DM566 Final Exam

The exam contains 17 questions summing up to 100 points. The point value of each question is stated at the beginning of its text in parentheses. All questions ask for evaluation of multiple statements with (yes/no) or (true/false) answers. You will gain the full point value of a question if you evaluate all of its statements correctly. Penalty applies to wrong evaluations. Hence, skipping to evaluate a statement is very likely to be more advantageous for you than making a random guess. The points earned from an X-point question is calculated as below:

X*(C-W)/A

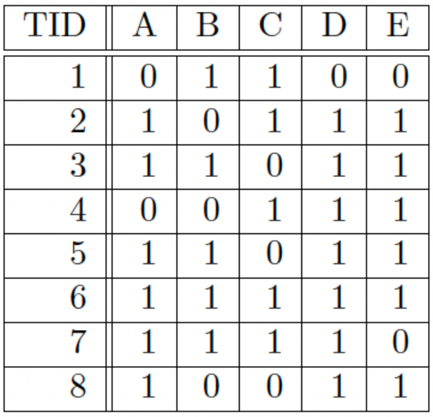
where

C : Number of correctly evaluated statements,
W: Number of wrongly evaluated statements,
A: Total number of statements in a question.

Note that C+W = A if you evaluate all statements and C+W < A if you skip evaluating at least one statement.

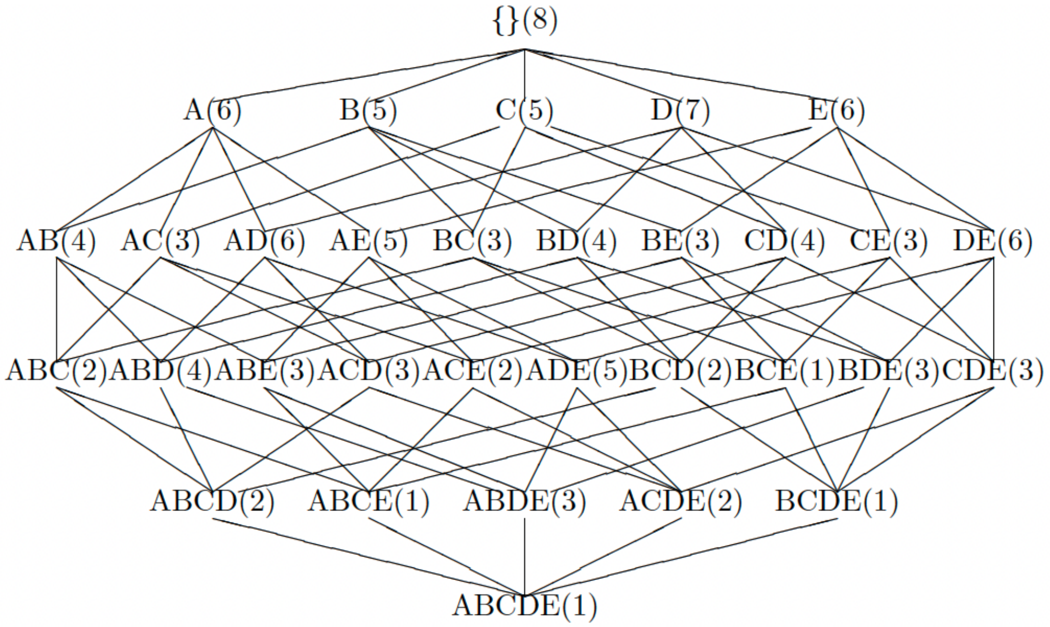
The point value distribution is as follows: X=6 for 16 of the questions and X=4 for one question.

(6 points) For a list of items $I = \{A, B, C, D, E\}$ and the transaction database below



TID	A	B	C	D	E
1	0	1	1	0	0
2	1	0	1	1	1
3	1	1	0	1	1
4	0	0	1	1	1
5	1	1	0	1	1
6	1	1	1	1	1
7	1	1	1	1	0
8	1	0	0	1	1

we computed the lattice below that lists the support of all itemsets:


{}(8)
A(6) B(5) C(5) D(7) E(6)
AB(4) AC(3) AD(6) AE(5) BC(3) BD(4) BE(3) CD(4) CE(3) DE(6)
ABC(2) ABD(4) ABE(3) ACD(3) ACE(2) ADE(5) BCD(2) BCE(1) BDE(3) CDE(3)
ABCD(2) ABCE(1) ABDE(3) ACDE(2) BCDE(1)
ABCDE(1)

Are of the following itemsets "closed frequent itemsets" for the support threshold $\sigma = 4$?

Select the correct answers

	Yes	No
AB	○	○
CE	○	○
BD	○	○
AD	○	○
ADE	○	○
ABCD	○	○
CD	○	○
ABD	○	○
CDE		

	○	○
AE	○	○
DE	○	○

(4 points) Consider a discrete random variable $\theta \in \{a, b\}$ that can take only two real values a and b within the $(0, 1)$ interval. Hence $\Pr(\theta = a) + \Pr(\theta = b) = 1$. Also consider a likelihood function $\Pr(X|\theta) = \theta^X(1 - \theta)^{(1-X)}$ which is applied to the data set $\square = \{X_1 = 0, X_2 = 1, X_3 = 1\}$.

Evaluate the correctness of the equalities below.

Select the correct answers

	True	False
$\Pr(\theta = a|\square) = \frac{0.5^a 0.5^{1-a}\Pr(\theta=a)}{0.5^a 0.5^{1-a}\Pr(\theta=a)+0.5^b 0.5^{1-b}\Pr(\theta=b)}$	○	○
$\Pr(\theta = a|\square) = \frac{a^2(1-a)^3\Pr(\theta=a)}{a^2(1-a)^3\Pr(\theta=a)+b^2(1-b)^3\Pr(\theta=b)}$	○	○
$\Pr(\theta = b|\square) = \frac{b^2(1-b)\Pr(\theta=b)}{a^2(1-a)\Pr(\theta=a)+b^2(1-b)\Pr(\theta=b)}$	○	○
$\Pr(\theta = a|\square) = \frac{a^2(1-a)\Pr(\theta=a)}{a^2(1-a)\Pr(\theta=a)+b^2(1-b)\Pr(\theta=b)}$	○	○
$\Pr(\theta = a|\square) = \frac{ab^2\Pr(\theta=a)}{ab^2\Pr(\theta=a)+ba^2\Pr(\theta=b)}$	○	○
$\Pr(\theta = a|\square) = \frac{a^2(1-a)^3\Pr(\theta=a)}{a^2(1-a)^3\Pr(\theta=a)+b^2(1-b)^3\Pr(\theta=b)}$	○	○

(6 points) Evaluate the following statements about the Expectation Maximization (EM) and k-means clustering algorithms.

Select the correct answers

	True	False
Choosing the optimal k is possible after training EM once on a data set. The same is not possible for $k-$means.	○	○
When trained on the same cluster count k, EM has more parameters to fit than $k-$means.	○	○
EM can learn covariances between feature dimensions. $k-$means cannot.	○	○
After being trained with two different choices of k, only $k-$means can provide a score for comparing which choice of k fits better to data.	○	○
The Lloyd-Forgy version of the k-means output is dependent on the processing order of the data points. The EM output is not.	○	○
The Mac-Queen version of the $k-$means output is dependent on the processing order of the data points. The EM output is not.	○	○

(6 points) Consider the case that the distance measure:

$$\mathrm{dist}(x, y) = \sqrt{(x_1 - y_1, x_2 - y_2)\begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}(x_1 - y_1, x_2 - y_2)^\top}$$

for two-dimensional points is chosen to be used for the Lloyd-Forgy implementation of the k−means algorithm with $k = 2$. The algorithm has calculated the cluster centroids $\mu_1 = (0, 2)$ and $\mu_2 = (4, 7)$ in the last computation step. Evaluate the statements below about the next computation step of the k−means algorithm.

Select the correct answers

	True	False
Point $(2, 5)$ will be assigned to Cluster 1.	○	○
Point $(4, 0)$ will be assigned to Cluster 1.	○	○
Point $(3, 4)$ will be assigned to Cluster 1.	○	○
Point $(3, 3)$ will be assigned to Cluster 1.	○	○
Point $(4, 4)$ will be assigned to Cluster 1.	○	○

(6 points) Consider two independent random variables

$X \in \{1, 2, 3, 4, 5, 6\}$
and
$Y \in \{1, 2, 3, 4, 5, 6, 7, 8\}$

with probability distributions

$\Pr[X = k] = 1/6$ for all $k \in \{1, 2, 3, 4, 5, 6\}$

and

$\Pr[Y = k] = 1/8$ for all $k \in \{1, 2, 3, 4, 5, 6, 7, 8\}$.

Evaluate the following statements about the random variable $Z = X + Y$.

Select the correct answers

	True	False
$\Pr[Z > 5] = 21/24$	○	○
$\Pr[Z = 7] = 1/8$	○	○
$\Pr[Z = 4] = 1/18$	○	○
$\Pr[Z < 7] = 3/16$	○	○

(6 points) We have a classification problem with two classes "+" and "−" and three trained classifiers h_1, h_2, and h_3 with the following probabilities of the classifiers, given the training data D:

$$\begin{aligned} \Pr(h_1|D) &= 0.4 \\ \Pr(h_2|D) &= 0.2 \\ \Pr(h_3|D) &= 0.4 \end{aligned}$$

For the three test instances o_1, o_2, o_3, the classifiers give the following class probabilities:

$$\begin{aligned}
o_1 &: \Pr(+|h_1) = 0.5 && \Pr(-|h_1) = 0.5 \\
& \Pr(+|h_2) = 0.6 && \Pr(-|h_2) = 0.4 \\
& \Pr(+|h_3) = 0.3 && \Pr(-|h_3) = 0.7 \\
o_2 &: \Pr(+|h_1) = 0.2 && \Pr(-|h_1) = 0.8 \\
& \Pr(+|h_2) = 0.9 && \Pr(-|h_2) = 0.1 \\
& \Pr(+|h_3) = 0.7 && \Pr(-|h_3) = 0.3 \\
o_3 &: \Pr(+|h_1) = 0.3 && \Pr(-|h_1) = 0.7 \\
& \Pr(+|h_2) = 0.0 && \Pr(-|h_2) = 1.0 \\
& \Pr(+|h_3) = 0.4 && \Pr(-|h_3) = 0.6
\end{aligned}$$

We combine the three classifiers to get a Bayes optimal classifier. Evaluate the correctness of the class probability calculations below for test instances o_1, o_2, o_3 we will get from this Bayes optimal classifier.

Select the correct answers

	True	False
o_3 : $\Pr(+|\text{Bayes optimal}) = 0.38$	○	○
o_2 : $\Pr(+|\text{Bayes optimal}) = 0.54$	○	○
o_3 : $\Pr(-|\text{Bayes optimal}) = 0.72$	○	○
o_1 : $\Pr(-|\text{Bayes optimal}) = 0.56$	○	○
o_2 : $\Pr(-|\text{Bayes optimal}) = 0.35$	○	○
o_1 : $\Pr(+|\text{Bayes optimal}) = 0.54$	○	○

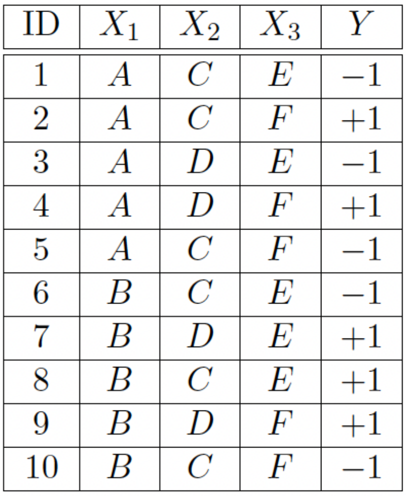
(6 points) Consider the results table

	Train Error	Test Error	Number of Model Parameters
Model 1	27%	27%	K
Model 2	12%	15%	2K
Model 3	5%	7%	3K
Model 4	5%	7%	4K
Model 5	3%	30%	5K

for a model whose prediction-time computational cost is linearly proportional to its number of parameters. Evaluate the the following statements about the performances of the models reported in the table above.

Select the correct answers

	True	False
Model 4 has over-fitted most	○	○
Model 5 has under-fitted compared to Model 1	○	○
Model 3 gives the highest test accuracy per unit prediction-time computational cost	○	○
Model 5 has over-fitted most	○	○
Model 1 has under-fitted compared to Model 2	○	○
Model 1 has over-fitted most	○	○
Model 4 gives the highest test accuracy per unit prediction-time computational cost	○	○



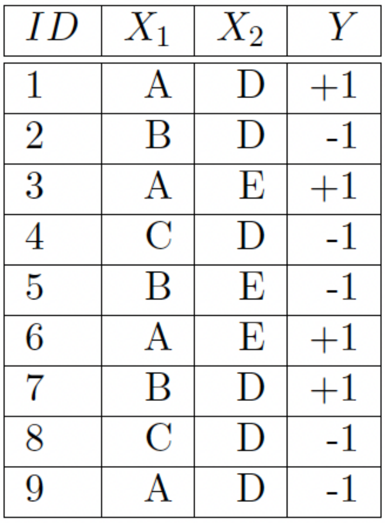
ID	X_1	X_2	X_3	Y
1	A	C	E	-1
2	A	C	F	$+1$
3	A	D	E	-1
4	A	D	F	$+1$
5	A	C	F	-1
6	B	C	E	-1
7	B	D	E	$+1$
8	B	C	E	$+1$
9	B	D	F	$+1$
10	B	C	F	-1

(6 points) A decision tree is being trained on the above data set, which uses the Gini index as the attribute selection criterion for splitting. As the root of the tree, the attribute X_1 was already selected for branching. Evaluate the statements below about the next attribute this decision tree will select for branching.

Select the correct answers

	True	False
For the branch of $X_1 = B$, it will select X_3.	○	○
For the branch of $X_1 = B$, it will select X_2.	○	○
For the branch of $X_1 = A$, it will select X_3.	○	○
For the branch of $X_1 = A$, it will select X_2.	○	○

(6 points) Consider a classification problem where the goal is to map two features $X_1 \in \{A, B, C\}$ and $X_2 \in \{D, E\}$ to class labels $Y \in \{-1, +1\}$. We train a naive Bayes classifier on the data set below



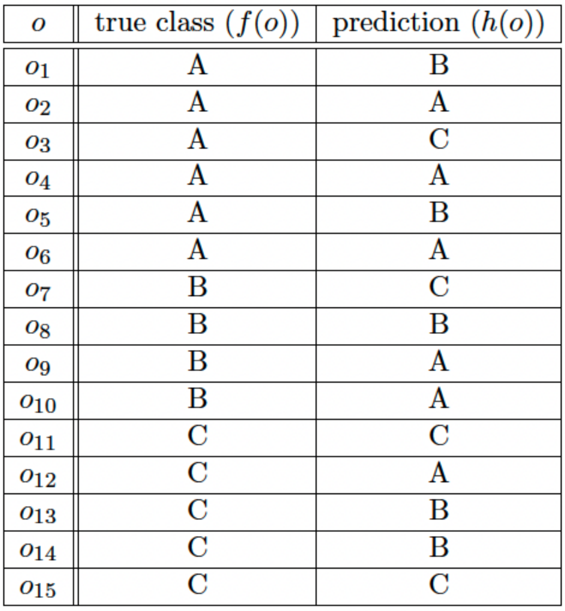
ID	X_1	X_2	Y
1	A	D	+1
2	B	D	-1
3	A	E	+1
4	C	D	-1
5	B	E	-1
6	A	E	+1
7	B	D	+1
8	C	D	-1
9	A	D	-1

Evaluate the following statements for this naive Bayes classifier.

Select the correct answers

	True	False
For test query (B, E), it predicts higher probability for class -1 than class $+1$.	○	○
For test query (C, E), it predicts higher probability for class $+1$ than class -1.	○	○
For test query (A, D), it predicts higher probability for class $+1$ than class -1.	○	○
For test query (B, D), it predicts higher probability for class $+1$ than class -1.	○	○

(6 points) The true class of 15 test objects and the corresponding predictions of some classifier h are given in the table below.



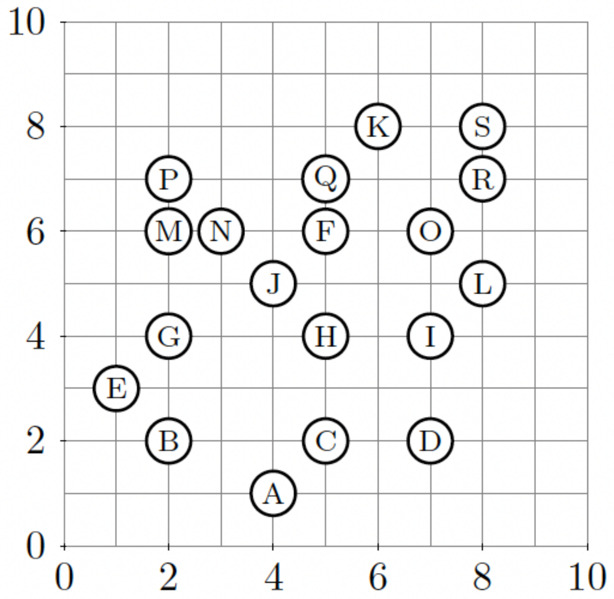
o	true class $(f(o))$	prediction $(h(o))$
o_1	A	B
o_2	A	A
o_3	A	C
o_4	A	A
o_5	A	B
o_6	A	A
o_7	B	C
o_8	B	B
o_9	B	A
o_{10}	B	A
o_{11}	C	C
o_{12}	C	A
o_{13}	C	B
o_{14}	C	B
o_{15}	C	C

Evaluate statements below about the class-specific recall and precision scores for this classifier.

Select the correct answers

	True	False
precision for class B is larger than precision for class C	○	○
recall for class B is larger than precision for class B	○	○
recall for class A is larger than recall for class B	○	○
precision for class C is larger than recall for class C	○	○
precision for class A is larger than precision for class C	○	○
precision for class A is larger than recall for class A	○	○
recall for class B is larger than recall for class C	○	○

(6 points) Consider the data set below consisting of observations on a two-dimensional feature space denoted as circles marked with capital letters:



Evaluate the following statements for the DBSCAN algorithm that uses the Manhattan distance and counts the query object as a member of its neighborhood, where ϵ denotes the distance threshold and MinPts denotes the minimum number of points used in the determination of a core point.

Select the correct answers

	True	False
M is core point for $\varepsilon = 1$ and $\mathrm{MinPts} = 2$	○	○
L is core point for $\varepsilon = 2$ and $\mathrm{MinPts} = 5$	○	○
B is core point for $\varepsilon = 2$ and $\mathrm{MinPts} = 3$	○	○
C is core point for $\varepsilon = 2$ and $\mathrm{MinPts} = 5$	○	○
O is core point for $\varepsilon = 3$ and $\mathrm{MinPts} = 5$	○	○
A is core point for $\varepsilon = 2$ and $\mathrm{MinPts} = 3$	○	○
S is core point for $\varepsilon = 3$ and $\mathrm{MinPts} = 4$	○	○
Q is core point for $\varepsilon = 1$ and $\mathrm{MinPts} = 4$	○	○

(6 points) Evaluate the below statements about the bias of a $k-$nearest neighbor classifier.

Select the correct answers

	True	False
Enlarging the data set and maintaining the train-test split ratio reduces bias.	○	○
Reducing the learning rate reduces bias.	○	○
Reducing k reduces bias.	○	○
Building a majority-voting ensemble of five $k-$nearest neighbor classifiers reduces bias.	○	○
Increasing the size of the test set reduces bias.	○	○
Increasing k reduces bias.	○	○

(6 points) Consider a neural network $f(x) = W_3^T\sigma(W_2^T\sigma(W_1^Tx))$ with three hidden layers connected to each other via synaptic weights W_1, W_2, and W_3, and the rectified linear unit (ReLU) activation function $\sigma(u) = \max(u, 0)$. Assume that we train this neural network on a regression data set $\square_{\text{train}} = \{(x_n, y_n) | n = 1, \ldots, N\}$consisting of N data points using the mean squared error loss

$$\mathbb{L}(W_1, W_2, W_3, \square_{\text{train}}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - f(x_n)\right)^2.$$

Evaluate the following statements about the optimization problem below

$$\underset{\text{W1,W2,W3}}{\text{argmin}}\ \mathbb{L}(W_1, W_2, W_3, \square_{\text{train}})$$

and a choice of values on neural network weights $(\widehat{W_1}, \widehat{W_2}, \widehat{W_3})$ that satisfy the equation below

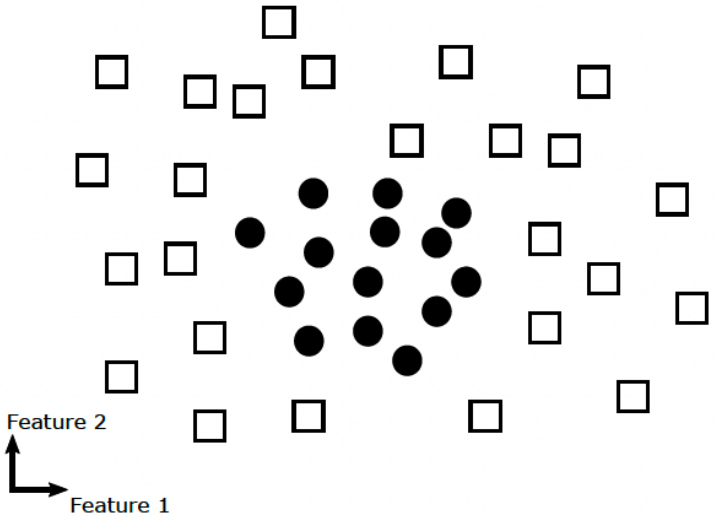
$$\nabla_{\text{W1,W2,W3}}\mathbb{L}(W1, W2, W3, \square_{\text{train}}) = 0.$$

Note: $\mathbb{L}(\widehat{W_1}, \widehat{W_2}, \widehat{W_3}, \square_{\text{train}})$ denotes the value the loss $\mathbb{L}$ gets when evaluated with weight values $(\widehat{W_1}, \widehat{W_2}, \widehat{W_3})$.

Select the correct answers

	True	False
Initializing the weights (W_1, W_2, W_3) to random values and updating them according to $(W_1, W_2, W_3) := (W_1, W_2, W_3) + \alpha\nabla_{W_1,W_2,W_3}\mathbb{L}(W_1, W_2, W_3, \square_{\text{train}})$ for learning rate $\alpha > 0$ until convergence finds a local minimum for $\mathbb{L}$. Remark: Pay attention to the sign in front of α.	○	○
$\mathbb{L}(\widehat{W_1}, \widehat{W_2}, \widehat{W_3}, \square_{\text{train}})$ is the unique global minimum for $\mathbb{L}$ when $\sigma(u) = 1/(1 + e^{-u})$.	○	○
$\mathbb{L}(\widehat{W_1}, \widehat{W_2}, \widehat{W_3}, \square_{\text{train}})$ is the unique global minimum for $\mathbb{L}$ when $\sigma(u) = \beta u$ for some $\beta > 0$.	○	○
Initializing the weights (W_1, W_2, W_3) to random values and updating them according to $(W_1, W_2, W_3) := (W_1, W_2, W_3) - \alpha\nabla_{W_1,W_2,W_3}\mathbb{L}(W_1, W_2, W_3, \square_{\text{train}})$ for learning rate $\alpha > 0$ until convergence finds a local minimum for $\mathbb{L}$. Remark: Pay attention to the sign in front of α.	○	○
$\mathbb{L}(\widehat{W_1}, \widehat{W_2}, \widehat{W_3}, \square_{\text{train}})$ is the unique global minimum for $\mathbb{L}$ when $\sigma(u) = u$.	○	○
$\mathbb{L}(\widehat{W_1}, \widehat{W_2}, \widehat{W_3}, \square_{\text{train}})$ is one of the many global minima for $\mathbb{L}$.	○	○
$\mathbb{L}(\widehat{W_1}, \widehat{W_2}, \widehat{W_3}, \square_{\text{train}})$ is the unique global minimum for $\mathbb{L}$.	○	○

(6 points) The figure below plots a set of query points $\mathrm{x} \in \mathbb{R}^2$ coming from an unknown data distribution of two classes $\{-1, +1\}$. The ground-truth labels of the data points are denoted as circles for class $+1$ and rectangles for class -1.



Is it possible for the classifiers below to reach zero prediction error on this data set after being trained on an arbitrarily large data set collected from the same data distribution for arbitrarily long time using an arbitrarily accurate optimization algorithm?

Select the correct answers

	Yes	No
Neural network with seven hidden layers that use the activation function $\sigma(u) = u$.	○	○
Support vector machine that uses a linear kernel function, i.e. $k(x, x') = x^T x'$.	○	○
Logistic regression, i.e. $\Pr(\text{class} = +1 \mid x) = 1/(1 + e^{-w^T x})$.	○	○
Decision tree with axis-aligned splits and unlimited maximum depth.	○	○
Decision tree with axis-aligned splits and a maximum depth of two.	○	○
Neural network with two hidden layers that use the activation function $\sigma(u) = \max(0, u)$.	○	○
Support vector machine that uses a squared exponential kernel function, i.e. $k(x, x') = e^{-\gamma \|\|x - x'\|\|_2^2}$.	○	○

(6 points) For a list of items $\mathrm{I} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{F}\}$, consider the following transaction database:

TransID	Items
1	A B F
2	B C
3	A E F
4	B D E
5	A C E
6	B C E F
7	A D E
8	A B C
9	A B C D F
10	C E F

When the APRIORI algorithm is applied to the transaction database above for a support threshold of 3, it outputs the frequent 2-itemsets below

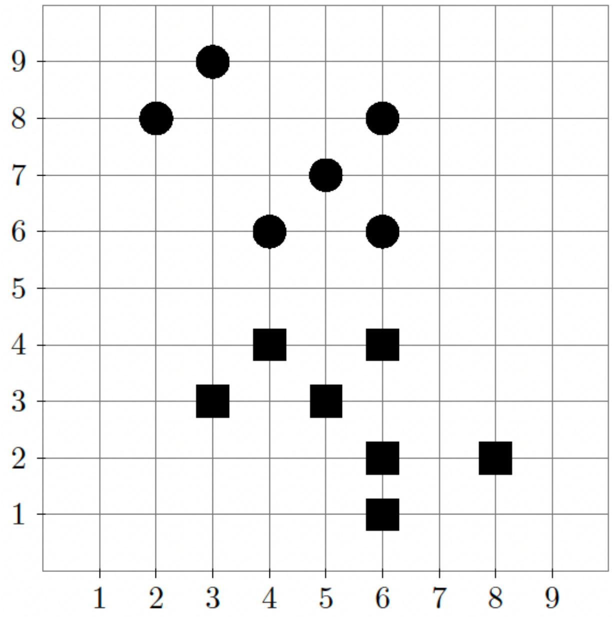
$$\mathrm{L}_2 = \{\mathrm{AB}, \mathrm{AC}, \mathrm{AE}, \mathrm{AF}, \mathrm{BC}, \mathrm{BF}, \mathrm{CE}, \mathrm{CF}, \mathrm{EF}\}.$$

Would the APRIORI algorithm output the following 3-itemsets after merging and before pruning?

Select the correct answers

	Yes	No
AEF	○	○
CDE	○	○
CEF	○	○
ABF	○	○
ABC	○	○
ABD	○	○
BCE	○	○
BDE	○	○

(6 points) The figure below plots a two-dimensional feature space in horizontal and vertical axes. The marks on the grid correspond to observations that are used by a binary hard-margin Support Vector Machine (SVM) classifier with linear kernel function $k(a, b) = a^T b$ for training. The labels of the positive class are denoted as circles and the negative class as rectangles. The data points are denoted as (X_1, X_2), where X_1 is the horizontal and X_2 is the vertical coordinate.

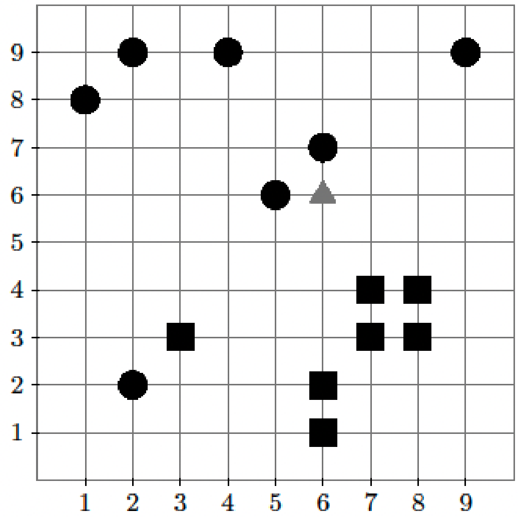


Evaluate the following statements for this SVM.

Select the correct answers

	True	False
The SVM will assign a test data point on (1,4) to the circle class	□	□
The SVM will assign a test data point on (4,1) to the circle class	□	□
The decision boundary passes through the coordinate (1,5)	□	□
The data points (2,8) is a support vector	□	□
The data point at (6,6) is a support vector	□	□
The data point at (3,3) is a support vector	□	□
The data point at (4,4) is a support vector	□	□

(6 points) The figure below shows a training set consisting of two-dimensional feature vectors plotted on a grid. The true class labels of the feature vectors are denoted as circles and squares. The triangle marked on coordinates (6, 6) is a test-time query input.



Consider a k−nearest neighbor classifier that uses Manhattan distance (L_1 norm) as distance function with non-weighted class counts, i.e. the object is assigned to the majority class within the k nearest neighbors. The classifier handles all cases where multiple data points have equal distance to the query point in such a way that its nearest neighbor list always contains exactly k neighbors. Evaluate the following statements about this classifier.

Select the correct answers

	True	False
For $k = 7$, the classifier will predict the label for the query point at (6, 6) as square.	○	○
For $k = 6$, the classifier will predict the label for the query point at (6, 6) as square.	○	○
For $k = 5$, the classifier will predict the label for the query point at (6, 6) as square.	○	○
For $k = 3$, the classifier will predict the label for the query point at (6, 6) as square.	○	○
For $k = 2$, the classifier will predict the label for the query point at (6, 6) as square.	○	○