[redacted]

Iniciada	sexta-feira, 3 de novembro de 2023 às 16:51
Estado	Terminada
Terminada	sexta-feira, 3 de novembro de 2023 às 17:29
Tempo gasto	38 minutos 5 segundos
Nota	**16,08** num máximo de 20,00 (**80,42**%)

[redacted]

Which of the following is not a step in spam detection using the Bag of Words Model?

- a. Training a machine learning model to classify emails as spam or non-spam based on their word count vectors.
- b. Removing stop words from the email text before representing it as a vector of word counts.
- c. Representing each email as a vector of word counts.
- d. Using the trained model to predict whether or not a new email is spam based on its word count vector.

Which of the following is the key difference between a collective anomaly and a point anomaly in anomaly detection?

a. A collective anomaly is an anomaly that occurs in a single data stream, while a point anomaly is an anomaly that occurs across multiple data streams.

b. A collective anomaly is an anomaly that is caused by a single underlying event, while a point anomaly is an anomaly that is caused by multiple underlying events.

c. A collective anomaly is a group of data points that deviate significantly from the expected behavior, while a point anomaly is a single data point that deviates significantly from the expected behavior.

d. All of the above.

Which of the following best describes the "curse of dimensionality" in the context of data analysis and machine learning?

a. The difficulty of finding patterns in high-dimensional data.
b. The requirement of more data to train machine learning models in high-dimensional spaces.
c. All of the options
d. The tendency of the volume of data to increase exponentially as the number of dimensions increases.

Pergunta **4**

Respondida

Nota: 1,00 em 1,00

Which of the following types of clustering algorithms is most suitable for identifying clusters of varying shapes and sizes in a dataset with noisy data points?

a. Spectral clustering
b. Partitioning clustering
c. Hierarchical clustering
d. Density-based clustering

Which of the following statements is true when comparing Principal Component Analysis (PCA) and Autoencoders for anomaly detection?

a. PCA is more robust to noise than Autoencoders.

b. PCA is a supervised learning technique, while Autoencoders are an unsupervised learning technique.

c. Autoencoders can learn more complex relationships between the data than PCA.

d. PCA is a linear dimensionality reduction technique, while Autoencoders are a non-linear dimensionality reduction technique.

How does FastText represent words differently from the BOW model?

Para o BOW, um texto como um email ou um SMS não possui estrutura semântica ou sintactica, é apenas um agrupamento sem ordem de palavras, e sobre as quais apenas vai analisar a frequência com que ocorrem. Por causa disto, apesar de textos que possam ser de teor dramaticamente diferente, se partilharem palavras iguais em grande proporção, serão considerados como similares.

No FastText, as palavras vão ser entendidas como vectores, o que resulta em que a posição e organização das palavras relativamente umas as outras, contexto, resulte em mudanças significativas do vetor e padrões sejam mais facilmente reconheciveis.

Which of the following is **not** a key difference between spam detection using embedded models (e.g., Word Embeddings) and the traditional Bag of Words Model?

a. Embedded models are more computationally expensive to train than the Bag of Words Model.

b. Embedded models can be used to detect spam emails that contain new or unseen words, while the Bag of Words Model cannot.

c. Embedded models require more data to train than the Bag of Words Model.

d. Embedded models capture the semantic and syntactic relationships between words, while the Bag of Words Model does not.

Which of the following is **not** a common technique used to reduce the false positive rate of spam filters?

- a. Training the spam filter on a large dataset of labeled emails.
- b. Using machine learning algorithms to identify patterns in spam emails.
- c. Using a whitelist to allow users to identify email addresses that should never be flagged as spam.
- d. Using a blacklist to block email addresses from known spammers.

Pergunta **9**

Respondida

Nota: 0,33 em 1,00

Which of the following is a key challenge in developing effective spam filters?

- a. To identify and prevent phishing emails from compromising users' accounts.
- b. All of the answers.
- c. To detect and remove malware from attachments to incoming emails.
- d. To block unsolicited emails from reaching users' inboxes.

[redacted]

Which of the following types of anomaly detection primarily relies on identifying data points that are farthest from the cluster centers?

- a. Density-based clustering
- b. Hierarchical clustering
- c. Distance-based clustering
- d. Spectral clustering

Pergunta **11**

Respondida

Nota: 1,00 em 1,00

Which of the following types of cybersecurity attacks is most likely to be detected using anomaly detection?

- a. Known attacks with a well-defined signature.
- b. Attacks that are carefully targeted at a specific organization or individual.
- c. Novel attacks that have never been seen before.
- d. Attacks that are disguised as legitimate activity.

[redacted]

Which of the following statements is true when comparing Isolation Forest and One-Class Support Vector Machine (One-Class SVM) for anomaly detection?

a. One-Class SVM can be used for novelty detection, while Isolation Forest cannot.

b. Isolation Forest is more computationally efficient to train than One-Class SVM.

c. Isolation Forest is more robust to outliers than One-Class SVM.

d. One-Class SVM is more sensitive to noise than Isolation Forest.

Which of the following is the primary role of the encoder component in anomaly detection using autoencoders?

- a. To reconstruct the input data from the latent space representation.
- b. To compress the input data into a latent space representation.
- c. To identify the anomalies in the input data by comparing the reconstruction error to a threshold.
- d. To identify the features of the input data that are most important for reconstructing the output data.

How can clustering algorithms be used to identify anomalies? What are the main differences between the Isolation Forest approach?

Algoritmos de clustering são usados para agrupar dados e detetar quais pontos não encaixam ou se distanciam da maioria, estes podem representar anomalias. Mas para isso funcionar, temos de assumir que anomalias são relativamente raras em relação aos dados normais. Isto pode tornar o treino de modelos dificeis se existir um grande numero de anomalias em relação ao que é considerado normal.

No caso das Isolation Forests, este algoritimo isola as anomalias em relação a forma como os seus caminhos divergem, geralmente com caminhos mais curtos que o que talvez fosse experado de dados não anomalos.

Which of the following statements about machine learning (ML) and security is **incorrect**?

a. ML can be used to develop security solutions that can detect and respond to threats more quickly and effectively than traditional methods.

b. ML can be used to automate security tasks, such as log analysis and threat detection.

c. ML can be used to identify and exploit security vulnerabilities in software and systems.

d. ML is a silver bullet that can solve all security problems.

[redacted]

Which of the following is a key component of a Bayesian spam filter?

- a. None of the above.
- b. A statistical model that calculates the probability of an email being spam based on the words and phrases it contains.
- c. A set of rules that identifies spam emails based on their sender, subject line, and other features.
- d. A training set of labeled emails that the filter uses to learn to identify spam emails.

[redacted]

Which of the following statements about bias and variance in machine learning is **incorrect**?

- a. Variance is the tendency of a machine learning model to make different predictions on the same data, depending on how the model is trained.
- b. Bias and variance are inversely proportional, meaning that reducing bias will increase variance, and vice versa.
- c. Bias and variance can be reduced by using a more complex model.
- d. Bias is the tendency of a machine learning model to make predictions that are consistently different from the true values.

Which of the following is **not** a goal of the ML process?

- a. To understand the underlying patterns in the data.
- b. To develop a model that can make accurate predictions on new data.
- c. To build a model that is robust to changes in the data.
- d. To minimize the cost of training and deploying the model.