



This is a copy of a conversation between ChatGPT & Anonymous.
[Report content](#)

nhập môn khoa học dữ liệu
Question types:
Define fundamental concepts
Choose true/false and justify the answer.
K-nn and k-means calculation
Tf-idf embeddings calculated from a given corpus
Compute evaluation metrics for classification, clustering, and regression
Draw ROC curve, PR curve
Visualize the data using basic and advanced visualization



Đây là một đề cương cơ bản cho môn học "Nhập môn Khoa học Dữ liệu". Dưới đây là các loại câu hỏi và hướng dẫn:

1. Define Fundamental Concepts (Định nghĩa các khái niệm cơ bản):

- Ví dụ câu hỏi:
 - Định nghĩa Khoa học Dữ liệu là gì?
 - Thế nào là supervised learning, unsupervised learning?
 - Sự khác biệt giữa classification và regression?
 - Giải thích khái niệm "overfitting" và "underfitting".
- Mục tiêu:

Đảm bảo hiểu rõ các khái niệm nền tảng của khoa học dữ liệu và học máy.

2. True/False Questions with Justification (Câu hỏi Đúng/Sai và giải thích):

- Ví dụ câu hỏi:
 - True/False: "KNN là một thuật toán không giám sát."
 - True/False: "K-means không yêu cầu xác định số cụm trước khi thực hiện thuật toán."
 - True/False: "ROC curve chỉ áp dụng được cho bài toán phân loại nhị phân."
- Mục tiêu:

Kiểm tra khả năng hiểu sâu và phân tích đúng/sai với lời giải thích.

3. KNN and K-means Calculation (Tính toán KNN và K-means):

- Ví dụ câu hỏi:
 - Với tập dữ liệu và giá trị K=3, hãy tìm nhãn của điểm mới (KNN).
 - Với dữ liệu đầu vào và K=2, thực hiện 2 bước của thuật toán K-means.
- Mục tiêu:

Đánh giá khả năng thực hành tính toán cơ bản bằng tay.

4. TF-IDF Embeddings Calculation:

- Ví dụ câu hỏi:
 - Cho tập văn bản sau:

vbnnet

D1: "machine learning is fun"
D2: "learning is essential"

Tính giá trị TF-IDF cho từ "learning" trong cả D1 và D2.
- Mục tiêu:

Hiểu và áp dụng phương pháp tính toán TF-IDF để biểu diễn văn bản.

5. Compute Evaluation Metrics (Tính các chỉ số đánh giá):

- Ví dụ câu hỏi:
 - Tính accuracy, precision, recall, F1-score từ ma trận nhầm lẫn (confusion matrix).
 - Với kết quả clustering, hãy tính Silhouette Score.
 - Tính R-squared và MAE trong bài toán regression.
- Mục tiêu:

Nắm vững các chỉ số đánh giá cho từng loại bài toán.

6. Draw ROC Curve, PR Curve (Vẽ đường cong ROC và PR):

- Ví dụ câu hỏi:
 - Với dữ liệu dự đoán:

```
yaml
Actual: [1, 0, 1, 1, 0]
Predicted Probabilities: [0.9, 0.4, 0.8, 0.7, 0.3]
```

Hãy vẽ đường cong ROC và tính AUC.
- Mục tiêu:

Hiểu cách vẽ và đánh giá mô hình dựa trên ROC, PR curve.

7. Visualize Data (Trực quan hóa dữ liệu):

- Ví dụ câu hỏi:
 - Dùng scatter plot để hiển thị phân phối của dữ liệu 2 chiều.
 - Vẽ heatmap để biểu diễn ma trận tương quan của tập dữ liệu.
 - Sử dụng biểu đồ nâng cao như box plot, violin plot, hoặc 3D scatter plot.
- Mục tiêu:

Thực hành sử dụng công cụ trực quan hóa như Matplotlib, Seaborn hoặc Plotly.

Nếu bạn cần bài tập mẫu hoặc hướng dẫn chi tiết cho từng phần, cứ cho mình biết nhé!

các câu hỏi sẽ liên quan đến :
ExploratoryDataAnalysis
ModelEvaluation
FeatureEngineering
DataVisualization



Dưới đây là danh sách các câu hỏi liên quan đến **Exploratory Data Analysis (EDA)**, **Model Evaluation**, **Feature Engineering**, và **Data Visualization**. Mỗi phần bao gồm các ví dụ cụ thể:

1. Exploratory Data Analysis (EDA)

- Câu hỏi:
 1. Hãy mô tả mục tiêu của quá trình EDA.
 2. Làm thế nào để xử lý các giá trị thiếu trong tập dữ liệu?
 3. Với tập dữ liệu sau:

```
r
Age: [25, 30, 35, NaN, 40, 30, NaN]
```

 - Điều giá trị thiếu bằng trung bình hoặc trung vị.
 4. Làm thế nào để phát hiện outlier trong tập dữ liệu số?
 5. Giải thích sự khác biệt giữa các loại biến: **categorical**, **numerical**, **ordinal**.

2. Model Evaluation

- Câu hỏi:

1. Classification:

- Giải thích các chỉ số **accuracy**, **precision**, **recall**, **F1-score**.
- Cho ma trận nhầm lẫn:

```
lua
```

Copy code

```
[[50, 10],  
 [5, 35]]
```

Tính các chỉ số trên.

2. Regression:

- Định nghĩa các chỉ số **MAE**, **MSE**, **R-squared**.
- Cho tập giá trị thực tế và dự đoán:

```
makefile
```

Copy code

```
Actual: [3, 5, 2.5, 7]  
Predicted: [2.5, 5, 4, 8]
```

Tính MAE và R-squared.

3. Clustering:

- Làm thế nào để đánh giá một bài toán phân cụm?
- Giải thích ý nghĩa của **Silhouette Score**.

4. Khi nào nên sử dụng **cross-validation**, và ưu điểm của nó là gì?

3. Feature Engineering

- Câu hỏi:

1. Feature engineering là gì? Tại sao nó quan trọng?
2. Làm thế nào để xử lý dữ liệu categorical (ví dụ: **one-hot encoding**, **label encoding**)?

3. Cho cột dữ liệu sau:

```
css
```

Copy code

```
Color: ["red", "blue", "green", "red", "blue"]
```

Biểu diễn dữ liệu này bằng one-hot encoding.

4. Làm thế nào để phát hiện và loại bỏ multicollinearity giữa các biến?
5. Giải thích kỹ thuật **Feature Scaling** (\downarrow Standardization, Min-Max Scaling).
6. Làm thế nào để chọn ra các biến quan trọng nhất cho mô hình?

4. Data Visualization

- Câu hỏi:

1. Các công cụ phổ biến để trực quan hóa dữ liệu là gì?
2. Hãy vẽ biểu đồ phân phôi cho một tập dữ liệu gồm các số:

```
csharp
```

Copy code

```
[1, 2, 2, 3, 3, 3, 4, 4, 4, 4]
```

3. Khi nào nên sử dụng biểu đồ scatter plot, bar chart, và box plot?
4. Với tập dữ liệu gồm 3 chiều (x, y, z), làm thế nào để trực quan hóa dữ liệu?
5. Vẽ heatmap để biểu diễn ma trận tương quan sau:

```
lua
```

Copy code

```
[[1, 0.8, 0.5],  
 [0.8, 1, 0.2],  
 [0.5, 0.2, 1]]
```

6. Giải thích cách biểu diễn dữ liệu phân loại trên biểu đồ 2D.

Nếu bạn cần tập dữ liệu hoặc code minh họa, mình có thể cung cấp cụ thể hơn!

hướng dẫn chi tiết
Question types:
Define fundamental concepts
Choose true/false and justify the answer.
K-nn and k-means calculation
Tf-idf embeddings calculated from a given corpus
Compute evaluation metrics for classification, clustering, and regression
Draw ROC curve, PR curve
Visualize the data using basic and advanced visualization
của các chủ đề
ExploratoryDataAnalysis
ModelEvaluation
FeatureEngineering
DataVisualization



Dưới đây là hướng dẫn chi tiết cho các loại câu hỏi và các chủ đề được đề cập:

1. Define Fundamental Concepts

Chủ đề: Exploratory Data Analysis (EDA)

- Câu hỏi: Định nghĩa EDA và nêu các bước chính trong EDA.
- Hướng dẫn trả lời:
 - EDA là gì? EDA là quá trình khám phá và phân tích dữ liệu ban đầu để hiểu rõ hơn về các đặc điểm và mối quan hệ giữa các biến.
 - Các bước chính trong EDA:
 1. Hiểu cấu trúc dữ liệu (kích thước, kiểu dữ liệu).
 2. Xử lý giá trị thiếu và dữ liệu ngoại lai.
 3. Tóm tắt dữ liệu bằng thống kê mô tả (mean, median, std).
 4. Trực quan hóa dữ liệu để tìm hiểu mối quan hệ.

Chủ đề: Model Evaluation

- Câu hỏi: Các chỉ số đánh giá mô hình trong bài toán phân loại?
- Hướng dẫn trả lời:
 - Accuracy, Precision, Recall, F1-Score.
 - Công thức:
 - Precision = $TP / (TP + FP)$.
 - Recall = $TP / (TP + FN)$.
 - F1-Score = $2 \times (Precision \times Recall) / (Precision + Recall)$.

Chủ đề: Feature Engineering

- Câu hỏi: Feature scaling là gì?
- Hướng dẫn trả lời:
 - Là quá trình chuẩn hóa các đặc trưng (features) để đảm bảo tất cả các giá trị nằm trong cùng một phạm vi.
 - Hai phương pháp chính:
 1. Standardization: Biến đổi dữ liệu về phân phối chuẩn ($mean = 0$, $std = 1$).
 2. Normalization (Min-Max Scaling): Đưa dữ liệu về phạm vi $[0, 1]$.

Chủ đề: Data Visualization

- Câu hỏi: Nêu các loại biểu đồ phổ biến và mục đích sử dụng?
- Hướng dẫn trả lời:
 - Scatter plot: Hiển thị mối quan hệ giữa hai biến số.
 - Bar chart: So sánh giá trị giữa các nhóm.

- Box plot: Tóm tắt dữ liệu qua các giá trị phân vị, phát hiện ngoại lai.

2. True/False and Justify the Answer

Chủ đề: EDA

- Câu hỏi: True/False: "EDA chỉ cần thực hiện một lần duy nhất trước khi xây dựng mô hình."
- Hướng dẫn trả lời:
 - False. EDA là một quá trình lặp lại. Bạn có thể cần quay lại EDA sau khi xây dựng mô hình để kiểm tra các giả định.

Chủ đề: Model Evaluation

- Câu hỏi: True/False: "Một mô hình có accuracy cao luôn là mô hình tốt nhất."
- Hướng dẫn trả lời:
 - False. Accuracy có thể không đáng tin cậy nếu dữ liệu không cân bằng. Precision và Recall quan trọng hơn trong các trường hợp như vậy.

Chủ đề: Feature Engineering

- Câu hỏi: True/False: "Tất cả các biến đều cần Feature Scaling."
- Hướng dẫn trả lời:
 - False. Các biến categorical hoặc biến đã được chuẩn hóa trước đó không cần scaling.

3. KNN and K-Means Calculation

KNN

- Câu hỏi: Cho điểm $(x, y) = (4, 4)$ và dữ liệu:

```
SCSS Copy code
A(1, 1, class=0), B(5, 5, class=1), C(3, 4, class=0), D(6, 7, class=1).
```

Hãy dự đoán lớp của điểm mới với $K = 3$.

- Hướng dẫn:

- Tính khoảng cách Euclidean giữa điểm mới và các điểm khác.
- Chọn $K = 3$ điểm gần nhất.
- Dựa vào đa số để quyết định lớp (0 hoặc 1).

K-Means

- Câu hỏi: Cho tập dữ liệu 2 chiều:

```
SCSS Copy code
(1, 2), (1, 4), (5, 6), (8, 8)
```

Thực hiện bước đầu tiên của thuật toán K-means với $K = 2$.

- Hướng dẫn:

1. Khởi tạo ngẫu nhiên 2 tâm cụm.
2. Tính khoảng cách từ mỗi điểm đến tâm cụm.
3. Gán điểm cho cụm gần nhất.

4. TF-IDF Embeddings

- Câu hỏi: Cho tập văn bản:

```
vbnet Copy code
Doc1: "data science is fun"
Doc2: "science is essential"
```

Tính TF-IDF cho từ "science".

- Hướng dẫn:

- Tính TF (Term Frequency) cho từ "science":

makefile

Copy code

```
TF_Doc1 = 1 / 4
TF_Doc2 = 1 / 3
```

- Tính IDF (Inverse Document Frequency):

mathematica

Copy code

```
IDF = log(Total Documents / Number of Documents containing "science").
```

- TF-IDF = TF × IDF.

5. Compute Evaluation Metrics

- Classification:

- Cho ma trận nhầm lẩn:

lua



Copy code

```
[[40, 10],
 [5, 45]]
```

Tính Precision, Recall, F1-Score.

- Regression:

- Với Actual = [2.5, 0.0, 2.1, 7.8] và Predicted = [3.0, -0.5, 2.0, 8.0], tính MAE và MSE.

- Clustering:

- Tính Silhouette Score từ phân cụm (cho dữ liệu cụ thể).

6. Draw ROC Curve, PR Curve

- ROC Curve:

- Cho dữ liệu:

yaml

Copy code

```
Actual: [1, 0, 1, 1, 0],
Predicted Probabilities: [0.9, 0.4, 0.8, 0.7, 0.3].
```

- Tính các giá trị TPR và FPR, sau đó vẽ ROC.

- PR Curve:

- Tính Precision và Recall ở các ngưỡng khác nhau, sau đó vẽ biểu đồ.

7. Data Visualization

- Scatter Plot:

- Dùng matplotlib để vẽ mối quan hệ giữa hai biến.

- Box Plot:

- Sử dụng Seaborn để vẽ biểu đồ box plot cho một tập dữ liệu.

- Heatmap:

- Tạo biểu đồ heatmap từ ma trận tương quan.

Nếu bạn muốn ví dụ code Python cụ thể, mình có thể cung cấp ngay!

cách làm các bài tập tính toán liên quan



Dưới đây là hướng dẫn chi tiết cách làm các bài tập tính toán liên quan đến các chủ đề đã đề cập:

1. KNN Calculation

Ví dụ:

Dự đoán lớp cho điểm $P(4, 4)$ dựa trên tập dữ liệu:

$$A(1, 1, \text{class} = 0), B(5, 5, \text{class} = 1), C(3, 4, \text{class} = 0), D(6, 7, \text{class} = 1)$$

với $K = 3$.

Các bước giải:

- Tính khoảng cách Euclidean từ $P(4, 4)$ đến các điểm:

Công thức:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Khoảng cách từ P đến A :

$$d(P, A) = \sqrt{(4 - 1)^2 + (4 - 1)^2} = \sqrt{9 + 9} = \sqrt{18} \approx 4.24$$

- Khoảng cách từ P đến B :

$$d(P, B) = \sqrt{(4 - 5)^2 + (4 - 5)^2} = \sqrt{1 + 1} = \sqrt{2} \approx 1.41$$

- Tương tự, tính $d(P, C) \approx 1.0$ và $d(\downarrow D) \approx 3.61$.

- Sắp xếp các khoảng cách:

- $C(1.0), B(1.41), D(3.61), A(4.24)$.

- Chọn $K = 3$ điểm gần nhất:

- $C(\text{class} = 0), B(\text{class} = 1), D(\text{class} = 1)$.

- Bầu chọn theo đa số:

- Lớp 1 xuất hiện 2 lần, lớp 0 xuất hiện 1 lần.
- Kết luận: $P(4, 4)$ thuộc lớp 1.

2. K-Means Calculation

Ví dụ:

Thực hiện một bước thuật toán K-means với tập dữ liệu:

$$(1, 2), (1, 4), (5, 6), (8, 8)$$

Và khởi tạo ngẫu nhiên tâm cụm:

$$C_1(2, 3), C_2(7, 7).$$

Các bước giải:



- Tính khoảng cách từ mỗi điểm đến các tâm cụm:

Sử dụng công thức Euclidean.

- Điểm $P_1(1, 2)$:

$$d(P_1, C_1) = \sqrt{(1 - 2)^2 + (2 - 3)^2} = \sqrt{2} \approx 1.41, d(P_1, C_2) = \sqrt{(1 - 7)^2 + (2 - 7)^2} \approx 7.81.$$

- Tương tự cho các điểm khác:

Điểm	C_1	C_2	Cụm
(1, 2)	1.41	7.81	C_1
(1, 4)	1.41	7.21	C_1
(5, 6)	5.0	2.24	C_2
(8, 8)	7.81	1.41	C_2

- Cập nhật tâm cụm:

Tính trung bình các điểm trong mỗi cụm:

- C_1 : Trung bình của $(1, 2), (1, 4)$:

$$C'_1 = \left(\frac{1+1}{2}, \frac{2+4}{2} \right) = (1, 3)$$

- C_2 : Trung bình của $(5, 6), (8, 8)$:

$$C'_2 = \left(\frac{5+8}{2}, \frac{6+8}{2} \right) = (6.5, 7)$$



3. TF-IDF Calculation

Ví dụ:

Cho tập văn bản:

Doc1: "data science is fun" Doc2: "science is essential"

Tính TF-IDF cho từ "science".

Các bước giải:

1. Tính TF (Term Frequency):

Công thức:

$$TF = \frac{\text{số lần xuất hiện của từ trong tài liệu}}{\text{tổng số từ trong tài liệu}}$$

- $TF_{Doc1}(\text{science}) = 1/4 = 0.25.$

- $TF_{Doc2}(\text{science}) = 1/3 \approx 0.333.$

2. Tính IDF (Inverse Document Frequency):

Công thức:

$$IDF = \log \left(\frac{\text{Tổng số tài liệu}}{\text{số tài liệu chứa từ đó}} \right)$$

- $IDF(\text{science}) = \log(2/2) = 0.$

3. Tính TF-IDF:

$$TF - IDF = TF \times IDF$$

- Vì $IDF = 0$, nên $TF - IDF = 0$ cho từ "science" trong cả hai tài liệu.

4. Compute Evaluation Metrics

Classification:

- Cho ma trận nhầm lẫn:

$$\begin{bmatrix} 40 & 10 \\ 5 & 45 \end{bmatrix}$$

Tính Precision, Recall, F1-Score.

Công thức và giải:

1. Precision (class=1):

$$Precision = \frac{TP}{TP + FP} = \frac{45}{45 + 10} = 0.818$$

2. Recall (class=1):

$$Recall = \frac{TP}{TP + FN} = \frac{45}{45 + 5} = 0.9$$

3. F1-Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{0.818 \cdot 0.9}{0.818 + 0.9} \approx 0.857$$

5. Draw ROC Curve

Ví dụ:

Dữ liệu:

Actual = [1, 0, 1, 1, 0] Predicted Probabilities = [0.9, 0.4, 0.8, 0.7, 0.3]

Các bước:

1. Sắp xếp xác suất dự đoán giảm dần:

$$[0.9, 0.8, 0.7, 0.4, 0.3]$$

2. Tính TPR (True Positive Rate) và FPR (False Positive Rate) tại mỗi ngưỡng:

- TPR:

$$TPR = \frac{TP}{TP + FN}$$

- FPR:

$$FPR = \frac{FP}{FP + TN}$$

3. Vẽ đồ thị TPR và FPR.

nhiều câu hỏi hơn 1 chút



Danh sách câu hỏi tính toán chi tiết hơn

1. Exploratory Data Analysis (EDA)

Câu hỏi:

- Tính các thống kê cơ bản (mean, median, variance, standard deviation) cho tập dữ liệu $[2, 4, 6, 8, 10]$.
- Cho bảng phân phối dữ liệu, hãy tính phân vị thứ 25 và phân vị thứ 75.
- Vẽ biểu đồ phân phối của một cột dữ liệu $\sim = [10, 20, 20, 30, 40]$ và chỉ ra bất kỳ giá trị ngoại lai nào.
- Cho ma trận tương quan giữa các cột trong tập dữ liệu, hãy xác định cặp biến có mối quan hệ mạnh nhất và yếu nhất.

2. Feature Engineering

Câu hỏi:

- Chuẩn hóa tập dữ liệu $x = [2, 4, 6, 8]$ theo chuẩn Z-score.
 - Công thức:
$$Z = \frac{x - \mu}{\sigma}$$
- Tính One-Hot Encoding cho cột dữ liệu $["red", "blue", "red", "green"]$.
- Tính TF-IDF cho từ "machine" trong tập văn bản:
 - Doc1: "machine learning is fun".
 - Doc2: "deep learning with machine learning".
- Thực hiện binning cho dữ liệu $x = [5, 15, 25, 35, 45]$ thành 3 nhóm.

3. Classification Metrics

Câu hỏi:

- Cho ma trận nhầm lẫn:

$$\begin{bmatrix} 50 & 5 \\ 10 & 35 \end{bmatrix}$$
 - Tính Precision, Recall và F1-Score cho lớp 1.
- Cho các xác suất dự đoán và nhãn thực tế:
 $y_{true} = [1, 0, 1, 0], \hat{y}_{prob} = [0.9, 0.3, 0.8, 0.4]$
 - Vẽ đường cong ROC và tính AUC.
- Dữ liệu sau biểu thị dự đoán nhị phân:
 $y_{true} = [1, 0, 1, 1, 0], \hat{y} = [1, 0, 1, 0, 1]$
 - Tính Accuracy, Precision, Recall.

4. Regression Metrics

Câu hỏi:

- Cho tập dự đoán và thực tế:
 $y_{true} = [3, -0.5, 7], y_{pred} = [2.5, 0.0, 2, 8]$
 - Tính Mean Absolute Error (MAE), Mean Squared Error (MSE), và Root Mean Squared Error (RMSE).
- Một mô hình hồi quy dự đoán như sau:
 $y_{true} = [10, 20, 30, 40], y_{pred} = [12, 18, 33, 39]$
 - Tính R-squared R^2 .

5. Clustering Metrics

Câu hỏi:

1. Cho tập dữ liệu 2D:

$$(1, 2), (2, 3), (6, 8), (7, 9)$$

- Khởi tạo ngẫu nhiên hai tâm cụm: $C_1(2, 3), C_2(6, 7)$.
- Thực hiện một bước K-Means clustering.

2. Tính Silhouette Score cho tập dữ liệu:

- Cụm 1: $(1, 2), (2, 3)$.

- Cụm 2: $(6, 8), (7, 9)$.

- Sử dụng công thức:

$$S(i) \downarrow \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Với $a(i)$: khoảng cách trung bình trong cụm, $b(i)$: khoảng cách trung bình đến cụm gần nhất.

6. Data Visualization

Câu hỏi:

1. Vẽ biểu đồ hộp (boxplot) cho tập dữ liệu $x = [5, 7, 8, 8, 10, 12, 14, 18]$.

2. Tạo biểu đồ tán xạ (scatter plot) với hai biến:

- $x = [1, 2, 3, 4], y = [2, 4, 6, 8]$.

3. Vẽ biểu đồ cột (bar chart) biểu diễn số lượng các nhãn:

- Labels = $[A, B, A, C, B, A]$.

7. KNN và K-Means

Câu hỏi:

1. Với KNN, cho tập dữ liệu sau:

$$(2, 3, \text{class} = A), (4, \downarrow, \text{class} = B), (3, 4, \text{class} = A)$$

Dự đoán lớp của $P(3, 5)$ với $K = 2$.

2. Thực hiện K-Means clustering cho tập:

$$(1, 2), (3, 4), (8, 8), (9, 10)$$

Với hai cụm ban đầu $C_1(2, 2), C_2(9, 9)$.

Nếu bạn muốn giải bài tập cụ thể nào hoặc cần hỗ trợ viết code Python để giải, hãy cho mình biết nhé!

giải các bài tập đó



Giải các bài tập đã liệt kê

1. Exploratory Data Analysis (EDA)

Bài 1:

Tính các thống kê cơ bản (mean, median, variance, standard deviation) cho tập dữ liệu $[2, 4, 6, 8, 10]$.



Giải:

- Mean:

$$\mu = \frac{\sum x_i}{n} = \frac{2 + 4 + 6 + 8 + 10}{5} = 6$$

- Median:

Dữ liệu đã sắp xếp, median là giá trị giữa:

$$\text{Median} = 6$$

- Variance (Phương sai):

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n} = \frac{(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2}{5} = 8$$

- Standard Deviation (Độ lệch chuẩn):

$$\sigma = \sqrt{\sigma^2} = \sqrt{8} \approx 2.83$$

2. Feature Engineering

Bài 1:

Chuẩn hóa tập dữ liệu $x = [2, 4, 6, 8]$ theo chuẩn Z-score.

Giải:



- Mean (μ):

$$\mu = \frac{\sum x_i}{n} = \frac{2+4+6+8}{4} = 5$$

- Standard Deviation (σ):

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}} = \sqrt{\frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4}} = \sqrt{5} \approx 2.236$$

- Z-score:

$$Z = \frac{x - \mu}{\sigma}$$

Tính lần lượt:

$$Z_1 = \frac{2-5}{2.236} \approx -1.34, Z_2 = \frac{4-5}{2.236} \approx -0.45$$

$$Z_3 = \frac{6-5}{2.236} \approx 0.45, Z_4 = \frac{8-5}{2.236} \approx 1.34$$

Kết quả: $[-1.34, -0.45, 0.45, 1.34]$.

3. Classification Metrics

Bài 1:



Cho ma trận nhầm lẫn:

$$\begin{bmatrix} 50 & 5 \\ 10 & 35 \end{bmatrix}$$

Tính Precision, Recall và F1-Score cho lớp 1.

Giải:

- Precision (class=1):

$$Precision = \frac{TP}{TP + FP} = \frac{35}{35 + 5} = 0.875$$

- Recall (class=1):

$$Recall = \frac{TP}{TP + FN} = \frac{35}{35 + 10} = 0.777$$

- F1-Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{0.875 \cdot 0.777}{0.875 + 0.777} \approx 0.823$$

4. Regression Metrics

Bài 1:



Cho tập dự đoán và thực tế:

$$y_{true} = [3, -0.5, 2, 7], y_{pred} = [2.5, 0.0, 2, 8]$$

Tính MAE, MSE, và RMSE.

Giải:

- Mean Absolute Error (MAE):

$$MAE = \frac{\sum |y_{true} - y_{pred}|}{n} = \frac{|3-2.5| + |-0.5-0| + |2-2| + |7-8|}{4} = 0.5$$

- Mean Squared Error (MSE):

$$MSE = \frac{\sum(y_{true} - y_{pred})^2}{n} = \frac{(3 - 2.5)^2 + (-0.5 - 0)^2 + (2 - 2)^2 + (7 - 8)^2}{4} = 0.375$$

- Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE} = \sqrt{0.375} \approx 0.612$$

5. Clustering Metrics

Bài 1:

Thực hiện một bước K-Means clustering cho t^*

$(1, 2), (3, 4), (8, 8), (9, 10)$

Với tâm cụm ban đầu: $C_1(2, 2), C_2(9, 9)$.

Giải:

1. Tính khoảng cách Euclidean:

- Điểm $(1, 2)$:

$$d((1, 2), C_1) = \sqrt{(1 - 2)^2 + (2 - 2)^2} = 1, d((1, 2), C_2) = \sqrt{(1 - 9)^2 + (2 - 9)^2} \approx 10.63$$

$(1, 2)$ thuộc C_1 .

- Tương tự:

Điểm	C_1	C_2	Cụm
$(1, 2)$	1	10.63	C_1
$(3, 4)$	2.24	8.06	C_1
$(8, 8)$	8.49	1.0	C_2
$(9, 10)$	10.2	1.0	C_2

2. Cập nhật tâm cụm:

- C'_1 : Trung bình của $(1, 2), (3, 4)$:

$$C'_1 = \left(\frac{1+3}{2}, \frac{2+4}{2} \right) = (2, 3)$$

- C'_2 : Trung bình của $(8, 8), (9, 10)$:

$$C'_2 = \left(\frac{8+9}{2}, \frac{8+10}{2} \right) = (8.5, 9)$$

Kết quả tâm cụm mới: $C_1(2, 3), C_2(8.5, 9)$.

Nếu cần chi tiết thêm bất kỳ bước nào hoặc triển khai mã Python để tính toán, hãy cho mình biết!

các câu hỏi lý thuyết, các câu hỏi so sánh giữa 2 cái gần liên quan



Danh sách câu hỏi lý thuyết và so sánh chi tiết

1. Exploratory Data Analysis (EDA)

Câu hỏi lý thuyết:

1. EDA là gì, và tại sao nó quan trọng trong quy trình phân tích dữ liệu?
2. Phân biệt giữa các loại biểu đồ:
 - Khi nào sử dụng biểu đồ hộp (boxplot) và khi nào sử dụng biểu đồ phân tán (scatter plot)?
3. So sánh mean và median trong việc mô tả trung tâm dữ liệu, và khi nào nên sử dụng từng loại.
4. Phân biệt giữa dữ liệu định tính (categorical) và dữ liệu định lượng (quantitative).
5. Giải thích vai trò của phát hiện giá trị ngoại lai trong EDA.

Câu hỏi so sánh:

1. So sánh các cách đo độ phân tán dữ liệu: range, variance, và standard deviation.
2. Khi nào sử dụng biểu đồ histogram, và khi nào sử dụng KDE plot?
3. So sánh giữa kiểm tra mối tương quan Pearson và kiểm tra Spearman.

2. Feature Engineering

Câu hỏi lý thuyết:

1. Feature Engineering là gì? Vai trò của nó trong mô hình học máy.
2. So sánh giữa các phương pháp chuẩn hóa (normalization) và chuẩn hóa theo Z-score (standardization).
3. Khi nào nên sử dụng One-Hot Encoding và khi nào sử dụng Label Encoding?
4. TF-IDF hoạt động như thế nào, và nó khác gì so với Count Vectorizer?
5. Tại sao lại cần xử lý giá trị bị thiếu (missing values) trước khi xây dựng mô hình?

Câu hỏi so sánh:

1. So sánh hiệu quả của việc giảm chiều dữ liệu (dimensionality reduction) bằng PCA và bằng chọn lọc đặc trưng (feature selection).
2. Tại sao sử dụng binning trong dữ liệu liên tục có thể giảm độ chính xác của mô hình?
3. So sánh giữa sử dụng các biến danh mục gốc (categorical features) và chuyển đổi chúng thành dạng số (numerical features).

3. Model Evaluation

Câu hỏi lý thuyết:

1. ROC Curve là gì, và tại sao nó quan trọng trong đánh giá mô hình phân loại?
2. Phân biệt giữa Precision và Recall, và đưa ra một ví dụ thực tế khi cần tối ưu từng chỉ số.
3. Tại sao Accuracy không phải là thước đo lý tưởng trong trường hợp dữ liệu không cân bằng?
4. Giải thích ý nghĩa của AUC trong bài toán phân loại nhị phân.
5. Mean Absolute Error (MAE) và Mean Squared Error (MSE) khác nhau như thế nào trong bài toán hồi quy?

Câu hỏi so sánh:

1. So sánh giữa Precision-Recall Curve và ROC Curve. Khi nào sử dụng từng loại?
2. So sánh các chỉ số F1-Score và Matthews Correlation Coefficient (MCC).
3. Trong hồi quy, tại sao R^2 không thể đánh giá chính xác cho tất cả các trường hợp?

4. Clustering

Câu hỏi lý thuyết:

1. K-means clustering là gì? Các bước cơ bản để thực hiện thuật toán này.
2. Giải thích Silhouette Score và ý nghĩa của giá trị này.
3. Phân biệt giữa clustering và classification.
4. Khi nào sử dụng K-means và khi nào sử dụng Hierarchical Clustering?
5. Các hạn chế của K-means clustering, và làm thế nào để khắc phục?

Câu hỏi so sánh:

1. So sánh K-means clustering và DBSCAN về cách hoạt động và ứng dụng.
2. Tại sao K-means không phù hợp với dữ liệu không hình cầu (non-spherical data)?
3. So sánh giữa thuật toán agglomerative và divisive trong Hierarchical Clustering.

5. Data Visualization

Câu hỏi lý thuyết:

1. Tại sao việc trực quan hóa dữ liệu là bước không thể thiếu trong phân tích dữ liệu?
2. Giải thích cách chọn biểu đồ phù hợp với từng loại dữ liệu.
3. So sánh giữa Heatmap và Pairplot trong việc biểu diễn tương quan dữ liệu.
4. Tại sao cần làm sạch dữ liệu trước khi trực quan hóa?
5. Vai trò của màu sắc trong việc tăng hiệu quả trực quan hóa dữ liệu.

Câu hỏi so sánh:

1. So sánh giữa biểu đồ cột (bar chart) và biểu đồ đường (line chart).
 2. Khi nào sử dụng biểu đồ hộp (boxplot) thay vì biểu đồ violin?
 3. Biểu đồ phân tán (scatter plot) có thể thay thế hoàn toàn biểu đồ đường xu hướng (trend line) không? Tại sao?
-

6. KNN và K-Means



Câu hỏi lý thuyết:

1. KNN hoạt động như thế nào, và tại sao nó là thuật toán "lazy"?
2. K-means có phụ thuộc vào cách khởi tạo tâm cụm ban đầu không? Giải thích.
3. Phân biệt giữa tham số K trong KNN và K-means.
4. Trong KNN, làm thế nào để xử lý trường hợp dữ liệu không cân bằng?
5. Ý nghĩa của khoảng cách Euclidean trong cả KNN và K-means.

Câu hỏi so sánh:

1. So sánh giữa KNN và K-means: mục tiêu và ứng dụng.
 2. Khi nào KNN không hoạt động tốt với dữ liệu lớn, và tại sao?
 3. Tại sao K-means không được sử dụng để phân loại trực tiếp mà chỉ dùng để phân cụm?
-

Nếu cần giải thích hoặc mở rộng thêm bất kỳ câu hỏi nào, hãy cho mình biết nhé!

giải cho tôi



Giải các câu hỏi lý thuyết và so sáºt tăa đưa ra

1. Exploratory Data Analysis (EDA)

Giải thích:

1. EDA là gì?

EDA là quá trình phân tích ban đầu để hiểu cấu trúc, sự phân phõi, và các đặc tính của dữ liệu.
Tại sao quan trọng?

- Xác định giá trị ngoại lai.
- Hiểu được xu hướng và mối tương quan giữa các biến.
- Hỗ trợ chọn lựa các bước tiền xử lý và mô hình phù hợp.

2. Phân biệt biểu đồ hộp (boxplot) và biểu đồ phân tán (scatter plot):

- Boxplot: Dùng để biểu diễn sự phân phõi và giá trị ngoại lai của một biến liên tục.
- Scatter plot: Dùng để biểu diễn mối quan hệ giữa hai biến liên tục.

3. So sánh mean và median:

- Mean (giá trị trung bình): Phù hợp khi dữ liệu không chứa giá trị ngoại lai.
- Median (trung vị): Ổn định hơn khi dữ liệu chứa ngoại lai.

4. Dữ liệu định tính vs. định lượng:

- Định tính (categorical): Dữ liệu rời rạc, không thể đo lường (VD: màu sắc, giới tính).
- Định lượng (quantitative): Dữ liệu đo được (VD: chiều cao, cân nặng).

2. Feature Engineering

Giải thích:

1. Feature Engineering là gì?

Quá trình chuyển đổi dữ liệu thô thành các đặc trưng hữu ích để tăng hiệu quả của mô hình học máy.

2. So sánh Normalization và Standardization:

- Normalization: Đưa dữ liệu về thang giá trị [0, 1].
- Standardization: Biến dữ liệu có mean = 0 và standard deviation = 1 (Z-score).

3. One-Hot Encoding vs. Label Encoding:
 - One-Hot Encoding: Biến các giá trị danh mục thành vector nhị phân (phù hợp khi không có thứ tự).
 - Label Encoding: Gán giá trị số nguyên cho các danh mục (phù hợp khi danh mục có thứ tự).
 4. TF-IDF hoạt động như thế nào?
 - TF: Số lần xuất hiện của một từ trong tài liệu.
 - IDF: Tần suất nghịch đảo tài liệu, để \downarrow cung trọng số của từ phổ biến.
 5. Xử lý missing values:
 - Lấp bằng giá trị trung bình/median.
 - Loại bỏ dòng chứa missing values (nếu không ảnh hưởng đến dữ liệu).
-

3. Model Evaluation

Giải thích:

1. ROC Curve:
 - Đường cong biểu diễn mối quan hệ giữa TPR (True Positive Rate) và FPR (False Positive Rate).
 - Quan trọng vì nó cho thấy hiệu năng của mô hình với mọi ngưỡng phân loại.
 2. Precision và Recall:
 - Precision: Tỷ lệ dự đoán đúng trong tổng số dự đoán là Positive.
 - Recall: Tỷ lệ dự đoán đúng trong tổng số Positive thực tế.
- Ví dụ: Trong y tế, tối ưu Recall để phát hiện bệnh sớm; tối ưu Precision để giảm sai sót khi chẩn đoán sai.
3. Accuracy không lý tưởng với dữ liệu không cân bằng:
 - Accuracy có thể cao dù mô hình chỉ \downarrow đoán lớp chiếm đa số.
 4. AUC:
 - Giá trị AUC càng lớn (gần 1) thì mô hình càng tốt.
 5. MAE vs. MSE:
 - MAE: Không bị ảnh hưởng bởi giá trị ngoại lai.
 - MSE: Nhạy cảm với giá trị ngoại lai, phù hợp khi muốn phạt lỗi lớn.
-

4. Clustering

Giải thích:

1. K-means clustering:

Thuật toán phân cụm dữ liệu thành k cụm dựa trên khoảng cách Euclidean.
 2. Silhouette Score:

Đánh giá chất lượng phân cụm:

 - Giá trị gần 1: Phân cụm tốt.
 - Giá trị gần 0: Các cụm chồng chéo nhau.
 3. Clustering vs. Classification:
 - Clustering: Không có nhãn, mục tiêu \downarrow tìm cấu trúc ẩn.
 - Classification: Dữ liệu có nhãn, mục tiêu là dự đoán nhãn.
 4. Khi nào dùng K-means vs. Hierarchical Clustering?
 - K-means: Khi dữ liệu lớn, cần tốc độ.
 - Hierarchical: Khi dữ liệu nhỏ, cần mô hình trực quan.
-

5. Data Visualization

Giải thích:

1. Tại sao cần trực quan hóa dữ liệu?
 - Giúp phát hiện các mẫu (patterns), xu hướng (trends), và giá trị ngoại lai.
2. Chọn biểu đồ phù hợp:

- Histogram: Biểu diễn phân phối dữ liệu.
- Scatter plot: Mô tả quan hệ giữa hai biến liên tục.

3. Heatmap vs. Pairplot:

- Heatmap: Tương quan tổng thể giữa các biến.
- Pairplot: Tương quan từng cặp biến kèm phân phối.

4. Làm sạch dữ liệu trước khi vẽ:

Tránh lỗi hiển thị, cải thiện độ chính xác biểu diễn.

5. Vai trò của màu sắc:

Màu sắc tốt giúp người xem dễ dàng hiểu ý nghĩa và xu hướng.

6. KNN và K-Means

Giải thích:

1. KNN là gì?

- Thuật toán phân loại dựa trên khoảng cách giữa điểm cần dự đoán và các điểm kề cận.

2. K-means phụ thuộc khơi tạo tâm cụm:

- Nếu khởi tạo không tốt, kết quả phân cụm có thể kém chính xác.

3. Phân biệt K :

- K trong KNN: Số láng giềng gần nhất.
- K trong K-means: Số cụm cần tìm.

4. KNN và dữ liệu không cân bằng:

- Cần dùng trọng số (weighted distance) để giảm ảnh hưởng của lớp lớn.

5. Khoảng cách Euclidean:

- Dùng để xác định sự tương tự hoặc sự gần nhau giữa các điểm.

Nếu cần thêm giải thích hoặc mã minh họa, hãy cho mình biết nhé!

✉️ Messages beyond this point are only visible to you

Message ChatGPT



ChatGPT can make mistakes. Check important info.

?