Exam II

Due No due date **Points** 15.6 **Questions** 52
Available Nov 14 at 11:30am - Nov 14 at 1pm about 2 hours **Time Limit** 75 Minutes

Instructions

Note, the questions are shown sequentially one at a time, and your answers are locked after you submit each question, so be sure of your answer to a question before you move on to the next one. Please pace appropriately.

This quiz was locked Nov 14 at 1pm.

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	69 minutes	10.35 out of 15.6

Score for this quiz: **10.35** out of 15.6
Submitted Nov 14 at 12:41pm
This attempt took 69 minutes.

Question 1

0 / 0.3 pts

If 5% of your samples have incorrect labels in your available labelled data, which option is likely best to improve model accuracy?

○ Derive/predict new features from current features in your data set

ou Answered ● Get more samples (even if they are 1% in error)

○ Add/remove features

orrect Answer ○ Change your hyperparameter to avoid overfitting

Question 2

0.3 / 0.3 pts

Someone wants to build a classifier with 1,000 samples and 100 features. You know that is a large number of features for learning from so few samples. Which one of the following would you NOT want to suggest?

○ Consider regularized logistic regression with a very high lasso (L1) penalty.

Correct!

● Use a decision tree classifier because decision trees always work well with large numbers of features relative to samples

○ Perform feature selection prior to building the model

○ Suggest collecting more data

Question 3

0.3 / 0.3 pts

When cross-validation is performed in the validation set, the score of the best fitted model hyperparameters in that set is on average lower than the the score of that best fitted model on a separate test set.

○ True

Correct!

● False

Question 4

0.3 / 0.3 pts

Logistic regression can be used to classify data and, as a regression, also provide probability estimates of that classification. The output is a number between 0 and 1 which represents the probability of belonging to the class represented by 1 in the training data.

Correct!

● True

○ False

Question 5

0 / 0.3 pts

In a given binary classification problem, Out of all the positive samples in the test set, the proportion of those which are correctly identified as positive by the classifier is called...

ou Answered

● Precision

orrect Answer

○ Recall

○ F1 Score

○ Specificity

Question 6

0 / 0.3 pts

The proportion of correctly identified samples from the test samples that were identified as belonging to a particular class by the classifier is called...

orrect Answer

○ Precision

○ F1 Score

○ Recall

ou Answered

● Sensitivity

Question 7

0.3 / 0.3 pts

An error of 110 instead of 100 is weighted equally to an error giving 11 instead of 10 for which type of error metric?

○ Mean Square Error

○ Mean Absolute Error

Correct! ● Root Mean Square Logarithmic Error

Question 8

0.3 / 0.3 pts

When you want to know how well a product will work on a new person without any individual-specific training, it is better to use subject-wise cross-validation than K-fold cross-validation, because K-fold cross-validation may have an individual's data in both the training and test sets, which would contaminate the training data.

Correct! ● True

○ False

Question 9

0.3 / 0.3 pts

When you use cross-validation to select the right hyperparameters, you do not need a separate set of test data to properly measure the quality of the model because cross-validation already separates training from testing.

○ True

Correct! ● False

Question 10

0 / 0.3 pts

Support vector machines are designed to minimize the margin when finding a linear separation between classes because a linearly separating hyperplane with a smaller margin leads to better separating using new test data.

'ou Answered ● True

orrect Answer ○ False

Question 11

0.3 / 0.3 pts

The Random Forest classifier uses boosting with multiple decision trees to create a better model than a single decision tree alone.

○ True

Correct! ● False

Question 12

0.3 / 0.3 pts

Boosting combine the predictions of all models in the ensemble with equal weight

○ True

Correct!

● False

Question 13

0.3 / 0.3 pts

Boosting is applied to learners that are more likely to overfit, while bagging is often applied to weak learners (to avoid overgeneralization).

○ True

Correct!

● False

Question 14

0.3 / 0.3 pts

Here is an analogy:

"Rose" is to "Flower" as "Porsche" is to "Automobile", because the first word is a type of the second word.

"North" is to "South" as "Black" is to "White" because second word is the opposite of the first word.

and so on...

The following is analogy can be said for four important concepts in machine learning. Fill in the blank.

Classification is to _______________ in supervised learning as clustering is to dimensionality reduction in unsupervised learning.

Or more succinctly

Classification is to ________________ as clustering is to dimensionality reduction

○ Clustering

Correct!

● Regression

○ Factor Analysis

○ PCA

Question 15

0 / 0.3 pts

Select all the model hyperparameters where a smaller value leads to overfitting/higher model complexity rather than overgeneralization/simpler models

□ the maximum depth parameter for decision trees

Correct!

☑ the k in k nearest neighbors

orrect Answer

□ the slack variable in support vector machine (larger means more slack or acceptance of errors)

'ou Answered

☑ the degree of the polynomial in polynomial regression

Question 16

0.3 / 0.3 pts

Which of the following is just an ensemble method applied to a simpler classifier?

Correct!

● Random Forest

○ Support Vector Machines

○ Regularized Logistic Regression

○ K Nearest Neighbors

Question 17

0.3 / 0.3 pts

Asking a thousand people hundreds of questions about their personalities, you can use which technique to find a small set of values which may approximate personality characteristics like the "Big 5".

○ Linear regression

Correct!

● PCA

○ Support Vector Machines

○ K-Means

Question 18

0.3 / 0.3 pts

After determining the best k value for a k nearest neighbors prediction, how might the best fitting k value change if we changed the training set by incorrectly labeling 10% of all examples?

○ best k value would on average be lower

Correct!

● best k value would on average be higher

○ mathematically, the best fitting k value would stay the same regardless of adding noise

Question 19

0 / 0.3 pts

P(features) = P(feature1) * P(feature2) * P(feature 3) ...

is an assumption in which model?

orrect Answer ○ Naive Bayes

'ou Answered ● Random Forest

○ SVM

Question 20

0 / 0.3 pts

A friend in your machine learning class created a movie rating prediction system that judges how many stars (out of 5) a person would rate a movie they haven't seen yet given their ratings for other movies. They stated their rating system is 100% accurate according to their data. What is the best question to ask them?

'ou Answered ● Did you consider both sensitivity and specificity?

○ Did you use random forest or SVMs?

orrect Answer ○ Did you remember to separate your training set from your test set?

Question 21

0.3 / 0.3 pts

There are three kinds of people who build machine learning models. Person A doesn’t separate training from testing, and just fits the model to all the data, Person B uses cross-validation over the entire data set to pick the best hyperparameters and reports the quality of the model on that data set.

Person C uses cross-validation on a validation set for hyperparameters and uses a separate test set for evaluating the model.

If enough data is available, which person should you be?

○ Person B

Correct!

● Person C

○ Person A

Question 22

0 / 0.3 pts

In Gaussian Naive Bayes, select all the parameters that have to be learned from the data to create a predictive model

'ou Answered

☑ the prior probability of each feature value's likelihood

orrect Answer

□ the proportion of training data in each class

□ The mean and standard deviation for each feature, combining all classes together

Correct!

☑ The mean and standard deviation for each feature for each class

Question 23

0 / 0.3 pts

Specificity is...

'ou Answered

● Recall for the positive case

○ Precision for the negative case

orrect Answer

○ Recall for the negative case

○ Precision for the positive case

Question 24

0.3 / 0.3 pts

Accuracy is

Correct!

● the average recall across classes if the number of items in each class is the same

○ The arithmetic mean of precision and recall

○ The geometric mean of precision and recall

Question 25

0 / 0.3 pts

K-fold cross-validation will lead to lower accuracies than expected with the full training set because only (K-1)/K % of the data is being used for training (e.g. 4/5ths for K=5). The way to improve this is by increasing K.

But what is a problem with increasing K?

orrect Answer

○ K models have to be trained which takes more time as K increases

○ The number of samples in the data set may not be perfectly divisible by K

'ou Answered

● The separated test set is getting small and may bias results of the cross-validation

Question 26

0.3 / 0.3 pts

If I want to test my voice recognition software to see how well it will works on a new person it has not yet been trained for, what type of cross-validation would give me the best sense of accuracy?

○ Leave one out cross-validation

○ Stratified K-fold cross-validation

Correct! ● Subject-wise cross-validation

○ K fold cross-validation

Question 27

0.3 / 0.3 pts

Which metric is best as a single number for evaluating a terrorist detection system for airport screening?

○ Sensitvity

○ Accuracy

○ Specificity

Correct! ● F1 score (geometric mean of Sensitivity and Specificity)

Question 28

0.3 / 0.3 pts

If a potential feature does not necessarily correlate with a target, it should not necessarily be removed because

○ correlation does not imply causation

○ the best fitting line in a scatter plot with the feature and target may have a non-zero slope for a line in linear regression

Correct! ● It may still have a dependent relationship with the target

○ lack of correlation does not imply lack of causation

Question 29

0.3 / 0.3 pts

Why are new features created by sums of features or differences of features not useful in most machine learning models?

○ Summed (and similarly, subtracted) independent features tend toward a gaussian distribution according to the central limit theorem

Correct! ● Most models already add and subtract features to arrive as predictions - such a feature would be redundant

Question 30

0.3 / 0.3 pts

In a classification problem using high dimensional data (e.g. greater than 10 features) a PCA dimensionality reduction to two PCA components was performed to visually observe how separable two classes are on a scatter plot with X as PCA component 1 and Y as PCA component 2 for each data point.

If the classes are not visibly separate in the 2D plot, what does that mean for a classifier trained on all the features?

○ Overlaps in the PCA plot indicate the classes are separable when all features are used

○ They cannot be distinguished by a classifier

Correct!

● They may be separable with more features, it is inconclusive

Question 31

0.15 / 0.3 pts

Check which of the following are associated with Bagging instead of Boosting

Correct!

☑ Random forest classifiers use this technique

Correct!

☑ This is more likely to be used for models which have the potential to overfit, like decision trees with no restrictions.

□ This is more likely to be used for models which are weak learners, like decision stumps - decision trees with only one level.

'ou Answered

☑ This is a common strategy to combine multiple learners, even if they are from completely different modeling strategies (e.g. combining logistic regression and naive bayes)

Question 32

0.3 / 0.3 pts

Check which of the following are associated with Bagging instead of Boosting

Correct!

☑ the features (commonly the columns in a data set) and samples/observations (commonly the rows in a data set) may be resampled. And this can be done with or without replacement.

Correct!

☑ All estimators are weighted equally.

□ This technique is one of the reasons that some Kaggle competitions don’t allow teams to merge during competitions (e.g. team #2 and #3 join together)

Question 33

0.3 / 0.3 pts

It is important to not remove features that are statistically independent of target values because they might be correlated to the target, and thus useful for prediction.

○ True

Correct!

● False

Question 34

0 / 0.3 pts

Which is NOT a reasonable metric for measuring the quality of clustering techniques

○ Compare the distances of samples from the centroid of a cluster to the average distance between centroids

○ Compare what fraction of sample pairs that are known to be in the same supervised group end up on the same cluster

'ou Answered

● Compare the distances between pairs of samples within cluster to pairs of samples between clusters

orrect Answer

○ Compare the number of samples in each cluster. All clusters should have roughly the same number of samples

Question 35

0.3 / 0.3 pts

Bayes rule can be straightforward to use to iteratively update estimates as more data comes in. This is because the likelihood produced from previously acquired data can be used as a posterior for estimates using newly acquired data.

○ True

Correct!

● False

Question 36

0.15 / 0.3 pts

If variable A has 4 options, B has 3 options, and C has 2 options, match the following probability functions with the number of independent probability values necessary to represent it as a table (ones that are not "1 - other values").

Hint: Think about the size of likelihood and prior tables in Bayes net examples.

'ou Answered **P(A|B) P(B|C) P(C)** 23 ▾

Correct Answer 14

Correct! **P(A|B,C)** 18 ▾

Correct! **P(A|C) P(C|B) P(B)** 11 ▾

'ou Answered **P(A,B,C)** 14 ▾

Correct Answer 23

Question 37

0.3 / 0.3 pts

When fully specifying a Bayesian network, priors and likelihood of discrete variables require functions while those using continuous variables can be defined with shorter tables of probabilities.

○ True

Correct! ● False

Question 38

0.3 / 0.3 pts

Which of the following Bayes nets does not represent a dependency between A and C (assuming the state of B is unknown)

Correct!

● A --> B <-- C

○ A <-- B --> C

○ A <-- B <-- C

○ A --> B --> C

Question 39

0.3 / 0.3 pts

Which of the following Bayes nets implies a conditional dependency between A and C when the state of B is known?

○ A <-- B --> C

○ A <-- B <-- C

○ A --> B --> C

Correct!

● A --> B <-- C

Question 40

0 / 0.3 pts

Which of the following are true of Markov models as opposed to Bayes nets

Correct!

☑ Links represent transition probabilities 'ou Answered

☑ Links represent dependent relationships

Correct!

☑ Nodes are discrete states of a variable

'ou Answered

☑ Nodes are variables

orrect Answer

☐ Generally used for sequential data

Question 41

0 / 0.3 pts

The Q in Q-learning for reinforcement learning is best described as

○ The discount factor

orrect Answer

○ The sum of future expected rewards

'ou Answered

● the reward prediction error quotient

○ The reward signal from the environment

Question 42

0 / 0.3 pts

Why do epsilon policies and softmax policies exist in reinforcement learning? Why not always just pick the action with the highest expected future reward?

'ou Answered

● Because future rewards are not as valuable as current rewards

○ Because learning happens too quickly if only the best options are chosen each time orrect Answer

○ It concerns the tradeoff between exploration and exploitation

Question 43

0.3 / 0.3 pts

In reinforcement learning, what is the nature of the relationship between the state value function, V(s), and the state-action value function, Q(s,a)?

Correct!

● It is possible to relate Q and V mathematically if you know the probabilities of actions given the states

○ V can be derives from Q regardless of the policy

○ Q and V are related, but there is no clear mathematical relationship

○ Q in unrelated to V

Question 44

0.3 / 0.3 pts

In which situation would the reward prediction error be negative

Correct!

● You received a worse reward than you expected

○ You receive a lighter punishment than you expected

○ You received a better reward than you anticipated

○ You received a reward when you expected a punishment

Question 45

0 / 0.3 pts

Which reinforcement learning parameter should gradually decrease as more is learned about the environment to make learning more stable?

○ Q(s,a)

○ slack variable

orrect Answer ○ Learning rate

○ reward

'ou Answered ● Discount factor

Question 46

0.3 / 0.3 pts

Which of the following is not an explicit part of the standard Q-learning equation?

○ Reward prediction error

Correct! ● the policy function

○ Temporal discounting

○ a state-action value function

○ a learning rate

Question 47

0.15 / 0.3 pts

Which is true of DBSCAN and other density-based clustering techniques and not of K-mean clustering

☐ It is very sensitive to starting conditions

Correct!

☑ Does not work well with clusters that differ greatly in density of samples

Correct!

☑ Clusters can be of arbitrary shapes

'ou Answered

☑ Cluster are expected to be spherical in shape

Question 48

0.3 / 0.3 pts

The adjusted RAND index is a useful method to score the quality of a clustering algorithm because it does not require knowing ahead of time which pairs of samples belong in the same cluster

○ True

Correct!

● False

Question 49

0.3 / 0.3 pts

Label spreading and label propagation are semisupervised learning techniques. In particular they are most useful when...

○ Most useful when the frequency of classes in a classifier is imbalanced (e.g. fall detection, terrorist detection, etc)

Correct!

● There is a great deal of unlabeled samples but only a few labeled samples

○ When there are an excessively large number of features compared to samples

○ When there is a large amount of error in the class labels in the training set

Question 50

0 / 0.3 pts

If you want to use PCA to preprocess the pixel value when performing digit recognition for classification using gaussian naive bayes, which is a more likely value to use to get the highest classification accuracy? (assume 8x8 pixel images)

Correct!

☑ 10-63 PCA dimensions - enough to capture the structure of the signal, and throw out the noise

'ou Answered

☑ 2 PCA dimensions - also the right amount to visualize on a 2D graph

□ 64 dimensions - you will always do better with all the dimensions of your data set represented

Question 51

0.3 / 0.3 pts

Dimensionality reduction is useful to lower the number of features in a systematic way. Which is NOT a reason why it may be useful to reduce the dimensionality of your feature set?

○ collapse redundant features to simplify the model

○ Remove noise

○ Visualize your state space in 2D or 3D

○ to transform features to understand the "latent variables" or underlying causes in your observations

○ Speed model learning by using fewer features

Correct!

● To project the data into a higher dimensional space to create a linear separating hyperplane

○ to compress the signal

Question 52

0.3 / 0.3 pts

In a PCA analysis of 100 questions related to basketball ability, it would not be possible to perfectly pick out factors like "height" and "weight" because they are not orthogonal, and PCA requires that vectors be orthogonal

Correct!

● True

○ False

Quiz Score: **10.35** out of 15.6