KHAI PHÁ DỮ LIỆU

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 4

PHÂN LỚP
(Một số lý thuyết hỗ trợ)

Suy luận thống kê

- Population
  - Tập hợp các đối tượng cần phân tích thống kê
  - Số lượng
    - Hữu hạn hoặc vô hạn
    - Lớn hoặc vô cùng lớn

    → Dù là hữu hạn, nhưng số lượng lớn, nên vẫn được xem là vô hạn
  - VD: một nhóm người, một nhóm đối tượng, một nhóm sự kiện, …
- Dataset
  - Tập con (chọn ngẫu nhiên) của population
  - Xây dựng mô hình thống kê để phân tích, suy luận trên dataset → để hiểu population (vì số lượng của population là vô hạn)

Suy luận thống kê

- Suy luận thống kê
  - Là một trong các phương pháp phổ biến dùng để thực hiện phân tích dữ liệu
  - Có thể chia ra làm 2 nhóm chính
    - Estimation
    - Test of hypotheses
- Estimation
  - Xét dataset $T = \{(x_{11}, \ldots, x_{1n}), (x_{21}, \ldots, x_{2n}), \ldots, (x_{m1}, \ldots, x_{mn})\}$ gồm một bộ có thứ tự của các dữ liệu $X = \{X_1, X_2, \ldots, X_n\}$ (lấy từ population)
  - Mục đích của estimation là ước lượng một hoặc nhiều tham số w thuộc model của bài toán thực tế $f(X, w)$.

Suy luận thống kê

- Estimation
  - Tác dụng
    - Dùng w ước lượng được để tạo các dự đoán cho dữ liệu mới (thuộc cùng phân bố với X)
  - Prediction error
    - Nhãn thực tế của data là Y
    - Mà model đoán nhãn của Y là $f(X^*, w)$.

    → thì một trong các phương pháp để đánh giá mức độ sai số là expected mean squared error đối với các data của toàn bộ dataset T

$$E_{\mathrm{T}}\left[(Y - f(X^*, w))^2\right]$$

- Testing
  - Hypotheses được đánh giá để chấp nhận hay loại bỏ hypotheses đó

Suy luận thống kê

- Một số công cụ thống kê thường dùng trong các model
  - Mean:

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Weighted mean:

$$\text{mean} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Median:

$$\text{median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \dfrac{\left(x_{n/2} + x_{(n/2)+1}\right)}{2} & \text{if } n \text{ is even} \end{cases}$$

  - Mode

Suy luận thống kê

- Một số công cụ thống kê thường dùng trong các model
  - Variance

$$\sigma^2 = \left(\frac{1}{(n-1)}\right) \sum_{i=1}^{n} (x_i - \text{mean})^2$$

  - Standard deviation
    - Là căn bậc 2 của variance

Suy luận Bayesian

- Prior distribution
  - Phân bố của dữ liệu trong dataset trước khi phân tích
- Posterior distribution
  - Ngoài dataset ra, còn có thông tin khác cũng được dùng
  - Lý thuyết Bayesian cung cấp những idea để kết hợp các thông tin bên ngoài này với dataset
  - Phân bố của dataset sau khi đã tích hợp các thông tin bên ngoài = posterior distribution
- Bayesian classifier
  - Dùng lý thuyết Bayesian để giải thích quá trình phân lớp

Suy luận Bayesian

- Lý thuyết Bayesian
  - H : hypothesis được sử dụng để xây dựng model
  - X : data thuộc class C nào đó
  - P(H|X) :
    - Là xác suất mà hypothesis H thỏa dựa trên data X quan sát được
    - Là xác xuất hậu nghiệm biểu diễn độ tin cậy của hypothesis H xảy ra khi data X xảy ra
  - P(H) : là xác suất tiên nghiệm của hypothesis H đối với bất kỳ data nào trong dataset xảy ra
  - Lý thuyết Bayesian cung cấp idea để tính P(H|X)

$$P(H/X) = [P(X/H) \times P(H)]/P(X)$$

Suy luận Bayesian

- Naïve Bayesian classifier:
  - Giả sử có training data $S = \{S_1, S_2, \ldots, S_m\}$
  - Các training data thuộc k class $C_1, C_2, \ldots, C_k$
  - X : là new data (chưa biết thuộc class nào)
  - Đoán class của new data X như sau:

$$P(C_i/X) = [P(X/C_i) \times P(C_i)]/P(X)$$

Hồi qui dự đoán

- Hồi qui:
  - Là kỹ thuật dùng lý thuyết thống kê để dự đoán cho các giá trị liên tục
  - Mục đích của phân tích hồi qui là là xác định model tốt nhất để biểu diễn sự liên hệ giữa một biến đầu ra Y với một hoặc nhiều biến đầu vào $x_1, x_2, \ldots, x_n$
- Công thức hồi qui dự đoán:

$$y_j = \alpha + \beta_1 \cdot x_{1j} + \beta_2 \cdot x_{2j} + \beta_3 \cdot x_{3j} + \cdots + \beta_n \cdot x_{nj} + \varepsilon_j \quad j = 1, \ldots, m$$

  - α, β là các hệ số hồi qui
  - ε_j : sai số hồi qui

Phân tích phương sai

- ANOVA
  - ANOVA = ANalysis-Of-Variance
  - Là phương pháp phân tích độ tốt của đường hồi qui tìm được
  - Là một công cụ rất hiệu quả
  - Được sử dụng trong nhiều ứng dụng data-mining
- Phân tích phương sai
  - Nhãn thật sự của training data là y_i
  - Nhãn đoán được bởi đường hồi qui là $f(x_i)$
  - Sai số được đánh giá

$$R_i = y_i - f(x_i)$$

Phân tích phương sai

- Phân tích phương sai
  - Mức độ sai số của một tập m sample trong dataset có tổng cộng n samples có thể đánh giá bằng variance như sau:

$$S^2 = \frac{\left[\sum_{i=1}^{m} (y_i - f(x_i))^2\right]}{(m-(n-1))}$$

Hồi qui Logistic

- Hồi qui tuyến tính:
  - Được dùng để model những continuous-value functions
- Hồi qui Logistic:
  - Là một trong các phương pháp dùng để model những non-continuous-value functions

Q / A