Các phương pháp Ensemble trong Machine Learning

Mục lục

1 Giới thiệu

Ensemble là các kỹ thuật trong machine learning sử dụng nhiều mô hình, thường được gọi là "weak learners,"để cải thiện hiệu suất dự đoán tổng thể. Ý tưởng chính của ensemble là bằng cách kết hợp các dự đoán từ nhiều mô hình, kết quả đạt được sẽ tốt hơn so với từng mô hình riêng lẻ.

2 Boosting

Boosting là một kỹ thuật ensemble kết hợp các mô hình weak learner theo trình tự nhằm tạo ra một mô hình mạnh. Mỗi mô hình mới tập trung vào việc sửa lỗi của các mô hình trước đó. Các thuật toán boosting phổ biến bao gồm:

2.1 AdaBoost

AdaBoost (Adaptive Boosting) là một trong những thuật toán boosting đầu tiên được phát triển và vẫn rất phổ biến. Nó hoạt động bằng cách kết hợp nhiều mô hình "weak learner"(thường là cây quyết định nông, gọi là stumps) để tạo ra một mô hình mạnh hơn. Các mô hình này được xây dựng theo một trình tự và trọng số của mỗi mẫu dữ liệu được cập nhật để các mô hình sau tập trung vào những mẫu khó hơn

- **Thuật toán:**
  1. Khởi tạo trọng số cho tất cả các mẫu bằng nhau.
  2. Huấn luyện một weak learner và tính toán tỷ lệ lỗi.
  3. Cập nhật trọng số mẫu để tập trung vào các mẫu bị phân loại sai.
  4. Kết hợp các weak learner thành một mô hình mạnh thông qua cách bỏ phiếu có trọng số.
- **Ứng dụng:** Các bài toán phân loại như phát hiện spam, phát hiện gian lận, v.v.

2.1.1 Nguyên lý hoạt động

:

Mỗi mẫu trong tập huấn luyện được gán một trọng số bằng ban đầu, thường là $w_i = \frac{1}{n}$ với n là số lượng mẫu dữ liệu.

Huấn luyện một week learner trên tập dữ liệu hiện tại, sử dụng trọng số để xac định tầm quan trọng của từng mẫu

Sai số của mô hình được tính như sau:

$$Error = \sum_{i=1}^{n} w_i \cdot f_x(y_i \neq \hat{y}_i) \tag{1}$$

Trong đó f_x là hàm dự đoán cho giá trị đầu vào x với các giá trị $\hat{y}_i$ để so sánh với y_i.

Cập nhật trọng số của mô hình:

$$\alpha = \frac{1}{2} ln(\frac{1 - Error}{Error}) \tag{2}$$

Cập nhật trọng số của mẫu:

$$w_i \leftarrow w_i \cdot e^{\alpha f_x} \tag{3}$$

Kết hợp các week learner với nhau bằng cách lấy tổng trọng số của các dự đoán

$$F(x) = \sum_{t=1}^{T} \alpha_t \cdot h_t(x) \tag{4}$$

2.2 Gradient Boosting

Gradient Boosting xây dựng các mô hình theo trình tự, tối ưu hóa dựa trên hàm mất mát có thể đạo hàm. Mỗi mô hình tiếp theo giảm thiểu lỗi dư của các mô hình trước đó.

- **Các biến thể:** XGBoost, LightGBM, CatBoost.
- **Ứng dụng:** Các bài toán dữ liệu cấu trúc như bảng dữ liệu.

2.2.1 Nguyên lý hoạt động

Hàm mất mát: Gradient Boosting sử dụng một hàm mất mát $L(y, F(x))$, trong đó y là giá trị thực và $F(x)$ là giá trị dự đoán.

$$\min_F \sum_{i=1}^{n} L(y_i, F(x_i)) \tag{5}$$

Residuals: Tại mỗi bước, tính toán gradient (residuals) của hàm mất mát:

$$r_i^{(m)} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \tag{6}$$

Cập nhật mô hình: Mô hình được cập nhật bằng cách thêm vào một phần $\eta h_m(x)$:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{7}$$

3 Bagging

Bagging (Bootstrap Aggregating) giảm phương sai bằng cách huấn luyện nhiều mô hình trên các tập con khác nhau của dữ liệu. Dự đoán được tổng hợp thông qua trung bình hoặc bỏ phiếu.

3.1 Random Forest

Random Forest là một thuật toán dựa trên Bagging sử dụng các cây quyết định làm mô hình cơ bản. Mỗi cây được huấn luyện trên một tập dữ liệu bootstrap và các đặc trưng được chọn ngẫu nhiên để chia nhánh.

- **Ưu điểm:** Xử lý tốt các tập dữ liệu lớn, giảm overfitting.
- **Ứng dụng:** Các bài toán phân loại và hồi quy.

4 K-Fold Cross-Validation

K-Fold Cross-Validation là một kỹ thuật để đánh giá hiệu suất của mô hình bằng cách chia dữ liệu thành k tập con (folds). Mô hình được huấn luyện trên $k-1$ folds và kiểm tra trên fold còn lại. Quá trình này được lặp lại k lần.

- **Ứng dụng:** Đánh giá mô hình để tránh overfitting.
- **Kết hợp với phương pháp Ensemble:** Dự đoán từ các fold khác nhau có thể được trung bình để tạo ra mô hình mạnh hơn.

5 Kết hợp Mô hình (Merging Models)

Kết hợp mô hình bao gồm việc gộp các dự đoán từ các mô hình khác nhau, thường với các kiến trúc hoặc phương pháp huấn luyện khác nhau, để cải thiện độ chính xác.

- **Trung bình có trọng số:** Gán trọng số cho các mô hình dựa trên hiệu suất của chúng.
- **Stacking:** Sử dụng một meta-model để học cách kết hợp tối ưu các dự đoán của các mô hình cơ bản.

6 Các phương pháp Ensemble khác

6.1 Voting Classifier

Một phương pháp ensemble đơn giản, trong đó dự đoán từ nhiều mô hình được tổng hợp thông qua bỏ phiếu đa số (đối với phân loại) hoặc trung bình (đối với hồi quy).

6.2 Bagged Boosting

Kết hợp Bagging và Boosting để tăng cường sự ổn định bằng cách giảm cả bias và phương sai.

7 Ưu điểm của Ensemble Learning

- Cải thiện độ chính xác và tính ổn định.
- Giảm nguy cơ overfitting (đặc biệt với Bagging).
- Ứng dụng được cho nhiều bài toán machine learning khác nhau.

8 Thách thức của Ensemble Learning

- Tăng độ phức tạp tính toán.
- Khó khăn trong việc diễn giải dự đoán của mô hình.
- Cần tinh chỉnh siêu tham số cẩn thận.

9 Kết luận

Ensemble là các kỹ thuật mạnh mẽ trong machine learning, có thể cải thiện đáng kể hiệu suất dự đoán. Bằng cách tận dụng sức mạnh của nhiều mô hình, các phương pháp này giúp giảm thiểu các điểm yếu và tạo ra các giải pháp mạnh mẽ cho các bài toán phức tạp. Tuy nhiên, cần cân nhắc kỹ giữa chi phí tính toán và hiệu suất đạt được.