TRƯỜNG ĐẠI HỌC VINH

MỘT SỐ LỖI TRONG CÁC MÔ HÌNH HỌC MÁY

Phan Anh Phong, PhD.
Vinh University

1

Nội dung

- Hiểu các lỗi khi xây dựng mô hình trong học máy
- Thẩm định chéo (Cross-validation)
- Phân lớp Naïve Bayes
- Ưu điểm và nhược điểm phân lớp Naïve Bayes
- Thảo luận

2

Lựa chọn siêu tham số

- Mỗi mô hình học máy đều cần tinh chỉnh các siêu tham số (Hyperparameter) của mô hình
- Ví dụ trong thuật toán phân lớp KNN
  - Chọn tham số k
  - Chọn một độ đo khoảng cách
- Mục tiêu: Muốn chọn các giá trị cho các siêu tham số để mang lại hiệu suất tốt nhất của mô hình học máy trên dữ liệu thử nghiệm
- Dữ liệu thử nghiệm = Dữ liệu huấn luyện + Dữ liệu kiểm tra
- Dữ liệu huấn luyện GIAO dữ liệu kiểm tra = RỖNG

3

Khái quát hóa về lỗi trong ML

- Mô hình đã học có khả năng khái quát hóa tốt như thế nào từ dữ liệu đã được huấn luyện sang tập thử nghiệm mới?
- LỖI trong các mô hình học máy:
  - Độ chệch (Bias): sự chênh lệch giữa giá trị trung bình mà mô hình dự đoán và giá trị thực tế của dữ liệu trên tập dữ liệu huấn luyện
  - Phương sai (Variance): độ phân tán của các giá trị mà mô hình dự đoán so với giá trị thực tế trên tập dữ liệu kiểm tra/xác thực
- Nguyên nhân của bị chệch thường là do mô hình quá đơn giản trong khi dữ liệu có mối quan hệ phức tạp hơn và thậm chí nằm ngoài khả năng biểu diễn của mô hình. Tuy nhiên, một mô hình quá phức tạp lại có khả năng xảy ra lỗi phương sai lớn.

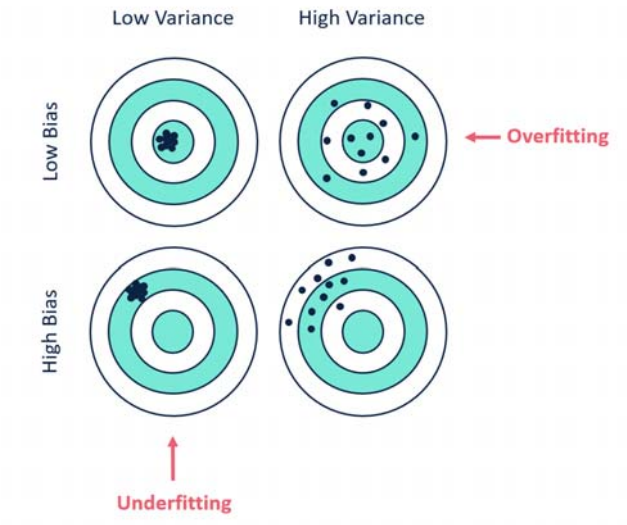
4

Overfiting và Underfiting

- Mô hình quá khớp (overfiting):
  - Mô hình quá "phức tạp" và phù hợp với các đặc điểm không liên quan (nhiễu) trong dữ liệu
  - **Độ chệch thấp và phương sai cao,** tức là lỗi huấn luyện thấp và lỗi kiểm tra cao
- Mô hình chưa khớp (underfitting):
  - Mô hình quá "đơn giản" để thể hiện tất cả các đặc điểm của dữ liệu liên quan
  - **Độ chệch cao và phương sai thấp,** tức là lỗi huấn luyện cao và lỗi kiểm tra thấp

5

Overfiting và Underfiting



6

Giảm variance – tránh quá khớp

- Chọn thuật toán phân lớp (bộ phân lớp) đơn giản hơn
- Bổ sung dữ liệu huấn luyện (training data)
- Kiểm tra chéo (Cross-validate): là một phương pháp thống kê được sử dụng để ước lượng hiệu quả của các mô hình học máy. Nó thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán

Slide credit: D. Hoiem

7

Tránh quá khớp – Cơ bản

- Kiểm tra chéo:
  - Mô hình quá "đơn giản" để thể hiện tất cả các đặc điểm của dữ liệu liên quan
  - **Độ chệch thấp và phương sai cao,** tức là lỗi huấn luyện thấp và lỗi kiểm tra cao
  - ***Kỹ thuật k-folds***: chia tập huấn luyện thành k tập con kích thước gần bằng nhau và không giao nhau. Tại mỗi lần thử một trong k tập con đó fdwocj lagmf tập xác thực, k-1 tập con còn lại dung để huấn luyện. Như vậy mỗi bộ tham số mô hình ta có k mô hình khác nhau. Sai số huấn luyện và sai số kiểm tra là trung bình cộng của các giá trị tương ứng trong k mo hình.
  - Thường chọn k=10, k=5. Nếu k bằng số lượng phần tử của tập huấn luyện thì gọi là là kỹ thuật ***leave one out***

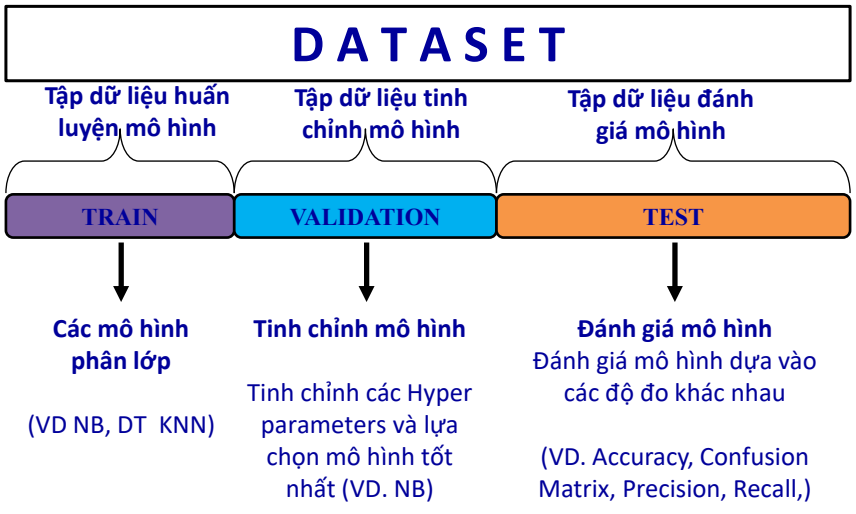
8

Đánh giá mô hình phân lớp

- Trả lời những câu hỏi :
  - Mô hình đã được huấn luyện thành công hay chưa?
  - Mức độ thành công của mô hình tốt đến đâu?
  - Khi nào nên dừng quá trình huấn luyện?
  - Khi nào nên cập nhật mô hình?
- Phương pháp đánh giá
  - Cách chia dữ liệu
  - Tiêu chí đánh giá
  - Độ đo (thước đo) sử dụng

9

Cách chia dữ liệu



10

Cách chia dữ liệu

- Một bộ dữ liệu có thể chia 2 hoặc 3 : huấn luyện mô hình, tinh chỉnh mô hình (fine-tuning) và đánh giá mô hình.
- Tùy thuộc tính chất của bộ dữ liệu (số mẫu dữ liệu lớn hay nhỏ, có cân bằng hay không…) => có nhiều cách để phân chia khác
  - Hold-out/Repeated Hold-out (thường được sử dụng khi dataset lớn)
  - K-Fold
  - Leave-one-out
  - Statified sampling (dùng cho imbalanced dataset)
  - Bootstrap sampling
  - …

11

Tiêu chí đánh giá

- Tính chính xác (Accuracy)
- Tính hiệu quả (Efficiency): Chi phí về thời gian và tài nguyên (bộ nhớ) cần thiết cho việc huấn luyện và kiểm thử hệ thống
- Khả năng xử lý nhiễu (Robustness): Khả năng xử lý (chịu được) của hệ thống đối với các ví dụ nhiễu (lỗi) hoặc thiếu giá trị
- Khả năng mở rộng (Scalability): Hiệu năng của hệ thống (vd: tốc độ học/phân lớp) thay đổi như thế nào đối với kích thước của tập dữ liệu tăng
- Khả năng diễn giải (Interpretability): Mức độ dễ hiểu (đối với người sử dụng) của các kết quả và hoạt động của hệ thống
- Mức độ phức tạp (Complexity): Mức độ phức tạp của mô hình hệ thống (hàm mục tiêu) học được

12

Độ đo đánh giá mô hình

- Accracy (độ chính xác): là tỉ lệ giữa số mẫu dữ liệu được dự đoán đúng và tổng số mẫu được kiểm tra.
  - Nhược điểm của Accuracy là chỉ cho ta biết độ chính xác khi dự báo của mô hình, nhưng không thể hiện mô hình đang dự đoán sai như thế nào
  - Accuracy lộ rõ hạn chế khi được sử dụng trên bộ dữ liệu không cân bằng (imbalanced dataset)
- Ma trận nhầm lẫn (Confusion Matrix)

13

Độ đo đánh giá mô hình

- Precision cho biết trong số các mẫu dữ liệu được mô hình phân lớp vào lớp Positive, có bao nhiêu mẫu thực sự thuộc lớp Positive.
- Recall cho biết được có bao nhiêu mẫu dữ liệu thực sự ở lớp Positive được mô hình phân lớp đúng trong mọi mẫu dữ liệu thực sự ở lớp Positive.
- Precision và Recall có giá trị trong [0,1], hai giá trị này càng gần với 1 thì mô hình càng chính xác. Precision càng cao đồng nghĩa với các điểm được phân loại càng chính xác. Recall càng cao cho thể hiện cho việc ít bỏ sót các điểm dữ liệu đúng.

14

Độ đo đánh giá mô hình

- Ma trận nhầm lẫn (Confusion Matrix)

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

15

Độ đo đánh giá mô hình

		Nhãn lớp thực tế	
		Positive	Negative
Phân lớp dự đoán	Positive	20	70
	Negative	80	930

- Accuracy = (20+930)/(20+80+70+930) = 86.36%
- TP_{rate} = Recall = 20/(20+80) = 20%
- Accuracy = (10+930)/(10+90+70+930) = 85.45%
- TP_{rate} = Recall = 10/(10+80) = 10%
- Accuracy = (00+930)/(00+100+70+930) = 84.54%
- TP_{rate} = 00%

16