Mô tả bài toán cho cả 4 thuật toán RF, XGBoost, Ada Boost, Gradient Boost

Trong bài toán này, chúng ta sẽ được cho một bộ dataset mô tả thông tin về nhân viên trong một công ty, bao gồm các features liên quan đến nhân viên và mức lương của họ. Nhiệm vụ của chúng ta là phân tích, xử lí bộ data dưới đây và trả lời các câu hỏi yêu cầu người làm phải thực hiện coding.
employee_data.csv

Tham khảo Link code gốc: Ada+Gradient

Thực hiện các yêu cầu sau đây

1. **Đọc dữ liệu**
   Sử dụng pandas, đọc file csv được cung cấp, sau đó hiển thị ra màn hình để hiểu các trường dữ liệu.
2. **Label Encoding**
   Chuyển đổi các cột dữ liệu dạng chữ (cụ thể là cột "Gender" và "Position") sang dạng số bằng cách sử dụng `LabelEncoder` từ thư viện `sklearn`.
3. **Tách dữ liệu thành bộ feature (X) và label (y)**

- Sử dụng các cột "Gender", "Experience (Years)" và "Position" làm features đầu vào (`X`).
- Sử dụng cột "Salary" làm biến đầu ra (`y`).

4. **Tách tập dữ liệu thành tập train và test**

- Chia dữ liệu thành tập huấn luyện (`X_train, y_train`) và tập kiểm tra (`X_test, y_test`) với tỷ lệ `80:20`.
- Đảm bảo rằng việc chia tách dữ liệu là ngẫu nhiên nhưng tái lập (reproducibility) được với `random_state=42`

Thư viện

[ ] ↳ 1 ô bị ẩn

Tải và đọc dữ liệu

Câu hỏi trắc nghiệm yêu cầu coding

Câu 1: Điểm khác biệt chính giữa AdaBoost và Gradient Boosting trong cách mà chúng cải thiện mô hình là gì?

A) AdaBoost tập trung vào việc sửa lỗi của các mẫu dữ liệu có lỗi cao nhất, còn Gradient Boosting tập trung vào giảm thiểu giá trị lỗi toàn bộ bằng cách sử dụng đạo hàm.
B) AdaBoost sử dụng các mô hình con yếu, trong khi Gradient Boosting chỉ sử dụng mô hình con mạnh.
C) AdaBoost không thể dẫn đến overfitting, trong khi Gradient Boosting dễ bị overfitting.
D) AdaBoost và Gradient Boosting có cùng cách tiếp cận trong việc cải thiện mô hình qua các bước lặp.
Đán án: A

Câu 2: Điều gì xảy ra khi bạn tăng số lượng mô hình con (estimators) trong AdaBoost hoặc Gradient Boosting? (Thực hiện thay đổi tham số estimators để kiểm tra)

A) Hiệu suất mô hình luôn tăng khi tăng số lượng mô hình con.
B) Hiệu suất có thể tăng, nhưng nếu quá cao sẽ gây overfitting.
C) Hiệu suất giảm dần khi tăng số lượng mô hình con.
D) Hiệu suất không bị ảnh hưởng bởi số lượng mô hình con.

Đáp án: B

Huấn luyện mô hình AdaBoost

```
ada_regressor = AdaBoostRegressor(n_estimators=50, random_state=42)
ada_regressor.fit(X_train, y_train)
```

```
AdaBoostRegressor
AdaBoostRegressor(random_state=42)
```

Huấn luyện mô hình GradientBoost

```
gb_regressor = GradientBoostingRegressor(n_estimators=50, random_state=42)
gb_regressor.fit(X_train, y_train)
```

```
GradientBoostingRegressor
GradientBoostingRegressor(n_estimators=50, random_state=42)
```

Câu 3: Khi nào overfitting có thể xảy ra trong AdaBoost và Gradient Boosting? (Hãy thử nghiệm với code các trường hợp dưới đây và đưa ra kết luận)

A) Khi sử dụng quá ít mô hình con.
B) Khi sử dụng giá trị learning rate quá cao và số lượng mô hình con quá nhiều.
C) Overfitting không xảy ra trong Gradient Boosting.
D) Khi mô hình không có đủ dữ liệu để huấn luyện.
Đáp án: B

Câu 4: AdaBoost và Gradient Boosting cho phép đánh giá tầm quan trọng của các đặc trưng. Tầm quan trọng của đặc trưng nào sẽ có khả năng cao nhất trong bài toán dự đoán lương nhân viên? (Dùng phương thức feature_importances_ có sẵn trong model)

A) Gender.
B) Experience (Years).
C) Position.
D) ID.

Đáp án: B

Check important feature AdaBoost

[ ] ↳ 2 ô bị ẩn

Important Features Gradient Boost

[ ] ↳ 2 ô bị ẩn

Câu 5: Sử dụng các hàm mean_squared_error, r2_score của thư viện sklearn.metrics để tính toán giá trị MSE và R^2 (sử dụng tham số của 2 mô hình là n_estimators=50, random_state=42), trả lời câu hỏi: Dựa trên 2 giá trị trên, hiệu suất của 2 mô hình Ada & Gradient Boost như thế nào?

A. Với giá trị R^2 từ 0.7-0.8, mô hình đang hoạt động khá tốt, giá trị MSE khá tương đối không cao không thấp.

B. Với giá trị R^2 từ 0.6-0.7, mô hình đang hoạt động ở mức khá, bao quát được một phần nhưng cũng nhiều sai sót, giá trị MSE khá cao.

C. Với giá trị R2 từ 0.4-0.5, mô hình đang hoạt động ở mức tệ, không nắm bắt được phương sai trong dữ liệu. Giá tri MSE lớn.

D. Với giá trị R2 từ 0.1-0.2, mô hình dường như không học được từ dữ liệu. Giá tri MSE siêu lớn thể hiện sự sai sót khi mô hình không thể học.

Đáp án: B

Câu 6: (Cho cả 3 phần) Hãy xem đây như là một bài toán thực nghiệm, thay thế từ mô hình Random Forest, XGBoost, AdaBoost, Gradient Boost vào dữ liệu trên. Sau đó đánh giá bộ dữ liệu trên tập test và đưa ra mô hình tốt nhất được chọn tối ưu cho bộ dữ liệu này. (Thực hiện với các siêu tham số chung: n_estimators = 50, random_state=42)

A. Random Forest
B. XGBoost
c. AdaBoost
D. Gradient Boost

Đáp án:

Compare 4 models

[ ] ↳ 3 ô bị ẩn