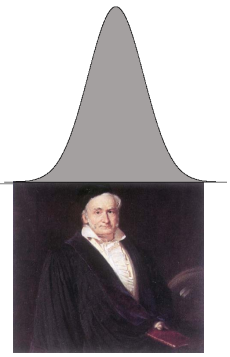
Dữ liệu liên tục



- Giả sử các thuộc tính có phân phối *Gaussian*
- hàm mật độ xác suất ***f(x)*** được tính như sau

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Karl Gauss, 1777-1855
great German
mathematician

- *Mean* μ $\quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- *Phương sai (Variance)* σ^2 $\quad \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$
- *Độ lệch chuẩn -standard deviation: căn bậc 2 của phương sai*

$$\sigma = \sqrt{\sigma^2}$$

https://www.mathsisfun.com/data/standard-deviation.html

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	70	96	False	**Yes**
Rainy	68	80	False	**Yes**
Rainy	65	70	True	**No**
Overcast	64	65	True	**Yes**
Sunny	72	95	False	**No**
Sunny	69	70	False	**Yes**
Rainy	75	80	False	**Yes**
Sunny	75	70	True	**Yes**
Overcast	72	90	True	**Yes**
Overcast	81	75	False	**Yes**
Rainy	71	91	True	**No**

Bước 1: huấn luyện mô hình

Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	70	96	False	**Yes**
Rainy	68	80	False	**Yes**
Rainy	65	70	True	**No**
Overcast	64	65	True	**Yes**
Sunny	72	95	False	**No**
Sunny	69	70	False	**Yes**
Rainy	75	80	False	**Yes**
Sunny	75	70	True	**Yes**
Overcast	72	90	True	**Yes**
Overcast	81	75	False	**Yes**
Rainy	71	91	True	**No**

36

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	????				False	6	2	9	5
Overcast	4	0					True	3	3		
Rainy	3	2									
Sunny	2/9	3/5					False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5					True	3/9	3/5		
Rainy	3/9	2/5									

Temp	Play
85	**No**
80	**No**
83	**Yes**
70	**Yes**
68	**Yes**
65	**No**
64	**Yes**
72	**No**
69	**Yes**
75	**Yes**
75	**Yes**
72	**Yes**
81	**Yes**
71	**No**

37

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Temp	Play
85	No
80	No
83	Yes
70	Yes
68	Yes
65	No
64	Yes
72	No
69	Yes
75	Yes
75	Yes
72	Yes
81	Yes
71	No

The numeric weather data with summary statistics

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						

38

Dữ liệu liên tục

- *mean – trung bình* $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

  $\mu = (83+ 70 + 68 + 64 + 69 + 75 + 75 + 72 +81)/9 = \mathbf{73}$

- *Phương sai* $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$

  $\sigma^2 = 1/8*[(83-73)^2+(70-73)^2+(68-73)^2+(64-73)^2+(69-73)^2$
  $+(75-73)^2+(75-73)^2+(72-73)^2+(81-73)^2) = \mathbf{38.44}$

- *standard deviation – độ lệch chuẩn* $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma^2}} = \mathbf{6.2}$

- hàm mật độ xác suất *f(x)* tính khi có phần tử mới xuất hiện

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

39

39

Temp	Play
85	No
80	No
83	Yes
70	Yes
68	Yes
65	No
64	Yes
72	No
69	Yes
75	Yes
75	Yes
72	Yes
81	Yes
71	No

The numeric weather data with summary statistics

	utlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no		yes	no
	2	3		83	85		86	85	false	6	2		9	5
	4	0		70	80		96	90	true	3	3			
	3	2		68	65		80	70						
				64	72		65	95						
				69	71		70	91						
				75			80							
				75			70							
				72			90							
				81			75							
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5		9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5			
rainy	3/9	2/5	σ^2	**38..44**										

40

The numeric weather data with summary statistics

outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	σ^2	**38..44**									

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$P(H|E) = \frac{P(E_1|H).P(E_2|H)....P(E_n|H).P(H)}{P(E)}$$

P[Yes| E] =(P(Outlook=*Sunny*|Play=*Yes*)
x **P(Temp.=*66*|Play=*Yes*)**
x **P(Hum.=*90*|Play=*Yes*)**
x P(Wind=True|Play=*Yes*)
x P(Play=*Yes*)) /**P[E]**

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = ??
P(Hum.=*90*|Play=*Yes*) = ??
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

41

The numeric weather data with summary statistics												
outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**Mean (μ)**	**73**	**74.6**	**Mean (μ)**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev (σ)**	**6.2**	**7.9**	**std dev (σ)**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	**σ^2**	**38..44**	**62.41**	**σ^2**	**104.04**	**86.2**					

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

P(Outl=*Sunny* | Play=*Yes*) = 2/9

P(Temp.=*66* | Play=*Yes*) = ??

P(Hum.=*90* | Play=*Yes*) = ??

P(Wind=True | Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

42

The numeric weather data with summary statistics												
outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	**σ^2**	**38..44**									

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

P(Outl=*Sunny* | Play=*Yes*) = 2/9

P(Temp.=*66* | Play=*Yes*) = 0.034

P(Hum.=*90* | Play=*Yes*) = ??

P(Wind=True | Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}6.2} e^{\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

43

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	**σ^2**	**38..44**									

P(Outl=*Sunny* | Play=*Yes*) = 2/9
P(Temp.=*66* | Play=*Yes*) = 0.034
P(Hum.=*90* | Play=*Yes*) = ??
P(Wind=True | Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

f(temp=66/Yes) =?

f(temp=66/No) =?

f(humidity=90/Yes) =?

f(humidity=90/No) =?

44

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	**σ^2**	**38..44**									

P(Outl=*Sunny* | Play=*Yes*) = 2/9
P(Temp.=*66* | Play=*Yes*) = 0.034
P(Hum.=*90* | Play=*Yes*) = 0.0221
P(Wind=True | Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

f(humidity=90/Yes) =

45

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	σ^2	**38..44**									

P(Outl=*Sunny* | Play=*Yes*) = 2/9
P(Temp.=*66* | Play=*Yes*) = 0.034
P(Hum.=*90* | Play=*Yes*) = 0.0221
P(Wind=True | Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

f(temp=66/Yes) = 0.034 f(humidity=90/Yes) = 0.0221
f(temp=66/No) = 0.0279 f(humidity=90/No) = 0.0380

46

Nhãn????

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$P(H|E) = \frac{P(E_1|H).P(E_2|H)....P(En|H).P(H)}{P(E)}$$

f(temp=66/Yes) = 0.034 f(humidity=90/Yes) = 0.0221
f(temp=66/No) = 0.0279 f(humidity=90/No) = 0.0380

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**mean**	**73**	**74.6**	**mean**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev**	**6.2**	**7.9**	**std dev**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	σ^2	**38..44**									

47

47

Dữ liệu liên tục

- **Bước 2- dự đoán**

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood(yes) = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036
Likelihood(no) = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136
Likelihood(yes) = 0.000036 / (0.000036 + 0. 000136) = 20.9%
Likelihood(no) = 0.000136 / (0.000036 + 0. 000136) = 79.1%

48

48

outlook			temperature			humidity			windy			play	
sunny	2/9	3/5	**Mean(μ)**	**73**	**74.6**	**Mean (μ)**	**79.1**	**86.2**	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	**std dev (σ)**	**6.2**	**7.9**	**std dev (σ)**	**10.2**	**9.7**	true	3/9	3/5		
rainy	3/9	2/5	**σ^2**	**38.44**	**62.41**	**σ^2**	**104.04**	**86.2**					

- **Bước 2- dự đoán**

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

P[Yes| E] = (2/9*0.034*0.022*3/9*9/14) /P[E] = 0.000036 /P[E]
P[No| E] = (3/5 *0.0279*0.038*3/5*5/14)/P[E]=0.000136 /P[E]

Likelihood(yes) = 0.000036 / (0.000036 + 0. 000136) = 20.9%
Likelihood(no) = 0.000136 / (0.000036 + 0. 000136) = 79.1%

49

49

Multinomial Naive Bayes

- Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bags of Words.
- Mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển.
- Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó

$$p(x_i|c) = \frac{N_{ci}}{N_c}$$

•Nci là tổng số lần từ thứ i xuất hiện trong các văn bản của class c, nó được tính là tổng của tất cả các thành phần thứ i của các feature vectors ứng với class c.
•Nc là tổng số từ (kể cả lặp) xuất hiện trong class c. Nói cách khác, nó bằng tổng độ dài của toàn bộ các văn bản thuộc vào class c.

50

Bernoulli Naive Bayes

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bẳng **0 hoặc 1.** Ví dụ: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không

Khi đó, $p(x_i|c)$ được tính bằng:

$$p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c)^{1-x_i}$$

p(i|c) có thể được hiểu là xác suất từ thứ "**i**" xuất hiện trong các văn bản của lớp **"c"**

51

Nội dung

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- Kết luận và hướng phát triển

Kết luận

- naïve Bayes
  - cho kết quả tốt trong thực tế mặc dù chịu những giả thiết về tính độc lập thống kê của các thuộc tính
  - phân lớp không yêu cầu phải ước lượng một cách chính xác xác suất
  - dễ cài đặt, học nhanh, kết quả dễ hiểu
  - sử dụng trong phân loại text, spam, etc
  - tuy nhiên khi dữ liệu có nhiều thuộc tính dư thừa thì naïve Bayes không còn hiệu quả
  - dữ liệu liên tục có thể không tuân theo phân phối chuẩn (=> kernel density estimators)

Hướng phát triển

- naïve Bayes
  - chọn thuộc tính con từ các thuộc tính ban đầu
  - chỉ sử dụng các thuộc tính con để học phân lớp
  - mạng Bayes : mối liên quan giữa các thuộc tính

54

55