Introduction to Machine learning - Quiz

If three dice are rolled, find the probability that the numbers add up to 10?

1/4

1/6

1/8

2/36

The sample space of rolling a die three times has 6*6*6= 216 outcomes out of which only 27 outcomes add up 10, these outcomes are:

(1,3,6), (1,4,5), (1,5,4), (1,6,3)

(2,2,6), (2,3,5), (2,4,4), (2,5,3),(2,6,2)

(3,1,6), (3,2,5),(3,3,4), (3,4,3), (3,5,2), (3,6,1)

(4,1,5), (4,2,4), (4,3,3), (4,4,2), (4,5,1)

(5,1,4), (5,2,3), (5, 3,2),(5,4,1)

(6,1,3), (6,2,2),(6,3,1)

Therefore, the probability is 27/216= 1/8

Question 2

Which of the following statements is True about hypothesis testing?

Type I error occurs when a True null hypothesis is rejected

Type I error occurs when a false null hypothesis is not rejected

Type II error occurs when the null hypothesis is rejected.

Type II error occurs when a true null hypothesis is rejected

Question 3

A survey asked 100 people if they thought women in the armed forces should be permitted to participate in combat. The results of the survey are shown. Use the formula for conditional probability to find the conditional probability that a respondent answered yes, given that the respondent was a male.

Gender		Yes	No	Total
Male	32	18	50	
Female		8	42	50
Total	40	60	100	

8/100
32/100
32/50
8/50

Question 4

Which of the following is not included in a boxplot?
Q2
The median
The mean
Q1

Question 5

In which distribution the mean is to the left of the median, and the mode is to the right of the median?
Symmetric
Right-skewed
Uniform
Left-skewed

Question 6

In a CS class there are 18 juniors and 10 seniors; 6 of the seniors are females and the rest are male, and 12 of the juniors are males and the rest are female. If a student is selected at random, find the probability of selecting a junior or a female.

16/28
24/28
6/28
22/28

p(j)=18/28

p(female)=12/28

p(j &f)=6/28

p(j or f) = p(j)+p(f)-p(j & f)= (18+12-6)/28=24/28

Question 7

For which distribution approximately 99.7% of data points fall within 3 standard deviations from the mean?

Normal
Uniform
Binomial
Poisson

Question 8

Find the median of 209, 223, 211, 227, 225, 240, 240, 211, 229, 212.

224
215
218
213

Question 9

Find the mode of 23, 9, 27, 11, 40, 13, 40, 11, 9, 40, 11, 29, 12, 40.

18
40
11
23

Question 10

Which of the following uses vertical bars to graphically represent the frequencies of data classes?

Ogive
Boxplot
Frequency polygon
Histogram

Question 11

A researcher wishes to test the claim that the average cost of tuition and fees at a four-year public college is greater than $5700. She identifies the hypotheses: H0: $\mu = \$5700$ and H1: $\mu > \$5700$ (claim). She selects a random sample of 36 four-year public colleges and finds the mean to be $5950. The P-value is 0.0113 and $\alpha = 0.01$. Based on the P-value, the researcher should reject the null hypothesis.

True
False

The p-value 0.0113 is greater than alpha=0.01 so the researcher cannot reject the null hypothesis.

Question 12

Suppose that we have a dataset -16,1,4,8,10, 15, 19, 23, 44, 55

Using the interquartile range method, -16 is an outlier.

True
False

Question 1

What type of learning is classification?
unsupervised
reinforcement
none of the above
supervised

Question 2

What type of learning is clustering?
unsupervised
supervised
reinforcement
none of the above

Question 3

When is a ML algorithm biased?
When it performs well on the training set but does poorly on the test set
When it is biased against some data features
When it uses too many heuristics
When its conclusions are systematically erroneous due to a priori information

Question 4

Attempting to model noise in training data might result in:
Bias

Overfitting
A model that generalizes well
A model that does not do well on training data

Question 5

What type of attribute is the school grade attribute with values A, B, C, D and F?
Ordinal
Ratio-scaled
Nominal
Interval-scaled

Question 6

Which process includes fitting a model to a dataset
Generalization
Testing
Training
Model evaluation

Question 7

The neural network model is inspired by:
Adaptive control theory
Neuroscience
Evolutionary models
Psychology

Question 8

What is the type of the response variable in regression?
Nominal
Continuous (real valued)
Interval-scaled
Ordinal

Question 9

Heuristics are guaranteed to find an optimal solution.
True
False

Question 10

Suppose that Illinois wants to use a ML model to predict total Illinois Sale's tax in dollars for a future year based on the total sale tax values in the previous years as well as some other attributes such as State GDP, median income, etc.

This is an example of:

Market Basket (Association Rule) Analysis problem
A clustering problem
A classification problem
A regression problem

Question 11

Suppose that a wholesale company wants to group customers based on their purchase histories and put customer with similar purchasing habits in the same group.

This is an example of a

supervised learning problem
classificaiton problem
clustering problem
regression problem

Question 12

In R, comments are preceded by:
/*
#
%
//

Question 13

Let:
L <- list(first=1, second = 2.5, third = "Apple")
Which of the following commands is NOT a valid way to access the second component of list L?
L[[second]]
L$second
L[["second"]]
L$sec

Question 14

Which function loads a previously installed package into memory?
install.packages(<package name>)
library(<package name>)
load(<package name>)
vignette(<package name>)

Question 15

Which of the following is not assignment in R:
=
==
<-
->

Question 16

Which of the following statements about data frames is FALSE?
Each data frame is a list
The vectors of a data frame must be of the same type
The vectors of a data frame usually correspond to features or attributes
Data frames are displayed in matrix form

Question 17

Which of the following statements about vectors is FALSE?
The lowest index in a vector is 0
Strings are vectors of length 1
Vectors are homogeneous data structures, i.e., they contain elements of the same type.
Even single numbers, such as 1, are vectors

Question 18

Let:
y<-c("A","B","C","D","E","F")
x <- c(F, T, F)
What would the following command return?
y[x*2]
[1] "B" "E"

[1] "A" "C" "A"
[1] "B"

Question 19

Assuming that:
> i <- 0
> sum <- 0
Which of the following statements computes the sum of all odd numbers from 1 to 20?
repeat{ i <- i +1; if(i > 20) next else if (i %% 2 == 0) sum <- sum +i}
repeat{ i <- i +1; if(i > 20) next else if (i %% 2 == 1) sum <- sum +i}
repeat{ i <- i +1; if(i > 20) break else if (i %% 2 == 1) sum <- sum +i}
repeat{ i <- i +1; if(i > 20) break else if (i %% 2 == 0) sum <- sum +i}

Question 20

Suppose that a cellphone company wants to create an ML model that inspect cellphones for screen scratches. The model gets the image of a cellphone screen and flags it as defective- or none-defective. This is an example of a
regression problem
unsupervised learning
clustering problem
Classification problem

Question 21

Let:
y <- c("A","B","C","D","E","F")
x <- c(2,4,6)
What would the following command return?
> y[x]
[1] "B" "D" "F"
[1] "A" "C" "E"
[1] NA NA NA
An error

Question 22

Let:
y<-c("A","B","C","D","E","F")
x <- c(11, 1, 15)
What would the following command return?
> y[x > 10]
[1] "A" "C" "D" "F"
[1] "B" "E"
[1] "C" "F"
character(0)

In the usedcars dataset, find the exact number of cars priced greater than $20000.

49
98
2
27

Question 2

What is the most frequent used car color in the usedcars dataset?

Black
Yellow
Silver
Gray

Question 3

What is the median of price in the usedcar dataset?

12995
10995
13592
12962

Question 4

The mileage variable in the usedcars dataset is negatively skewed

True
False

Question 5

Find how many cars in the usedcars dataset have a mileage within one standard deviation from its mean?

53
107
155
143

Question 6

What percentage of flights are canceled in the flight dataset?

2.17
1.172
88

98.82

Question 7

Draw a scatterplot of Price vs. Year in the usedcars dataset. Is there a tendency for the price to increase with the year? (True=Yes, False=No)

True

False

Question 8

In the usedcars dataset, the mode for year is greater than the mean for year AND the median for year is equal to the mean of year. (round the numbers to whole years).

True

False

Question 9

How many 2008 cars are there in usedcars dataset?

11

6

14

3

Question 10

The OP_CARRIER (airline carrier) and DEP_DELAY (departure delay) are associated in the flight dataset (assuming significance level alpha=0.05).

True

False

Question 11

in the airline satisfaction dataset, gate.location is associated with satisfaction (assuming significance level alpha=0.05)?

True

False

satisfaction is categorical, gate.location is ordinal to test association between an ordinal and a categorical variable we can use kruskal-wallis test.

```
kruskal.test(airline$Gate.location~ airline$satisfaction)
Kruskal-Wallis rank sum test
data:  airline$Gate.location by airline$satisfaction
Kruskal-Wallis chi-squared = 0.0019626, df = 1, p-value = 0.9647
```

p-value is greater than alpha (0.05) so there is no evidence of association between the two variables.

Question 12

In the airline satisfaction dataset satisfaction is independent of Gender (assuming significance level alpha=0.05)

True
False

Question 1

Which of the following is not an advantage of k-NN?

It is simple
Has only few parameters to learn
Can learn complex functions
Makes no assumption about underlying data distribution

Question 2

Underfitting often leads to:

lower variance and lower bias
Higher variance and lower bias
lower variance and higher bias
Higher variance and higher bias

Question 3

Learning algorithm that can reduce the chance of fitting noise is called:

Underfitting
Eager
Lazy
Robust

Question 4

Which of the following is NOT a problem for distance measures:

Irrelevant attributes
Having a small number of features
Mixing numeric and categorical data
Different scales

Question 5

Large values of k in the knn algorithm tend to produce:
- More bias and more variability
- **More bias and less variability**
- Less bias and more variability
- Less bias and less variability

Question 6

Which of the following statements is FALSE?
- k-NN is a lazy learning algorithm
- k-NN is an example of instance-based learning
- **k-NN is an eager learning algorithm**
- k-NN is a non-parametric algorithm

Question 7

Diagnosing breast cancer with k-NN is an example of:
- Multi-label classification
- None of the above
- **Binary classification**
- Multiclass classification

Question 8

Adding more data (i.e., more observations) helps improve the performance of a high biased machine learning model.
- True
- **False**

Question 9

Which option can help improve a machine learning model with a high bias?
- get more data
- use regularization
- use less number of features (attributes)
- **using a more complex model**

Question 10

How can we diagnose overfitting in a machine learning model?
- **low error on the training data but a higher error on the testing data**
- high error on both testing and training data

high error on the training data but a lower error on the testing data
low error on both testing and training data

Question 11

Categorical input features must be converted to numeric before before applying KNN with any p-norm distance measures (p>=1)
True
False

Question 12

To avoid data leakage during normalization, we should first split the data then compute any statistics needed for normalization from training split only.
True
False

Question 13

Which statement is NOT true about a 5-fold cross validation:
The original training data is partitioned into 5 equally-sized partitions
five separate models are trained
Error is averaged across five models to estimate model's performance
All models are evaluated on the same partition
Each model is trained on four partitions and evaluated on the remaining partition

---

Question 1

Which statement about the mobile spam filter is FALSE?
The classifier counts all words with the same stem as occurrences of the same word
The classifier ignores numbers
The classifier uses word positions in sentences
The classifier uses word frequencies

Question 2

Which data structure is used to store word frequencies in the mobile spam filter?
Corpus
Histogram
Word cloud
Document term matrix

Question 3
1 / 1 pts
Which of the following types of data preparation was NOT applied in the mobile spam filter:
Removing stop words
Removing punctuation symbols
Converting uppercase characters to lowercase characters
Removing synonyms

Question 4

Why the Naïve Bayes classifier is called naïve?
Because it uses a class-conditional independence assumption
Because it is simple
Because it works with categorical attributes
Because it uses conditional probabilities

Question 5

Which technique can be used to prepare numeric data for Naïve Bayes?
z-score standardization
Normalization
Scaling
Binning

Question 7

Using the diagram on slide 15, find the conditional probability that Johnny likes a pie given that its filling size is NOT thick.
0.75
0.60
0.25
0.80
We need to find P(pos|not thick)=p(pos,not_thick)/p(not thick)

p(pos,not thick)= 3/12 this is because there are three pies that johny likes which are not thick

p(not thick=4/12 this is the prior probability of a pie not being thick which is 4/12 because there are 4 pies that are not thick out of the total 12 pies.

so p(pos|not thick)= (3/12)/(4/12)=3/4=0.75

Question 8

Why does the Naïve Bays classifier use Laplacian correction?

To deal with cases when a given class and feature value never occur together in the test data

To deal with cases when the DTM is sparse

To deal with cases when a given class and feature value never occur together in the training data

To deal with cases when an entry in the DTM is zero

Question 9

In the following formula, which expression is the prior probability?

P left parenthesis C space vertical line space D right parenthesis space equals space fraction numerator P left parenthesis D space vertical line space C right parenthesis space asterisk times space P left parenthesis C right parenthesis over denominator P left parenthesis D right parenthesis end fraction

P(D | C)

P(C)

P(D)

P(C | D)

Question 10

Which R function takes a corpus and a transformation function as attributes and applies the transformation function to each document in the corpus?

tm_map

lapply

VCorpus

apply

Question 11

consider the table in slide 22 of the lectures. What is the probability that an email is spam if it has the words Money, Groceries and unsubscribe but does NOT have the word viagra after applying the laplacian correction. P(spam|~w1,w2,w3,w4)?

0.85

0

0.09

0.25

After applying the Bayes rule:

P(spam|~w1,w2,w3,w4) =(p(~w1,w2,w3,w4|spam)*p(spam))/p(~w1,w2,w3,w4)

after applying the conditional independence assumption:
=(p(~w1|spam)*p(w2|spam)*p(w3|spam)*p(w4|spam) *p(spam)/ /p(~w1,w2,w3,w4)
Let's first compute the numerator:
p(~w1|spam)=(16+1)/(20+4)
p(w2|spam)=(10+1)/(20+4)
p(w3|spam)=(0+1)/(20+4)
p(w4|spam)=(12+1)/(20+4)
p(spam)=20/100
so (p(~w1,w2,w3,w4|spam)*p(spam))= 17/24*11/24*1/24*13/24*2/10=0.0014
Now let's compute the denominator: p(~w1,w2,w3,w4)
p(~w1,w2,w3,w4) = (p(~w1,w2,w3,w4|spam)*p(spam))+ (p(~w1,w2,w3,w4|ham)*p(ham))
(p(~w1,w2,w3,w4|ham)*p(ham)) = (p(~w1|ham)*p(w2|ham)*p(w3|ham)*p(w4|ham) *p(ham)
=(81/84)*(15/84)*(9/84)*(24/84)*(80/100)=0.0042
so p(~w1,w2,w3,w4) = (p(~w1,w2,w3,w4|spam)*p(spam))+ (p(~w1,w2,w3,w4|ham)*p(ham))= 0.0014+0.0042=0.0056
and
P(spam|~w1,w2,w3,w4) =(p(~w1,w2,w3,w4|spam)*p(spam))/p(~w1,w2,w3,w4)=0.0014/0.0056= 0.25

Question 1

Which attribute is selected by ID3 at each step of the tree construction?
The attribute with the lowest entropy
The attribute with the maximum information content
The attribute with the highest expected reduction in entropy caused by partitioning the examples according to this attribute
The attribute with the highest entropy

Question 2

Which of the following is NOT a reason for preferring small decision trees over large decision trees?
Larger trees are prone to overfitting
Smaller trees tend to get rid of irrelevant attribute
It is easier for a human to explain small trees
Larger trees have larger bias

Question 3

Which statement regarding Boolean classification is FALSE?
The entropy is always equal or greater than zero

The entropy is always equal or less than 1
Sets with low entropy are very diverse
The entropy is maximum when the set contains an equal number of positive and negative examples

Question 4

What is the name of the process that first builds the entire decision tree and then removes some of its branches?
Model fitting
Pre-pruning
Runtime pruning
Post-pruning

Question 5

The information gain of an attribute can be negative.
True
False

Question 6

How many bits is the information content of a message whose probability is 1/8?
2
3
1
4

Question 7

What is the main idea of boosting?
To boost the performance of a decision tree by reducing its size
To create several decision trees and choose the best one
To adapt the parameters of a decision tree recursively
To create a team of learners that is much stronger than any of the learners alone

Question 8

Why are decision trees greedy?
Because it used entropy as a measure of purity
Because it is recursive
Because at each step, the algorithm chooses the attribute that looks the best for the current moment, not the best attribute in the long run.
Because at each step the algorithm chooses only one attribute, not a subset of attributes.

Question 9

What is the entropy of the node to the left of the root node in the decision tree in slide #14? That is, the node consisting of one movie with Critical Success, one movie with Mainstream Hit, and 10 movies with Box Office Box

-1*1log(1)-1*log(1)-10*log(10)

-1/6*log(1/12)-5/6*log(5/6)

1

-1/12*log(1/12)-5/6*log(5/6)

Question 10

in the decision tree in slide 14, what is the information gain from splitting the root node on the "number of celebrities" attribute? (compute all logarithms in base 2 and round the final answer to two decimal points)

0.54

To compute information gain from splitting on an attribute A, we need to compute the entropy before splitting on A (E(S)) and the entropy after splitting on A (E(S,A)). then the reduction in entropy E(S)-E(S,A) will give us the information gain from splitting S on attribute A.

in slide 14, the entropy of the root node E(Root) is:

E(Root)= -(1/3)*log(1/3) -(1/3)*log(1/3) - (1/3)*log(1/3)= 1.584963

entropy after splitting on the number of celebrities attribute is:

E(Root,number_of_celebrities)= p(number_of_celebrities=low)*E(number of celebrities=low) + p(number_of_celebrities=high)*E(number of celebrities=high)

E(number of celebrities=low)= -(1/12)* log(1/12)-(1/12)*log(1/12)-(10/12)*log(10/12)=0.8166891

E(number of celebrities=high)=-(9/18)*log(9/18)-(9/18)*log(9/18)=1

p(number_of_celebrities=low)=12/30   p(number_of_celebrities=high=18/30

Hence,

E(Root,number_of_celebrities)=12/30*0.8166891+18/30*1=0.9266756

Gain(Root, number_of_celebrities)= E(Root)-E(Root,number_of_celebrities)=1.584963 - 0.9266756= 0.6582874

Question 1

The Pearson correlation coefficient tests for what type of relationship between two variables?

Linear
Quadratic
Exponential
Polynomial

Question 2

A decision tree with a real-valued feature, will split on every single value of that feature without overfitting. For example, suppose that a dataset contains the feature "weight" with 100 unique values. A decision tree splits the training data based on the "weight" attribute into 100 branches.

True
False

Question 3

Which regression method is used for predicting a categorical response variable?

Polynomial
Logistic
Simple linear
Multiple linear

Question 4

Linear regression requires that every feature be numeric.

True
False

Linear regression model can only handle numeric variables, all string variables are converted to numeric using dummy-coding before being fed to a linear regression model.

Question 5

Step-wise regression is more computationally intensive than standard multiple linear regression.
True
False

Question 6

What does the p-value of a coefficient in linear regression shows?
the difference between the observed target and the target predicted by linear regression model
The likelihood that the coefficient is significant
The fraction of the variance of the dependent variable that is explained by the regression model
The likelihood that the coefficient is not significant

Question 7

Which trees return the mean of the target values on their leaves?
Classification trees
regression trees
model trees
pruned trees

Question 8

Which trees use regression models on their leaves?
regression trees
pruned trees
model trees
ID3 trees

Question 9

Which is NOT an advantage of decision/regression tree?
Is a strong learner and typically has a high prediction accurracy
It requires minimal pre-processing
does not require feature normalization
Can handle both categorical and numerical features
has embedded variable selection

Question 10

In general, regression trees are computationally more complex than model trees.
True
False

Question 11

Regression trees are sensitive to features' scales and require that features be normalized.
True
False as
Regression trees are insensitive to predictors' scales; hence, feature normalization is not necessary for regression trees.

Question 12

What does the Ordinary Least Squares method minimize?
the variance of the residual sum
the residual sum
the mean of the residual sum
the residual sum of squares

Question 13

Which method of variable selection starts with a model containing all the variables and then at each step purges a variable whose removal causes the most improvement to the model performance?
backward elimination
embedded models
forward selection
filter models

Question 14

In logistic regression the probability that an observation
belongs to class encoded as 1 :
is equal to a linear function of the predictors/independent variables:
True
False

Question 15

Assume that the logistic regression model is used to predict whether a person defaults on their credit loan on the basis of their monthly balance, annual income, and some other variables. Suppose that for a given person the logistic regression model predicted that the log odds of default is 0.4. What is the predicted probability that this person defaults on his/her loan? (round the result)

(hint: use the formula in slide 7 of lecture 7.2 and sections 4.3.3 and 4.3.4 of the book "Introduction to Statistical Learning")

0.28
0.50
0.45
0.60

Question 1

The gradient descent algorithm may get stuck in a local optimal point because the gradient is near zero at these points and the parameters don't get updated.
True
False

Question 2

Which statement about ANNs is false?
The neurons in a convolutional layer are fully connected to the neurons/pixels in the previous layer.
Complex network topologies carry the risk of overfitting
Hidden layers allow the network to extract higher-order statistics from its input
The back-propagation algorithm uses the gradient descent method

Question 3

The backpropagation learning is unsupervised learning.
True
False

Question 4

Which statements are NOT true about RelU activation function?
it is a linear activation function
it's range is between [0,Inf]
relu(v)=max(-v,v)

A neuron with a Relu activation function does not learn from an example which makes its activation zero.

Question 5

One must normalize the input features before training a neural network model.

True

False

Question 6

In a mini-batch gradient descent with batch size 100, loss is computed and the weights are updated for every training example.

True

False

Question 7

Which statements are true about artificial neural networks?

if an ANN has one neuron in the last layer and $w^{[l-1]}$ is a connection weight to this neuron and $z^{[l]}$ and $a^{[l]}$ are the input and output values of this neuron, respectively then the gradient of the loss function with respect to $w^{[l-1]}$ is calculated as follows:

$$\frac{\partial loss}{\partial w^{[l-1]}} = \frac{\partial loss}{\partial a^{[l]}} \times \frac{\partial a^{[l]}}{\partial z^{[l]}} \times \frac{\partial z^{[l]}}{\partial w^{[l-1]}}$$

In backward pass of the back propagation algorithm, we compute the gradient of the error with respect to connection weights and apply the gradient descent algorithm to update the weights

an epoch is a pass through the training examples in a minibatch.

In a regression neural network, the neurons in the final layers have softmax activation funciton.

Question 8

Which option is NOT a termination condition for the backpropagation algorithm

Once connection weights are updated for a mini-batch.

Once a certain number of epochs has reached

Once the training error falls below a threshold

Once the validation error meet some requirement

Question 9

The hyper-parameters of a model must be tuned on the test data. In other words, model's performance on the test data is used to select the best hyperparameter combination.

True

False

Question 10

Suppose that we want to train a neural network model to classify images of handwritten digit (0-9). Suppose that the network outputs the vector [0, 0.1, 0.1, 0.4, 0.2, 0, 0, 0.1, 0, 0.1] for a given hand written image of digit 3 in the training data. What is the corss-entropy loss for this image(Note: use base 2 for logarithm, also conventionally 0*log(0)=0) (answers are rounded to one decimal point)

-1.3

1.3

-2.3

2.3

Question 11

Suppose I want to use a neural network to classify images into five categories (i.e., classes) which I have encoded using numeric indices from 0-4 Which loss function should I use in keras to train my network?

categorical_crossentropy

sparse_categorical_crossentropy

mse

accuracy

Question 12

In a binary classification problem, how many neurons should be in the final/output layer and what should be the activation function in that layer?

2, sigmoid

1, softmax
1, no activation
1, sigmoid

Question 1

Which statement is false?
A soft margin hyperplane depends only on the support vectors
A Support Vector classifier is not affected by the observations that are far away from the hyperplane
A soft margin classifier can misclassify a few training observations for a better generalization
Observations that lie on the correct side of the margin do not affect the support vector classifier
A maximal margin classifier is not prone to overfitting

Question 2

To classify a new data point
using a support vector classifier one needs to compute the inner products of
with each training example in the support vector.
True
False

Question 3

Suppose that for the ith training example in a support vector classifier, the slack variable 0 less than epsilon subscript i less than 1 Which statement is true about this training example?
It is on the correct side of the margin
It is on the wrong side of the hyper-plane but it is classified correctly by the classifier
It is on the correct side of the hyper-plane but the wrong side of the margin and is classified correctly by the classifier
it is on the wrong side of the hyper-plane and is miss-classified by the classifier

Question 4

A maximal margin classifier tries tries to find a separating hyper-plane with the largest minimum distance to the training observations
True
False

Question 5

An RBF kernel function has a small value for two data points that are far away

True
False

Question 6

Suppose that a support vector classifier uses the following hyperplane to separate data points in two classes (y=+1,y=-1).

f(x)=-3x1+2x2-x3-1

Which of the following points will be miss-classified by this classifier.

	x1	x2	x3	Y
point1	1	3	4	+1
point2	1	0.5	1	+1
point3	3	3	4	-1
point4	5	7	-3	-1

point1, point4
point2, point4, point1
point2,point4
point2, point3

Question 7

A data that is not linearly separable in a lower dimension might become linearly separable if transformed to a higher dimension.
True
False

Question 8

The bias of a support vector classifier increases as we widen the margins.
True
False

Question 9

Which Statements are True?
The equation of a Support Vector Classifier is written using a nonlinear kernel function
all kernel functions are linear
Using a kernel function free us from having to compute the coordinates of each data point in the enlarged feature space
kernel functions quantify the similarity between data points

Question 10

The larger the tuning parameter C in a support vector classifier, the wider the margins and the higher the variance.

True

False

Question 11

Which of the following are support vectors in support vector machine?(choose all that apply)

A point that is on the correct side of the hyperplane and margin.

a point that is on the wrong side of the margin

a point that is on the margin

a point that is on the correct side of the margin

a point that is on the wrong side of the margin but correct side of the hyperplane

---

Question 1

What is the sensitivity of the following model?

Predicted

Actual	yes	no
yes	40	8
no	2	50

0.9091

0.8333

0.8621

0.9523

tp=40

fn=8

sensitivity= tp/(tp+fn)=0.8333

Question 2

What is the precision of the following model?

Predicted

Actual	yes	no
yes	40	8
no	2	50

0.8333
0.9523
0.9091
0.8862

tp=40

fp=2

precision=tp/(tp+fp)=40/(40+2)=0.9523

Question 3

What is the specificity of the following model?

Predicted

Actual	yes	no
yes	40	8
no	2	50

0.9615
0.9524
0.8862
0.8620

tn=50

fp=2

specificity= tn/(tn+fp)=0.9615

Question 4

What is the accuracy of the following model?

Actual

Predicted		yes	no
yes	11	49	
no	9	50	

0.9
0.5
0.82
0.42

Question 5

What is the kappa statistic of the following model?

Predicted

Actual	no	yes
no	42	10
yes	3	56

0.9259
0.7629
0.9091
0.7473

Question 6

What is the F-measure of the following model?

Actual

Predicted		yes	No
yes	40	8	
No	4	50	

0.5992
0.888
0.8695
0.9091

Question 7

The elastic net regularization adds a combination of Lasso and Ridge penalty to the loss function. The hyper-parameter alpha is between [0,1] and controls the impact of lasso vs ridge penalty in elastic net.

True
False

Question 8

Suppose a model uses 3 parameters. The first one has 3 values, the second parameter has 4 values, and the third parameter has 2 values. What is the maximum number of candidate models (parameter combinations) that caret tests by default in automatic parameter tuning?

12

24

16

18

Caret automatic tuning only tries three values for each parameter so the number of models would be 3*3*2=18 (that means if you let it tune automatically instead of providing a tunegrid it will try only three values for the second parameter instead of four)

Question 9

Which method uses reweighted training data?

automatic parameter tuning

bagging

Adaptive boosting

random forests

Question 10

Oversampling should be done on the entire training dataset prior to cross validation

True

False

Reason:

the oversampling step should NOT be performed prior to cross validation. This may cause the training and validation data to share samples in each partition created by cross validation. As a result the cross-validation estimates of the model performance will be over-optimistic and biased, not reflecting the true out of sample performance of the model. Instead, the oversampling should be performed during cross validation where only training data in each partition is oversampled and the validation data remains unchanged.

Question 11

Which statement is Not true about Gradient Boosted Trees?

Each subsequent tree uses the residuals left from the previous trees as outcome variable and is fit to these residuals.

It is an additive model

The residuals are updated at each iteration

It is better to fit a deep tree instead of a shallow tree in each iteration of Gradient boosting to avoid overfitting

Boosting methods are prone to overfitting. As the number of iterations/trees is increased, the gradient boosted tree method may fit an overly complex function to the training data and hence, overfit. The hyper-parameter eta is a shrinkage factor which controls such overfitting.

Question 12

Which statement is Not true about Leave One Out Cross Validation (LOOCV)

The number of folds in LOOCV is equal to the number of training examples

A single observation is used for validation in each fold

it is computationally more expensive than k fold cross validation where k << number of training observations.

It reduces the variance of the true error estimate

It reduces the bias of the true error estimate

Question 13

Why random forests use a random subset of features?

to create less correlated trees

to create a random sample with replacement

to remove bias in feature selection

to perform cross validation

Question 14

Random forest that uses the full set of features is equivalent to bagging.

True

False

Question 15

Which method combines the predictions of several learners?

bagging

random forest

all of the above

boosting

Question 16

suppose that in ADA boost, a training example with weight 0.001 is classified correctly by a learner at stage t with an error value of e r r o r subscript t equals 0.005

What would be the updated weight for this example?

0.001

0.00007
0.0009
0.0141

Question 17

Which statements are true about LASSO linear regression?
adds the L2 norm of the coefficients as penalty to the loss function to penalize larger coefficients
if there are multiple correlated predictors lasso will select all of them
h**as embedded variable selection by shrinking the coefficient of some variables to exactly zero.**
has one hyper-parameter lambda (The regularization coefficient) which needs to be tuned

Question 18

Which statement about k-fold cross-validation is FALSE?
All observations are used for both training and validation
partitions the data into k non-overlapping folds
is typically used to tune and select the best hyper-parameters for the model
The last step of the k-fold cross-validation is to compute the average performance estimate
On each step, one fold is used as the training data and the remaining k − 1 folds are used as testing data

Question 19

Data sampling technique with its appropriate description:
Random Oversampling --Replicates random samples of minority class to balance the training data
Random undersampling -Eliminates random samples from majority class to balance the training data
SMOTE -Creates syntactic samples that are interpolation between randomly chosen examples from the minority class and their nearest neighbors
ROSE -randomly draw an observation from majority or minority class (with equal probabilities) and generates a new example in its neighborhood (using smoothed bootstrap sampling).

Question 20

Which method uses uniform sampling with replacement?
holdout
10-fold cross-validation
repeated k-fold cross-validation
bootstrapping

Question 21

A Ridge Linear Regression adds the sum of the squared values of the coefficients to the loss function to penalize large coefficients.

True

False

Which statements are true about embedding vectors

Embedding vectors preserve semantic relationship between different levels of a categorical variable.

Embedding vectors are used to create a dense/lower dimensional representation of high dimensional sparse vectors.

embedding layer is essentially a dense layer with sigmoid activation function.

Embedding vectors can be learned jointly with the task at hand (e.g., classification or regression)

Word embeddings map each word in a document to a variable length floating point vector

Question 2

A silhouette point coefficient of zero means that:

A point is as close to the objects in the same cluster as it is to the objects in other clusters

A point is closer to the objects in the same cluster than to the objects in other clusters

A point is outlier

A point is closer to the objects in other clusters than to the objects in the same cluster

Question 3

Autoencoders generate a set of compressed features that are uncorrelated and orthogonal

True

False

Question 4

Autoencoders are special type of neural networks trained to estimate identity function.

True

False

Question 5

Which method continuously merges small clusters into bigger clusters?
Grid-based
Density-based
Hierarchical
Partitioning

Question 6

Which of the following statements are true about the standard K-means?
Initially, the cluster centers are chosen at random
It aims for high intracluster similarity and low intercluster similarity
Final clusters are not sensitive to initial cluster centers
It is based on a distance metric

Question 7

What type of algorithm is k-means?
Density-based
Partitioning
Grid-based
Hierarchical

Question 8

Which method uses both labeled and unlabeled data?
Semi-supervised clustering
Outlier detection
Unsupervised clustering
Standard clustering

Question 9

Which statements are true about PCA?
t**ries to find a small number of representative variables that collectively explain most of the variance in the original data.**
is an unsupervised learning algorithm
is a variable selection method , that is, select a subset of original variables with the most predictive power
is used for dimensionality reduction

Question 10

Autoencoders are a type of unsupervised learning.

True

False

Question 11

Which loss function is used in keras for a a classification problem with n classes coded as 0-(n-1)?

sparse_categorical_crossentropy

sparse_binary_crossentropy

categorical_crossentropy

binary_crossentropy

Question 12

Before performing PCA, one must center and scale the variables.

True

False

Question 13

K-means can discover clusters of arbitrary shape.

True

False

Question 14

The proportion of variance explained by a principal component (PVE) is the variance of this principal component over the total variance of the data.

True

False

Question 15

The output of the bottleneck layer in a simple autoencoder gives a compressed representation of the input.

True

False

Question 16

An autoencoder can capture none-linear relationship between features.

True

False

Question 17

Autoencoders' compressed features are much more simple and efficient to compute than principal components.

True

False

Question 18

The k-means algorithm always converges to the optimal cluster configuration.

True

False

Question 19

Which statements are true about using PCA as a preprocessing step to reduce the dimensionality of a supervised learning problem?

principal components cannot capture non-linear relationship between the predictors

using principal components as a preprocessing step reduces the memory/disk required to store the data and speeds up the supervised learning algorithm

Using PCA as preprocessing step always improves the performance of a supervised learning algorithm

Instead of the original raw features, the first M principal components which together explain most of the variance in the feature set are used as predictors

Question 20

The first principal component is a normalized linear combination of the original variables with the largest variance.

True

False

—----------THE END—----------