TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

IT3160
Nhập môn Trí tuệ nhân tạo

Artificial Intelligence

Th.S Ngô Văn Linh
Trường Công nghệ thông tin và Truyền thông
Đại học Bách khoa Hà Nội

ONE LOVE. ONE FUTURE.

Nội dung môn học

- Chương 1. Tổng quan
- Chương 2. Tác tử thông minh
- Chương 3. Giải quyết vấn đề
- Chương 4. Tri thức và suy diễn
- Chương 5. Học máy
  - Giới thiệu về học máy
  - K láng giềng gần
  - Phân lớp Naïve Bayes

Phân lớp Naïve Bayes

- Là các phương pháp học phân lớp có giám sát và dựa trên xác suất
- Dựa trên một mô hình (hàm) xác suất
- Việc phân loại dựa trên các giá trị xác suất của các khả năng xảy ra của các giả thiết
- Là một trong các phương pháp học máy thường được sử dụng trong các bài toán thực tế
- Dựa trên định lý Bayes (Bayes theorem)

Định lý Bayes

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

- `P(D)`: Xác suất trước (tiên nghiệm) của việc quan sát được dữ liệu `D`
- `P(h)`: Xác suất trước (tiên nghiệm) của giả thiết `h`
- `P(D|h)`: Xác suất xảy ra của dữ liệu `D`, nếu biết giả thiết `h` là đúng. (likelihood)
- `P(h|D)`: Xác suất (hậu nghiệm) của giả thiết `h` là đúng, nếu quan sát được dữ liệu `D`
  - ➢ **Nhiều phương pháp phân loại dựa trên xác suất sẽ sử dụng xác suất hậu nghiệm (*posterior probability*) này!**

Định lý Bayes: Ví dụ (1)

Giả sử chúng ta có tập dữ liệu sau (dự đoán 1 người có chơi tennis)?

Ngày	Ngoài trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
N1	Nắng	Nóng	Cao	Yếu	Không
N2	Nắng	Nóng	Cao	Mạnh	Không
N3	Âm u	Nóng	Cao	Yếu	Có
N4	Mưa	Bình thường	Cao	Yếu	Có
N5	Mưa	Mát mẻ	Bình thường	Yếu	Có
N6	Mưa	Mát mẻ	Bình thường	Mạnh	Không
N7	Âm u	Mát mẻ	Bình thường	Mạnh	Có
N8	Nắng	Bình thường	Cao	Yếu	Không
N9	Nắng	Mát mẻ	Bình thường	Yếu	Có
N10	Mưa	Bình thường	Bình thường	Yếu	Có
N11	Nắng	Bình thường	Bình thường	Mạnh	Có
N12	Âm u	Bình thường	Cao	Mạnh	Có

[Mitchell, 1997]

Định lý Bayes: Ví dụ (2)

- Dữ liệu `D`. *Ngoài trời* là *nắng* và *Gió* là *mạnh*
- Giả thiết (phân loại) `h`. Anh ta chơi tennis
- Xác suất trước `P(h)`. Xác suất rằng anh ta chơi tennis (<u>bất kể *Ngoài trời* như thế nào và *Gió* ra sao</u>)
- Xác suất trước `P(D)`. Xác suất rằng *Ngoài trời* là *nắng* và *Gió* là *mạnh*
- `P(D|h)`. Xác suất *Ngoài trời* là *nắng* và *Gió* là *mạnh*, nếu biết rằng anh ta chơi tennis
- `P(h|D)`. Xác suất anh ta chơi tennis, nếu biết rằng *Ngoài trời* là *nắng* và *Gió* là *mạnh*

Xác suất hậu nghiệm cực đại (MAP)

- Với một tập các giả thiết (các phân lớp) có thể H, hệ thống học sẽ tìm ***giả thiết có thể xảy ra nhất (the most probable hypothesis)*** h $(\in$ H$)$ đối với các dữ liệu quan sát được D
- Giả thiết h này được gọi là giả thiết có xác suất hậu nghiệm cực đại (**Maximum a posteriori – MAP**)

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$h_{MAP} = \arg\max_{h \in H} \frac{P(D \mid h).P(h)}{P(D)}$$

(bởi định lý Bayes)

$$h_{MAP} = \arg\max_{h \in H} P(D \mid h).P(h)$$

(P(D) là như nhau đối với các giả thiết h)

MAP: Ví dụ

- Tập H bao gồm 2 giả thiết (có thể)
  - h_1: Anh ta chơi tennis
  - h_2: Anh ta không chơi tennis
- Tính giá trị của 2 xác xuất có điều kiện: $P(h_1|D)$, $P(h_2|D)$
- Giả thiết có thể nhất $h_{MAP}=h_1$ nếu $P(h_1|D) \geq P(h_2|D)$; ngược lại thì $h_{MAP}=h_2$
- Vì vậy, cần tính 2 biểu thức: $P(D|h_1).P(h_1)$ và $P(D|h_2).P(h_2)$, và đưa ra quyết định tương ứng
  - Nếu $P(D|h_1).P(h_1) \geq P(D|h_2).P(h_2)$, thì kết luận là anh ta chơi tennis
  - Ngược lại, thì kết luận là anh ta không chơi tennis

Đánh giá khả năng có thể nhất (MLE)

- Phương pháp MAP: Với một tập các giả thiết có thể H, cần tìm một giả thiết cực đại hóa giá trị: P(D|h).P(h)
- Giả sử (assumption) trong phương pháp **đánh giá khả năng có thể nhất (Maximum likelihood estimation – MLE)**: Tất cả các giả thiết đều có giá trị xác suất trước như nhau: $P(h_i)=P(h_j), \forall h_i,h_j \in H$
- Phương pháp MLE tìm giả thiết cực đại hóa giá trị P(D|h); trong đó P(D|h) được gọi là *khả năng có thể (likelihood)* của dữ liệu D đối với h
- Giả thiết có khả năng nhất (maximum likelihood hypothesis)

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

MLE: Ví dụ

- Tập $\mathtt{H}$ bao gồm 2 giả thiết có thể
  - h_1: Anh ta chơi tennis
  - h_2: Anh ta không chơi tennis

  D: Tập dữ liệu (các ngày) mà trong đó thuộc tính *Outlook* có giá trị *Sunny* và thuộc tính Wind có giá trị *Strong*
- Tính 2 giá trị khả năng xảy ra (likelihood values) của dữ liệu D đối với 2 giả thiết: $P(D|h_1)$ và $P(D|h_2)$
  - $P(Outlook=Sunny, Wind=Strong|h_1) = 1/8$
  - $P(Outlook=Sunny, Wind=Strong|h_2) = 1/4$
- Giả thiết MLE $h_{MLE}=h_1$ nếu $P(D|h_1) \geq P(D|h_2)$; và ngược lại thì $h_{MLE}=h_2$

  → Bởi vì $P(Outlook=Sunny, Wind=Strong|h_1) < P(Outlook=Sunny, Wind=Strong|h_2)$, hệ thống kết luận rằng: *Anh ta sẽ không chơi tennis!*

Phân loại Naïve Bayes (1)

- Biểu diễn bài toán phân loại (classification problem)
  - Một tập học `D_train`, trong đó mỗi ví dụ học `x` được biểu diễn là một vectơ `n` chiều: $(x_1, x_2, ..., x_n)$
  - Một tập xác định các nhãn lớp: $C=\{c_1, c_2, ..., c_m\}$
  - Với một ví dụ (mới) `z`, thì `z` sẽ được phân vào lớp nào?
- Mục tiêu: Xác định phân lớp có thể (phù hợp) nhất đối với `z`

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid z)$$

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid z_1, z_2, ..., z_n)$$

$$c_{MAP} = \arg\max_{c_i \in C} \frac{P(z_1, z_2, ..., z_n \mid c_i).P(c_i)}{P(z_1, z_2, ..., z_n)}$$ (bởi định lý Bayes)

Phân loại Naïve Bayes (2)

- Để tìm được phân lớp có thể nhất đối với z

$$c_{MAP} = \arg\max_{c_i \in C} P(z_1, z_2, ..., z_n | c_i).P(c_i)$$

(P(z1,z2,...,zn) là như nhau với các lớp)

- **Giả thuyết (assumption) trong phương pháp phân loại Naïve Bayes**: Các thuộc tính là *độc lập có điều kiện (conditionally independent)* đối với các lớp

$$P(z_1, z_2, ..., z_n | c_i) = \prod_{j=1}^{n} P(z_j | c_i)$$

- Phân loại Naïve Bayes tìm phân lớp có thể nhất đối với z

$$c_{NB} = \arg\max_{c_i \in C} P(c_i).\prod_{j=1}^{n} P(z_j | c_i)$$

Phân loại Naïve Bayes: Giải thuật

- Giai đoạn học (training phase), sử dụng một tập học

  Đối với mỗi phân lớp có thể (mỗi nhãn lớp) $c_i \in C$

  - Tính giá trị xác suất tiên nghiệm: $P(c_i)$
  - Đối với mỗi giá trị thuộc tính x_j, tính giá trị xác suất xảy ra của giá trị thuộc tính đó đối với một phân lớp c_i: $P(x_j | c_i)$

- Giai đoạn phân lớp (classification phase), đối với một ví dụ mới
  - Đối với mỗi phân lớp $c_i \in C$, tính giá trị của biểu thức:

$$P(c_i).\prod_{j=1}^{n} P(x_j | c_i)$$

  - Xác định phân lớp của z là lớp có thể nhất c^*

$$c^* = \arg\max_{c_i \in C} P(c_i).\prod_{j=1}^{n} P(x_j | c_i)$$

Phân loại Naïve Bayes: Ví dụ (1)

Một sinh viên trẻ với thu nhập trung bình và mức đánh giá tín dụng bình thường sẽ mua một cái máy tính?

Rec. ID	Age	Income	Student	Credit_Rating	Buy_Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Medium	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Medium	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young 1	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young 2	Medium	Yes	Excellent	Yes
12	Medium	Medium	No	Excellent	Yes
13	Medium	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

http://www.cs.sunysb.edu/~cse634/lecture_notes/07classification.pdf

Phân loại Naïve Bayes: Ví dụ (2)

- Biểu diễn bài toán phân loại
  - z = (Age=Young,Income=Medium,Student=Yes,Credit_Rating=Fair)
  - Có 2 phân lớp có thể: c_1 (“Mua máy tính”) và c_2 (“Không mua máy tính”)
- Tính giá trị xác suất trước cho mỗi phân lớp
  - $P(c_1) = 9/14$
  - $P(c_2) = 5/14$
- Tính giá trị xác suất của mỗi giá trị thuộc tính đối với mỗi phân lớp
  - $P(Age=Young|c_1) = 2/9$; $P(Age=Young|c_2) = 3/5$
  - $P(Income=Medium|c_1) = 4/9$; $P(Income=Medium|c_2) = 2/5$
  - $P(Student=Yes|c_1) = 6/9$; $P(Student=Yes|c_2) = 1/5$
  - $P(Credit\_Rating=Fair|c_1) = 6/9$; $P(Credit\_Rating=Fair|c_2) = 2/5$

Phân loại Naïve Bayes: Ví dụ (3)

- Tính toán xác suất có thể xảy ra (likelihood) của ví dụ z đối với mỗi phân lớp
  - Đối với phân lớp c_1

    $P(z|c_1) = P(Age=Young|c_1).P(Income=Medium|c_1).P(Student=Yes|c_1).P(Credit\_Rating=Fair|c_1) = (2/9).(4/9).(6/9).(6/9) = 0.044$
  - Đối với phân lớp c_2

    $P(z|c_2) = P(Age=Young|c_2).P(Income=Medium|c_2).P(Student=Yes|c_2).P(Credit\_Rating=Fair|c_2) = (3/5).(2/5).(1/5).(2/5) = 0.019$
- Xác định phân lớp có thể nhất (the most probable class)
  - Đối với phân lớp c_1

    $P(c_1).P(z|c_1) = (9/14).(0.044) = 0.028$
  - Đối với phân lớp c_2

    $P(c_2).P(z|c_2) = (5/14).(0.019) = 0.007$

  →Kết luận: *Anh ta (z) sẽ mua một máy tính*!

Phân loại Naïve Bayes: Vấn đề (1)

- Nếu không có ví dụ nào gắn với phân lớp c_i có giá trị thuộc tính x_j…
  $P(x_j|c_i)=0$, và vì vậy:

$$P(c_i).\prod_{j=1}^{n} P(x_j \mid c_i) = 0$$

- Giải pháp: Sử dụng phương pháp Bayes để ước lượng $P(x_j|c_i)$

$$P(x_j \mid c_i) = \frac{n(c_i, x_j) + mp}{n(c_i) + m}$$

  - $n(c_i)$: số lượng các ví dụ học gắn với phân lớp c_i
  - $n(c_i,x_j)$: số lượng các ví dụ học gắn với phân lớp c_i có giá trị thuộc tính x_j
  - p: ước lượng đối với giá trị xác suất $P(x_j|c_i)$
    → Các ước lượng đồng mức: p=1/k, nếu thuộc tính f_j có k giá trị
  - m: một hệ số (trọng số)
    → Để bổ sung cho $n(c_i)$ các ví dụ thực sự được quan sát với thêm m mẫu ví dụ với ước lượng p

Phân loại Naïve Bayes: Vấn đề (2)

- Giới hạn về độ chính xác trong tính toán của máy tính
  - $P(x_j|c_i)<1$, đối với mọi giá trị thuộc tính x_j và phân lớp c_i
  - Vì vậy, khi số lượng các giá trị thuộc tính là rất lớn, thì:

$$\lim_{n\to\infty}\left(\prod_{j=1}^{n} P(x_j \mid c_i)\right) = 0$$

- Giải pháp: Sử dụng hàm lôgarit cho các giá trị xác suất

$$c_{NB} = \arg\max_{c_i\in C}\left(\log\left[P(c_i).\prod_{j=1}^{n} P(x_j \mid c_i)\right]\right)$$

$$c_{NB} = \arg\max_{c_i\in C}\left(\log P(c_i) + \sum_{j=1}^{n}\log P(x_j \mid c_i)\right)$$

Phân loại văn bản bằng NB (1)

- Biểu diễn bài toán phân loại văn bản
  - Tập học D, trong đó mỗi ví dụ học là một biểu diễn văn bản gắn với một nhãn lớp: D = {(d_k, c_i)}
  - Một tập các nhãn lớp xác định: C = {c_i}
- Giai đoạn học
  - Từ tập các văn bản trong D, trích ra tập các từ khóa T = {t_j}
  - Gọi D_{c_i} là tập các văn bản trong D có nhãn lớp c_i
  - Đối với mỗi phân lớp c_i
    - Tính giá trị xác suất trước của phân lớp c_i: $P(c_i) = \frac{|D_{c_i}|}{|D|}$
    - Đối với mỗi từ khóa t_j, tính xác suất từ khóa t_j xuất hiện đối với lớp c_i

$$P(t_j \mid c_i) = \frac{\left(\sum_{d_k \in D_{c_i}} n(d_k, t_j)\right) + 1}{\left(\sum_{d_k \in D_{c_i}} \sum_{t_m \in T} n(d_k, t_m)\right) + |T|}$$

($n(d_k, t_j)$: số lần xuất hiện của từ khóa t_j trong văn bản d_k)

Phân loại văn bản bằng NB (2)

- Giai đoạn phân lớp đối với một văn bản mới d
  - Từ văn bản d, trích ra tập T_d gồm các từ khóa (keywords) t_j đã được định nghĩa trong tập T
  - **Giả sử (assumption)**. Xác suất từ khóa t_j xuất hiện đối với lớp c_i là độc lập đối với vị trí của từ khóa đó trong văn bản

    $P(t_j \text{ ở vị trí } k|c_i) = P(t_j \text{ ở vị trí } m|c_i), \ \forall k,m$
  - Đối với mỗi phân lớp c_i, tính xác suất hậu nghiệm của văn bản d đối với c_i

$$P(c_i).\prod_{t_j \in T_d} P(t_j | c_i)$$

  - Phân lớp văn bản d thuộc vào lớp c^*

$$c^* = \arg\max_{c_i \in C} P(c_i).\prod_{t_j \in T_d} P(t_j | c_i)$$

Tài liệu tham khảo

- E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2010.
- T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- T. M. Mitchell. *The discipline of machine learning*. CMU technical report, 2006.
- H. A. Simon. *Why Should Machines Learn?* In R. S. Michalski, J. Carbonell, and T. M. Mitchell (Eds.): Machine learning: An artificial intelligence approach, chapter 2, pp. 25-38. Morgan Kaufmann, 1983.
- A. Kontorovich and Weiss. *A Bayes consistent 1-NN classifier*. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS). JMLR: W&CP volume 38, 2015.
- A. Guyader, N. Hengartner. *On the Mutual Nearest Neighbors Estimate in Regression.* Journal of Machine Learning Research 14 (2013) 2361-2376.
- L. Gottlieb, A. Kontorovich, and P. Nisnevitch. *Near-optimal sample compression for nearest neighbors*. Advances in Neural Information Processing Systems, 2014.