

TRƯỜNG ĐẠI HỌC VĂN LANG
ĐƠN VỊ: Khoa CNTT

ĐỀ THI VÀ ĐÁP ÁN
THI KẾT THÚC HỌC PHẦN
Học kỳ 2, năm học 2023-2024

I. Thông tin chung

Tên học phần:	Nhập môn Phân tích Dữ liệu và Học Sâu		
Mã học phần:	71ITDS30203	Số tín chỉ:	3
Mã nhóm lớp học phần:			
Hình thức thi: Trắc nghiệm kết hợp Tự luận	Thời gian làm bài:	60	phút
<i>Thí sinh được tham khảo tài liệu:</i>	<input type="checkbox"/> Có	<input checked="" type="checkbox"/> Không	

Cách thức nộp bài phân tự luận (Giảng viên ghi rõ yêu cầu):

- SV gõ trực tiếp trên khung trả lời của hệ thống thi;

II. Các yêu cầu của đề thi nhằm đáp ứng CLO

Ký hiệu CLO	Nội dung CLO	Hình thức đánh giá	Trọng số CLO trong thành phần đánh giá (%)	Câu hỏi thi số	Điểm số tối đa	Lấy dữ liệu đo lường mức đạt PLO/PI
(1)	(2)	(3)	(4)	(5)	(6)	(7)
CLO 1	Hiểu được nền tảng phân tích dữ liệu dựa trên các thuật toán, quy trình phát triển ứng dụng (thư viện phân mềm) dùng trong phân tích dữ liệu bằng ngôn ngữ lập trình Python	Trắc nghiệm	30%	1-20	3	PLO1/P I1.2
CLO 2	Làm quen với các khái niệm cơ bản của Học sâu, hiểu biết về các thư viện và công cụ sử dụng phổ biến trong Học sâu.	Tự luận	50%	Bài 1, Bài 2	5	PLO1/P I1.2
CLO 3	Có kiến thức cơ bản về quy trình phân tích dữ liệu, xây dựng và khai thác các thư viện và giải thuật trong Học sâu để giải quyết một số bài toán thực tế	Trắc nghiệm + Tự luận	20%	1-20	2	PLO1/P I1.2

III. Nội dung câu hỏi thi

PHẦN TRẮC NGHIỆM

Câu 1: Phương thức nào trong Pandas được sử dụng để đọc tệp Excel vào DataFrame?

- A) read_excel()
- B) import_excel()
- C) load_excel()
- D) open_excel()

Câu 2: Làm thế nào để tạo một mảng NumPy với các số từ 0 đến 9?

- A) np.range(10)
- B) np.linspace(0, 9, 10)
- C) np.numbers(10)
- D) np.arange(10)

Câu 3: Phương thức nào trong Pandas được sử dụng để nhóm dữ liệu theo một hoặc nhiều cột?

- A) aggregate()
- B) summarize()
- C) groupby()
- D) group()

Câu 4: Hàm nào trong NumPy được sử dụng để tính tổng của tất cả các phần tử trong mảng?

- A) add()
- B) sum()
- C) total()
- D) summing()

Câu 5: Làm thế nào để đổi tên các cột trong DataFrame của Pandas?

- A) rename(columns=new_names)
- B) change_columns(new_names)
- C) set_columns(new_names)
- D) update_columns(new_names)

Câu 6: Phương thức nào trong Pandas được sử dụng để lấy các thông kê mô tả của DataFrame?

- A) summary()
- B) statistics()
- C) describe()
- D) detail()

Câu 7: Làm thế nào để chuyển đổi một cột chuỗi thành kiểu số trong Pandas?

- A) `pd.to_number()`
- B) `pd.to_int()`
- C) `pd.to_float()`
- D) `pd.to_numeric()`

Câu 8: Phương thức nào trong Pandas được sử dụng để sắp xếp các giá trị trong DataFrame?

- A) `sort_values()`
- B) `order_values()`
- C) `arrange_values()`
- D) `sort_data()`

Câu 9: Làm thế nào để tính giá trị trung vị của một cột trong DataFrame của Pandas?

- A) `df['column_name'].middle()`
- B) `df['column_name'].mean()`
- C) `df['column_name'].median()`
- D) `df['column_name'].average()`

Câu 10: Hàm nào trong NumPy được sử dụng để tạo một mảng các số ngẫu nhiên tuân theo phân phối chuẩn?

- A) `np.random.randn()`
- B) `np.random.normal()`
- C) `np.random.standard()`
- D) `np.random.norm()`

Câu 11: Làm thế nào để lấy các hàng đầu tiên của DataFrame trong Pandas?

- A) `df.top()`
- B) `df.first()`
- C) `df.head()`
- D) `df.start()`

Câu 12: Làm thế nào để tính tổng các giá trị của một cột trong DataFrame của Pandas?

- A) `df['column_name'].total()`
- B) `df['column_name'].sum()`
- C) `df['column_name'].add()`
- D) `df['column_name'].accumulate()`

Câu 13: Làm thế nào để thêm một cột mới vào DataFrame trong Pandas?

- A) `df['new_column'] = values`
- B) `df.add_column('new_column', values)`

C) `df.insert_column('new_column', values)`

D) `df.create_column('new_column', values)`

Câu 14: Hàm nào trong NumPy được sử dụng để tính trị tuyệt đối của các phần tử trong mảng?

A) `np.abs()`

B) `np.absolute()`

C) `np.abs_value()`

D) `np.modulus()`

Câu 15: Hàm nào trong NumPy được sử dụng để tìm chỉ số của giá trị lớn nhất trong mảng?

A) `np.max_index()`

B) `np.max_loc()`

C) `np.index_of_max()`

D) `np.argmax()`

Câu 16: **Tình huống Tách dữ liệu thành tập huấn luyện và tập kiểm tra:** Bạn có một tập dữ liệu lớn và muốn chia nó thành tập huấn luyện và tập kiểm tra với tỉ lệ 80:20. Làm thế nào để bạn thực hiện điều này?

A) Sử dụng phương thức `split` của DataFrame.

B) Sử dụng `df.sample()` để chọn ngẫu nhiên.

C) Sử dụng hàm `train_test_split` của thư viện `scikit-learn`.

D) Sử dụng phương thức `divide()` của DataFrame.

Câu 17: **Tình huống: Xử lý dữ liệu bị thiếu:** Bạn có một DataFrame `df` với nhiều giá trị bị thiếu. Làm thế nào để bạn thay thế các giá trị bị thiếu này bằng giá trị trung bình của cột tương ứng?

A) Sử dụng phương thức `fillna(df.mean())`.

B) Sử dụng phương thức `replace_na(df.mean())`.

C) Sử dụng phương thức `fill_missing(df.mean())`.

D) Sử dụng phương thức `replace_missing(df.mean())`.

Câu 18: **Tình huống Chuẩn hóa dữ liệu:** Bạn có một tập dữ liệu với các đặc trưng có phạm vi giá trị khác nhau. Làm thế nào để bạn chuẩn hóa các đặc trưng này để đưa về cùng phạm vi?

A) Sử dụng phương pháp Standard Scaling.

B) Sử dụng phương pháp Robust Scaling.

C) Sử dụng phương pháp Log Transformation.

D) Sử dụng phương pháp Min-Max Scaling.

Câu 19: Tình huống Phát hiện và xử lý điểm ngoại lai (outlier): Bạn đang làm việc với một tập dữ liệu và phát hiện có một số điểm ngoại lai. Làm thế nào để bạn phát hiện và xử lý các điểm ngoại lai này?

- A) Sử dụng Z-score để phát hiện và loại bỏ các điểm ngoại lai.
- B) Sử dụng IQR (Interquartile Range) để phát hiện và loại bỏ các điểm ngoại lai.
- C) Sử dụng phương pháp k-means clustering để phát hiện và loại bỏ các điểm ngoại lai.
- D) Sử dụng phương pháp DBSCAN để phát hiện và loại bỏ các điểm ngoại lai.

Câu 20: Tình huống: Tạo các đặc trưng từ dữ liệu thời gian: Bạn có một cột thời gian timestamp trong DataFrame. Làm thế nào để bạn tạo các đặc trưng mới như year, month, và day từ cột timestamp này?

- A) Chuyển đổi cột timestamp sang dạng datetime và sử dụng các thuộc tính .dt.year, .dt.month, .dt.day.
- B) Sử dụng hàm split để tách cột timestamp.
- C) Sử dụng hàm extract để lấy ra các giá trị year, month, và day.
- D) Sử dụng phương thức datetime.split() để tách cột timestamp.

PHẦN TỰ LUẬN (tổng số câu hỏi + thang điểm từng câu hỏi)

Câu hỏi 1: (2.5 điểm)

Cho input là một ảnh kích thước 4x4 và kernel (bộ lọc) kích thước 2x2. Tính toán output của phép tích chập (convolution) với stride = 1 và padding = 0.

Input	Kernel
1 2 0 4	0 1
1 6 2 8	-1 0
2 1 1 2	
3 4 1 1	

Câu hỏi 2: (2.5 điểm)

Cho tập kết quả test gồm 10 phần tử và giá trị thực tế (ground truth) tương ứng. Tính số TP, TN, FN, FP và accuracy. Hãy giải thích ý nghĩa của TP, TN, FN, FP.

Predicted:	1	0	1	0	1	1	0	1	1	0
Actual (Ground truth):	0	1	1	0	1	0	1	1	0	1

ĐÁP ÁN PHẦN TỰ LUẬN VÀ THANG ĐIỂM

Phân câu hỏi	Nội dung đáp án	Thang điểm	Ghi chú
I. Trắc nghiệm		5.0	
Câu 1 – 10	1.A 2.D 3.C 4.B 5.A 6.C 7.D 8.A 9.C 10.B	2.5	
Câu 11 – 20	11.C 12.B 13.A 14.A 15.D 16.C 17.A 18.D 19.B 20.A	2.5	
		5.0	
II. Tự luận		5.0	
Nội dung a.	Mỗi giá trị tính đúng đạt 0.25 điểm, 1 -6 2 4 1 7 -2 -3 1	2.0	
Nội dung b.	Đếm đúng số lượng TP, FP, TN, FN đạt 0.25 điểm; tính đúng accuracy, recall, precision đạt 0.5 điểm. <input type="checkbox"/> TP: 3 <input type="checkbox"/> TN: 1 <input type="checkbox"/> FP: 3 <input type="checkbox"/> FN: 3 Accuracy = 40%, recall=50%, precision=50%	2.5	
		5.0	
	Điểm tổng	10.0	

TP. Hồ Chí Minh, ngày 21 tháng 07 năm 2024

Người duyệt đề



Trần Ngọc Việt

Giảng viên ra đề



Huỳnh Thái Học