TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

Mã đề thi: 100

ĐỀ THI CUỐI HỌC KỲ I (2023-2024)
MÔN: LẬP TRÌNH PYTHON CHO MÁY HỌC
MÃ LỚP: CS116.O11
Thời gian: 90 phút

(Đề thi gồm có 10 trang. Sinh viên không được sử dụng tài liệu. Sinh viên làm bài trực tiếp trên đề.)

HỌ VÀ TÊN SV: ..... **MSSV: .....** **STT: .....** **PHÒNG THI: .....**	**ĐIỂM**	**CÁN BỘ COI THI**

BẢNG TRẢ LỜI TRẮC NGHIỆM

Câu 1		**Câu 2**		**Câu 3**		**Câu 4**		**Câu 5**	
Câu 6		**Câu 7**		**Câu 8**		**Câu 9**		**Câu 10**	
Câu 11		**Câu 12**		**Câu 13**		**Câu 14**		**Câu 15**	
Câu 16		**Câu 17**		**Câu 18**		**Câu 19**		**Câu 20**	
Câu 21		**Câu 22**		**Câu 23**		**Câu 24**		**Câu 25**	

I. CÂU HỎI TRẮC NGHIỆM (5 điểm; 0.2 điểm/câu; sinh viên **chọn một hoặc nhiều đáp án đúng dựa theo yêu cầu của từng câu hỏi** và điền vào **BẢNG TRẢ LỜI TRẮC NGHIỆM**. *Đối với những câu hỏi có nhiều đáp án đúng, sinh viên cần chọn và chỉ chọn tất cả đáp án đúng để được trọn vẹn điểm. Nếu chọn thiếu hoặc sai đáp án, sinh viên sẽ không được tính điểm.*)

Câu 1. *[Nhiều đáp án đúng] (G2)* Việc chuẩn hóa (**scaling**) dữ liệu trước khi huấn luyện:

A. Luôn cần thiết cho các mô hình tuyến tính có chính quy hóa (**regularized model**)

B. Không ảnh hưởng đến sự lựa chọn siêu tham tham số chính quy hóa (**regularized parameter**) tối ưu

C. Có thể tăng tốc độ huấn luyện mô hình

D. Thường hiệu quả cho các mô hình tuyến tính có chính quy hóa (**regularized model**)

Câu 2. *[Nhiều đáp án đúng] (G2)* Các mô hình đa thức với tham số bậc cao:

A. Không khớp (**underfit**) hơn so với các mô hình hồi quy tuyến tính

B. Có lỗi huấn luyện thấp hơn so với các mô hình đa thức bậc thấp

C. Có khả năng quá khớp (**overfit**) hơn so với các mô hình đa thức bậc thấp

D. Luôn có lỗi kiểm thử tốt nhất, nhưng có thể quá trình huấn luyện chậm hơn

Câu 3. *[Nhiều đáp án đúng] (G2)* Kết hợp một hoặc nhiều bộ biến đổi đặc trưng trong một **pipeline** duy nhất:

A. Đảm bảo các mô hình dự đoán chính xác bất kể phân phối của dữ liệu (tính tổng quát của mô hình)

B. Tăng khả năng biểu diễn của mô hình

C. Ngăn chặn hiện tượng không khớp (**underfitting**)

D. Có thêm các siêu tham số để tinh chỉnh (**tuning**)

Câu 4. *[Một đáp án đúng] (G2)* Biên quyết định (**decision boundary**) của một mô hình hồi quy luận lý (**logistic**):

A. Phân chia các lớp sử dụng chỉ một trong các đặc trưng đầu vào

B. Phân chia các lớp sử dụng sự kết hợp của các đặc trưng đầu vào

C. Thường có hình dạng cong

Câu 5. *[Một đáp án đúng] (G2)* Một mô hình đang bị quá khớp (**overfitting**) khi:

A. Cả lỗi huấn luyện và lỗi kiểm thử đều thấp
B. Lỗi huấn luyện cao nhưng lỗi kiểm thử thấp
C. Cả lỗi huấn luyện và lỗi kiểm thử đều cao
D. Lỗi huấn luyện thấp nhưng lỗi kiểm thử cao

Câu 6. *[Nhiều đáp án đúng] (G2)* Bộ chuẩn hóa **StandardScaler** trong **scikit-learn** với tham số mặc định:

A. Có thể giúp hồi quy luận lý (**logistic**) hội tụ nhanh hơn (ít lần lặp hơn)

B. Biến đổi các đặc trưng sao cho chúng có phạm vi tương tự nhau

C. Biến đổi các đặc trưng để nằm trong phạm vi [0.0, 1.0]

D. Biến đổi các giá trị đặc trưng ban đầu chỉ dương thành các giá trị có thể âm hoặc dương

Câu 7. *[Một đáp án đúng] (G2)* Với một tập huấn luyện cố định, bằng cách tuần tự thêm các tham số để tăng tính linh hoạt cho mô hình, chúng ta có khả năng quan sát thấy:

A. Sự chênh lệch nhỏ hơn giữa lỗi huấn luyện và lỗi kiểm thử

B. Lỗi huấn luyện giảm xuống

C. Lỗi huấn luyện tăng lên hoặc ổn định

D. Sự chênh lệch lớn hơn giữa lỗi huấn luyện và lỗi kiểm thử

Câu 8. *[Nhiều đáp án đúng] (G2)* Hiệu suất tổng quát hóa của một mô hình **scikit-learn** có thể được đánh giá bằng cách:

A. Gọi hàm **cross_validate** bằng cách truyền vào mô hình, dữ liệu và **ground truth**

B. Gọi hàm **fit** để huấn luyện trên tập huấn luyện và hàm **score** để tính điểm số trên tập kiểm thử

C. Gọi hàm **fit** để huấn luyện trên tập huấn luyện, hàm **predict** trên tập kiểm thử để dự đoán, và tính điểm số bằng cách truyền các dự đoán và **ground truth** vào một hàm độ đo (**metric**) nào đó

D. Gọi hàm **fit_transform** trên dữ liệu huấn luyện và sau đó sử dụng hàm **score** để tính điểm số trên tập kiểm thử

Câu 9. *[Nhiều đáp án đúng] (G2)* Giả sử ta có một tập dữ liệu mà mỗi dòng mô tả một công ty. Các cột nào sau đây nên được coi là **đặc trưng số học có ý nghĩa** để huấn luyện một mô hình học máy phân loại công ty:

A. Lợi nhuận của quý cuối cùng

B. Số điện thoại của bộ phận kinh doanh

C. Lĩnh vực hoạt động ("xây dựng", "bán lẻ", "năng lượng", "bảo hiểm", v.v...)

D. Số lượng nhân viên

E. Mã bưu điện của trụ sở chính

Câu 10. *[Một đáp án đúng] (G2)* Hàm **make_pipeline** (cũng như **Pipeline**):

A. Thử nghiệm nhiều mô hình cùng một lúc

B. Kết hợp một hoặc nhiều bộ biến đổi và một mô hình dự đoán

C. Tự động vẽ biểu đồ **histogram** của các đặc trưng

D. Thực hiện kiểm định chéo sử dụng các bộ biến đổi và mô hình dự đoán được truyền vào làm tham số

Câu 11. *[Nhiều đáp án đúng] (G2)* Khi cố định các tham số của mô hình, nếu tăng số lượng mẫu huấn luyện, ta có khả năng quan sát thấy:

A. Sự chênh lệch nhỏ hơn giữa lỗi huấn luyện và lỗi kiểm thử

B. Lỗi huấn luyện tăng lên hoặc ổn định

C. Sự chênh lệch lớn hơn giữa lỗi huấn luyện và lỗi kiểm thử

D. Lỗi huấn luyện giảm xuống

Câu 12. *[Nhiều đáp án đúng] (G2)* Một mô hình bị hiện tượng quá khớp (**overfitting**) khi:

A. Mô hình thường dự đoán không chính xác ngay cả trên các mẫu huấn luyện

B. Mô hình quá hạn chế các tham số khi huấn luyện và do đó bị giới hạn về khả năng biểu diễn của mô hình

C. Mô hình tập trung quá nhiều vào chi tiết nhiễu của tập huấn luyện, dẫn đến không có tính tổng quát

D. Mô hình quá phức tạp (số lượng tham số nhiều, bậc của tham số cao, hoặc hàm phi tuyến) và do đó rất linh hoạt khi học/huấn luyện

Câu 13. *[Một đáp án đúng] (G2)* Xét mô hình hồi quy tuyến tính đơn giản: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Trong số các công thức sau, công thức nào là công thức đúng để tính toán ước lượng $\hat{\beta}_0$?

A. $\hat{\beta}_0 = \bar{X} - \hat{\beta}_1 \bar{Y}$ B. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ C. $\hat{\beta}_0 = \bar{X}\bar{Y} - \hat{\beta}_1 \bar{Y}$ D. $\hat{\beta}_0 = \bar{X} - \bar{Y}$

Câu 14. *[Một đáp án đúng] (G2)* Xét dữ liệu bài toán hồi quy tuyến tính, trong số các đẳng thức sau đây (với n là số mẫu dữ liệu huấn luyện, X_i là véc-tơ của mẫu huấn luyện thứ i, và Y_i là **ground truth** của mẫu huấn luyện thứ i, đẳng thức nào sau đây **sai**:

A. $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})(Y_i)$ B. $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i)(Y_i - \bar{Y})$
C. $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i Y_i)$ D. $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i Y_i) - n\bar{X}\bar{Y}$

Câu 15. *[Nhiều đáp án đúng] (G2)* Kiểm định chéo (**cross-validation**) cho phép:

A. Ước lượng sự biến thiên hay tính ổn định của mô hình

B. Đo lường hiệu suất của mô hình một cách tổng quát hơn

C. Huấn luyện mô hình nhanh hơn

Câu 16. *[Một đáp án đúng] (G2)* Xét mô hình hồi quy tổng thể sau đây: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Trong các phát biểu sau, phát biểu nào là một trong những tính chất đại số của OLS (**Ordinary Least Squares**)?

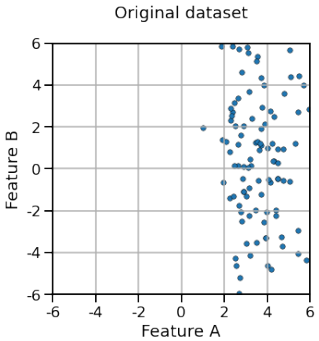
A. Hiệp phương sai mẫu (**sample covariance**) giữa biến X và phần dư (**residual**) là số dương

B. Tổng của các phần dư (**residuals**) là số dương, tức là $\sum \hat{\varepsilon}_i > 0$

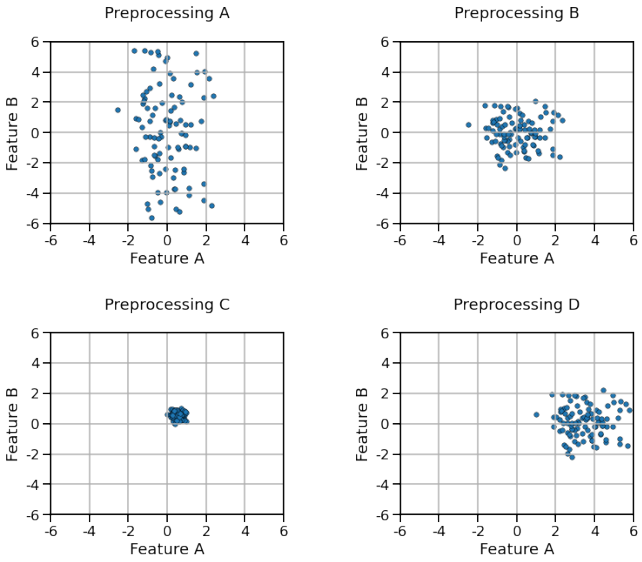
C. Tổng của các phần dư (**residuals**) là số âm, tức là $\sum \hat{\varepsilon}_i < 0$

D. Điểm $(\bar{X}, \bar{Y})$ luôn nằm trên đường hồi quy (**regression line**), trong đó $\bar{X}$ là giá trị trung bình mẫu của X và $\bar{Y}$ giá trị trung bình mẫu của Y

Câu 17. *[Một đáp án đúng] (G2)* Một tập dữ liệu 2 chiều được biểu diễn như sau:



Nếu tiền xử lý tập dữ liệu sử dụng bộ chuẩn hóa **StandardScaler** trong **scikit-learn** với các tham số mặc định, bạn hãy đoán kết quả nào sau đây:



A. Preprocessing C B. Preprocessing B C. Preprocessing D D. Preprocessing A

Câu 18. *[Nhiều đáp án đúng] (G2)* Mã hóa **one-hot**:

A. Biến đổi các biến dạng chuỗi thành dạng biểu diễn số học (có thể tính toán số học được)

B. Tạo thêm một cột cho mỗi loại giá trị phân loại (**categorical value**)

C. Biến đổi một biến số thành một biến phân loại

D. Biến đổi mỗi cột có giá trị chuỗi thành một cột có mã dạng số nguyên

Câu 19. *[Một đáp án đúng] (G2)* Nếu chúng ta huấn luyện mô hình **LinearRegression** của **scikit-learn** với **X** là một vector cột đơn và **y** là một vector, các thuộc tính **coef_** và **intercept_** của thuộc lớp **LinearRegression** sẽ lần lượt là:

A. Một mảng số thực có **shape** (1, 1) và một mảng số thực có **shape** (1,)

B. Một mảng số thực có **shape** (1,) và một số thực

C. Một mảng số thực có **shape** (1,) và một mảng số thực có **shape** (1,)

D. Một mảng số thực có **shape** (1, 1) và một số thực

Câu 20. *[Một đáp án đúng] (G2)* Với mô hình hồi quy tuyến tính đơn giản, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, dấu (dương hoặc âm) của ước lượng độ dốc ($\hat{\beta_1}$) là giống như dấu của mối tương quan (**correlation**) giữa X và Y.

A. Sai B. Đúng

Câu 21. *[Nhiều đáp án đúng] (G2)* Khi gọi hàm **cross_validate(estimator, X, y, cv=5)**, những điều sau đây sẽ được thực hiện:

A. **X** và **y** được chia thành năm phần riêng biệt với các tập kiểm thử không chồng lấn nhau.

B. **estimator.fit** được gọi 5 lần trên toàn bộ **X** và **y**

C. **estimator.fit** được gọi 5 lần, mỗi lần trên một tập huấn luyện khác nhau

Câu 22. *[Một đáp án đúng] (G2)* Chính quy hóa (**regularization**) cho phép:

A. Giảm thiểu quá khớp (**overfitting**) bằng cách ràng buộc trọng số gần với không

B. Tạo ra một mô hình bền vững trước các dữ liệu nhiễu (**outlier**) (các mẫu quá khác biệt với các mẫu khác trong tập huấn luyện)

C. Giảm thiểu không khớp (**underfitting**) bằng cách tuyến tính hóa bài toán

Câu 23. *[Nhiều đáp án đúng] (G2)* Việc sử dụng một mô hình với độ lệch (**bias**) cao:

A. Gây ra một mô hình quá khớp (**overfit**) với dữ liệu huấn luyện

B. Gây ra những lỗi mang tính hệ thống trong mô hình dự đoán

C. Mô hình dự đoán không tốt trên một mẫu huấn luyện nào đó

D. Gây ra một mô hình không khớp (**underfit**) với dữ liệu huấn luyện

Câu 24. *[Nhiều đáp án đúng] (G2)* Xét mô hình hồi quy tuyến tính đơn giản: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Trong số các công thức sau, công thức nào dưới đây là công thức đúng để tính ước lượng $\hat{\beta}_1$?

A. $\frac{\text{Sample\_Variance}(X)}{\text{Sample\_Covariance}(X,Y)}$ B. $\frac{\text{Sample\_Covariance}(X,Y)}{\text{Sample\_Variance}(X)}$ C. $\frac{\sum(X_iY_i)-n\bar{X}\bar{Y}}{\sum {X_i}^2-n\bar{X}^2}$ D. $\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$

Câu 25. *[Một đáp án đúng] (G2)* Một nút (**node**) phân chia trong một cây quyết định (**DecisionTreeClassifier** trong **scikit-learn**) thực hiện:

A. Phân làm hai nhánh quyết định dựa trên tất cả đặc trưng

B. Phân làm hai nhánh quyết định dựa trên một tổ hợp phi tuyến tính của tất cả đặc trưng

C. Phân làm hai nhánh quyết định dựa trên một đặc trưng duy nhất tại một thời điểm

D. Phân làm nhiều nhánh quyết định dựa trên một đặc trưng duy nhất

---

II. CÂU HỎI TỰ LUẬN (5 điểm)

Câu 1. (2 điểm) *(G1)* Hãy hoàn thiện đoạn mã nguồn dưới đây bằng cách áp dụng thư viện **scikit-learn** trong môi trường **Python 3**.

```
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import f1_score
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.pipeline import Pipeline
# Tải dữ liệu về hoa (iris), với X chứa các biến độc lập
# và y là biến phụ thuộc
X, y = load_iris(return_X_y=True)
# Chia dữ liệu thành 80% huấn luyện, 20% kiểm thử,
# với random_state=42 để kết quả nhất quán
X_train, X_test, y_train, y_test =




# Tạo pipeline theo trình tự gồm 3 phần liên tiếp sau:
```

```
#   i)   SimpleImputer để xử lý dữ liệu thiếu
#   ii)  StandardScaler để chuẩn hóa dữ liệu
#   iii) GradientBoostingClassifier như là thuật toán phân loại
```

```
# Thiết lập các siêu tham số cho GridSearchCV để tối ưu mô hình
#   - SimpleImputer có strategy 'mean' hoặc 'median'
#   - GradientBoostingClassifier:
#       + Số lượng cây (n_estimators): 50, 100, 150
#       + Tốc độ học (learning_rate): 0.01, 0.1, 0.2
#       + Độ sâu tối đa của cây (max_depth): 3, 4, 5
```

```
# Sử dụng GridSearchCV với phương pháp kiểm định chéo 5 lần,
# (cv=5) để chọn siêu tham số tốt nhất




# In ra màn hình bộ siêu tham số tối ưu từ GridSearchCV




# Dự đoán trên tập kiểm thử và tính toán F1-score,
# với average='weighted', và in ra màn hình giá trị F1-score
```

Câu 2. (3 điểm) *(G3)* Bạn được yêu cầu thiết kế một giải pháp để dự đoán sản phẩm phù hợp với từng người dùng trên một trang thương mại điện tử. Thông tin đầu vào sẽ bao gồm:

- Mã người dùng (*int*),
- Lịch sử mua hàng gần đây của người dùng (mảng các mã sản phẩm, *int*),
- Thời gian trung bình mỗi phiên truy cập trang web (đơn vị giây, *float*),
- Danh mục sản phẩm duyệt qua trong phiên trước đó (mảng các *string*),
- Đánh giá sản phẩm từ người dùng (thang điểm 1-5, *float*),
- Tần suất mua hàng (số lần mua hàng trong một khoảng thời gian nhất định, *int*),
- Thông tin demographic của người dùng (tuổi - *int*, giới tính - *string*, vị trí địa lý - *string*).

Trình bày quy trình xây dựng mô hình để thuyết phục khách hàng chấp thuận thực hiện dự án. Cần chú trọng chứng minh tính hợp lý và hiệu quả của mô hình, từ khâu thu thập dữ liệu đến quá trình đánh giá độ chính xác của mô hình.

HẾT

Bảng chuẩn đầu ra môn học Lập trình Python cho Máy học:

CĐRMH	Mô tả CĐRMH
G1	Làm việc ở mức độ cá nhân và cộng tác nhóm để trình bày và giải quyết một số thuật toán học không giám sát và có giám sát.
G2	Hiểu và giải thích được các khái niệm, thuật ngữ liên quan tới các quy trình xây dựng mô hình máy học, một số phương pháp phân tích, tiền xử lý dữ liệu, một số mô hình máy học có giám sát, không giám sát, đánh giá mô hình.
G3	Ứng dụng các lý thuyết, mô hình và thuật toán học có giám sát và không giám sát vào giải quyết các bài toán trong thực tế.