ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Khai thác dữ liệu

Phân lớp và đánh giá mô hình

Nguyễn Ngọc Đức

2022

Nội dung

k láng giềng gần nhất

Học chủ động và bị động

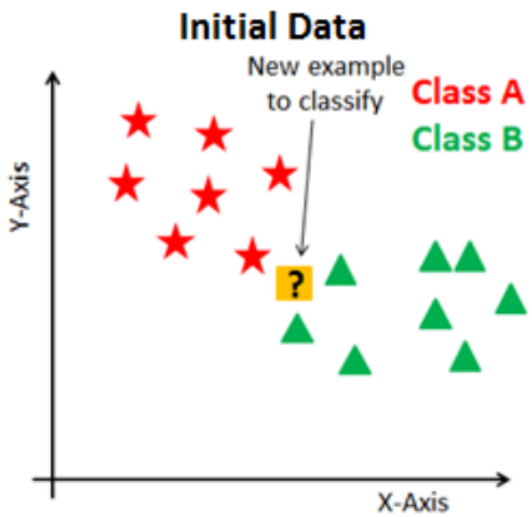
- Học chủ động: Xây dựng một mô hình phân lớp từ tập các dữ liệu huấn luyện trước khi nhận dữ liệu phân lớp mới.
- Học bị động: Lưu dữ liệu huấn luyện trì hoãn đến khi nhận một thể hiện dữ liệu kiểm tra.

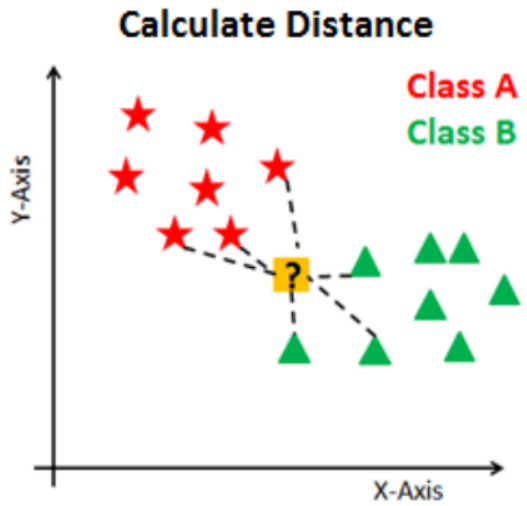
Học chủ động và bị động

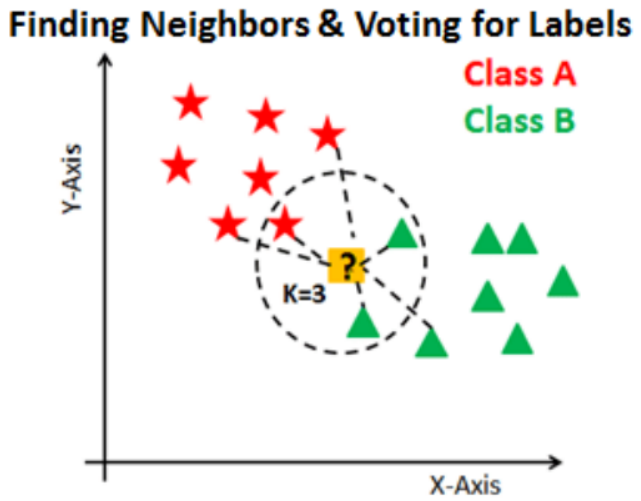
- Học chủ động: Xây dựng một mô hình phân lớp từ tập các dữ liệu huấn luyện trước khi nhận dữ liệu phân lớp mới.
- Học bị động: Lưu dữ liệu huấn luyện trì hoãn đến khi nhận một thể hiện dữ liệu kiểm tra.
  - Thời gian dự đoán cao

k láng giềng gần nhất

- Đồng thuận đa số: phân lớp một đối tượng vào lớp có k láng giềng gần nhất
  - k thường là các số nguyên nhỏ như 1, 3, 5,..
- Các láng giềng gần nhất được xác định dựa trên công thức tính khoảng cách (Euclidean, Manhattan,...)







Ví dụ

ID	Age	Income (K)	No. Cards	Response	Euclidean distance to unseen record
1	35	35	3	Yes	22.14
2	22	50	2	No	20.9
3	28	40	1	Yes	21.35
4	45	100	2	No	44.11
5	20	30	3	Yes	34.06
6	34	55	2	No	8.12
7	63	200	1	No	145.54
8	55	140	2	No	85.01
9	59	170	1	No	115.28
10	25	40	4	Yes	23.37
Unseen	**42**	**56**	**3**	?	

- **k = 5**: 3 "Yes" samples and 2 "No" samples → the class assigned is **Yes**

Chuẩn hóa dữ liệu k-NN

- Các thuộc tính của dữ liệu có thể có các miền giá trị khác nhau
- Tính khoảng cách trực tiếp cho các ước lượng không chính xác

ID	Age	Income (K)	No. Cards	Response	Distance
1	0.35	0.03	0.67	Yes	0.21
2	0.05	0.12	0.33	No	0.57
3	0.19	0.06	0	Yes	0.75
4	0.58	0.41	0.33	No	0.43
5	0	0	0.67	Yes	0.53
6	0.33	0.15	0.33	No	0.38
7	1	1	0	No	1.19
8	0.81	0.65	0.33	No	0.67
9	0.91	0.82	0	No	1.03
10	0.12	0.06	1	Yes	0.52
Unseen	0.51	0.15	0.67	?	

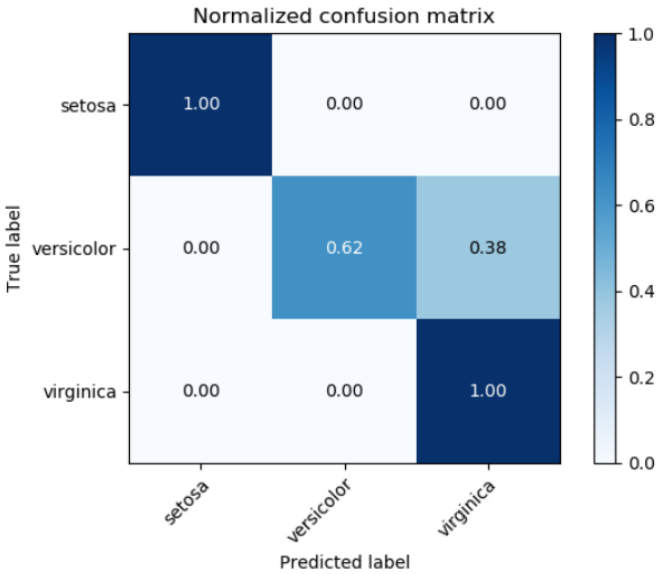
- **k = 5**: 2 "Yes" samples and 3 "No" samples
- The class assigned is **No**

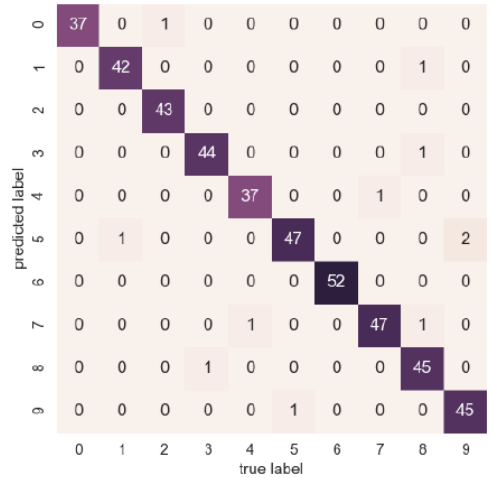
kNN: đánh giá thuật toán

- Ưu điểm:
  - Dễ dàng sử dụng và cài đặt
  - Khả năng chống nhiễu cao bằng cách tính trung bình k láng giềng gần nhất
- Nhược điểm
  - Tất cả các mẫu dữ liệu cần được lưu trữ: Tốn chi phí bộ nhớ, thời gian thực thi cao
  - Giá trị k có ảnh hưởng lớn đến hiệu suất của thuật toán
    - k quá bé: không đủ thông tin để đưa ra quyết định
    - k quá lớn: khả năng chứa dữ liệu nhiễu cao, các khu vực của các lớp có khả năng chồng lên nhau

Confusion matrix

- Một phần tử c_{ij} trong một confusion matrix là giá trị các lớp i được mô hình phân lớp vào j





		Predicted class		Total
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

Accuracy, Error rate, Sensitivity

		Predicted class		
		C_1	$\neg C_1$	
Actual Class	C_1	True Positives (TP)	False Negatives (FN)	P
	$\neg C_1$	False Positives (FP)	True Negatives (TN)	N
		P'	N'	All

- **Accuracy:** $Accuracy = (TP + TN)/All$
- **Error rate:** $Error\ rate = 1 - Accuracy = \frac{FP+FN}{All}$
- **Sensitivity** – TP recognition rate: $Sensitivity = TP/P$
- **Specificity** – TN recognition rate: $Specificity = TN/N$

Precision, Recall, F-score

- Precision:

$$P = \frac{TP}{TP + FP}$$

- Recall:

$$R = \frac{TP}{TP + FN}$$

- F1-score: **Harmonic mean**

$$F = \frac{2 \times P \times R}{P + R}$$

Ví dụ

		Predicted class		Total
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

Ví dụ

		Predicted class		Total
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

- Accuracy: $(90 + 9560)/10000 = 0.964$

Ví dụ

		Predicted class		Total
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

- Accuracy: $(90 + 9560)/10000 = 0.964$
- Precision: $90/230 = 0.391$

Ví dụ

		Predicted class		Total
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

- Accuracy: $(90 + 9560)/10000 = 0.964$
- Precision: $90/230 = 0.391$
- Recall: $90/300 = 0.3$

Ví dụ

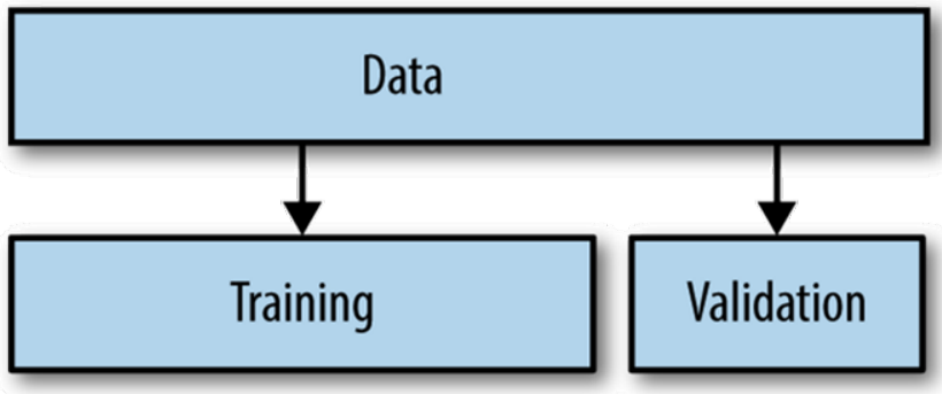
		Predicted class		Total
		cancer = yes	cancer = no	
Actual Class	cancer = yes	**90**	**210**	300
	cancer = no	**140**	**9560**	9700
Total		230	9770	10000

- Accuracy: $(90 + 9560)/10000 = 0.964$
- Precision: $90/230 = 0.391$
- Recall: $90/300 = 0.3$
- Mô hình dự đoán ung thư tốt?

Các chiến lược kiểm thử

Holdout

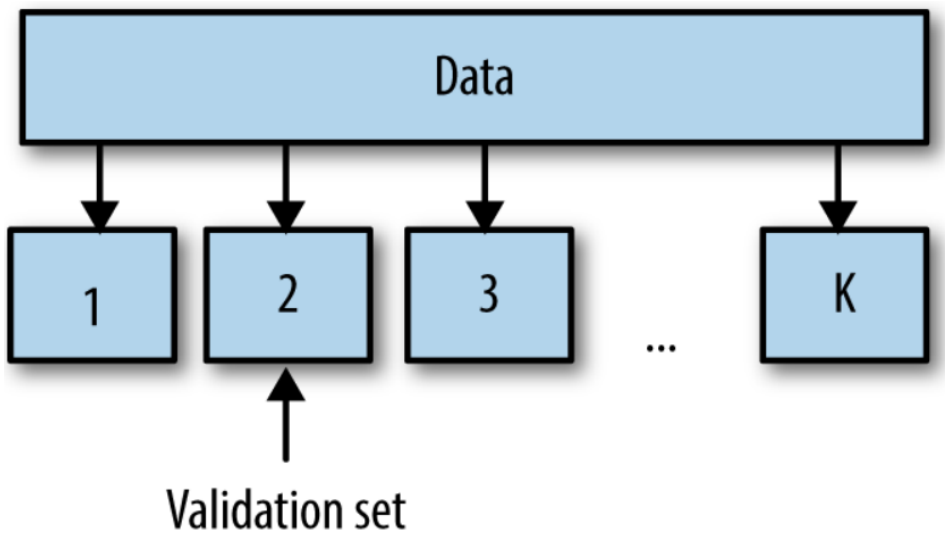
- Dữ liệu được chia ngẫu nhiên thành 2 tập
  - Ví dụ: 2/3 dữ liệu cho việc xây dựng mô hình (traning set) và 1/3 cho việc kiểm thử hiệu suất mô hình (validation set)



- Random sampling:
  - Sử dụng holdout k lần
  - Độ chính xác: Trung bình

Cross-validation

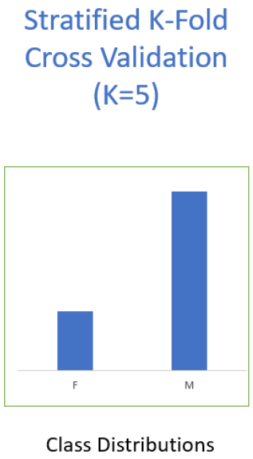
- Chia ngẫu nhiên dữ liệu thành k tập con riêng biệt với kích thước bằng nhau

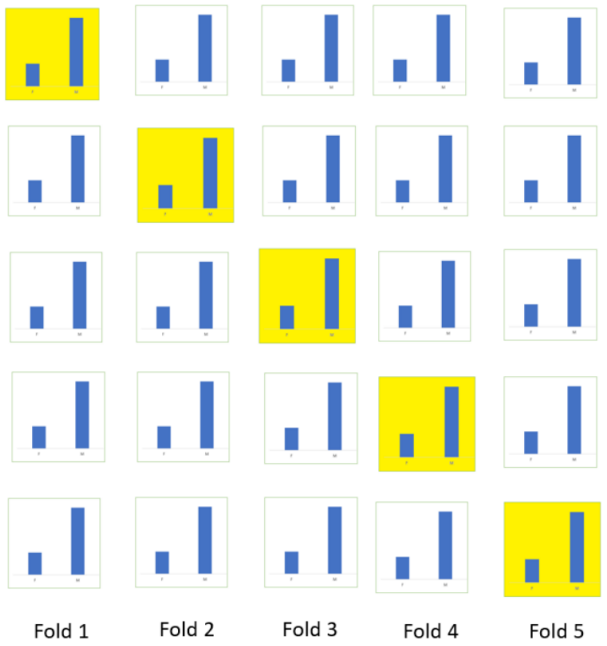


  - Thông thường người ta sẽ lựa chọn k=10
- Tại bước thứ i, sử dụng tập con thứ i làm tập kiểm thử và các tập còn lại làm tập huấn luyện

Cross-validation

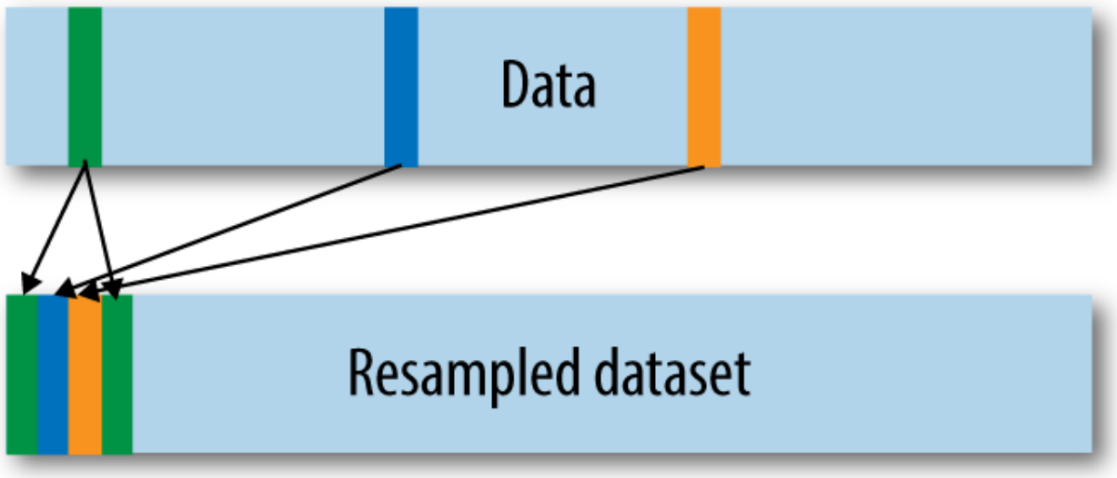
- Leave-one-out cross-validation: số lượng các fold tương ứng với số lượng thể hiện dữ liệu.
- Stratified cross-validation: các fold được phân tầng theo lớp phân phối dữ liệu tuân theo phân phối dữ liệu ban đầu.





Bootstrap

- Hoạt động tốt với các bộ dữ liệu nhỏ
- Các bộ huấn luyện được lấy mẫu đồng nhất và có khả năng thay thế



Kiểm định thống kê

Kiểm định thống kê I

- **Kiểm định thống kê (Significance test):** là phương pháp sử dụng dữ liệu để tổng hợp **minh chứng** cho một giả thuyết
- Trước khi kiểm định ý nghĩa của một giả thuyết, ta cần phải định nghĩa
  1. **Phương pháp đánh giá**
  2. **Tham số** của giả thuyết

Kiểm định thống kê II

Các bước thực hiện:

1. Giả định
   - Mỗi phép kiểm định ý nghĩa của giả thuyết đều được thực hiện dựa trên một số điều kiện cũng như giả định cụ thể.
   - Trước hết, giả định dữ liệu có được dựa trên các phương pháp ngẫu nhiên.
   - Một số giả định khác như kích thước mẫu cũng như dạng phân phối.

Kiểm định thống kê III

2. Giả thuyết:
   - Giả thuyết null (Null Hypothesis)
   - Giả thuyết thay thế (Alternative Hypothesis)

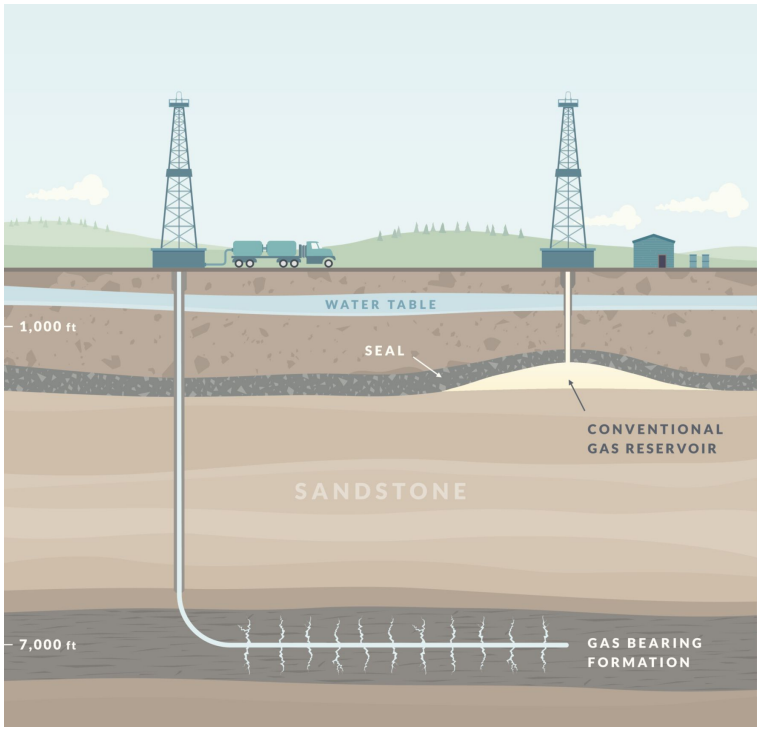
Giả thuyết null, giả thuyết thay thế

- Một **giả thuyết null** kết luận cụ thể tham số của giả thuyết.
- Một **giả thuyết thay thế** kết luận khoảng thay thế của tham số.
- Thông thường giả thuyết null được ký hiệu H_0, và giả thuyết thay thế được ký hiệu H_a.

Kiểm định thống kê IV

Ví dụ:

- Cắt thủy lực là một phương pháp khoan sử dụng áp lực nước và hóa chất để khai thác dầu khí từ các hóa thạch dưới lòng đất.



Kiểm định thống kê V

- Mặc dù lợi ích kinh tế cao, ngày càng có nhiều quan ngại do những tác động tiềm tàng của cắt thủy lực đối với môi trường.
- Một số ở các tiểu bang Mỹ và các quốc gia khác đã cấm sử dụng phương pháp này.
- Khảo sát những người phản đối việc gia tăng sử dụng cắt thủy lực ở Hoa Kỳ vẫn chỉ chiếm thiểu số.
- Giả sử p biểu thị tỷ lệ người dân Hoa Kỳ phản đối tăng cường sử dụng cắt thủy lực.

Kiểm định thống kê VI

- Ở Hoa Kỳ, tỷ lệ người phản đối tăng cường sử dụng cắt thủy lực thấp hơn 50%.
  a. Kết luận trên là giả thuyết null hay giả thuyết thay thế?
  b. Tỷ lệ phản đối là 50%?

Kiểm định thống kê VII

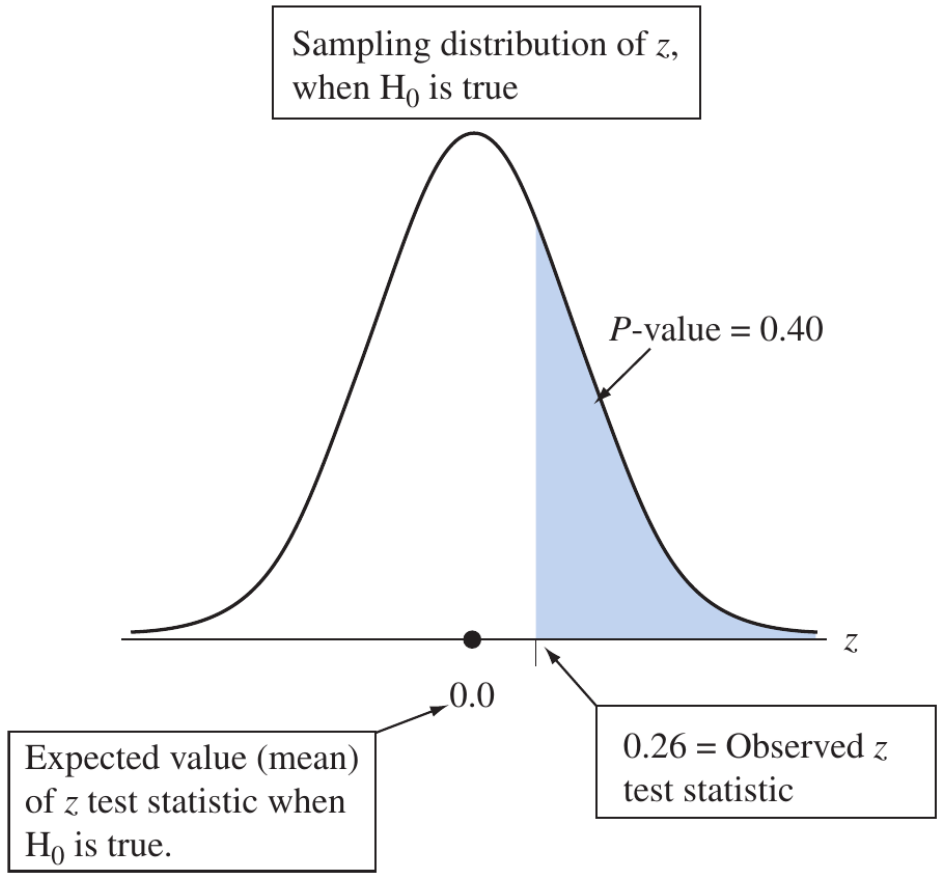
3. Kiểm định thống kê: mô tả cách biệt giữa tham số ước lượng và tham số của giả thuyết null
   - Cụ thể: quan sát phân phối Z

$$z = \frac{\overline{x} - \mu_0}{se_0}$$

   - Ví dụ: xác suất chiêm tinh gia dự đoán đúng trong mẫu dữ liệu có kích thước 116 là $\hat{p} = \frac{40}{116} = 0.345$. Giá trị $z = 0.26$
   - Sai số của giá trị ước tính với giá trị của null hypothesis là 0.26

Kiểm định thống kê VIII

4. Tính p-value:
   - P-value mô tả sự bất thường của dữ liệu so với giả thuyết null hay xác suất mà kiểm định thống kê cho giá trị giống hoặc hiếm gặp hơn.
5. Kết luận.



Tổng kết

Tổng kết

- Thuật toán phân lớp: kNN
- Độ đo đánh giá hiệu suất: Accuracy, Precision, Recall
- Kiểm thử mô hình: Holdout, Cross Validation, Bootstrap
- Kiểm định thống kê

References I

[1] Nguyễn Ngọc Thảo

Slides bài giảng môn Khai thác dữ liệu

Bộ môn Khoa học máy tính.