Microsoft: DAT210x Programming with Python for Data Science

Bookmarks

- Start Here
- 1. The Big Picture
- 2. Data And Features
- 3. Exploring Data
- 4. Transforming Data
- 5. Data Modeling
- 6. Data Modeling II
- 7. Evaluating Data
- Course Wrap-up

Final Quiz
Quiz

Final Project

Course Wrap-up > Final Quiz > Final Quiz

Bookmark

Final Quiz Notes

Take a deep breath, you earned it!

This is a comprehensive quiz, designed to coarsely test your knowledge of the numerous topics covered during this course. Good luck!

Quiz Question

(1/1 point)

If you were given a higher dimensionality data set you wanted to visualize, which of the following techniques might you use to carry that out (select all that apply):

☐ K-Nearest Neighbors

☐ Cross Validation

☑ Principal Component Analysis ✓

☑ Parallel Coordinates ✓

Wrap-up

Post-course Survey
Survey

□ Histograms

✓

EXPLANATION

Histograms aren't going to help you visualize a higher dimensionality data set.

Cross validation doesn't come until later in the pipeline.

And K-Neighbors is a technique used for modeling.

You have used 2 of 2 submissions

Quiz Question

(1/1 point)
The main advantage of random forest over decision trees are that:

● They allow you to bootstrap, thereby reducing the potential of overfitting ✓

○ They are a lot faster, since you're growing more trees

○ They are more configurable than regular decision trees because Sci-Kit learn exposes more optional parameters

○ They allow you to use your entire dataset for training and testing

EXPLANATION

Random forest were created for the sole purpose of reducing the potential of decision tree overfitting. There might be other benefits, however that is the most important advantage they offer.

You have used 1 of 2 submissions

Quiz Question

(1/1 point)
If successful, Panda's read_html() method returns:

○ A dataframe containing the contents of the html table(s) on the webpage

○ A dataframe containing the first html table on the webpage

○ An NDArray of dataframes, one per table, as seen on the webpage

● A Python List of dataframes, one per table, as seen on the webpage ✓

EXPLANATION

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_html.html

You have used 1 of 2 submissions

Quiz Question

(1/1 point)

Confusion matrices help you calculate (check all that apply):

☑ How many predictions of a target your model guessed ✓

☑ The number of true positives + false negatives ✓

☑ The number of false negatives + false positives ✓

☑ How many observations per target exist in your dataset ✓

✓

EXPLANATION

Confusion matrices display states on all of this for you.

You have used 2 of 2 submissions

Quiz Question

(1/1 point)

Given the feature CardSuits that contained Hearts, Spades, Diamonds, and Clubs, before running SciKit-Learn modeling, which of the following is most applicable?

● Use pandas to crate dummy features ✓

○ Use the values as a textual feature

○ Convert them using CardSuits.astype('category').cat.codes

○ Properly encode them for processing using an ordered category

○ Use CardSuits.Map() to represent them as numerals, so that SciKit-Learn can understand and process them correctly

You have used 1 of 2 submissions

Quiz Question

(1/1 point)

Every one of the steps used in your pipelines must:

○ Implement .fit() and .transform(), so that one step's results are fed into the next.

○ Have parameters defined for it using the .set_params() method of the pipeline

● Implement .fit() to process incoming data ✓

○ Alter the incoming data in some way

EXPLANATION

The steps don't need to alter anything. Nor do they all need .transform(), for example the last step does not use that. They also do not all need parameters.

You have used 2 of 2 submissions

Quiz Question

(1/1 point)

Under what circumstances would you use principal component analysis instead of isomap?

○ When it appears that linearity is preserved in local neighborhoods within your dataset.

○ When you have a very large dataset and need to quickly reduce its dimensionality.

○ When you want to preserve covariance, and you want preserve--as much as possible--the interpoint distances in your reduced dimensionality representation.

● When it appears that linearity is preserved over your entire dataset. ✓

EXPLANATION

Both PCA and Isomap preserve covariance, Isomap just does it in a non-linear way.

In fact, the two algorithms are really quite similar. The source of Isomap's nonlinearity is the way interpoint distances are calculated. Rather than generally using Euclidean distances between samples like PCA, Isomap uses those distances only for points considered neighbors. The rest of interpoint distances are calculated by finding the shortest path through the graph on the manifold.

You have used 2 of 2 submissions

Quiz Question

(1/1 point)

Support vector machines are:

● A algorithm that relies on linear decision boundaries to do potentially non-linear regression or classification. ✓

○ A dimensionality reduction algorithm similar to PCA, as they both reduce the distance from your samples to the best-fit line

○ A clustering technique for unsupervised learning

○ A pre-processing technique

You have used 1 of 2 submissions

Quiz Question

(1/1 point)

Select the major categories of machine learning from the list below

□ Deep Model Learning

☑ Reinforcement Learning ✓

□ Reverse Learning

☑ Supervised Learning ✓

□ Unstructured Learning

☑ Unsupervised Learning ✓

✓

You have used 1 of 2 submissions

Quiz Question

(1/1 point)

Which of the following statements about K-Means and K-Neighbors is the most accurate?

○ The K-Means and K-Nearest algorithms group records by similarity.

○ One is a supervised clustering algorithm, the other is an unsupervised classification algorithm.

● They are distance based algorithms susceptible to feature scaling ✓

○ Their training times are proportional to the number of samples in your dataset

EXPLANATION

"One is a supervised clustering algorithm, the other is an unsupervised classification algorithm."

This was particularly tricky. To correct, it'd have to read, one is a supervised classification algorithm, and the other is an unsupervised clustering algorithm.

You have used 1 of 2 submissions

Quiz Question

(1/1 point)

First, figure out what **type** of features the items in the following list fall into:

1. Dish Type
2. Yelp Star Rank
3. Altitude

Order the items above--if necessary--so that they are sorted by their ***feature type*** in alphabetical order:

● Altitude, Dish Type, Yelp Star Rank ✓

○ Altitude, Yelp Star Rank, Dish Type

○ Dish Type, Altitude, Yelp Star Rank

○ Dish Type, Yelp Star Rank, Altitude

○ Yelp Star Rank, Dish Type, Altitude

EXPLANATION

Dish Type: Nominal - Cup, Bowl, Plate, etc.

Yelp Star Rank: Ordinal - 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5

Altitude: 400.75 meters above sea-level.

You have used 1 of 2 submissions

Quiz Question

(1/1 point)
Under which set of circumstances would it make more sense to use K-Neighbors over Decision Trees?

● You have a lot of data, it has very irregular decision boundaries, and execution speed isn't of concern. ✓

○ You have a lot of data, the satellite executing the algorithm doesn't have a lot of memory, and execution speed isn't of concern.

○ You have a small amount of noisy data, but it has a well-formed decision boundary, and execution speed is of great concern.

○ You have a small amount of data, it has a well-formed decision boundary, and execution speed is of great concern.

EXPLANATION

If you have a lot of data with irregular decision boundaries, and execution speed isn't of concern, K-Neighbors is your algorithm.

Since K-Neighbors stores all the data, if you're limited by space, a large dataset will not do.

If you have a small amount of data, K-Neighbors will decrease in accuracy; but the well-formed decision boundary counters that. However execution speed being of great concern, walking a short decision tree of IF-Statements would actually be faster.

You have used 1 of 2 submissions