Unsupervised Exam [Timed] (Version : 0)

TEST

Correct Answer

Answered in 96.166666666667 Minutes

Uploaded File : No File Uploaded

Question 1/11

The practical questions of this Exam should be answered using the attached UFO.csv file. This file contains information pertaining to various UFO sightings over the last 70 years. Use default parameters for the practical questions unless otherwise specified.

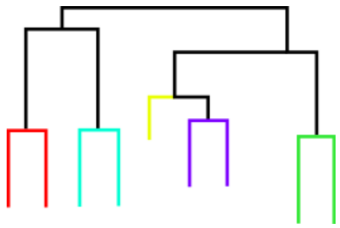
Which of the following models is best suited to classify 10,000 rows of unlabelled Twitter data?

1. Naïve Bayes
2. Logistic regression
3. SVM
4. Linear regression

- [x] None of these
- [ ] 1, 3
- [ ] 1, 2, 3
- [ ] 1, 2, 3, and 4

Question 2/11

What is the appropriate choice for the number of clusters, given this dendrogram?



- [ ] 2
- [ ] 5
- [x] 4
- [ ] 3

Question 3/11

Using K-means and the Elbow method, what is an appropriate number of geospatial clusters for this data?

- [x] 4
- [ ] 3
- [ ] 5
- [ ] 6

Question 4/11

Cluster the data geospatially using K-means, specifying 5 clusters. The largest cluster contains approximately what portion of the data?

- [x] 40%
- [ ] 35%

- [ ] 30%
- [ ] 50%

Question 5/11

If we plot these clusters geographically - which country appears to have the densest population of UFO sightings?

- [ ] UK
- [x] USA
- [ ] Canada
- [ ] New Zealand

Question 6/11

Plot a dendrogram of the years of UFO sightings using the ward method, using only the first 1000 entries. What is the optimal number of clusters judging by this dendrogram?

- [x] 2
- [ ] 5
- [x] 4
- [ ] 3

Question 7/11

If we attempt to visualise a dendrogram of the years of UFO sightings of the full data set, what error will we most likely get?

- [ ] ValueError
- [ ] ParseError
- [x] MemoryError
- [ ] TypeError

Question 8/11

Using a TfidVectorizer (with English stopwords), what are the 3 “most important”*unique words* found in the comments column?

- [ ] sky, moving, craft
- [ ] bright, light, moving
- [x] light, object, sky
- [ ] bright, object, orange

Question 9/11

Using the following vectorizer:

vectorizer = TfidfVectorizer(max_features=20, stop_words='english'),

fit-transform it to the comments column. Then use PCA (with 10 components and a random_state of 1) to determine the percentage of the variance that the first two principal components explain.

- [x] 22%

- [ ] 25%
- [ ] 34%
- [ ] 17%

Question 10/11

Which of the following is not a feature selection/extraction technique?

- [ ] PCA
- [ ] Multidimensional scaling
- [x] Pipelines
- [ ] Term frequency-inverse document frequency

Question 11/11

In K-means clustering, the k parameter does not need to be defined initially.

- [ ] True
- [x] False