

Applied Machine Learning

Lecture 09

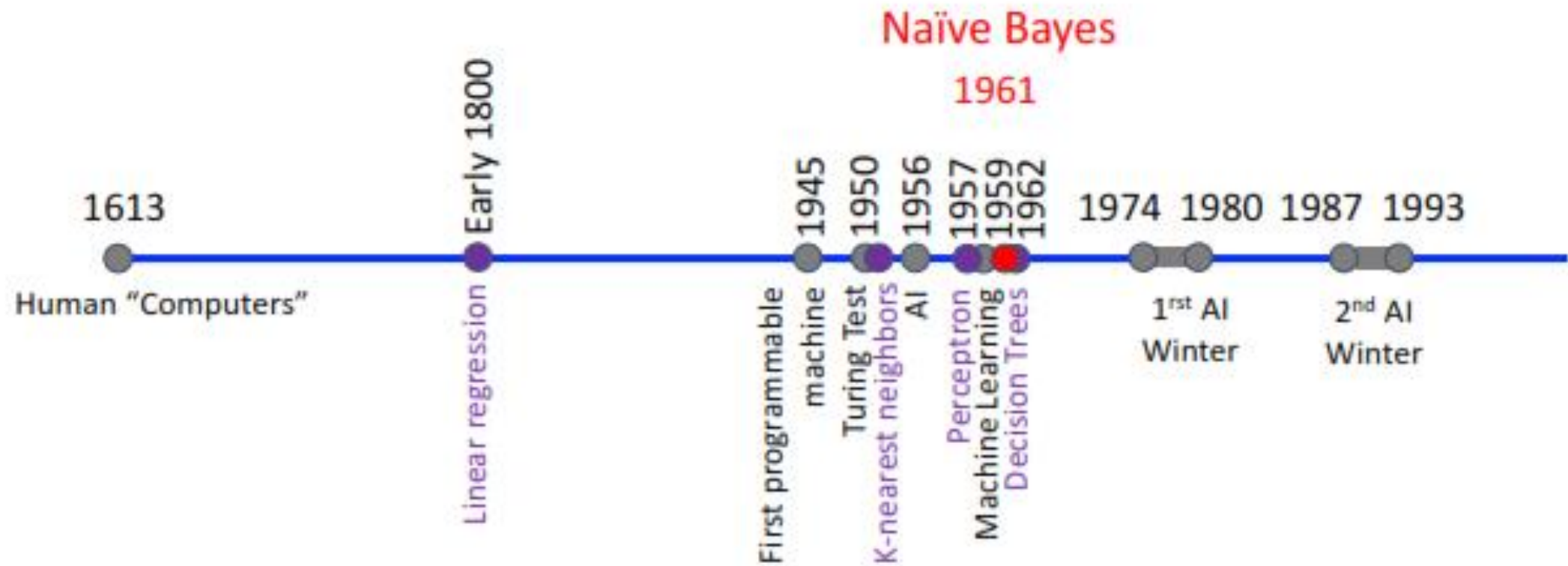
Naïve Bayes

Dr Muhammad Gufran Khan

Today's Topics

- Naïve Bayes

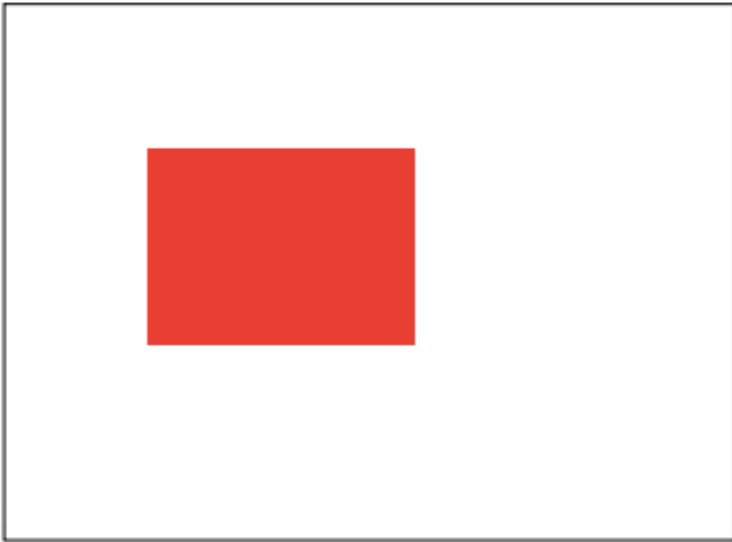
Historical Context of ML Models



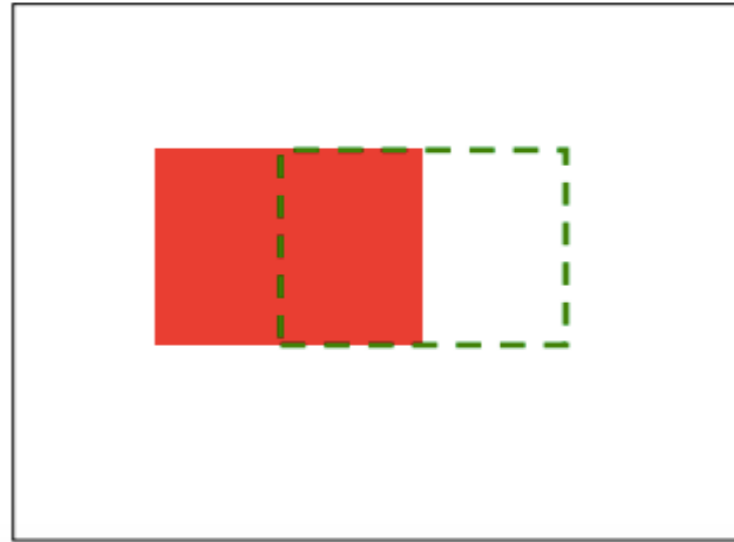
Background: Conditional Probability

- $P(A = 1 \mid B = 1)$: fraction of cases where A is true if B is true

$P(A = 0.2)$



$P(A|B = 0.5)$



Background: Conditional Probability

- Knowledge of additional random variables can improve our prior belief of another random variable
- $P(\text{Slept in movie}) = ?$
 - 0.5
- $P(\text{Slept in movie} \mid \text{Like Movie}) = ?$
 - $1/4$
- $(\text{Didn't sleep in movie} \mid \text{Like Movie}) = ?$
 - $3/4$

Slept	Liked
1	0
0	1
1	1
1	0
0	0
1	0
0	1
0	1

Background: Joint Distribution

- $P(A, B)$: probability a set of random variables will take a specific value

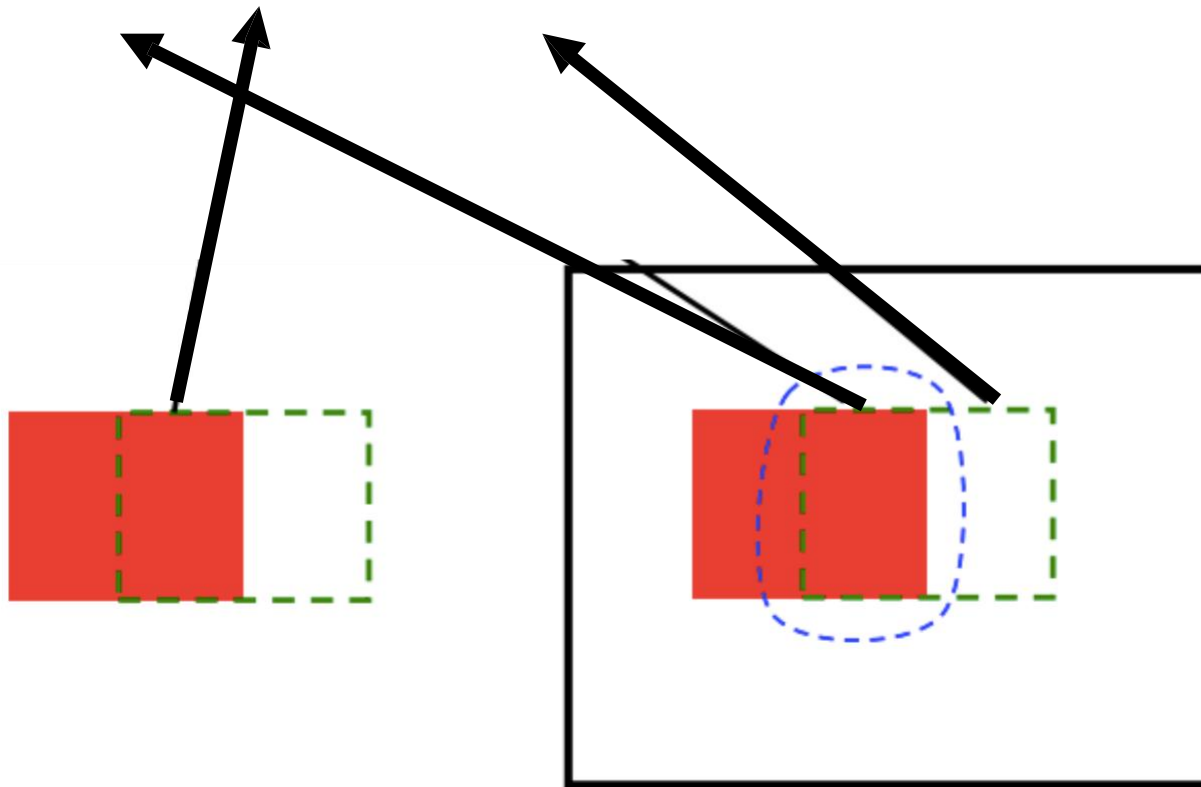
If we assume independence then

$$P(A, B) = P(A)P(B)$$

However, in many cases such an assumption maybe too strong (more later in the class)

Background: Chain Rule

- Joint probability can be represented with conditional probability
- $P(A, B) = P(A | B) * P(B)$



Bayes' Theorem: Derivation of Formula

- Recall Chain Rule:
 - $P(A, B) = P(A|B) * P(B)$
 - $P(A, B) = P(B|A) * P(A)$
- Therefore:
 - $P(A|B) * P(B) = P(B|A) * P(A)$
- Rearranging:
 - $P(A|B) = (P(B|A) * P(A))/P(B)$
- Rewriting:

$$P(C_i|features) = \frac{P(features|C_i)*P(C_i)}{P(features)}$$

Need to solve this...
more to follow

Need to solve this...
but how?

Want to find class with the largest probability

Constant for all classes... so can ignore this!

Naïve Bayes

- Learns a model of the joint probability of the input features and each class, and then picks the most probable class

Naïve Bayes: Naively Assumes Features Are Class Conditionally Independent

- Recall:

$$P(C_i | \text{features}) = P(\text{features} | C_i) * P(C_i)$$

$$P(\text{features} | C_i) = \prod_{j=1}^m P(x_j | C_i)$$

$$P(\text{features} | C_i) = P(x_1 | C_i) * P(x_2 | C_i) * \dots * P(x_m | C_i)$$

$$P(C_i | \text{features}) = P(x_1 | C_i) * P(x_2 | C_i) * \dots * P(x_m | C_i) * P(C_i)$$

If we assume independence then

$$P(A, B) = P(A)P(B)$$

However, in many cases such an assumption maybe too strong (more later in the class)

Naïve Bayes: Different Generative Models Can Yield the Observed Features

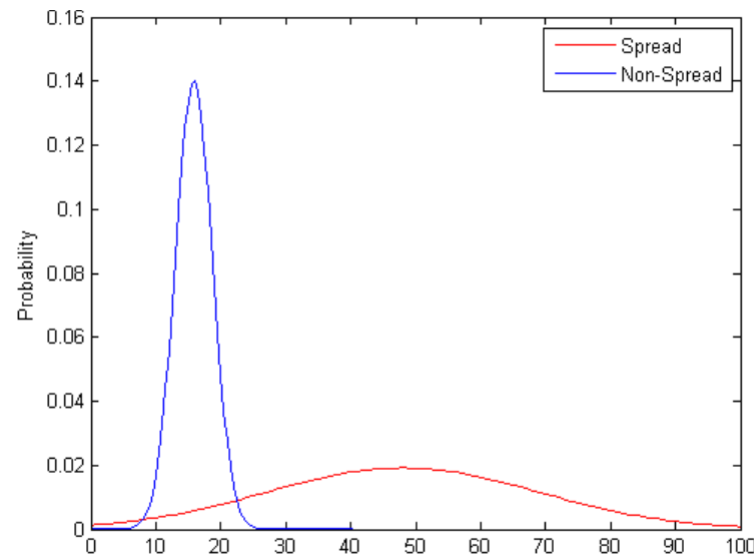
Recall: Want to find class with the largest probability

Key Decision: How to compute probability of each feature given the class?

$$P(C_i | features) = P(x_1 | C_i) * P(x_2 | C_i) * \dots * P(x_m | C_i) * P(C_i)$$

Naïve Bayes: Different Generative Models Can Yield the Observed Features

- **Gaussian** Naïve Bayes (typically used for “continuous”-valued features)
 - Assume data drawn from a Gaussian distribution: mean + standard deviation



$$P(C_i | features) = P(x_1 | C_i) * P(x_2 | C_i) * ... * P(x_m | C_i) * P(C_i)$$

Naïve Bayes: Different Generative Models Can Yield the Observed Features

- **Multinomial** Naïve Bayes (typically used for “discrete”-valued features)
 - Assume count data and computes fraction of entries belonging to the category

e.g.,

Movie	Type	Length	Liked?
m1	Comedy	Short	Yes
m2	Drama	Medium	Yes
m3	Comedy	Medium	No
m4	Drama	Long	No
m5	Drama	Medium	Yes
m6	Drama	Short	No
m7	Comedy	Short	Yes
m8	Drama	Medium	Yes

$$P(C_i | \text{features}) = P(x_1 | C_i) * P(x_2 | C_i) * \dots * P(x_m | C_i) * P(C_i)$$

Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = ?$
 - $5/8 = 0.625$

$$P(C_i | \text{features}) = P(x_1 | C_i) * P(C_i)$$

Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = ?$
 - $5/8 = 0.625$
- $P(\text{Not Liked}) = ?$
 - $3/8 = 0.375$

$$P(C_i | \text{features}) = P(x_1 | C_i) * P(C_i)$$

Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = 5/8 = 0.625$
- $P(\text{Not Liked}) = 3/8 = 0.375$
- $P(\text{IMDb Rating} \mid \text{Liked})$: Mean and Standard Deviation?
 - Mean = 8.06
 - Standard Deviation = 0.81

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(C_i)$$

Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = 5/8 = 0.625$
- $P(\text{Not Liked}) = 3/8 = 0.375$
- $P(\text{IMDb Rating} \mid \text{Liked})$
 - Mean = 8.06
 - Standard Deviation = 0.81
- $P(\text{IMDb Rating} \mid \text{Not Liked})$: Mean and Standard Deviation?
 - Mean = 5.5
 - Standard Deviation = 1.25

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(C_i)$$

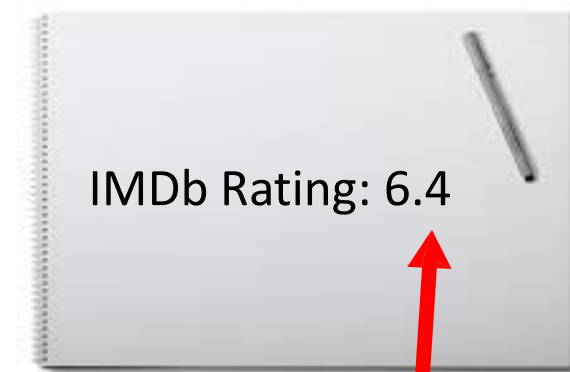
Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = 5/8 = 0.625$
- $P(\text{Not Liked}) = 3/8 = 0.375$
- $P(\text{IMDb Rating} \mid \text{Liked})$
 - Mean = 8.06
 - Standard Deviation = 0.81
- $P(\text{IMDb Rating} \mid \text{Not Liked})$
 - Mean = 5.5
 - Standard Deviation = 1.25

Test Example



- $P(\text{Liked} \mid \text{Features})$

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

(Can Use: <https://planetcalc.com/4986/>)

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(C_i)$$


Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = 5/8 = 0.625$
- $P(\text{Not Liked}) = 3/8 = 0.375$
- $P(\text{IMDb Rating} \mid \text{Liked})$
 - Mean = 8.06
 - Standard Deviation = 0.81
- $P(\text{IMDb Rating} \mid \text{Not Liked})$
 - Mean = 5.5
 - Standard Deviation = 1.25

Test Example



IMDb Rating: 6.4

- $P(\text{Liked} \mid \text{Features})$
 - $= 0.06 * 0.625$

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(C_i)$$

Gaussian Naïve Bayes: Example

e.g.,

x_1	
IMDb Rating	Liked?
7.2	Yes
9.3	Yes
5.1	No
6.9	No
8.3	Yes
4.5	No
8.0	Yes
7.5	Yes

- $P(\text{Liked}) = 5/8 = 0.625$
- $P(\text{Not Liked}) = 3/8 = 0.375$
- $P(\text{IMDb Rating} \mid \text{Liked})$
 - Mean = 8.06
 - Standard Deviation = 0.81
- $P(\text{IMDb Rating} \mid \text{Not Liked})$
 - Mean = 5.5
 - Standard Deviation = 1.25

Test Example



- $P(\text{Liked} \mid \text{Features})$
 - $= 0.06 * 0.625$
 - $= 0.0375$
- $P(\text{Not Liked} \mid \text{Features})$
 - $= 0.25 * 0.375$
 - $= 0.09$

Which class is the most probable?

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(C_i)$$

Multinomial Naïve Bayes: Example

	x_1	x_2	
Movie	Type	Length	Liked?
m1	Comedy	Short	Yes
m2	Drama	Medium	Yes
m3	Comedy	Medium	No
m4	Drama	Long	No
m5	Drama	Medium	Yes
m6	Drama	Short	No
m7	Comedy	Short	Yes
m8	Drama	Medium	Yes

- $P(\text{Liked}) = 5/8 = 0.625$
- $P(\text{Not Liked}) = 3/8 = 0.375$
- $P(\text{Comedy} \mid \text{Liked}) = ?$
 - $2/5 = 0.4$
- $P(\text{Comedy} \mid \text{Not Liked}) = ?$
 - $1/3 = 0.333$
- $P(\text{Drama} \mid \text{Liked}) = ?$
 - $3/5 = 0.6$
- $P(\text{Drama} \mid \text{Not Liked}) = ?$
 - $2/3 = 0.666$

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * P(C_i)$$

Multinomial Naïve Bayes: Example

	x_1	x_2	
Movie	Type	Length	Liked?
m1	Comedy	Short	Yes
m2	Drama	Medium	Yes
m3	Comedy	Medium	No
m4	Drama	Long	No
m5	Drama	Medium	Yes
m6	Drama	Short	No
m7	Comedy	Short	Yes
m8	Drama	Medium	Yes

- $P(\text{Short} \mid \text{Liked}) = ?$
 - $2/5 = 0.4$
- $P(\text{Short} \mid \text{Not Liked}) = ?$
 - $1/3 = 0.333$
- $P(\text{Medium} \mid \text{Liked}) = ?$
 - $3/5 = 0.6$
- $P(\text{Medium} \mid \text{Not Liked}) = ?$
 - $1/3 = 0.333$
- $P(\text{Long} \mid \text{Liked}) = ?$
 - $0/5 = 0$
- $P(\text{Long} \mid \text{Not Liked}) = ?$
 - $1/3 = 0.333$

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * P(C_i)$$

Test Example

Multinomial Naïve Bayes: Example



	x_1	x_2	
Movie	Type	Length	Liked?
m1	Comedy	Short	Yes
m2	Drama	Medium	Yes
m3	Comedy	Medium	No
m4	Drama	Long	No
m5	Drama	Medium	Yes
m6	Drama	Short	No
m7	Comedy	Short	Yes
m8	Drama	Medium	Yes

Which class is the most probable?

- $P(\text{Liked}) = 0.63$
- $P(\text{Not Liked}) = 0.38$
- $P(\text{Comedy} \mid \text{Liked}) = 0.4$
- $P(\text{Comedy} \mid \text{Not Liked}) = 0.33$
- $P(\text{Drama} \mid \text{Liked}) = 0.6$
- $P(\text{Drama} \mid \text{Not Liked}) = 0.67$
- $P(\text{Short} \mid \text{Liked}) = 0.4$
- $P(\text{Short} \mid \text{Not Liked}) = 0.33$
- $P(\text{Medium} \mid \text{Liked}) = 0.6$
- $P(\text{Medium} \mid \text{Not Liked}) = 0.33$
- $P(\text{Long} \mid \text{Liked}) = 0$
- $P(\text{Long} \mid \text{Not Liked}) = 0.33$

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * P(C_i)$$

$$P(\text{Liked} \mid \text{Features}) = 0.4 \times 0.6 \times 0.63 = 0.15$$

$$P(\text{Not Liked} \mid \text{Features}) = 0.33 \times 0.33 \times 0.38 = 0.04$$

Multinomial Naïve Bayes: Example

Test Example



	x_1	x_2	
Movie	Type	Length	Liked?
m1	Comedy	Short	Yes
m2	Drama	Medium	Yes
m3	Comedy	Medium	No
m4	Drama	Long	No
m5	Drama	Medium	Yes
m6	Drama	Short	No
m7	Comedy	Short	Yes
m8	Drama	Medium	Yes

Which class is the most probable?

- $P(\text{Liked}) = 0.63$
- $P(\text{Not Liked}) = 0.38$
- $P(\text{Comedy} \mid \text{Liked}) = 0.4$
- $P(\text{Comedy} \mid \text{Not Liked}) = 0.33$
- $P(\text{Drama} \mid \text{Liked}) = 0.6$
- $P(\text{Drama} \mid \text{Not Liked}) = 0.67$
- $P(\text{Short} \mid \text{Liked}) = 0.4$
- $P(\text{Short} \mid \text{Not Liked}) = 0.33$
- $P(\text{Medium} \mid \text{Liked}) = 0.6$
- $P(\text{Medium} \mid \text{Not Liked}) = 0.33$
- $P(\text{Long} \mid \text{Liked}) = 0$
- $P(\text{Long} \mid \text{Not Liked}) = 0.33$

To avoid zero, assume training data is so large that adding one to each count makes a negligible difference

$$P(C_i \mid \text{features}) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * P(C_i)$$

Additional Slides



NAÏVE BAYES CLASSIFIER



Naïve Bayes Classifier

30

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \dots a_n \rangle$.

Most probable value of $f(x)$ is:

$$\begin{aligned} v_{MAP} &= \operatorname{argmax}_{v_j \in V} P(v_j | a_1, a_2 \dots a_n) \\ v_{MAP} &= \operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2 \dots a_n | v_j) P(v_j)}{P(a_1, a_2 \dots a_n)} \\ &= \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j) \end{aligned}$$

Naïve Bayes Classifier

31

Assumption:

- The naive Bayes classifier is based on the assumption that the attribute values are *conditionally independent* given the target value.

$$P(a_1, a_2 \dots a_n | v_j) = \prod_i P(a_i | v_j)$$

- Substituting this into

$$v_{MAP} = \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n | v_j) P(v_j)$$

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

Naïve Bayes Classifier

32

Naive Bayes Algorithm

Naive_Bayes_Learn(*examples*)

For each target value v_j

- $\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$
- For each attribute value a_i of each attribute a
 $\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance(x)

$$v_{NB} = \operatorname{argmax}_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Example

33

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Weekday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time

Air-Traffic Data

Example

34

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Air-Traffic Data

Example

35

- In this database, there are **four attributes** with **20 tuples**.

A = [Day, Season, Fog, Rain]

- The **categories of classes** are:

C = [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, the target is to find most likely classification for any other unseen instance, for example:

Week Day	Winter	High	None	???
-----------------	---------------	-------------	-------------	------------

- Classification technique eventually to map this tuple into an accurate class.

Example

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

36

		Class			
Attribute		On Time(14)	Late(2)	Very Late(3)	Cancelled(1)
Day	Weekday	9/14 = 0.64	2/2 = 1	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	0/2 = 0	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
Season	Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	1/1 = 1
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Example

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

37

		Class			
Attribute		On Time(14)	Late(2)	Very Late(3)	Cancelled(1)
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	6/14 = 0.43	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	6/14 = 0.43	1/2 = 0.5	0/3 = 0	0/1 = 0
	Heavy	2/14 = 0.14	0/2 = 0	2/3 = 0.67	1/1 = 1
Prior Probability		14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Example

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

38

Instance:

Week Day	Winter	High	None	???
----------	--------	------	------	-----

Case 1: Class = On Time:

$$0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.43 = 0.0078$$

Case 2: Class = Late:

$$0.10 \times 1.0 \times 1.0 \times 0.50 \times 0.50 = \mathbf{0.025}$$

Case 3: Class = Very Late:

$$0.15 \times 1.0 \times 0.67 \times 0.33 \times 0.33 = 0.0109$$

Case 4: Class = Cancelled:

$$0.05 \times 0.0 \times 0.0 \times 1.0 \times 0.0 = 0.0$$

Case 2 is the strongest; hence the correct classification is **Late**

Example 2

39

X					Y
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example 2

39

Consider *PlayTennis* again, and new instance

$\langle Outlk = sun, Temp = cool, Humid = high, Wind = strong \rangle$

Want to compute:

$$v_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

$$P(y) P(sun|y) P(cool|y) P(high|y) P(strong|y) = .005$$

$$P(n) P(sun|n) P(cool|n) P(high|n) P(strong|n) = .021$$

$$\rightarrow v_{NB} = n$$

Naïve Bayes Classifier

40

- Highly practical Bayesian learning method
 - ▣ In some domains its performance can be comparable to that of neural network and decision-tree learning
- **When to use,**
 - ▣ Moderate or large training dataset is available
 - ▣ Attributes that describe instances are conditionally independent given classification
- **Application**
 - ▣ Diagnosis systems (expert systems)
 - ▣ Classifying text documents

Reading Material

41

- **Artificial Intelligence, A Modern Approach**

Stuart J. Russell and Peter Norvig

- ▣ Chapter 13.

- **Machine Learning**

Tom M. Mitchell

- ▣ Chapter 6.

