

Review: Máy học

Chương 1: Giới thiệu

1) Trình bày một ứng dụng của máy học và đề xuất phương pháp để giải quyết

- Project môn học : bài toán là gì? Data được sử dụng? Input là gì? Output là gì? Phương pháp sử dụng

2) Sự khác biệt giữa phân lớp (classification) và gom nhóm (clustering). Cho ví dụ minh họa

- Phân lớp: quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp. Tập dữ liệu được xây dựng trước đó có gán nhãn. Là kỹ thuật dựa vào học có giám sát.

- Gom nhóm: quá trình gom các đối tượng dữ liệu vào thành từng nhóm (cluster) sao cho các đối tượng trong cùng một nhóm có sự tương đồng theo một tiêu chí nào đó. Tập dữ liệu chưa được đánh nhãn. Là kỹ thuật dựa vào học không giám sát

- Cho ví dụ:

3) Sự khác biệt giữa phân lớp (classification) và hồi quy (regression). Cho ví dụ minh họa.

- Phân lớp: quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp. Tập dữ liệu được xây dựng trước đó có gán nhãn. Biến đầu ra dạng định tính (kiểu rời rạc/thứ bậc/định danh)

- Hồi quy: quá trình mô hình hóa, định lượng hóa mối quan hệ giữa biến đầu vào (các biến độc lập) và biến đầu ra (biến mục tiêu, biến phụ thuộc) để xác định được sự thay đổi của biến đầu ra khi biến đầu vào thay đổi. Biến đầu ra là định lượng (liên tục/dạng số/có thứ tự)

- Cho ví dụ:

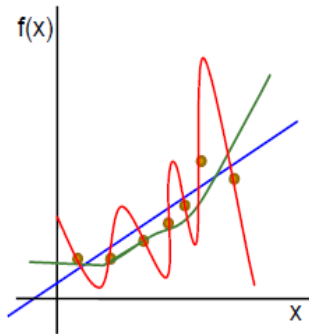
4) Overfitting và underfitting. Cho ví dụ minh họa. Các kỹ thuật để giảm overfitting.

- Overfitting: là hiện tượng mô hình hoạt động tốt trên tập huấn luyện nhưng đạt kết quả kém trên tập kiểm tra.

- Underfitting: là hiện tượng mô hình không đạt được độ chính xác cao trên tập huấn luyện và cũng như tổng quát trên cả tập dữ liệu. Có thể hiểu là mô hình không học được gì cả

- Ví dụ.

Hàm mục tiêu $f(x)$ nào
đạt độ chính xác cao nhất
đối với các ví dụ sau này?



- Các kỹ thuật để giảm overfitting: kỹ thuật regularization, kỹ thuật cross-validation hoặc VC dimension để đo độ phức tạp của mô hình, đơn giản hóa mô hình học (kỹ thuật pruning trong cây quyết định; weight decay hoặc dropout trong mạng neural và deep learning), ...

Chương 2: Hồi quy tuyến tính

5) Định nghĩa hồi quy tuyến tính. Cho ví dụ. Phương pháp ước lượng tham số để giải bài toán hồi quy tuyến tính đa biến.

- Xem slide bài giảng (định nghĩa, ví dụ, phương pháp ước lượng tham số đa biến)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d$$

$$X^T X \hat{\beta} = X^T Y$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

6) Phương pháp gradient descent và ứng dụng của nó để giải bài toán hồi quy tuyến tính.

- Xem slide bài giảng (phương pháp gradient descent cho trường hợp tổng quát)
- Áp dụng cho hồi quy tuyến tính (định nghĩa được loss function của hồi quy tuyến tính).
Trình bày gradient descent cho loss function này.

Gradient descent cho hàm nhiều biến: tìm global minimum cho hàm $f(\theta)$ trong đó θ là một vector.

- Thuật toán Gradient Descent

Khởi tạo ngẫu nhiên θ_0 .

Lặp (cho đến khi hội tụ hoặc số lượng vòng lặp vượt quá một ngưỡng)

{

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} f(\theta_t)$$

}

Trình bày loss function của hồi quy tuyến tính

Đạo hàm của hàm mất mát là:

$$\begin{aligned} J(\mathbf{w}) &= \frac{1}{2N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \\ &= \frac{1}{2N} \sum_{i=1}^N (\mathbf{x}_i \mathbf{w} - y_i)^2 \end{aligned}$$

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i^T (\mathbf{x}_i \mathbf{w} - y_i)$$

- Trình bày một trong 3 phương pháp :

Batch Gradient Descent

Stochastic Gradient Descent

Mini-batch Gradient Descent

Chương 3: Neural Network – Perceptron

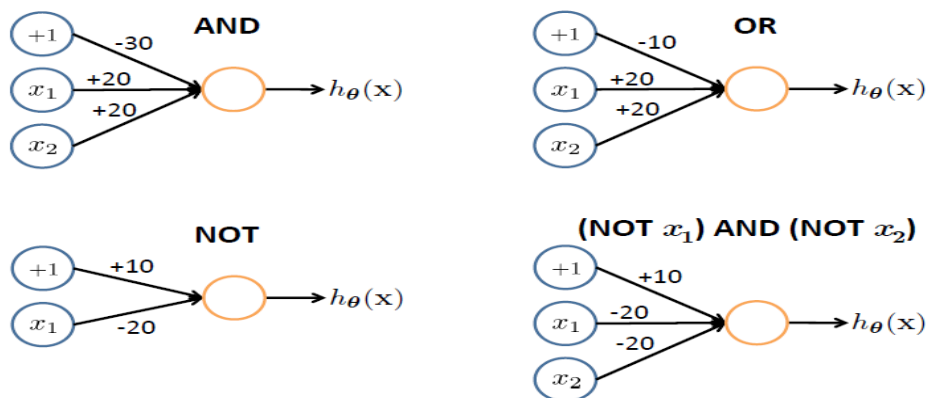
7) Định nghĩa Perceptron. Vẽ hình minh họa Áp dụng Perceptron để biểu diễn các hàm boolean như AND, OR, NOT.

- Xem slide bài giảng

- Perceptron: là một mạng neural đơn giản, bao gồm input layer (lớp đầu vào), output layer (lớp đầu ra). Perceptron được dùng trong các bài toán phân lớp nhị phân. Perceptron được xem như là bộ phân lớp tuyến tính.

- Vẽ hình minh họa (xem slide bài giảng)

- Dùng Perceptron biểu diễn AND, OR, NOT (xem slide bài giảng)

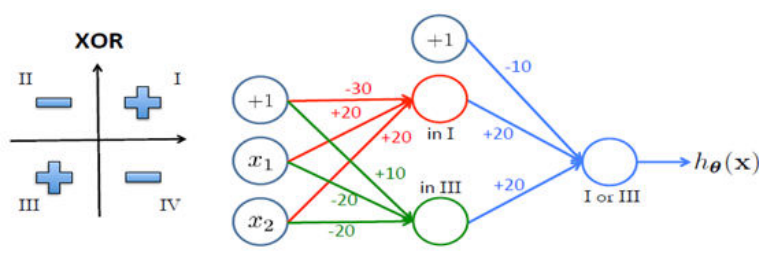


8) Tại sao Perceptron không thể biểu diễn phép toán XOR? Trình bày cách khắc phục vấn đề này.

- Với hàm XOR là dữ liệu phi tuyến, không thể phân tách tuyến tính, tức không thể tìm được 1 đường thẳng giúp phân chia hai lớp riêng biệt.

- Kết hợp các Perceptron thành Multi-layer Perceptron để biểu diễn được hàm XOR.

- Trình bày ví dụ (xem slide bài giảng)



Chương 4: Mô hình học máy kết hợp (ensemble model)

9) Định nghĩa mô hình học máy kết hợp. Ví dụ?

- Xem slide bài giảng
- Ví dụ:

10) Định nghĩa Bagging. Trình bày thuật toán Bagging

- Xem slide bài giảng
- Trình bày thuật toán Bagging có 5 bước

11) Định nghĩa Boosting. Trình bày thuật toán Adaboost.

- Xem slide bài giảng
- Trình bày thuật toán Adaboost có 3 bước

Chương 5: Gom nhóm k-means và phân cấp (hierarchical clustering)

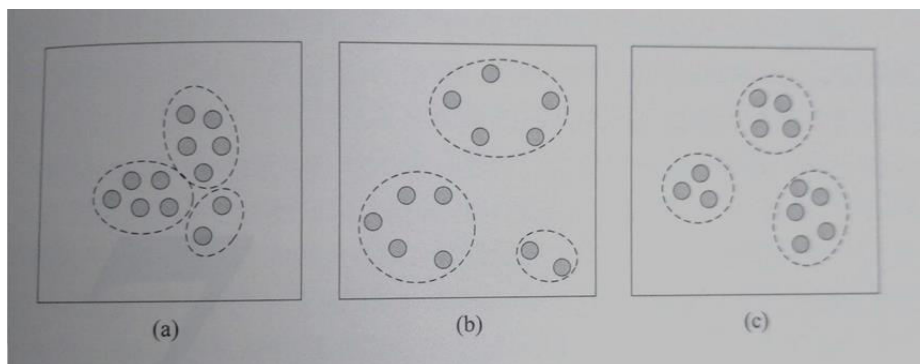
12) Thuật toán gom nhóm k-means, ưu và khuyết điểm của k-means

- Xem slide bài giảng .

13) Kết quả của k-means phụ thuộc vào quá trình khởi tạo trọng tâm ban đầu. Trình bày ví dụ minh họa và hướng tiếp cận để khắc phục vấn đề này.

- Xem ví dụ trong slide bài giảng. dùng k-means++ để khắc phục.

14) Cho biết kết quả gom nhóm nào tốt nhất? Giải thích?



- Kết quả gom nhóm trong hình (c) là tốt nhất vì các nhóm phân biệt nhau rất rõ ràng. Khoảng cách các phần tử trong mỗi nhóm gần nhau, tập trung xung quanh trọng tâm của mỗi nhóm ; khoảng cách giữa các nhóm xa nhau.

15) Thuật toán gom nhóm bằng phân cấp (hierarchical clustering). Trình bày ví dụ minh họa.

- Xem slide bài giảng 2 phương pháp Agglomerative và Divisive. Ví dụ: xem hình trong slide

16) Cho hai nhóm (clusters) c_1 và c_2 như sau:

$$- c_1: (1, 1)^T, (2, 2)^T, (2, 3)^T$$

$$- c_2: (4, 1)^T, (4, 2)^T, (5, 4)^T$$

Tính : $D_{\max}$ (Complete-linkage), $D_{\min}$ (Single-linkage) và D_{ave} (Average-linkage) giữa hai nhóm c_1 và c_2 (giả sử $D(x_1, x_2)$ dựa vào khoảng cách Euclid).

- Tính khoảng cách giữa các điểm trong nhóm c_1 và các điểm trong nhóm c_2

$$d_{11} = d((1,1)^T, (4,1)^T) = 3 \quad ; \quad d_{12} = d((1,1)^T, (4,2)^T) = 3.1623 \quad ; \quad d_{13} = d((1,1)^T, (5,4)^T) = 5$$

$$d_{21} = d((2,2)^T, (4,1)^T) = 2.2361 \quad ; \quad d_{22} = d((2,2)^T, (4,2)^T) = 2 \quad ; \quad d_{23} = d((2,2)^T, (5,4)^T) = 3.6056$$

$$d_{31} = d((2,3)^T, (4,1)^T) = 2.8284 \quad ; \quad d_{32} = d((2,3)^T, (4,2)^T) = 2.2361 \quad ; \quad d_{33} = d((2,3)^T, (5,4)^T) = 3.1623$$

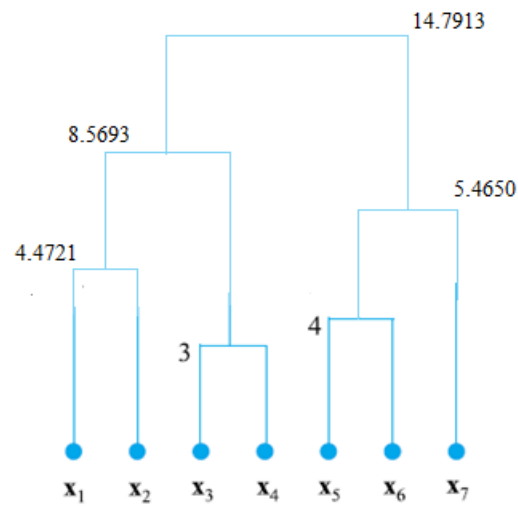
Ta được :

$$D_{\max} = 5$$

$$D_{\min} = 2$$

$$D_{\text{ave}} = (1/9) * (d_{11} + d_{12} + d_{13} + d_{21} + d_{22} + d_{23} + d_{31} + d_{32} + d_{33}) = 27.2307 * (1/9) = 3.0256$$

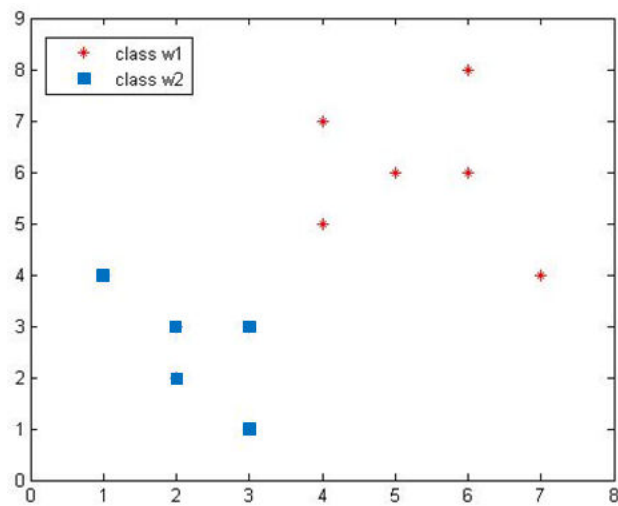
17) Quá trình gom nhóm được biểu diễn bằng cấu trúc cây (dendrogram). Trình bày cách tạo ra 3 nhóm từ cây phân cấp trên



- Cắt ngang cây dendrogram tại level có ngưỡng trong khoảng (5.4650, 85.693) ta được 3 nhóm (x_1, x_2) , (x_3, x_4) , (x_5, x_6, x_7)

Chương 6: SVM và phân lớp tuyến tính

18) Định nghĩa phân lớp tuyến tính. Cho hình sau . Các mẫu của hai lớp này có thể phân tách tuyến tính không?



- Phân lớp tuyến tính là mô hình dựa trên giá trị của sự kết hợp tuyến tính của các đặc tính. Các phân lớp tuyến tính tạo ra được các biên quyết định là các siêu phẳng để phân tách dữ liệu ra thành các nhóm riêng biệt nhau.

- Các mẫu dữ liệu như trong hình có thể phân tách tuyến tính vì có thể dùng biên quyết định đường thẳng $d_2(x) = x_1 + x_2 - 7.5$ để phân tách thành hai nhóm riêng biệt nhau.

19) Sử dụng data trong câu (a), cho trước hai biên quyết định (decision boundary) $d_1(x) = 3.5x_1 - 12.25$ và $d_2(x) = x_1 + x_2 - 7.5$.

Biên quyết định nào tốt hơn (có margin lớn nhất)? Giải thích? Tìm các support vectors tương ứng với biên quyết định đó.

- Tính :

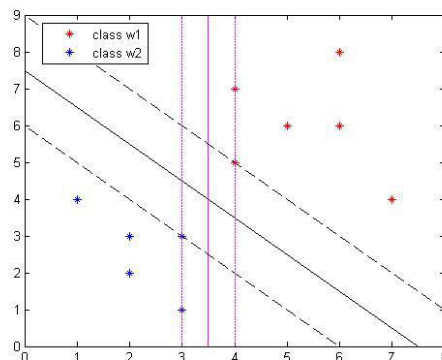
$$\frac{1}{\|w_1\|} = \frac{1}{3.5} = \frac{2}{7} = 0.285$$

$$\frac{1}{\|w_2\|} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2} = 0.7$$

Vì $\frac{1}{\|w_2\|} > \frac{1}{\|w_1\|}$, decision boundary $d_2(x)$ tốt hơn.

The support vectors for $d_1(x)$ là (3, 1), (3, 3), (4, 5) và (4, 7)

The support vectors for $d_2(x)$ là (3, 3) và (4, 5)



Chương 7: Feature Selection và giảm số chiều

20) Định nghĩa Giảm số chiều là gì? Tại sao lại cần thiết? Ví dụ

- Xem slide bài giảng: Phương pháp giảm số chiều tạo ra các biến mới (có số chiều nhỏ hơn) từ các biến đầu vào ban đầu sao cho bảo toàn được các thông tin quan trọng, giúp giảm nhiễu và các đặc trưng dư thừa. Thuộc nhóm học không giám sát.

Phương pháp giảm số chiều giúp tránh hiện tượng “curse of dimensionality”, giảm chi phí tính toán, lưu trữ và biểu diễn đặc trưng

- Trình bày được ví dụ:

21) Phương pháp lựa chọn đặc trưng (feature selection) dùng wrapper model.

- Xem slide bài giảng

22) Fisher Score cho 4 đặc trưng của IRIS dataset như trong bảng sau . Chọn ra 2 đặc trưng tốt nhất. Giải thích

	Feature f1	Feature f2	Feature f3	Feature f4
Fisher Score	1.6226	0.6688	16.0380	13.0613

- Hai đặc trưng tốt nhất là f3 và f4 vì chúng có giá trị Fisher score cao hơn hai đặc trưng còn lại.

- Giá trị Fisher score của đặc trưng càng cao thì đặc trưng đó có tính rời rạc càng cao và tăng hiệu quả cho quá trình phân lớp