

🔗 Google Discussions



Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

[Go to Exam](#)

📄 EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 14 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 14

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

- A. There are very few occurrences of mutations relative to normal samples.
- B. There are roughly equal occurrences of both normal and mutated samples in the database.
- C. You expect future mutations to have different features from the mutated samples in the database.
- D. You expect future mutations to have similar features to the mutated samples in the database.
- E. You already have labels for which samples are mutated and which are normal in the database.

[Show Suggested Answer](#)

by [jvg637](#) at March 11, 2020, 6:24 p.m.

Comments

Type your comment...

[Submit](#)

🗨️ [jvg637](#) Highly Voted 5 years, 1 month ago

I think that AD makes more sense. D is the explanation you gave. In the rest, A makes more sense, in any anomaly detection algorithm it is assumed a priori that you have much more "normal" samples than mutated ones, so that you can model normal patterns and detect patterns that are "off" that normal pattern. For that you will always need the no. of normal samples to be much bigger than the no. of mutated samples.

   upvoted 73 times

  **AmitK121981** 4 months, 3 weeks ago

as per chatGPT, it can be different (C) - that's how unsupervised anomaly detection works - as long as they are different than "normal" tissues, they would be detected

   upvoted 1 times

  **BigQuery** 3 years, 5 months ago




Guys its A & C.

Anomaly detection has two basic assumptions:

->Anomalies only occur very rarely in the data. (a)

->Their features differ from the normal instances significantly. (c)

link -> <https://towardsdatascience.com/anomaly-detection-for-dummies-15f148e559c1#:~:text=Unsupervised%20Anomaly%20Detection%20for%20Univariate%20%26%20Multivariate%20Data.&text=Anomaly%20detection%20has%20two%20basic,from%20the%20normal%20instances%20significantly.>

   upvoted 22 times




  **szefco** 3 years, 4 months ago

I don't agree on C. Anomaly detection assumes "Their features differ from the NORMAL INSTANCES significantly" and in the C option you have:

"You expect future mutations to have different features from the MUTATED SAMPLES IN THE DATABASE".

IMHO Answer D fits better: "D. You expect future mutations to have similar features to the mutated samples in the database." - in other words: Expect future anomalies to be similar to the anomalies that we already have in database

   upvoted 27 times

  **jvg637** Highly Voted  5 years, 1 month ago

A instead of B:

"anomaly detection (also outlier detection[1]) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data

   upvoted 21 times

  **vosang5299** Most Recent  2 weeks, 2 days ago

Selected Answer: AD

A. There are very few occurrences of mutations relative to normal samples. This is a strong characteristic supporting unsupervised anomaly detection. Anomaly detection methods often work well when the "normal" class is the majority and anomalies are rare outliers. The mutated samples, being few, would likely appear as anomalies relative to the cluster of normal samples.

D. You expect future mutations to have similar features to the mutated samples in the database. This characteristic supports unsupervised anomaly detection. If future mutations are expected to share similar characteristics with the existing mutated samples (even if they are rare), an anomaly detection method could potentially learn the pattern of the normal samples and flag anything significantly different (including the mutated samples) as an anomaly.

   upvoted 1 times

  **Parandhaman_Margan** 1 month, 3 weeks ago

Selected Answer: AC

Unsupervised anomaly detection. Characteristics are few anomalies and future mutations different. So A and C.

   upvoted 2 times

  **LP_PDE** 3 months, 1 week ago

Selected Answer: AC

AC. Answer C. Unsupervised anomaly detection methods are particularly useful when you don't have labeled examples of the anomalies you're trying to detect. Why not D, if future mutations are similar to existing ones, a supervised model trained on labeled examples of known mutations would likely be more accurate in classifying new samples.

   upvoted 3 times

  **cqrm3n** 3 months, 2 weeks ago

Selected Answer: AC

The answer should be A and C.

Unsupervised anomaly detection is useful when labels are unavailable, or when anomalies are rare and distinct.

Hence A is definitely correct because anomaly detection excels when anomalies are rare compared to normal data.

I think C is correct because by adding new mutation data that is similar to the existing mutation data, the model will learn in a

broader sense on what constitutes to 'mutation', and it leads to a better generalization. If the new data is too similar to the existing mutation data (answer D), the model might overfit to those specific examples. However, the new data should still share some fundamental characteristic to the existing data so that the model can recognize them as belonging to the same anomaly category.


   upvoted 2 times

  **kumar34** 4 months, 1 week ago

Selected Answer: AC

I think it's A&C. For an anomaly detection model, the ratio of normal vs abnormal is expected to be high. 'C' because the model is expected to be adaptive meaning the model detects the abnormal features that can be different from the abnormal features currently being trained on.

   upvoted 2 times

  **SamuelTsch** 6 months, 2 weeks ago

Selected Answer: AC

The keyword is unsupervised anomaly detection. So A is correct. We think and should ensure the majority of data represents 'normal'. Unsupervised methods are good for detecting unknown patterns. Thus C could be correct.

   upvoted 2 times

  **SamuelTsch** 6 months, 2 weeks ago

I correct my answer. AD should be better. Unsupervised method is usually used for grouping the data. So, if the future mutations have similar features to the mutated samples, our trained model should group it into anomalies even though no label exists.

   upvoted 1 times

  **hendrixlives** 7 months, 1 week ago

Selected Answer: AD

AD: to use unsupervised anomaly detection the anomalies a) must be rare b) they must differ from the NORMAL. So...

A: mutated samples must be scarce compared to normal tissue.

D: yes, we expect the future mutated samples to have similar features to the mutated samples currently in the database.

Why not C? If I train my model with mutated samples with specific characteristics, I do not expect it to find different mutations. In the future, when new mutations appear, I would retrain my model including those new samples.

   upvoted 4 times

  **MaxNRG** 7 months, 1 week ago

Selected Answer: AD

Anomaly detection has two basic assumptions:

*Anomalies only occur very rarely in the data.

*Their features differ from the normal instances significantly.

Anomaly detection involves identifying rare data instances (anomalies) that come from a different class or distribution than the majority (which are simply called "normal" instances). Given a training set of only normal data, the semi-supervised anomaly detection task is to identify anomalies in the future. Good solutions to this task have applications in fraud and intrusion detection.

The unsupervised anomaly detection task is different: Given unlabeled, mostly-normal data, identify the anomalies among them.

<https://www.science.gov/topicpages/u/unsupervised+anomaly+detection>

A because "Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal", B is for Supervised anomaly detection

https://en.wikipedia.org/wiki/Anomaly_detection

   upvoted 4 times

  **gudiking** 7 months, 1 week ago

Selected Answer: AD

A - anomaly detection is used for detecting rare events, meaning it is expected that there are much less of those than of normal ones.

D - you expect the future mutations to be similar to the mutations you already have, so that you can detect them (pattern recognition)

   upvoted 2 times

  **jkhong** 7 months, 1 week ago

Selected Answer: AD

A makes sense

C and D compares future mutations to mutated samples in database

The question is pretty badly worded... If we were to run a full unsupervised anomaly detection over the entire dataset, C and D will be true, since some future mutations may be similar to current mutations and some will be significantly different to current mutations.

The question is concerning "What will not always occur in unsupervised anomaly detection, and which method is best for

The question is suggesting "labelling" tissue samples using unsupervised anomaly detection, and subsequently using the labels with a supervised algorithm to classify future samples. If this interpretation of the question is correct, then D makes sense

   upvoted 4 times

  **korntewin** 7 months, 1 week ago

Selected Answer: AD

The answer should be AD.

A, anomaly should have a little amount, if there are many samples then we should do classification instead, because unsupervised will give a lot of false positive.

D, the future anomaly should be of the same distribution as present anomaly! or else our anomaly detection will not be generalize to the future feature.



   upvoted 2 times

  **samdhimal** 7 months, 1 week ago

A. There are very few occurrences of mutations relative to normal samples. This characteristic is supportive of using an unsupervised anomaly detection method, as it is well suited for identifying rare events or anomalies in large amounts of data. By training the algorithm on the normal tissue samples in the database, it can then identify new tissue samples that have different features from the normal samples and classify them as mutated.

D. You expect future mutations to have similar features to the mutated samples in the database. This characteristic is supportive of using an unsupervised anomaly detection method, as it is well suited for identifying patterns or anomalies in the data. By training the algorithm on the mutated tissue samples in the database, it can then identify new tissue samples that have similar features and classify them as mutated.

   upvoted 2 times

  **azmiozgen** 7 months, 1 week ago

Selected Answer: AD

D should be correct. You expect future samples will correlate with the training samples. That's the whole point of learning procedure. If you do not expect that they have similar features, then why would you use features in the training samples in the first place? A is also correct, since anomaly labels would be seen rarely.

   upvoted 5 times

  **rocky48** 7 months, 1 week ago

Selected Answer: AD

A. There are very few occurrences of mutations relative to normal samples. This characteristic is supportive of using an unsupervised anomaly detection method, as it is well suited for identifying rare events or anomalies in large amounts of data. By training the algorithm on the normal tissue samples in the database, it can then identify new tissue samples that have different features from the normal samples and classify them as mutated.

D. You expect future mutations to have similar features to the mutated samples in the database. This characteristic is supportive of using an unsupervised anomaly detection method, as it is well suited for identifying patterns or anomalies in the data. By training the algorithm on the mutated tissue samples in the database, it can then identify new tissue samples that have similar features and classify them as mutated.

   upvoted 2 times

  **Nittin** 9 months ago

Selected Answer: AC

A. There are very few occurrences of mutations relative to normal samples.

Anomaly detection is well-suited for situations where anomalies (in this case, mutations) are rare compared to the normal cases. When the dataset is highly imbalanced, with far fewer mutated samples than normal samples, anomaly detection can be used to identify these rare cases as outliers or anomalies.

C. You expect future mutations to have different features from the mutated samples in the database.

Unsupervised anomaly detection works under the assumption that anomalies (mutations) will differ significantly from the majority of the data (normal samples). If future mutations are expected to exhibit different features, this method can help detect those anomalies as deviations from the normal samples.

   upvoted 3 times

[Load full discussion...](#)

Platform

- > Home
- > Examtopics PRO
- > All Exams
- > Training Courses

