. The samples of records and features do not overlap among the trees
- ○ (e) A sample of records from X and a sample of features from F. The samples of records can overlap among the threes but the samples of features do not
- ● (f) A sample of records from X and all the features F. However, only a subset of F is considered at each split

A KNN classifier is trained on the training set in the table below, which only contains categorical features.

	width	weight	speed	Class
x1	big	light	fast	A
x2	big	heavy	fast	B
x3	small	heavy	slow	B
x4	big	heavy	slow	A
x5	small	light	slow	A

With categorical features, the distance between two samples *a* and *b* can be computed as the number of features with different values:

$dist(a, b) = \sum_i \ell(a_i, b_i)$

Where:

$$\ell(m, n) = \begin{cases} 0 & \text{if } m = n \\ 1 & \text{otherwise} \end{cases}$$

For each neighbor x_i of x the vote is weighted as follows:

$weight(x, x_i) = \frac{1}{dist(x, x_i) + 0.5}$

Given the test point below, write in the answer box:

1. The list of the 3 neighbors of t1
2. The class assigned to t1 with K = 3

	width	weight	speed	Class
t1	big	heavy	fast	?

Use the following notation:
neighbors={list of neighbors for t1}
class=class assigned

Example:
neighbors={point1 point2 point3}
class=A

1. $n = \{x_1, x_2, x_4\}$
2. A: $0.66 + 0.66 = 1.32$; B: 2 ⇒ B

A classifier is incremental if, when adding a single new point in the training set, the training algorithm analyzes the new data point and does not require the availability of the training set.

Inference time is the time required to classify a new test object.

Which of these statements is correct?

- (a) Decision tree and Naive Bayes classifiers are not incremental. However, when adding a data point to the training set, the inference time of a decision tree does not increase.
- (b) Decision tree is incremental, while Naive Bayes is not. However, when adding a data point to the training set, the inference time of Naive Bayes does not increase.
- (c) Naive Bayes is incremental. When adding a data point to the training set, the inference time of Naive Bayes does not increase.
- (d) None of the answers is correct.
- (e) Decision tree is incremental. Moreover, when adding a data point to the training set, the inference time of the decision tree does not increase.
- (f) Naive Bayes is incremental, while Decision tree is not. When adding a data point to the training set, the inference time of Decision tree always increases.

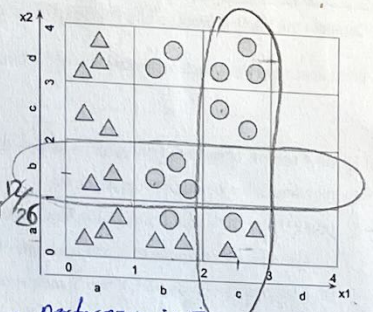
Let X be a feature vector, $X \in \mathbf{R}^N$, i.e. X = (x1, x2, ..., xn).

Naive Bayes classifiers compute the output class based on the following definition:

P(class|X)=P(x1|class)*P(x2|class)*...*P(xn|class)*P(class)

It is given the dataset represented in figure, where x1 and x2 are two continuous features, while triangle and circle are two class labels.

The two features are discretized in four bins [a,b,c,d], depicted in the figure with a grid.



Suppose that a Naive Bayes classifier has been trained on the provided dataset, which contains 14 triangles and 12 circles.

Classify the test data sample X with features x1=2.5, x2=1.5

Write in the box the following values (**one answer per row**):

1. P(triangle|X) = ?
2. P(circle|X) = ?
3. The class assigned to X (circle or triangle?)

1) $P(\Delta|X) = P(x_1|\Delta)\,P(x_2|\Delta)\,P(\Delta) = \frac{2}{14} \times \frac{2}{14} \times \frac{14}{26} = 0.01$

2) $P(\circ|X) = P(x_1|\circ)\,P(x_2|\circ)\,P(\circ) = \frac{6}{12} \times \frac{3}{12} \times \frac{12}{26} = 0.69$

You are given a dataset containing 300,000 records divided into 3 classes with the following schema:

```
Id | value_0 | ... | value_100 | Label
```

You have to build a classification model to avoid overfitting. Which is the best strategy to tune and test the classifier?

- (a)

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
kf = KFold(n_splits=4)
for train, test in kf.split(X_train):
    # Train and Validate
# Test the model
```

- (b)

```
loo = LeaveOneOut()
for train_index, test_index in loo.split(X):
    # Train and Test
```

- (c)

```
tscv = TimeSeriesSplit(n_splits=4)
for train, test in tscv.split(X):
    # Train and Test
```

- (d)

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1)
# Train and test the model
```

You are given a decision tree trained on a dataset with categorical features only.

Which statement is true?

- (a) Each internal node of a decision tree can have at most C children, where C is equal to the number of attributes of the training set
- (b) Each internal node of a decision tree can have at most C children, where C is equal to the number of distinct values of the attribute associated with the internal node
- (c) Each internal node of a decision tree can have at most C children, where C is equal to the number of training records
- (d) Each internal node of a decision tree can have at most C children, where C is equal to the number of distinct class labels

The matrix in the figure below represents the distances between data points, while the table on the right lists the ground truth labels. Points a, b, c and d represent training samples, while x_1 and x_2 are test samples.

	a	b	c	d	x1	x2
a	0	1	2	3	2	2
b	1	0	2	1	2	5
c	2	2	0	3	4	2
d	3	1	3	0	1	4
x1	2	2	4	1	0	4
x2	2	5	2	4	4	0

Question 1) If the value of K is set to 2, what is the number of distances computed by the K-Nearest- Neighbor algorithm to obtain the label for x_1?

- The number of computed distances is 2
- The number of computed distances is 4
- The number of computed distances is at most 2
- The number of computed distances is 5

sample	ground truth
a	True
b	True
c	False
d	False

Question 2) Apply a K-Nearest-Neighbor classifier and use K=3 neighbors as hyperparameter. Use a weighted voting with the following schema:

$$vote(C, x) = \frac{1}{K_c} \sum_{n_c \in C} \frac{1}{d^2(n_c, x)}$$

Where C is a class label, x is the sample being labeled, K_c is the number of neighbors belonging to class C ($Kc \leq K$), n_c is one of the K_c neighbors of x belonging to class C, $d(n_c, x)$ is the distance between n_c and x.

What is the value of $vote(True, x_2)$?

$vote(True, x_2) =$ [ ]

Question 3) Apply a K-Nearest-Neighbor classifier and use $K = 3$ neighbors as hyperparameter. Use the voting schema of the previous question and label the samples x_1 and x_2 with the class having the highest vote.

x_1: [ ]

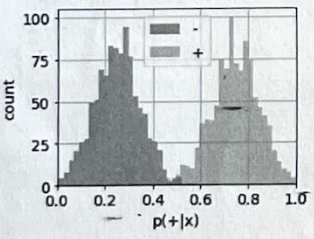
x_2: [ ]

A dataset is comprised of 2,000 points assigned to either of two labels: positive (+) or negative (-). The dataset is well-balanced, with 1,000 points belonging to the positive class, and 1,000 points belonging to the negative class.

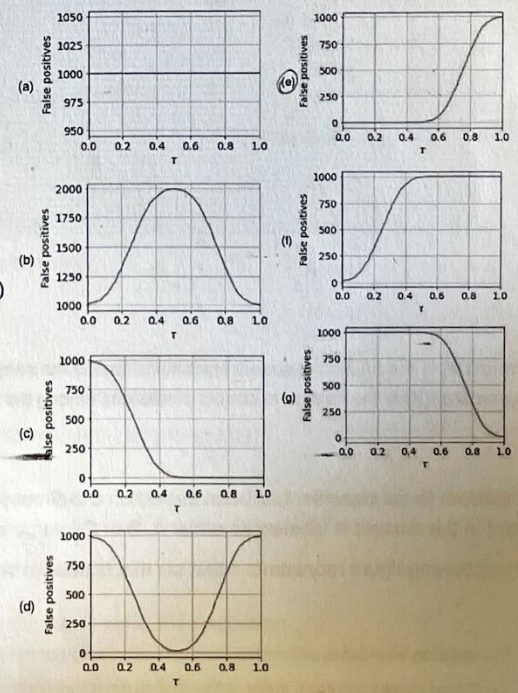
A classifier trained on this dataset produces, for a given point x, a probability of belonging to the positive class p(+|x). The probability p(-|x) can be computed as 1 - p(+|x).

A threshold value τ can be defined to assign a class to any given point x. In particular, if p(+|x) > τ the point is labelled as positive; negative otherwise.

The figure below shows how the positive and negative points are distributed w.r.t. p(+|x). In particular, the points known to have a negative ground truth are shown in the blue (left-most) distribution, whereas the points having a positive ground truth are shown in the orange (right-most) one.



Which of the following figures correctly represents the number of false positives obtained as τ varies between 0 and 1?



A random forest (RF) and a decision tree (DT) classifiers have been trained on a dataset D.

You accidentally train the random forest using only 1 decision tree.

Which of the following statements is correct? Choose all correct answers (multiple answers may be correct)

Note: when referring to "RF" in the answers, it means the specific 1-tree random forest mentioned above.

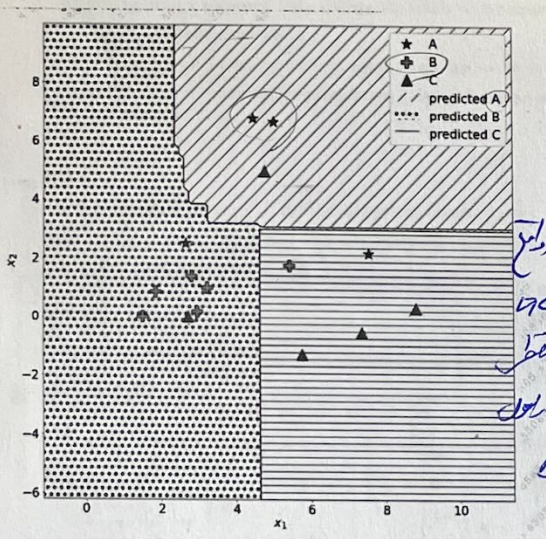
Scegli una o più alternative:

- ☐ (a) The RF and the DT will produce different results, since the voting scheme of the RF will affect its predictions
- ☐ (b) None of the other statements is correct
- ☑ (c) The RF and the DT will produce different results, since the pool of features used changes for each split of RF ✓
- ☐ (d) The RF and the DT will produce different results, since the data used for training them will differ
- ☐ (e) The RF and the DT will produce the same results
- ☐ (f) The RF and the DT will produce different results, since the pool of features used changes for each split of DT

- recall(X) is the fraction of correct predictions among the samples with actual class X
- precision(X) is the fraction of correct predictions among the samples predicted with class X

A random forest classifier has been trained on a 2-dimensional dataset (features x_1, x_2). Each point in the dataset is labeled as either A, B or C (star, cross, triangle respectively).

The following figure represents a test set that is used to validate the classifier.



The decision boundaries of the model are shown in the figure:

- Diagonal lines represent areas of the input space where the model predicts class A
- Small circles represent areas of the input space where the model predicts class B
- Horizontal lines represent areas of the input space where the model predicts class C

Write in the box below:

- accuracy(B)
- accuracy(C)
- recall(C)

Per(A) = 2/3

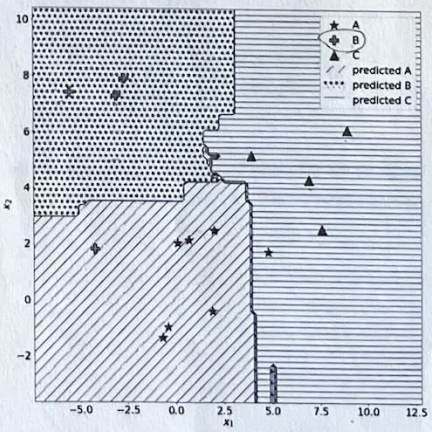
Recall(A) = 2/4 = 1/2 = 0.5

Recall(B) = 5/6

recall(X) is the fraction of correct predictions among the samples with actual class X
precision(X) is the fraction of correct predictions among the samples predicted with class X

A random forest classifier has been trained on a 2-dimensional dataset (features x_1, x_2). Each point in the dataset is labelled as either A, B or C (star, cross, triangle respectively).

The following figure represents a test set that is used to validate the classifier.



The decision boundaries of the model are shown in the figure:

- Diagonal lines represent areas of the input space where the model predicts class A
- Small circles represent areas of the input space where the model predicts class B
- Horizontal lines represent areas of the input space where the model predicts class C

Write in the box below:

precision(B)

precision(C)

recall(B)

P(B) = 3/3 = 1

P(C) = 4/5 = 0.8

R(B) = 3/4 = 0.75

Rand index is defined as follows, given a ground truth y_true and a cluster assignment y_pred.

$$RI(y\_true, y\_pred) = \frac{a+b}{\binom{n}{2}}$$

Where

- a is the number of pairs of objects with different labels in y_true and different labels in y_pred
- b is the number of pairs of objects having the same label in y_true and the same label in y_pred
- n = number of elements in y_true and y_pred

You are given an **unlabelled** dataset X (n rows, m columns) and are required to run a cluster analysis using DBSCAN. Which of the following statements is correct?

- ○ (a) The number of clusters to be extracted is a hyperparameter that can be fine-tuned based on the knee of the curve of the sum of squared errors
- ○ (b) The number of clusters extracted is independent of the hyperparameter ϵ (`eps`, maximum distance between two points for them to be considered neighbors)
- ○ (c) DBSCAN may extract at most n clusters
- ○ (d) You can use the Rand index to assess the quality of the resulting clusters
- ○ (e) DBSCAN may extract at most m clusters
- ○ (f) None of the other answers is correct

- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C
- Accuracy is the fraction of correct predictions among all samples

Let y_pred, y_true be the prediction vector and the ground truth vector respectively.

```
y_true: [A A B C B A A A B A]
y_pred: [C A C A A B A B A B]
```

The Gini index of a node is computed as follows:

$$gini(node) = 1 - \sum_j P(j|t)^2$$

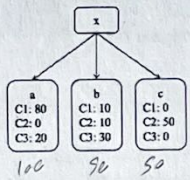
where $P(j|t)$ is the relative frequency of class j at node t

The Gini index of a *split* with parent p and children C_i is computed as follows:

$$gini(split) = \sum_i \frac{n_i}{n} gini(c_i)$$

where n_i is the number of records at child C_i and n is the number of records in p.

In the figure below it is shown a split x, with three children (a, b, c). For each child you are given the number of elements belonging to each of three classes (C1, C2, C3).



Write in the box below:
1. the Gini index of child a
2. the Gini index of the split x

Use the following notation:
gini(a)=value
gini(x)=value

Example
gini(a)=1/2
gini(x)=1/3

A binary classifier uses the following spherical decision boundary:

$$(x-3)^2 + (y-4)^2 + (z-1)^2 = r^2$$

Points strictly inside the sphere are classified with "Reject", otherwise they are classified with "Accept".

Assuming $r^2 = 15$, report the overall precision and recall for both the classes on the following test set.

x	y	z	Ground truth
2	1	3	Accept
0	2	-1	Accept
-3	1	1	Accept
1	3	-2	Reject
0	0	0	Accept

The Fowlkes-Mallows score* is a quality index designed for evaluating clustering predictions against a ground truth.

It is defined as follows:

$$FM_{score} = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}$$

Where:

- TP = number of pairs of elements that are in the same set in the ground truth and in the same set in the prediction
- FP = number of pairs of elements that are in different sets in the ground truth and in the same set in the prediction
- FN = number of pairs of elements that are in the same set in the ground truth and in different sets in the prediction

Given the vectors:

```
y_pred = [1 1 2 3 2] # prediction
y_true = [1 3 2 3 2] # ground truth
```

Write in the box below the value of Fowlkes-Mallows score computed for y_pred and y_true.

- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C
- Accuracy is the fraction of correct predictions among all samples

Let y_pred, y_true be the prediction vector and the ground truth vector respectively.

```
y_true: [C A A B C B A B A A]
y_pred: [A A C A C C C B A C]
```

Write in the answer box:

- Accuracy = 4/10
- Precision(A) = 2/4 = 1/2 = 0.5
- Recall(A) = 2/5 = 0.4

Use the following notation:

```
accuracy=value
precision(A)=value
recall(A)=value
```

Example:

```
accuracy=1
precision(A)=1/3
recall(A)=0.25
```

The true positive rate (TPR) and false postive rate (FPR) are defined as:

$TPR = \frac{TP}{TP+FN}$

$FPR = \frac{FP}{FP+TN}$

Where TP = true positive, FP = false positive, TN = true negative, FN = false negative

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the TPR (y axis) against the FPR (x axis) at various threshold settings.

A binary classifier (output: positive, negative) produces a posterior probability x = P(class=positive | t) for any test instance t. The class label is obtained based on a discrimination threshold applied to *x*. Values above the threshold are predicted *positive*, values below the threshold are predicted *negative*.

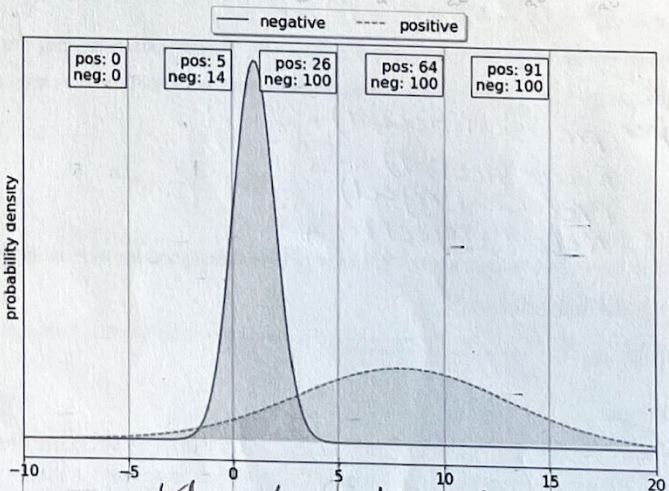
The following plot shows how the posterior probabilities are distributed within a dataset comprised of 200 records (100 positive records, 100 negative records).

In particular, the 5 boxes in the upper part of the figure show the cumulative number of records that belong to the positive (pos) and negative (neg) classes, from $-\infty$ to the threshold defined by the respective vertical line.

For example, there are 26 positive records and 100 negative records that have $x \leq 5$.

Consider the ROC curve for this classifier. Which of the following pairs of points belong to the curve? (each of the points is provided as a pair of (x, y) coordinates).

- ○ (a) (1, 1), (5/100, 14/100)
- ○ (b) The information provided is not sufficient to answer the question
- ○ (c) (0, 1), (86/100, 95/100)
- ● (d) (86/100, 95/100), (0, 74/100)
- ○ (e) (5/100, 86/100), (0, 1)
- ○ (f) (5/100, 86/100), (91/100, 0)



TN	0	14	100	100	100
FP	100	100−14=86	0	0	0
FN	0	5	26	64	91
TP	100	95	100−26=74	36	100−91=9
TPR	1	0.95	0.74	0.36	0.8
FPR	1	0.86	0	0	0

The softmax activation function is defined as:

$$softmax(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

for $i = 1, \ldots, K$ and $\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$

Where $\mathbf{z}$ is the K-dimensional input vector.

Which of the following statements is true?

- ○ (a) The output $softmax(\mathbf{z})_i$ can be computed using only z_i
- ● (b) None of the other statements is correct
- ○ (c) Softmax is typically used for lasso regression
- ○ (d) Softmax is typically used as an activation function for hidden layers of a neural network
- ○ (e) Softmax is idempotent, i.e. $softmax(\mathbf{z})_i = softmax(softmax(\mathbf{z})_i)_i$
- ○ (f) The output of softmax is a probability distribution of a continuous variable

- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C
- Accuracy is the fraction of correct predictions

Let y_pred, y_true be the prediction vector and the ground truth vector respectively.

```
y_pred = [A B C A A B B A]
y_true = [B B C B A B C C]
```

Write in the answer box:

- Accuracy = 4/8 = 0.5
- Precision(A) = 1/4 = 0.25
- Recall(A) = 1/1 = 1

- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C

After training two models (**M1**, **M2**), we obtain the confusion matrices shown below. Consider class **b** only. Which of the following statements is true?

M1	predicted class: a	b	c	d
true class: a	15	10	1	0
b	0	50	5	5
c	5	20	70	0
d	5	10	0	60

$R(b_{M1}) = \frac{50}{60}$

$P(b_{M1}) = \frac{50}{90}$

M2	predicted class: a	b	c	d
true class: a	14	10	0	2
b	5	40	5	10
c	5	5	80	5
d	10	5	10	50

$R(b_{M1}) = \frac{40}{60}$

$P(b_{M1}) = \frac{40}{60} \times \frac{3}{3} = \frac{120}{90}$

- ● (a) M1 has higher recall than M2 and lower precision.
- ○ (b) M1 has higher recall than M2 and higher precision.
- ○ (c) M1 has lower recall than M2 and lower precision.
- ○ (d) M1 has lower recall than M2 and higher precision.
- ○ (e) M1 has the same recall as M2 and lower precision.
- ○ (f) None of the other answers is correct

Truth \ pred	1	2	3
1	0	2	0
2	2	0	0
3	2	0	0

You are given the ground truth labels of a dataset and the predicted values obtained with a clustering algorithm (`y_true` and `y_pred` respectively).

```
y_true = [1,1,2,2,3,3]
y_pred = [2,2,1,1,1,1]
```

$$RI = \frac{TP + TN}{\binom{n}{2}} = \frac{3+12}{15} = \frac{15}{15} = 1$$

	Truth	pred	TP	TN
ab	1	2	✓	
ac	1	2		✓
ad	1	2		✓
ae	1	3		✓
af	1	3		✓
bc	1	2		✓
bd	1	2		✓
be	1	3		✓
bf	1	3		✓
cd	2	2	✓	
ce	2	3		✓
cf	2	3		✓
de	2	3		✓
df	2	3		✓
ef	3	3	✓	

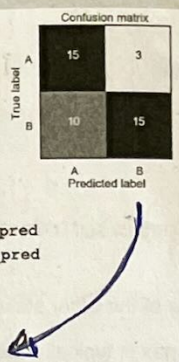
Compute the Rand Index (RI) between `y_true` and `y_pred`.

Remember that the Rand Index is given by:

$$RI = \frac{a+b}{DP} = \frac{TP+TN}{DP}$$

Where:

- a is the number of pairs of elements (without ordering) that are in the same set in `y_true` and in the same set in `y_pred`
- b is the number of pairs of elements (without ordering) that are in different sets in `y_true` and in different sets in `y_pred`
- DP is the total number of possible pairs in the dataset (without ordering)



You are given the confusion matrix in the figure below obtained after the validation phase of a classifier.

Question 1) Compute the accuracy of the classifier.

Accuracy: [ ] × $ACC = \frac{30}{43} = 0.6977$

Question 2) Compute the precision for class with label A.

Precision for A: [ ] × $P(A) = \frac{15}{25} = 0.6$

Given a set of n ground truth values y_i and predictions made by a regressor, $\hat{y}_i$, the Mean Absolute Error (MAE) is defined as:

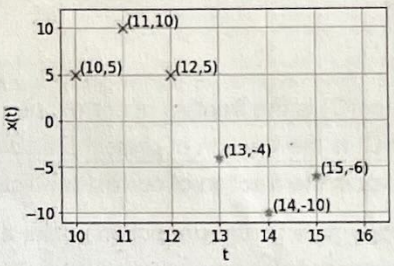
$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

You are given a 1-dimensional time series. You are interested doing time series forecasting (i.e. predicting the next values of the time series, based on the past ones).

Using a window containing the past 3 readings of the time series, x(t-2), x(t-1), x(t), you build the following model:

$$x(t+1) = a \cdot x(t-2) + b \cdot x(t-1) + c \cdot x(t) + d$$

Where a = -0.68, b = 0.26, c = 0.4, d = 0



The following figure represents the ground truth values for x(t), ($t \in [10, 15]$).

You use x(10), x(11), x(12) to predict x(13), x(14), x(15).
After acquiring the ground truth values for x(13), x(14), x(15), you validate your results by computing the mean absolute error.

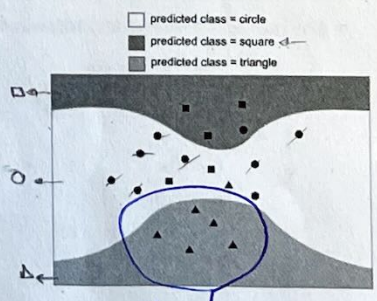
Answer the following question:

1. What are the values predicted for x(13), x(14), x(15)?
2. What is the MAE of the regressor, as evaluated on x(13), x(14), x(15)?

To assess the quality of a classification result the following measures were defined:

- accuracy: the percentage of correctly classified items
- precision(C): the percentage of correctly classified items among those predicted with class C
- recall(C): the percentage of correctly classified items among those with actual class C

Let C be a supervised classifier and I a two-dimensional input space. The points drawn from I belong to one of the three classes: *Circle*, *Square*, or *Triangle*. After the training stage, C learns the following decision boundaries:



Where the label assigned by C is given by the background color of the region in I.

Suppose that the symbols in the image are test points, whose shape represents the respective ground truth label (either a *Circle*, a *Square*, or a *Triangle*).

Which of the following statements is true?

- (a) The precision for Square is ⅓
- (b) The precision for Triangle is 1
- (c) The precision for Circle is 1
- (d) None of the answers is correct
- (e) The recall for Square is ¾
- (f) The accuracy is 1

a) $P_\square = \frac{3}{4}$

b) $P_\triangle = \frac{5}{5} = 1$

c) $P_\circ = \frac{8}{11}$

e) $R_\square = \frac{3}{5}$

f) $ACC = \frac{16}{20}$

$$precision = \frac{\text{تعداد نمونه‌های درست پیش‌بینی شده}}{\text{تعداد کل نمونه‌های پیش‌بینی شده}}$$

$$Recall = \frac{\text{تعداد نمونه‌های درست پیش‌بینی شده}}{\text{تعداد کل نمونه‌های واقعی}}$$

$$F_1 = \frac{2RP}{P+R}, \quad ACC =$$