Naive Bayes classifier

Procheta Sen

UNIVERSITY OF LIVERPOOL

Bayes' Rule

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where H and E are events and $P(E) \neq 0$

Terminology

- $P(H|E)$: The probability of **hypothesis** H, given **evidence** E
- $P(E)$: **Marginal** probability of the evidence E
- $P(H)$: **Prior** probability of hypothesis H
- $P(E|H)$: Likelihood of the evidence given hypothesis
- $P(H|E)$: **Posterior** probability of the hypothesis H

Bayes' Rule

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

where H and E are events and $P(E) \neq 0$

Bayes' Rule is useful for estimating $P(H \mid E)$ when it is hard to estimate $P(H \mid E)$ directly from the training data, but $P(E \mid H)$, $P(H)$, and $P(E)$ can be estimated more easily.

Bayes' Rule: derivation

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

where H and E are events and $P(E) \neq 0$

By the definition of conditional probability, we have

$$P(H \mid E) = \frac{P(H, E)}{P(E)} \qquad \text{and} \qquad P(E \mid H) = \frac{P(H, E)}{P(H)},$$

from which we derive Bayes' Rule.

Example (single feature)

- **Meningitis** causes a **stiff neck** 50% of the time.
- Meningitis occurs 1/50000 and stiff neck occurs 1/20.
- Compute the probability of meningitis, given that the patient has a stiff neck.
- H = meningitis, E = stiff neck
- $P(H)$ = 1/50000, $P(E)$ = 1/20, $P(E \mid H)$ = 0.5
- From Bayes' rule we have

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} = 0.0002$$

Example

- **Meningitis** causes a **stiff neck** 50% of the time.
- Meningitis occurs 1/50000 and stiff neck occurs 1/20.
- Compute the probability of meningitis, given that the patient has a stiff neck.
- H = meningitis, E = stiff neck
- $P(H)$ = 1/50000, $P(E)$ = 1/20, $P(E \mid H)$ = 0.5
- From Bayes' rule we have

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} = 0.0002$$

- If we have 1-dimensional space (only one feature: $\overline{X} = (a)$), then we can estimate $P(H \mid \overline{X})$ directly from the training data set.
- It becomes more problematic if we have higher dimension. Say $\overline{X} = (a_1, a_2, \ldots, a_d)$ for $d > 1$.

Example (2 features)

- Let H=**engine-does-not-start**, and
- Evidences A = **weak-battery** and B = **no-gas**
- To estimate $P(H|A,B)$ directly, we need to restrict our consideration only to those cars (objects) in the dataset that had a **weak battery** and **no gas**. Among those we need to count the cars with **non-working engine**. Such cases could be rare in our dataset making estimate of $P(H|A,B)$ unreliable or zero (in the worst case).

$$P(H|A,B) = \frac{\text{\# weak-bat. \& no-gas \& eng.-not-working}}{\text{\# weak-bat. \& no-gas}}$$

- Bayes' rule provides a way of expressing $P(H|A,B)$ directly in terms of $P(A,B|H)$:

$$P(H|A,B) = \frac{P(A,B|H)P(H)}{P(A,B)}$$

and estimate the latter using **naive Bayes approximation** (which is much easier to do).

Naive Bayes approximation

- Let C be a random variable representing the class of an unseen d-dimensional test object $\overline{X} = (x_1, x_2, \ldots, x_d)$, where $x_1, x_2, \ldots, x_d$ denote random variables of individual dimensions

- Given a specific test object $(a_1, a_2, \ldots, a_d)$ the goal is to estimate $P(C = c \,|\, \overline{X} = (a_1, a_2, \ldots, a_d)) = P(C = c \,|\, x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d)$

- By Bayes' rule

$$P(C = c \,|\, x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d) = \frac{P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \,|\, C)}{P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d)}$$

Naive Bayes approximation

$$P(C = c \mid x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d) = \frac{P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c)}{P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d)}$$

Does not depend on the class variable C

The class c with the largest numerator

$$P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c)$$

has the largest **posterior** probability

$$P(C = c \mid x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d)$$

Naive Bayes approximation

$$P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \,|\, C = c)$$

- $P(C = c)$: Can be **estimated** as the fraction of the training data objects that belong to class c.
- How to estimate $P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \,|\, C = c)$?

Naive assumption

The values of different features $x_1, x_2, \ldots, x_d$
are **independent** of one another conditional on the class

Independent events (2 events)

Joint probability of two events A and B

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

this holds for ANY pair of events A and B, irrespective of whether they are independent or not.

If A is independent of B, then B's occurrence has no "consequence" on A

$$P(A \mid B) = P(A)$$

Thus, when A and B are **independent**, their joint probability

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

Independent events (> 2 events)

A finite set $\{A_1, A_2, \ldots, A_n\}$ of events is **mutually independent** if for any $2 \leq k \leq n$ and for any subset of k events

$$\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\} \subseteq \{A_1, A_2, \ldots, A_n\}$$

we have

$$P(A_{i_1}, A_{i_2}, \ldots, A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j})$$

Naive Bayes approximation

How to estimate $P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c)$?

Naive assumption

The values of different features $x_1, x_2, \ldots, x_d$ are **independent** of one another conditional on the class

$$P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c) = \prod_{i=1}^{d} P(x_i = a_i \mid C = c)$$

$P(x_i = a_i \mid C = c)$: Can be estimated as the fraction of training objects in class c that have feature $x_i = a_i$.

Naive Bayes approximation

$P(x_i = a_i \mid C = c)$: Can be estimated as the fraction of training objects in class c that have feature $x_i = a_i$

This is much easier to estimate from the training data than
$P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c)$

Hence, under the independence assumption, the estimation of

$$P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c) = P(C = c)\prod_{i=1}^{d} P(x_i = a_i \mid C = c)$$

reduces to the estimations of $P(C), P(x_1 = a_1 \mid C = c), P(x_2 = a_2 \mid C = c), \ldots, P(x_d = a_d \mid C = c)$

Being naive makes life easy

Example (2 features)

- Let H=**engine-does-not-start**, and
- Evidences A = **weak-battery** and B = **no-gas**
- Direct estimation

$$P(H \mid A, B) = \frac{\text{\# weak-bat. \& no-gas \& eng.-not-working}}{\text{\# weak-bat. \& no-gas}}$$

- Using Bayes' rule + **Naive Bayes approximation**

$$P(H \mid A, B) = \frac{P(A, B \mid H)P(H)}{P(A, B)} = \frac{P(A \mid H)P(B \mid H)P(H)}{P(A, B)}$$

- $P(A \mid H)$ can be estimated as the fraction of cars with **weak battery** among cars with **engine not working**
- $P(B \mid H)$ can be estimated as the fraction of cars with **no gas** among cars with **engines not working**

Making the independence assumption makes estimates possible in practice

Bayes' rule: Proportional Form

- Assume $\bar{X} = (a_1, a_2, \ldots, a_k)$ is the input test object
- We need to select/predict the class of $\bar{X}$ from the set $\{c_1, c_2, \ldots, c_k\}$.

$$P(C = c \mid x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d) = \frac{P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c)}{P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d)}$$

$$P(C = c \mid x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d) \propto P(C = c)P(x_1 = a_1, x_2 = a_2, \ldots, x_d = a_d \mid C = c)$$

$$P(C \mid X) \propto P(C)\, P(X \mid C)$$

posterior $\propto$ prior $\times$ likelihood

Example: predicting whether to play or not

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Test instance $\bar{X} = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

$$\begin{aligned}
P(\text{Play} = \text{yes} \mid \bar{X}) &\propto P(\bar{X} \mid \text{Play} = \text{yes}) P(\text{Play} = \text{yes}) \\
&= P(\text{Outlook} = \text{sunny} \mid \text{Play} = \text{yes}) \times P(\text{Temp} = \text{cool} \mid \text{Play} = \text{yes}) \\
&\times P(\text{Humidity} = \text{high} \mid \text{Play} = \text{yes}) \times P(\text{Windy} = \text{true} \mid \text{Play} = \text{yes}) \times P(\text{Play} = \text{yes}) \\
&= 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529
\end{aligned}$$

Example: predicting whether to play or not

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Test instance $\bar{X} = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

$$P(\text{Play} = \text{no} \,|\, \bar{X}) \propto P(\bar{X} \,|\, \text{Play} = \text{no}) P(\text{Play} = \text{no})$$

$$= P(\text{Outlook} = \text{sunny} \,|\, \text{Play} = \text{no}) \times P(\text{Temp} = \text{cool} \,|\, \text{Play} = \text{no})$$

$$\times P(\text{Humidity} = \text{high} \,|\, \text{Play} = \text{no}) \times P(\text{Windy} = \text{true} \,|\, \text{Play} = \text{no}) \times P(\text{Play} = \text{no})$$

$$= 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.020$$

Example: predicting whether to play or not

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Test instance $\bar{X} = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

$$P(\text{Play} = \text{yes} \,|\, \bar{X}) \propto P(\bar{X} \,|\, \text{Play} = \text{yes}) \times P(\text{Play} = \text{yes}) = 0.00529$$

$$P(\text{Play} = \text{no} \,|\, \bar{X}) \propto P(\bar{X} \,|\, \text{Play} = \text{no}) \times P(\text{Play} = \text{no}) = 0.020$$

Therefore, $\text{Play} = \text{no}$.

Naive Bayes: classification task

In the previous setting (**choose the most probable class**):

- Given a test object $\overline{X} = (a_1, a_2, \ldots, a_k)$, we wanted to predict its class C
- For this, we used the proportional form

$$P(C = c \,|\, X = \overline{X}) \propto P(C = c) \prod_{i=1}^{d} P(x_i = a_i \,|\, C = c)$$

and it was enough to find $c \in \{c_1, c_2, \ldots, c_k\}$ that maximises

$$P(C = c) \prod_{i=1}^{d} P(x_i = a_i \,|\, C = c)$$