TRƯỜNG ĐẠI HỌC MỞ TP. HỒ CHÍ MINH
HO CHI MINH CITY OPEN UNIVERSITY

Buổi 5: Phân Lớp bằng KNN và Naive Bayes

Môn: KHAI PHÁ DỮ LIỆU

(DATA MINING)

Một số hình trong bài được lấy từ 02 giáo trình

- Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition, Elsevier, 2012.
- Max Bramer, Principles of data mining, Springer, 2007.

Phân lớp (Classification)

- Gán 1 đối tượng chưa biết vào một lớp (class) trong số các lớp đã biết.
- Binary classification và multi-class classification
- Phân lớp: Ứng dụng rất nhiều trong cuộc sống hàng ngày.
- Là công cụ hữu ích cho các hệ hỗ trợ ra quyết định:
  - Quyết định cho vay tín dụng.
  - Quyết định khi chẩn đóan y khoa
  - Ví dụ khác??

Phân lớp

- Ví dụ
  - Lĩnh vực bảo hiểm.
  - Lĩnh vực điều tra (bầu cử,…)
  - Lĩnh vực thời tiết.
- Khi phân lớp → khả năng xảy ra sai sót → giả thiết H_0, sai lầm lọai 1, sai lầm lọai 2.
- Giả thiết H_0, sau khi kiểm định thì không chấp nhận H_0 nhưng thực sự H_0 đúng!!!!→ loại 1
- Giả thiết H_0, sau khi kiểm định thì chấp nhận H0 nhưng thực sự H_0 sai!! → loại 2

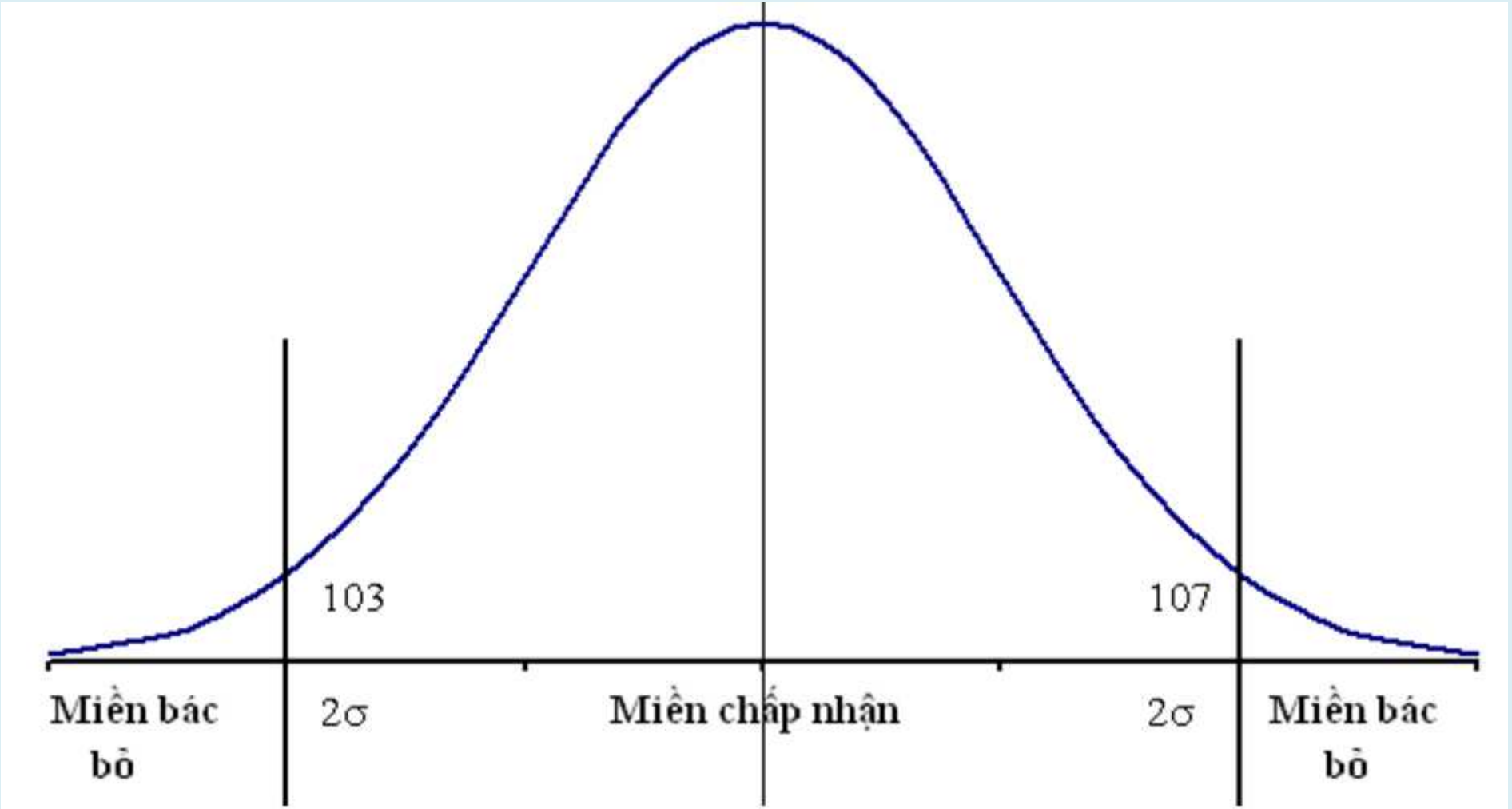
Phân lớp

		Thực hiện kiểm định	
	Giả thiết H_0	Chấp nhận giả thiết H_0	Không chấp nhận giả thiết H_0
Thực tế	Đúng	OK	Sai lầm loại 1
	Sai	Sai lầm loại 2	OK

		Thực hiện kiểm định	
	Ngày mai mưa	Ngày mai mưa	Ngày mai không mưa
Thực tế	Đúng	OK	Sai lầm loại 1
	Sai	Sai lầm loại 2	OK

- Sai lầm loại 1, loại 2: mang ý nghĩa tương đối do phụ thuộc vào cách đặt giả thiết H_0
- Sai lầm gây tổn thất lớn hơn➔ đặt giả thiết H_0 để sai lầm đó là loại 1 + qui định mức sai sót khi mắc sai lầm không vượt quá giá trị α nào đó (VD α = 5%, α = 1%)

Phân lớp



Phân lớp

- Ví dụ: Dự đoán thời tiết
  - Nếu kiểm định và kết luận trời sẽ mưa ➔ mang theo áo mưa, Nếu kết luận này SAI ➔ ???
  - Nếu kiểm định và kết luận trời không mưa ➔ không mang theo áo mưa, Nếu kết luận này SAI ➔ ???
- Ví dụ: Chẩn đoán bệnh
  - Nếu kiểm định bệnh nhân và kết luận có bệnh.

    Nếu kết luận này SAI➔ ???
  - Nếu kiểm định bệnh nhân và kết luận không có bệnh. Nếu kết luận này SAI➔ ???

Phân lớp

- Có nhiều phương pháp phân lớp.
- Các khái niệm:
  - Training dataset: tập huấn luyện.
  - Testing data set: tập kiểm tra.
  - Classifier: bộ phân lớp.
  - Evaluate: Spit xx%, k-cross validation
- Chủ đề trong bài:
  - Naïve Bayes
  - K-Neighbour (K láng giềng gần nhất KNN: K Nearest Neighbour)

Phân lớp bằng K-Neighbour

- Thường áp dụng cho những dataset có thuộc tính dạng giá trị liên tục.
- Tuy vậy hoàn tòan cũng có thể áp dụng cho các thuộc tính dạng định danh (nominal) với 1 số biến đổi?
- Ý tưởng phân lớp:
  - Like father, like son!!!
  - Gần mực thì đen, gần đèn thì sáng!!!

Phân lớp bằng K-Neighbour

- Tóm tắt thuật tóan
  - Tìm k phần tử trong bộ dữ liệu huấn luyện (training data set) "gần nhất" với phần tử cần dự đóan.
  - Căn cứ vào lớp chiếm đa số của k phần tử này để dự báo lớp của phần tử mới.

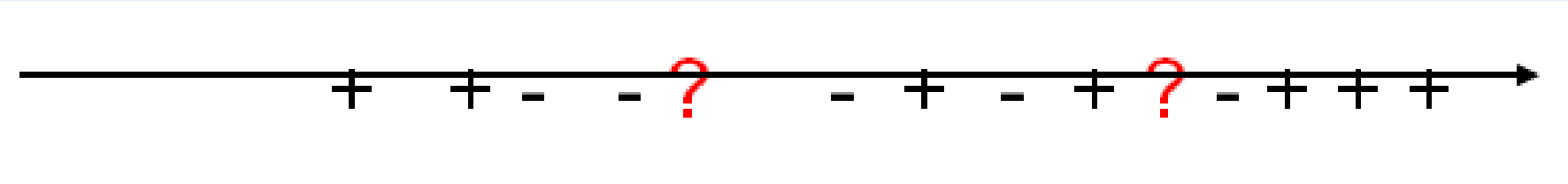
a	b	c	d	e	f	Class
Yes	No	No	6.4	8.3	Low	Negative
Yes	Yes	Yes	19.2	4.7	High	Positive

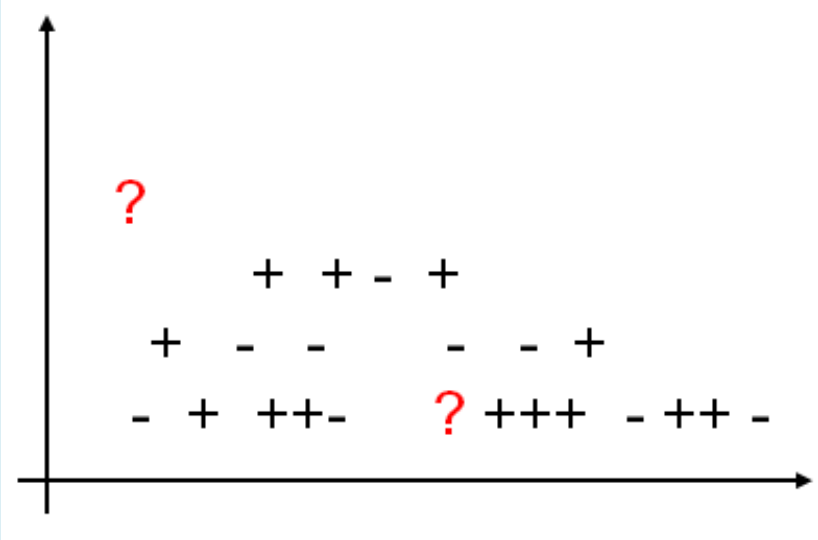
a	b	c	d	e	f	Class
Yes	No	No	6.6	8	Low	**???**

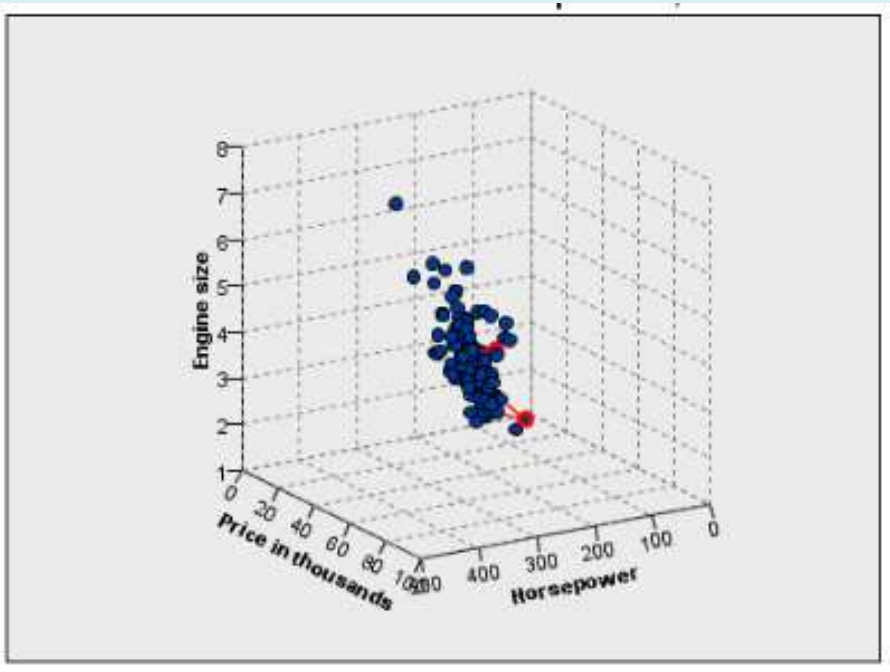
Phân lớp bằng K-Neighbour

- Triển khai thực tế:
  - K là bao nhiêu là vừa? Chẵn hay lẻ?
  - Phương thức bình chọn?
    - Đa số thắng thiểu số
    - Sử dụng trọng số.
    - …
  - Cách tính khoảng cách?
    - Không gian 1 chiều.
    - Không gian 2 chiều
    - Không gian 3 chiều

Phân lớp bằng K-Neighbour



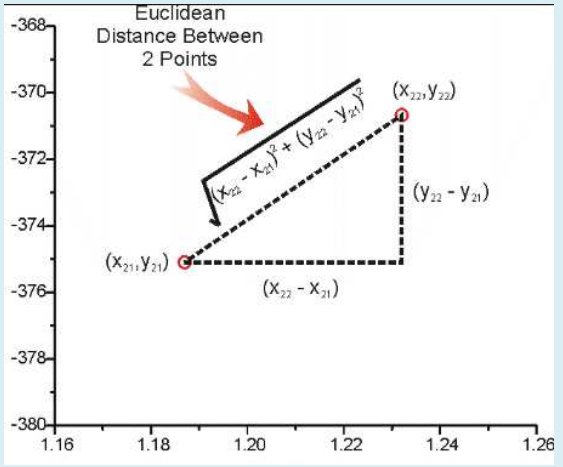


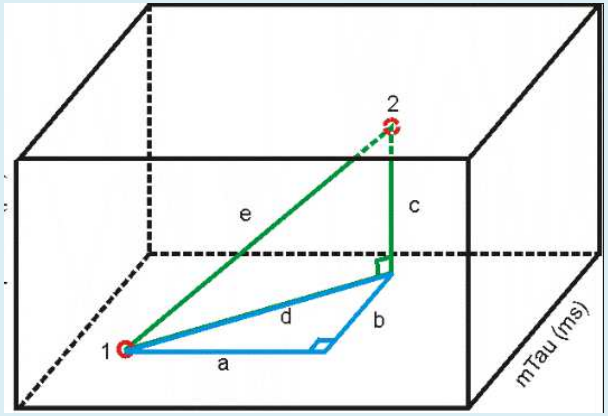


- Tính khỏang cách ra sao?
- Data set n thuộc tính ➔Không gian n-chiều

Phân lớp bằng K-Neighbour

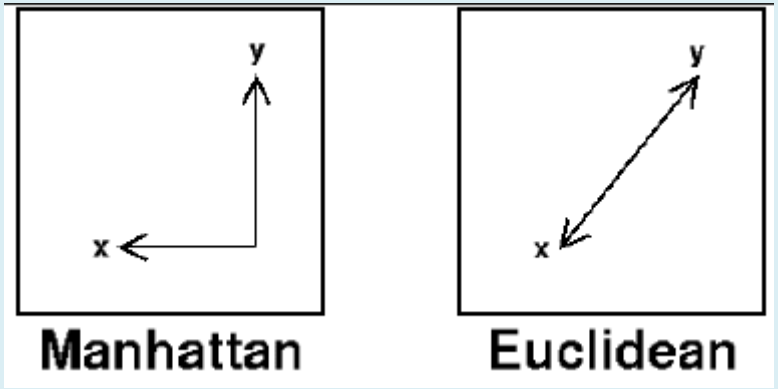
- Xây dựng 1 độ đo (metric): thỏa 3 điều kiện
  - d(X,X) = 0
  - d(X,Y) = d (Y,X)
  - Bất đẳng thức tam giác!!!
    - d(X,Y) <= d (X,Z) + d(Z,Y)
- Dễ hình dung trong 2D và 3D
- Thường gặp nhất là khoảng cách Euclide

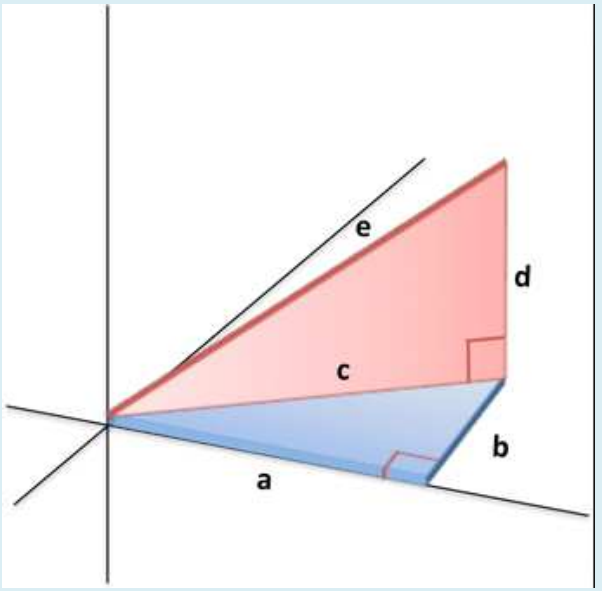




Phân lớp bằng K-Neighbour

- Còn những độ đo khác
  - Mahattan
  - Khoảng cách tối đa
- Chứng minh 2 định nghĩa trên là 1 độ đo??!!





Phân lớp bằng K-Neighbour

- Mở rộng

	2D	3D	N Dimension
Euclide	$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$	$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$	???
Mahattan	$\lvert x_1 - x_2\rvert + \lvert y_1 - y_2\rvert$	$\lvert x_1 - x_2\rvert + \lvert y_1 - y_2\rvert + \lvert z_1 - z_2\rvert$	???
Max Dimension	???	???	???

- Trực quan
  - Euclide: đường tròn, khối cầu.
  - Mahattan: hình chữ nhật, khối chữ nhật.
  - Khỏang cách tối đa: hình vuông, khối vuông

Phân lớp bằng K-Neighbour

- Ví dụ minh họa các độ đo khoảng cách:
  - X(4,12) và Y(12,9)
  - Euclide:
    - $((4-12)^2 + (12-9)^2)^{½} = 8.544$
  - Mahattan
    - $|4-12| + |12-9| = 11$
  - Khoảng cách lớn nhất
    - $Max(|4-12|, |12-9|) = 8$
- Nên sử dụng độ đo nào??

Phân lớp bằng K-Neighbour

- Tóm tắt thuật tóan
  - Tìm k phần tử trong bộ dữ liệu huấn luyện (training data set) "gần nhất" với phần tử cần dự đóan.
  - Căn cứ vào **lớp chiếm đa số** của k phần tử này để dự báo lớp của phần tử mới.
- Thử thực hành phương pháp K láng giềng.
- Yêu cầu: k=3/5/7, dùng độ đo Euclide/Mahattan
- Training data set: tập tin bảng tính
- Bao gồm 4 sheet

Chuẩn hóa (Normalisation)

- Sử dụng khoảng cách Euclide: khuyết điểm?
- Sự "áp đảo" của 1 vài thuộc tính khi tính tóan khoảng cách Euclide.
- Lý do xuất hiện tình trạng này?

Mileage	# Doors	Years	# Owner
18,475	2	12	8
26,292	4	8	1

X (Mileage=24,500, Doors=2, Years= 8, Owners=9)

Chuẩn hóa (Normalisation)

- Giải quyết: Chuẩn Hóa!!
- Chuyển phạm vi giá trị của 1 thuộc tính về phạm vi 0..1.
- Nhắc lại: Chuẩn hóa trong hồi qui tuyến tính
- Phương pháp:
  - Căn cứ vào giá trị nhỏ nhất min , lớn nhất max của thuộc tính trong training data set.
  - Chuẩn hóa v ➔ v'= (v-min)/(max-min)

Chuẩn hóa (Normalisation)

- Tình huống:
  - Cần phân lọai một thể hiện nhưng trong 1/nhiều thuộc tính lại chứa giá trị gốc ngòai phạm vi chuẩn hóa (min .. Max)?
  - Căn cứ vào tầm quan trọng của thuộc tính để điều chỉnh công thức tính khỏang cách cùng với trọng số?
  - Cần xử lý ra sao trong tình huống data set có chứa các thuộc tính dạng nominal?

Bộ phân lớp Naïve Bayes

- Naïve là gì
- Reverend Thomas Bayes (1702-1761).
- Sử dụng một vài khái niệm về xác suất cơ bản.
- Nhắc lại vài kiến thức cơ bản:
  - Xe lửa khởi hành lúc 6.30pm từ London đến ga chờ.
  - Sự kiện (Event): xe lửa đến ga đúng giờ
  - Xác suất của sự kiện này (probability of event)
    - Sử dụng giá trị từ 0 đến 1
    - Ý nghĩa của giá trị 0? Giá trị 1?
    - Ý nghĩa của giá trị 0.5? 0.86 ?

Bộ phân lớp Naïve Bayes

- Xác suất của sự kiện xe lửa đến ga đúng giờ là 0.86:
  - Nếu quan sát thống kê trong N ngày thì số ngày tàu lửa đến đúng giờ (ta hy vọng) sẽ là 0.86xN.
- Nhắc lại khái niệm: đám đông và mẫu.
- Thực tế: Quan tâm đến 1 bộ/tập hợp các sự kiện với điều kiện các sự kiện này lọai trừ nhau ➔ không thể có hơn 1 sự kiện xảy ra cùng lúc
  - E1: chuyến xe lửa bị hủy, E2: chuyến xe lửa tới sớm hay đúng giờ, E3: chuyến xe lửa đến trễ dưới 10 phút, E4: chuyến xe lửa đến trễ hơn 10 phút

Bộ phân lớp Naïve Bayes

- Các kết quả
  - Xác xuất của 1 sự kiện P(E1), P(E2),..
  - Tổng các xác suất bằng 1
- Trong thực tế: không thể xác định chính xác xác suất của từng biến cố!!!
- Kỹ thuật: thu thập mẫu, thông qua mẫu để đưa ra giá trị xác suất của từng sự kiện.
- Nhắc lại
  - Độ lớn mẫu, tỷ lệ mẫu,…
  - Ước lượng điểm, ước lượng khỏang,…

Bộ phân lớp Naïve Bayes

- Quan sát 1 ví dụ (xem tập tin pdf đính kèm email)
- Cỡ mẫu: 20.
- Số lượng thuộc tính: 5 (tính cả thuộc tính lớp).
- Các sự kiện (lớp):
  - Đúng giờ.
  - Hủy chuyến.
  - Trễ chuyến.
  - Rất trễ.
- Thử cho biết kết quả khi
  - (weekday, winter, high, heavy) ➔ ????

Bộ phân lớp Naïve Bayes

#	Day	Season	Wind	Rain	Class
1	weekday	spring	none	none	on time
2	weekday	winter	none	slight	on time
3	weekday	winter	none	slight	on time
4	weekday	winter	high	heavy	late
5	Saturday	summer	normal	none	on time
6	weekday	autumm	normal	none	very late
7	holiday	summer	high	slight	on time
8	Sunday	summer	normal	none	on time
9	weekday	winter	high	heavy	very late
10	weekday	summer	none	slight	on time
11	Saturday	spring	high	heavy	cancelled
12	weekday	summer	high	slight	on time
13	Saturday	winter	normal	none	late
14	weekday	summer	high	none	on time
15	weekday	winter	normal	heavy	very late
16	Saturday	autumm	high	slight	on time
17	weekday	autumm	none	heavy	on time
18	holiday	spring	normal	slight	on time
19	weekday	spring	normal	none	on time
20	weekday	spring	normal	slight	on time

(weekday, winter, high, heavy) ➔ ????

Dựa vào
-Các thể hiện tương tự??
➔ late
-Dựa vào xác suất của từng sự kiện?
➔ on time (0.7 max!)

Bộ phân lớp Naïve Bayes

- Thực tế các sự kiện on time, late, very late, cancelled : phụ thuộc vào các thuộc tính Day, Season, Wind và Rain.
- Khái niệm: xác suất có điều kiện.
- Ký hiệu
  - Ví dụ: xác suất xe lửa đúng giờ nếu đang mùa đông
  - P(class = on time | season = winter).
  - Cách tính: “đơn giản” như “sự cảm nhận”
    - Tính P(class= on time| season = winter)
    - Tính P (class= late | season = winter)
    - Tính P (class= late | wind = high)

Bộ phân lớp Naïve Bayes

- P(class = on time | season = winter)?
- Đếm số instance có season = winter → N
- Đếm số instance (trong các instance ở bước trên) có class= on time → M
- P(class = on time | season = winter) = M/N

Bộ phân lớp Naïve Bayes

- Công thức xác suất có điều kiện
  - P(A|B)xP(B) = P(A and B) = P(B|A)xP(A)
- Thử lại với bộ dữ liệu Train
  - P(Days=weekday)?
  - P(Train=on time)?
  - P(Days=weekday And Train=on time)?
  - P(Days=weekday|Train=on time)?
  - P(Train=on time|Days=weekday)?

Bộ phân lớp Naïve Bayes

- Giả sử cần phán đoán Train=?? khi Season=winter
- Nhận xét:
  - Có các khả năng của Train={on time, cancelled, late, very late}
  - Có ảnh hưởng của Season đến Train.
  - Cần tính xác suất sau
    - P(Train=on time|Season=Winter)
    - P(Train=late|Season=Winter)
    - P(Train=very late|Season=Winter)
    - P(Train=cancelled|Season=Winter)
  - Kết quả là ???

Bộ phân lớp Naïve Bayes

- Thử mở rộng
  - Phán đoán Train=?? Khi Days=weekday và Season=winter
  - Thực hiện tương tự
    - P(Train=on time| Days=weekday And Season=winter)
    - P(Train=late| Days=weekday And Season=winter)
    - P(Train=very late| Days=weekday And Season=winter)
    - P(Train=cancelled| Days=weekday And Season=winter)
  - Nhận xét ?

Bộ phân lớp Naïve Bayes

- Thay đổi cách tính toán
- Lưu ý P(A|B)xP(B) = P(A and B) = P(B|A)xP(A)
- Như vậy ???
  - P(Train=on time| Days=weekday And Season=winter) x P(Days=weekday And Season=winter)
  - P(Train=on time And Days=weekday And Season=winter)
  - P(Days=weekday And Season=winter| Train= on time) x P(Train=on time)
- Sử dụng một cách tính đơn giản hơn cho việc đánh giá
    - P(Train=on time| Days=weekday And Season=winter)
    - P(Train=late| Days=weekday And Season=winter)
    - P(Train=very late| Days=weekday And Season=winter)
    - P(Train=cancelled| Days=weekday And Season=winter)

Bộ phân lớp Naïve Bayes

- P(on time | weekday And winter) =

  1/P(weekday And winter) x P(weekday And winter| on time) x P(on time)

- P(late| weekday And winter)=

  1/P(weekday And winter) x P(weekday And winter| late) x P(late)

- P(very late| weekday And winter)=

  1/P(weekday And winter) x P(weekday And winter| very late) x P(very late)

- P(cancelled| weekday And winter)=

  1/P(weekday And winter) x P(weekday And winter| cancelled) x P(cancelled)

Bộ phân lớp Naïve Bayes

➔ Tính toán và sau đó chọn giá trị Max

- P(on time | weekday And winter) =

  P(weekday And winter| on time) x P(on time)

- P(late| weekday And winter)=

  P(weekday And winter| late) x P(late)

- P(very late| weekday And winter)=

  P(weekday And winter| very late) x P(very late)

- P(cancelled| weekday And winter)=

  P(weekday And winter| cancelled) x P(cancelled)

Bộ phân lớp Naïve Bayes

- Vấn đề đặt ra
  - weekday And winter And high And Heavy
  - Saturday And spring And normal And Slight
  - Mở rộng: dk1 And dk2 And ….. And dkN
  - Khó tính toán!!!
  - Ý nghĩa từ NAÏVE
  - Đặt giả thiết: Tất cả các thuộc tính đều độc lập nhau
  - Đây là nhược điểm của thuật toán Naïve Bayes

Bộ phân lớp Naïve Bayes

- Áp dụng

P(on time | weekday And winter) =

P(weekday And winter| on time) x P(on time)=

P(weekday|on time) x P(winter|on time) x P(on time)

P(late| weekday And winter)=

P(weekday And winter| late) x P(late)=

P(weekday|late) x P(winter| late) x P(late)

P(very late| weekday And winter)=

P(weekday And winter| very late) x P(very late)=

P(weekday|very late) x P(winter| very late) x P(very late)

P(cancelled| weekday And winter)=

P(weekday And winter| cancelled)x P(cancelled)=

P(weekday|cancelled) x P(winter| cancelled)x P(cancelled)

Bộ phân lớp Naïve Bayes

		Class=On time	Class = late	Class=very late	Class= cancelled
Day	Day=weekday				
	Day=Saturday				
	Day=Sunday				
	Day=Holiday				
Season	Season=Spring				
	Season=Summer				
	Season=Autumm				
	Season=winter				
Wind	Wind=None				
	Wind=High				
	Wind=Normal				
Rain	Rain=none				
	Rain=Slight				
	Rain=Heavy				
	Prior Probability				

Nếu dùng bảng tính: Có thể sắp xếp theo Class + thuộc tính để đếm cho dễ dàng

Bộ phân lớp Naïve Bayes

		Class=On time	Class = late	Class=very late	Class= cancelled
Day	Day=weekday	9	1	3	
	Day=Saturday	2	1		1
	Day=Sunday	1			
	Day=Holiday	2			
Season	Season=Spring	4			1
	Season=Summer	6			
	Season=Autumm	2		1	
	Season=winter	2	2	2	
Wind	Wind=None	5			
	Wind=High	4	1	1	1
	Wind=Normal	5	1	2	
Rain	Rain=none	5	1	1	
	Rain=Slight	8			
	Rain=Heavy	1	1	2	1
	Prior Probability	14	2	3	1

Bộ phân lớp Naïve Bayes

		Class=On time	Class = late	Class=very late	Class= cancelled
Day	Day=weekday	0.64	0.50	1.00	-
	Day=Saturday	0.14	0.50	-	1.00
	Day=Sunday	0.07	-	-	-
	Day=Holiday	0.14	-	-	-
Season	Season=Spring	0.29	-	-	1.00
	Season=Summer	0.43	-	-	-
	Season=Autumm	0.14	-	0.33	-
	Season=winter	0.14	1.00	0.67	-
Wind	Wind=None	0.36	-	-	-
	Wind=High	0.29	0.50	0.33	1.00
	Wind=Normal	0.36	0.50	0.67	-
Rain	Rain=none	0.36	0.50	0.33	-
	Rain=Slight	0.57	-	-	-
	Rain=Heavy	0.07	0.07	0.67	1.00
	Prior Probability	0.7	0.05	0.2	0.05

- (weekday, winter, high, slight) ➔ ??

Bộ phân lớp Naïve Bayes

- Thuật toán Naïve Bayes
- Cho bộ dữ liệu huấn luyện :
  - Có m lớp (C_i với i=1-m)
  - Có n thuộc tính (F_j với j=1-n)
  - Ký hiệu mỗi giá trị thuộc tính F_j là V_{jk}
- Bước 1: Xây dựng bảng kết quả xác suất
  - Tính toán tất cả các xác suất $P(C_i)$ và $P(F_j=V_{ik}|C_i)$
- Bước 2: Dự đoán lớp của phần tử mới
  - Tính xác suất mà phần tử này thuộc về từng lớp.
  - Chọn lớp có giá trị xác suất cao nhất.

Bộ phân lớp Naïve Bayes

- Công thức tính toán

$$P(C_i) \times P(F_1=V_1 \text{ and } F_2=V_2 \ldots \text{ and } F_n=V_n | C_i)$$

- Do chấp nhận giả thiết các thuộc tính dộc lập nhau nên công thức tính toán trên trở thành

$$\mathbf{P(C_i) \times P(F_1=V_1|C_i) \times P(F_2=V_2|C_i) \times \ldots \times P(F_n=V_n|C_i)}$$

- Tính toán xác suất phần tử đang xét thuộc từng lớp. Sau đó chọn lớp có giá trị xác suất lớn nhất.

Eager Learning/Lazy Learning

- Eager Learning
  - Học sớm!!!
  - Tính toán, xử lý trên bộ dữ liệu huấn luyện ➔ bảng quyết định, bảng số liệu, hệ thống luật, cây quyết định.
  - Phân lớp cho phần tử mới➔ đưa vào
- Lazy Learning
  - Học muộn!!!

Tổng kết

- Bài toán phân lớp (Classification)?
- Thuật toán KNN: K láng giềng gần nhất.
  - Ý tưởng
  - Các ưu điểm, nhược điểm.
- Thuật toán Naïve Bayes: sử dụng lý thuyết xác suất
  - Ý tưởng
  - Ưu và nhược điểm
- Học “sớm” và “muộn”

Câu hỏi

- Áp dụng KNN cho các bộ dữ liệu chứa các thuộc tính có giá trĩ dạng Nominal?
- Áp dụng Naïve Bayes cho các bộ dữ liệu chứa các thuộc tính có giá trị dạng Continuous?
- Thử tính độ phức tạp của thuật toán KNN và Naïve Bayes.
- Trong thuật toán Naïve Bayes, giả sử ta cần xác định lớp của một phần tử mà trong đó xuất hiện 1 giá trị trong 1 thuộc tính chưa từng tồn tại trong bộ dữ liệu huấn luyện. Cách giải quyết?