Final Exam

Due No due date	Points 20	Questions 50	
Available Dec 10 at 1	0:30am - Dec 10) at 11:42am about 1 hour	Time Limit 110 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 Dec 10 at 11:42am.

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	21 minutes	15.4 out of 20

Score for this quiz: **15.4** out of 20 Submitted Dec 10 at 10:55am This attempt took 21 minutes.

	Question 1	0 / 0.4 pts
	If 5% of your samples have incorrect labels in your available labe which option is likely best to improve model accuracy?	lled data,
	Add/remove features	
	Derive/predict new features from current features in your data set	
orrect Answer	Change your hyperparameter to avoid overfitting	
ou Answered	Get more samples (even if they are 1% in error)	

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

	Question 3 0 / 0	.4 pts
	When cross-validation is performed in the validation set, the score of the fitted model hyperparameters in that set is on average higher than the t score of that best fitted model on a separate test set.	
orrect Answer	○ True	
ou Answered	False	

Question 4 0.4 / 0.4 pts

"Logistic regression" is used for predicting numeric targets rather than for performing classification because regression is used for predicting numeric

Correct!

information while classification is for predicting discrete classes. The commonly used term is "logistic classifier" for predicting classes. True False

0.4 / 0.4 pts **Question 5** In a given binary classification problem, Out of all the negative samples in the test set, the proportion of those which are correctly identified as negative by the classifier is called... Specificity Recall F1 Score

0.4 / 0.4 pts **Question 6** The proportion of correctly identified samples from the test samples that were identified by the model as belonging to a particular class by the classifier is called... F1 Score Recall Sensitivity

Precision

Precision

An error of 110 instead of 100 is weighted equally to an error giving 11 instead of 10 for which type of error metric? Root Mean Square Logarithmic Error Mean Absolute Error Mean Square Error

Question 8 0 / 0.4 pts

When you want to know how well a product will work on a person after it has been trained specifically with that person's data, subject-wise cross-validation is superior to K-fold cross-validation with the individual's data, because K-fold cross-validation may have an individual's data in both the training and test sets, which would contaminate the training data.

'ou Answered

True

orrect Answer

False

Question 9 0.4 / 0.4 pts

When you use cross-validation to select the right hyperparameters, you still need a separate set of test data outside of that used for hyperparameter selection to properly measure the quality of the model.

Question 10 Support vector machines are designed to maximize the margin when finding a linear separation between classes because a linearly separating hyperplane with a larger margin leads to better separating of classes when using new test data. Orrect Answer True True False

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

Question 12 0.4 / 0.4 pts

Boosting combines the predictions of all models but unlike bagging does not weight each model equally

Correct!

Question 13	0.4 / 0.4 pts
Boosting is applied to learners that are roften applied to weak learners (to avoid	, , , , , , , , , , , , , , , , , , , ,
True	
False	

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

12/18/2019

Regression

Question 15 Select all the model hyperparameters where a larger value leads to overfitting/higher model complexity rather than overgeneralization/simpler models orrect Answer the k in k nearest neighbors the maximum depth parameter for decision trees the degree of the polynomial in polynomial regression the slack variable in support vector machine (larger means more slack or acceptance of errors)

Which of the following is just an ensemble method applied to a simpler classifier? K Nearest Neighbors Support Vector Machines

- Random Forest
- Regularized Logistic Regression

Question 17

0.4 / 0.4 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".

K-Means

Correct!

- PCA
- Support Vector Machines
- Linear regression

Question 18

0 / 0.4 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?

'ou Answered

- best k value would on average be lower
- mathematically, the best fitting k value would stay the same regardless of adding noise

orrect Answer

best k value would on average be higher

Question 19 O.4 / 0.4 pts P(features) = P(feature1) * P(feature2) * P(feature 3) ... is an assumption in which model? Random Forest SVM Naive Bayes

Question 20 0.4 / 0.4 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?

Did you consider both sensitivity and specificity?

Correct!

- Did you remember to separate your training set from your test set?
- Did you use random forest or SVMs?

Question 21 0.4 / 0.4 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

Person A

Person C

Question 23

Sensitivity is...

Recall for the negative case

Precision for the positive case

Precision for the negative case	
Recall for the positive case	

Question 24 0.4 / 0.4 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?

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

Correct!

- K models have to be trained which takes more time as K increases
- The separated test set is getting small and may bias results of the cross-validation

Question 25 0.4 / 0.4 pts

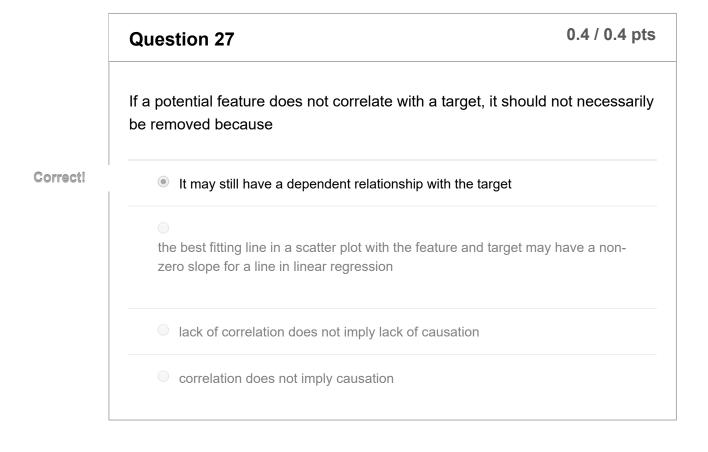
If I want to test my voice recognition software to see how well it will works on a new person it has not yet been trained for, what type of cross-validation would give me the best sense of accuracy?

- Leave one out cross-validation
- K fold cross-validation

Correct!

- Subject-wise cross-validation
- Stratified K-fold cross-validation

Which metric is best as a single number for evaluating a terrorist detection system for airport screening? Accuracy Specificity Sensitvity F1 score (geometric mean of Sensitivity and Specificity)



Question 28 0.4 / 0.4 pts

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

Correct!



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



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

Question 29 0.4 / 0.4 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 <u>trained on all the features</u>?



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

Correct!

- They may be separable with more features, it is inconclusive
- They cannot be distinguished by a classifier

Question 30

0.4 / 0.4 pts

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

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

Random forest classifiers use this technique

Correct!



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)

Correct!



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

Question 31 0.4 / 0.4 pts

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

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.

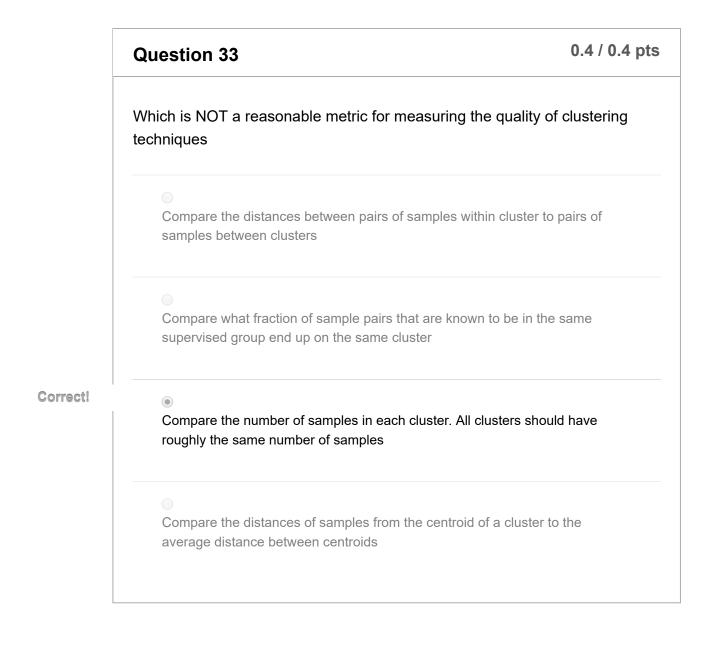
All estimators are weighted equally.

Correct!

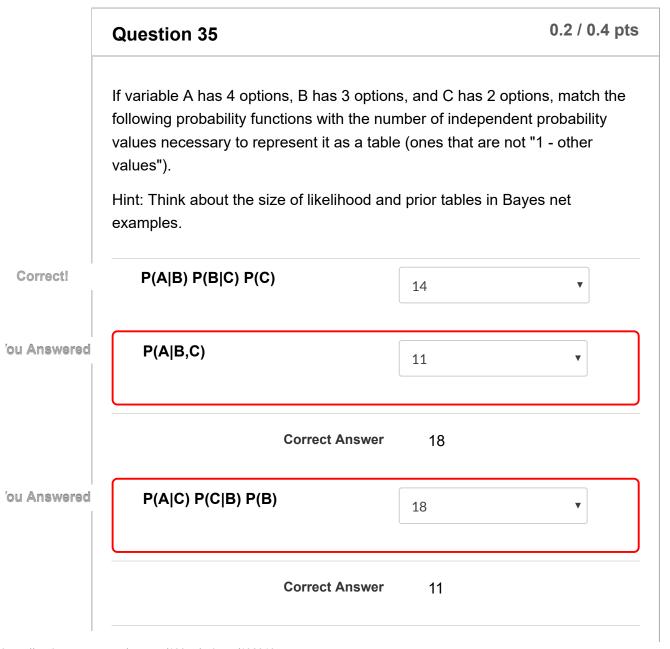


This technique is one of the reasons that some Kaggle competitions don't allow teams to merge during competitions (e.g. team #2 and #3 join together)

Question 32 It is important to not remove features that are uncorrelated to target values because they might still be statistically dependent to the target values in a nonlinear way, and thus useful for prediction. Orrect Answer True False



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 True False



P(A,B,C)

23

Question 36

0 / 0.4 pts

When fully specifying a Bayesian network, priors and likelihood of discrete variables require probability tables while those using continuous variables must use functional forms for their definition.

orrect Answer

True

'ou Answered

False

Question 37

0.4 / 0.4 pts

Which of the following Bayes nets represents a dependency between A and C (assuming the state of B is unknown)

Correct!

- ✓ A <-- B --> C
- A --> B <-- C

Correct!

✓ A <-- B <-- C
</p>

Correct!

✓ A --> B --> C

Question 38

0.4 / 0.4 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</p>
- A <-- B <-- C
- A <-- B --> C
- A --> B --> C

Question 39
0 / 0.4 pts

Which of the following are true of Bayes nets as opposed to Markov models
'ou Answered
Nodes are discrete states of a variable
Links represent transition probabilities
orrect Answer
Nodes are variables
'ou Answered
Generally used for sequential data
Correct!
Links represent dependent relationships

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

The reward signal from the environment

The discount factor

The sum of future expected rewards

the reward prediction error quotient

Why do epsilon policies and softmax policies exist in reinforcement learning? Why not always just pick the action with the highest expected future reward? Because learning happens too quickly if only the best options are chosen each time Because future rewards are not as valuable as current rewards It concerns the tradeoff between exploration and exploitation

	Question 42	0.2 / 0.4 pts
	In which situation would the reward prediction error be positive	
	You received a worse reward than you expected	
Correct!	✓ You receive a lighter punishment than you expected	
orrect Answer	You received a reward when you expected a punishment	
	You received a worse reward than you anticipated	

Question 43	0.4 / 0.4 pts

	Which reinforcement learning parameter should gradually decrease as more is learned about the environment to make learning more stable?
	Discount factor
Correct!	Q(s,a)
	Learning rate
	 slack variable
	reward

	Question 44	0.4 / 0.4 pts
	Which of the following is not an explicit part of the standard equation?	Q-learning
	a learning rate	
	a state-action value function	
	Reward prediction error	
	Temporal discounting	
Correct!	the policy function	

Question 45 O.2 / 0.4 pts Which is true of K-means clustering as opposed to DBSCAN and other density-based clustering techniques?

Question 46 The adjusted RAND index is a useful method to score the quality of a clustering algorithm however it requires knowing ahead of time which pairs of samples belong in the same cluster orrect Answer orrect Answer False

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

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

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

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

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

Question 48 0.4 / 0.4 pts

On a limited set of data, 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)

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

Correct!

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

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

Question 49 0.4 / 0.4 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?

Remove noise

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

Speed model learning by using fewer features

	Question 50	0.4 / 0.4 pts
	In a PCA analysis of 100 questions related to basket be possible to perfectly pick out factors like "height" they are not orthogonal, and PCA requires that vector	and "weight" because
orrect!	True	
	False	

Quiz Score: 15.4 out of 20