Bayesian Learning

MACHINE LEARNING APPROACH

Outline

Conditional Probability

- Definition
- Example

Bayes' Theorem

- Example

Conditional Probability

Definition:

- Probability of an event happening has some relationship to one or more other events.

Example:

- Probability of getting a parking space is connected to the time of day you park, where you park, and what conventions are going on at any time.
- Bayes' theorem gives you the actual probability of an **event** given information about **tests**.

Conditional Probability - Events and Tests

Test:

There is a **test** for liver disease

Event:

Actually having liver disease or not.

Bayes' Theorem

The Theorem was named after English mathematician Thomas Bayes (1701-1761).

Bayes' theorem takes the test results and calculates your *real probability* that the test has identified the event.

Bayes' Theorem (also known as Bayes' rule) is a simple formula used to calculate conditional probability.

The formal definition for the rule is:

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

Bayes Theorem

m1

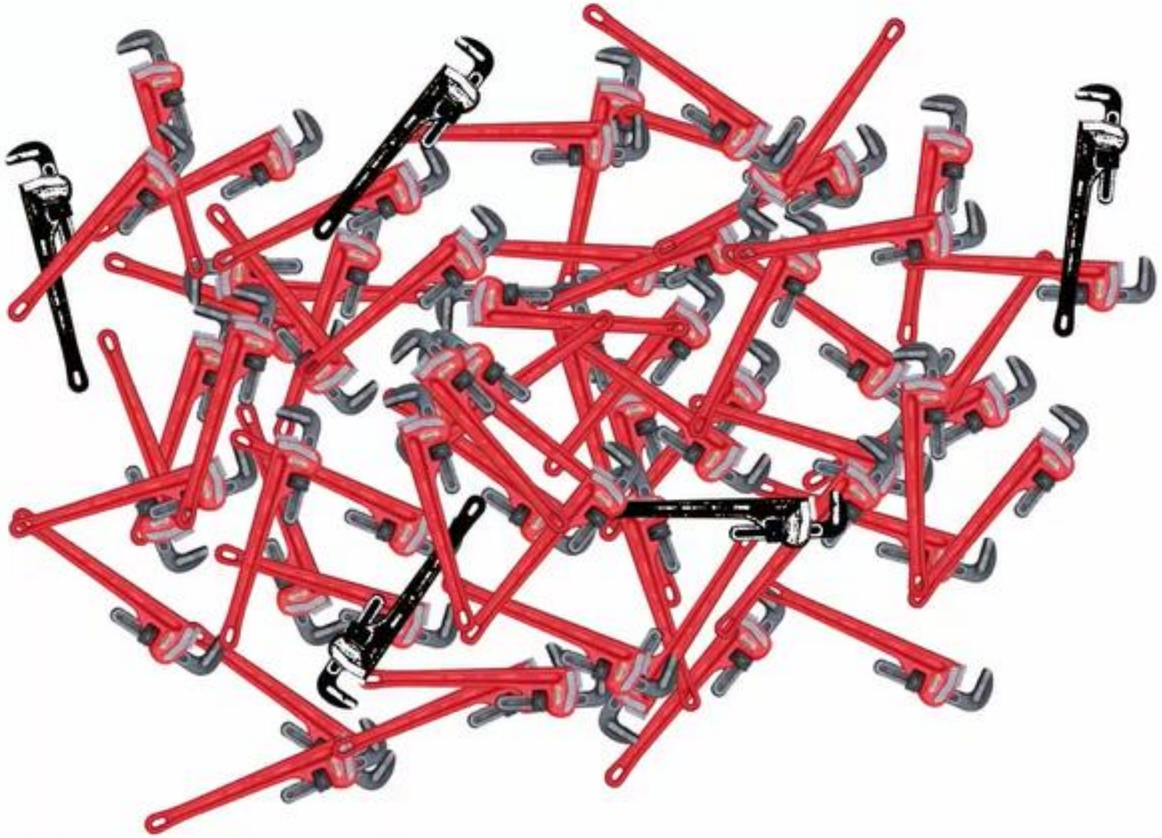
m1 m1 m1 m1 m1 m1 m1 m1 m1 m1 m1 m1 m1

m2

m2 m2 m2 m2 m2 m2 m2 m2 m2 m2

Bayes Theorem

m1



m2

Bayes Theorem

What's the probability?

m2

Bayes Theorem

Mach1: 30 wrenches / hr
Mach2: 20 wrenches / hr

-> P(Mach1) = 30/50 = 0.6
-> P(Mach2) = 20/50 = 0.4

Out of all produced parts:
We can SEE that 1% are defective

-> P(Defect) = 1%

Out of all defective parts:
We can SEE that 50% came from mach1
And 50% came from mach2

-> P(Mach1 | Defect) = 50%
-> P(Mach2 | Defect) = 50%

Question:
What is the probability that a part produced by mach2 is defective = ?

-> P(Defect | Mach2) = ?

Bayes Theorem

Mach1: 30 wrenches / hr
Mach2: 20 wrenches / hr
Out of all produced parts:
We can SEE that 1% are defective
Out of all defective parts:
We can SEE that 50% came from mach1
And 50% came from mach2
Question:
What is the probability that a part produced by mach2 is defective = ?

-> P(Mach2) = 20/50 = 0.4
-> P(Defect) = 1%
-> P(Mach2 | Defect) = 50%
-> P(Defect | Mach2) = ?

$$P(Defect \mid Mach2) = \frac{P(Mach2 \mid Defect) * P(Defect)}{P(Mach2)}$$

Bayes Theorem

Mach1: 30 wrenches / hr
Mach2: 20 wrenches / hr
Out of all produced parts:
We can SEE that 1% are defective
Out of all defective parts:
We can SEE that 50% came from mach1
And 50% came from mach2
Question:
What is the probability that a part produced by mach2 is defective = ?

-> P(Mach2) = 20/50 = 0.4
-> P(Defect) = 1%
-> P(Mach2 | Defect) = 50%
-> P(Defect | Mach2) = ?

$$P(\text{Defect} \mid \text{Mach2}) = \frac{0.5 \quad * \quad 0.01}{0.4} = 0.0125$$

Bayes Theorem

Quick exercise:

$P(Defect \mid Mach1) = ?$

Bayes Theorem

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

P(h) *prior probability of* ***h***, reflects any background knowledge about the chance that h is correct

P(D) *prior probability of* ***D***, probability that D will be observed

P(D|h) probability of observing **D** given a world in which **h** holds

P(h|D) *posterior probability of* ***h***, reflects confidence that **h** holds after **D** has been observed

MAP Hypothesis

In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis h ϵ H given the observed training data D

Any maximally probable hypothesis is called *maximum a posteriori* (**MAP**) hypotheses

$$
\begin{aligned}
h_{MAP} &= \underset{h \in H}{argmax}\ P(h|D) \\
&= \underset{h \in H}{argmax}\ \frac{P(D|h)P(h)}{P(D)} \\
&= \underset{h \in H}{argmax}\ P(D|h)P(h)
\end{aligned}
$$

Note that P(D) can be dropped, because it is a constant independent of h

ML Hypothesis

Sometimes it is assumed that every hypothesis in H is equally probable a priori

In this case, the equation above can be simplified, we need only P(D|H) to find most probable hypothesis.

P(D|h) is often called the *likelihood of D given h*

Any hypothesis that maximizes P(D|h) is called *maximum likelihood* (ML) hypothesis h_{ML}

$$h_{ML} = \underset{h \in H}{argmax}\, P(D|h)$$

note that in this case P(h) can be dropped,

because it is equal for each h Є H

Example

Consider a medical diagnosis problem in which there are two alternative hypotheses:
(1) the patient has a particular (denoted by *cancer*).
(2) the patient does not (denoted by $\neg$*cancer)*
Prior knowledge: over the entire population of people only .008 have this disease.

The available data is from a particular laboratory test with two possible outcomes:

$\oplus$ (positive) and $\ominus$ (negative)

The lab test is only an **imperfect** indicator of the disease. The test returns a **correct positive result** in only 98% of the cases in which the disease is actually present and a **correct negative result** in only 97% of the cases in which the disease is not present.

Example

$\oplus$ (positive) and $\ominus$ (negative)

$$P(cancer) = .008 \qquad P(\neg cancer) = 0.992$$

$$P(\oplus|cancer) = .98 \qquad P(\ominus|cancer) = .02$$

$$P(\oplus|\neg cancer) = .03 \qquad P(\ominus|\neg cancer) = .97$$

Example

Suppose, a new patient is observed for whom the lab test returns a positive result.

Should we diagnose the patient as having cancer or not?

$$P(\oplus|cancer)P(cancer) = (.98).008 = .0078$$
$$P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$$
$$\Rightarrow h_{MAP} = \neg cancer$$

Example

the exact posterior probabilites can be determined by normalizing the above properties to 1

$$P(cancer|\oplus) = \frac{.0078}{.0078 + 0.0298} = .21$$

$$P(\neg cancer|\oplus) = \frac{.0298}{.0078 + 0.0298} = .79$$

$\Rightarrow$ the result of Bayesian inference depends strongly on the prior probabilities, which must be available in order to apply the method directly

Bayes Optimal Classifier

Bayes Optimal Classifier

- Consider a hypothesis space containing three hypotheses: h_1, h_2, and h_3.
- Posterior probabilities of $\mathbf{h_1}$, $\mathbf{h_2}$, and $\mathbf{h_3}$ given the training data are **0.4, 0.3,** and **0.3** respectively.
- Thus, **$\mathbf{h_1}$ is the MAP** hypothesis.

What is the most probable ***classification*** of the new instance given the training data?

$\mathbf{h_1 = 0.4}$
$\mathbf{h_2 = 0.3}$
$\mathbf{h_3 = 0.3}$

If the new instance (x) is classified as +ve by h1, then x is +ve.

Bayes Optimal Classifier

- New instance x is classified positive by h_1, but negative by $h2$ and h3.
- Taking all hypotheses into account, the probability that x is positive is .4, a the probability that it is negative is therefore .6.
- The most probable classification (negative) in this case is different from the classification generated by the MAP hypothesis

Bayes Optimal Classifier

- Most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities.
- If the possible classification of new example take any value v_j from some set V, then the probability **$P(v_j|D)$** for new instance is,

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Bayes Optimal Classifier

The optimal classification of the new instance is the value $\boldsymbol{v_j}$, for which $\mathbf{P}\ (\mathbf{v_j}|\mathbf{D})$ is maximum.

Bayes optimal classification:

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Any system that classifies new instances according to above equation is called a ***Bayes*** optimal ***classifier*** or Bayes optimal learner.

Bayes Optimal Classifier

To illustrate in terms of the above example, the set of possible classifications of the new instance is $V = \{\oplus, \ominus\}$, and

Posterior probabilities of these hypotheses

$$\boxed{\begin{aligned} P(h_1|D) &= .4, \\ P(h_2|D) &= .3, \\ P(h_3|D) &= .3, \end{aligned}} \boxed{\begin{aligned} P(\ominus|h_1) &= 0, \quad P(\oplus|h_1) = 1 \\ P(\ominus|h_2) &= 1, \quad P(\oplus|h_2) = 0 \\ P(\ominus|h_3) &= 1, \quad P(\oplus|h_3) = 0 \end{aligned}}$$

New instance V is classified as **+**ve and **−**ve.

therefore

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = .4$$

(1*.4 + 0*.3 + 0*.3 = 0.4)

$$\sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = .6$$

(0*.4+1*.3+1*.3 = 0.6)

Bayes Optimal Classifier

$$\underset{v_j \in \{\oplus, \ominus\}}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

Bayesian Learning

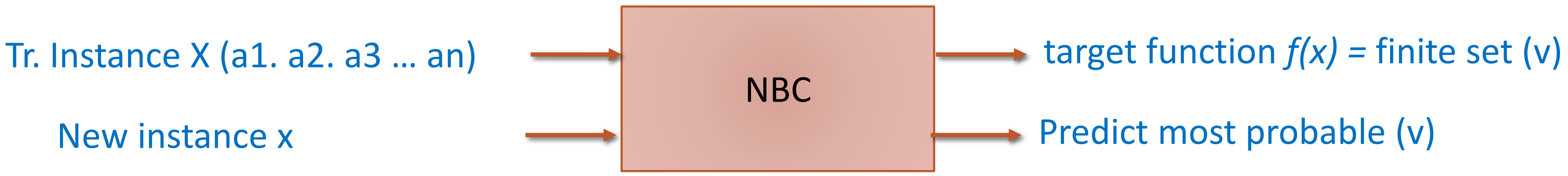
IN MACHINE LEARNING

Naïve Bayes Classifier

MACHINE LEARNING APPROACH

Naive Bayes Classifier

- This is a practical learning method.
- Its performance is comparable to that of neural network and decision tree learning.
- In naive Bayes classifier learning tasks
  - Each instance x is described by a conjunction of attribute values and
  - The target function *f(x)* can take on any value from some finite set V.



Naive Bayes Classifier

- Bayesian approach classify the new instance by assigning the most probable target value, V_{MAP} given the attribute values <a_1, a_2.. .a_n> that describe the instance.

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

Naive Bayes Classifier

We can use Bayes theorem to rewrite this expression as

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

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

Now we could attempt to estimate the two terms in above Equation based on the training data.

Naive Bayes Classifier

- Estimate each of the $P(v_j)$: by counting the frequency of each target value v_j occurs in the training data.

- Estimating different $\boldsymbol{P(a_1, a_2 .. . a_n | v_j)}$ in this fashion we need a very, very large set of training data.

Naive Bayes Classifier

The naive Bayes classifier assume that the attribute values are conditionally independent given the target value.

- Probability of observing the conjunction $a_1, a_2 .. . a_n$ is just the product of the probabilities for the individual attributes:

$$P(a_1, a_2 \ldots a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

Substituting this into Equation of V_{MAP},

Naive Bayes Classifier

Naive Bayes classifier:

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

where $\boldsymbol{V_{NB}}$ ***is*** the target value output by the naive Bayes classifier.

Naive Bayes Classifier

One interesting difference between the naive Bayes learning method and other learning methods:

- There is no explicit search through the space of possible hypotheses
- The space of possible hypotheses is the space of possible values that can be assigned to the various $P(v_j)$ and $P(a_i|v_j)$ terms.
- The hypothesis is formed without searching, simply by counting the frequency of various data combinations within the training examples.

Naïve Bayes Classifier

Numerical Example

Example

- Apply the naive Bayes classifier to a concept learning problem we considered for decision tree learning: classifying days according to whether someone will play tennis.

- This table provides a set of 14 training examples of the target concept **PlayTennis**, where each day is described by the attributes ***Outlook, Temperature, Humidity, and Wind***.

- Use naive Bayes classifier and the training data from this table to classify the following novel instance:

(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

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

Our task is to predict the target value (***yes*** or ***no***) of the target concept ***PlayTennis*** for this new instance.

The target value V_{NB} is given by

$$v_{NB} = \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

$$= \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} P(v_j) \quad P(Outlook = sunny | v_j) P(Temperature = cool | v_j)$$

$$P(Humidity = high | v_j) P(Wind = strong | v_j)$$

(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

Example

To calculate V_{NB} we now require 10 probabilities that can be estimated from the training data.

First, the probabilities of the different target values can easily be estimated based on their frequencies over the **14** training examples

P(PlayTennis = yes) = **9/14** = **.64**

P(PlayTennis = no) = **5/14** = **.36**

An Illustrative Example

Similarly, we can estimate the conditional probabilities. For example, those for

Wind = ***strong*** are

P(Wind = ***stronglPlayTennis*** = ***yes)*** = 3/9 = **.33**

P(Wind = ***strongl PlayTennis*** = ***no)*** = 3/5 = ***.60***

Using these probability estimates and similar estimates for the remaining attribute values, we calculate V_{NB} as follows

Example

$$P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

$$P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

Thus, the naive Bayes classifier assigns the target value ***PlayTennis*** = ***no*** to this new instance, based on the probability estimates learned from the training data.

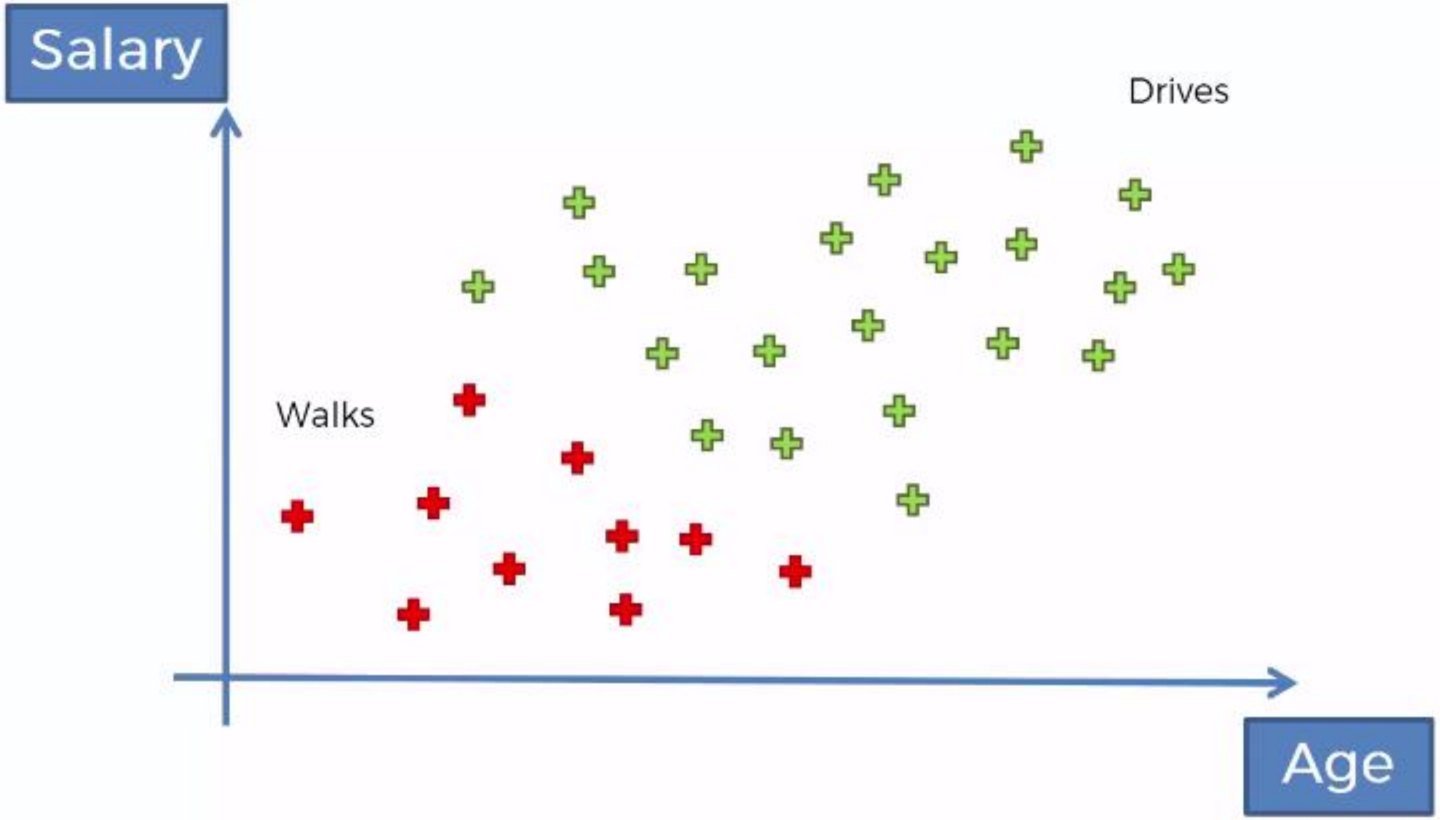
By normalizing the above quantities to sum to one we can calculate the conditional probability that the target value is ***no,*** given the observed attribute values. For the current example, this probability

$$\frac{.0206}{.0206+.0053} = .795.$$

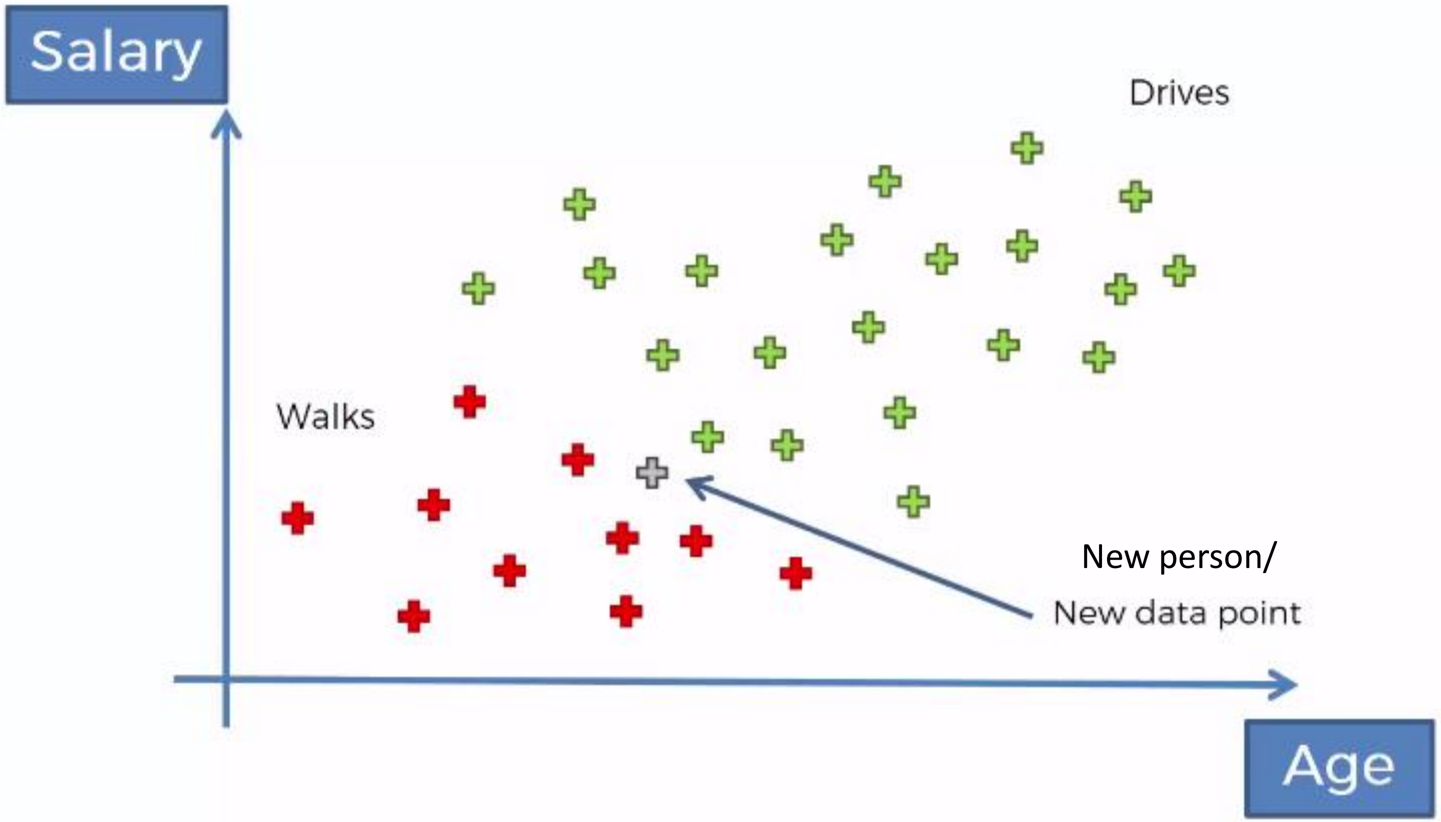
Naïve Bayes Classifier

Model Example

Naïve Bayes



Naïve Bayes



Step 1

$$P(Walks|X) = \frac{P(X|Walks) * P(Walks)}{P(X)}$$

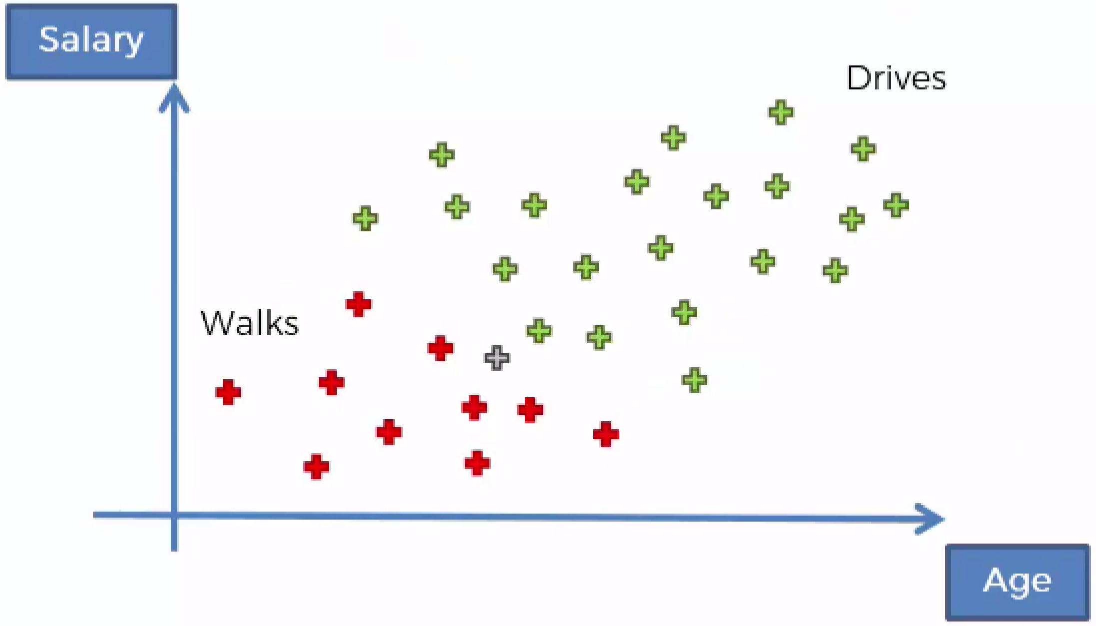
Step 2

$$P(Drives|X) = \frac{P(X|Drives) * P(Drives)}{P(X)}$$

Step 3

$$P(Walks|X) \ v.s. \ P(Drives|X)$$

Naïve Bayes: Step 1

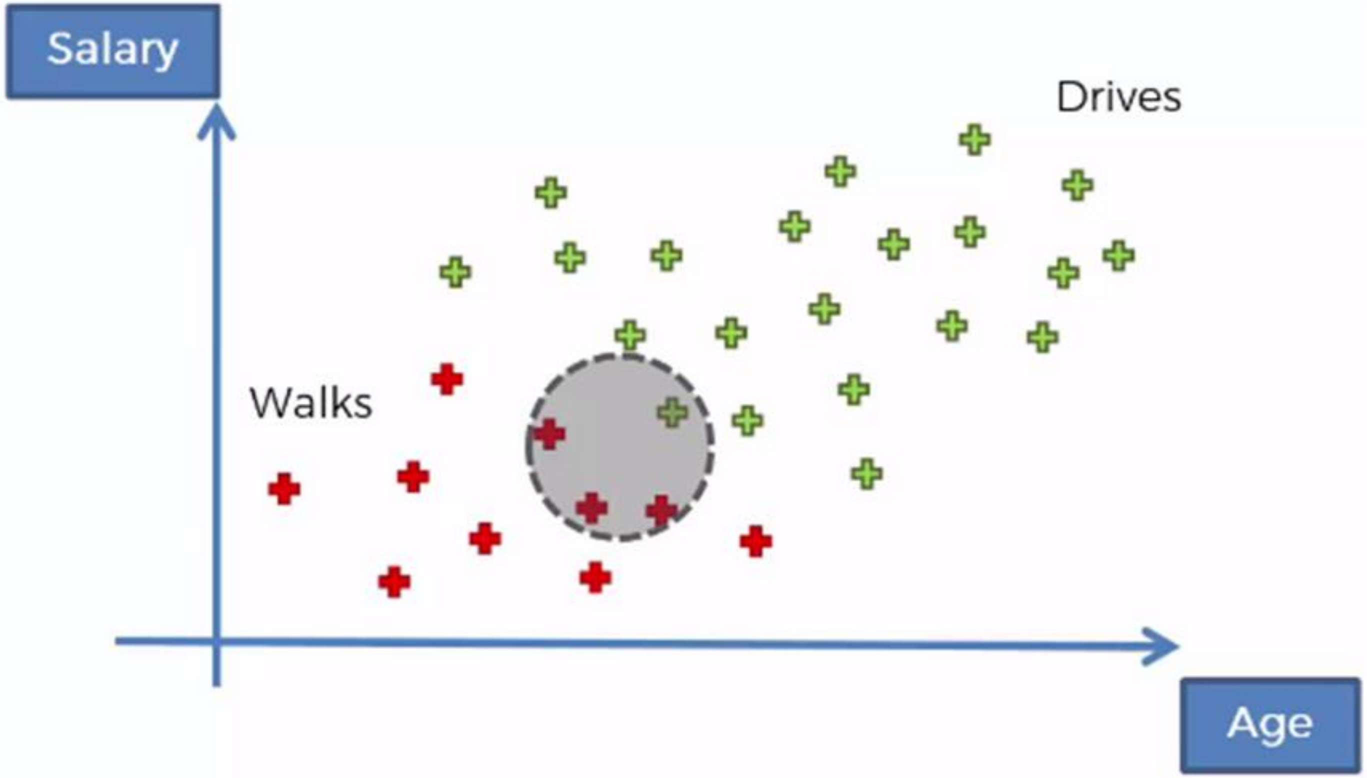


#1. P(Walks)

$$P(Walks) = \frac{Number\ of\ Walkers}{Total\ Observations}$$

$$P(Walks) = \frac{10}{30}$$

Naïve Bayes: Step 1

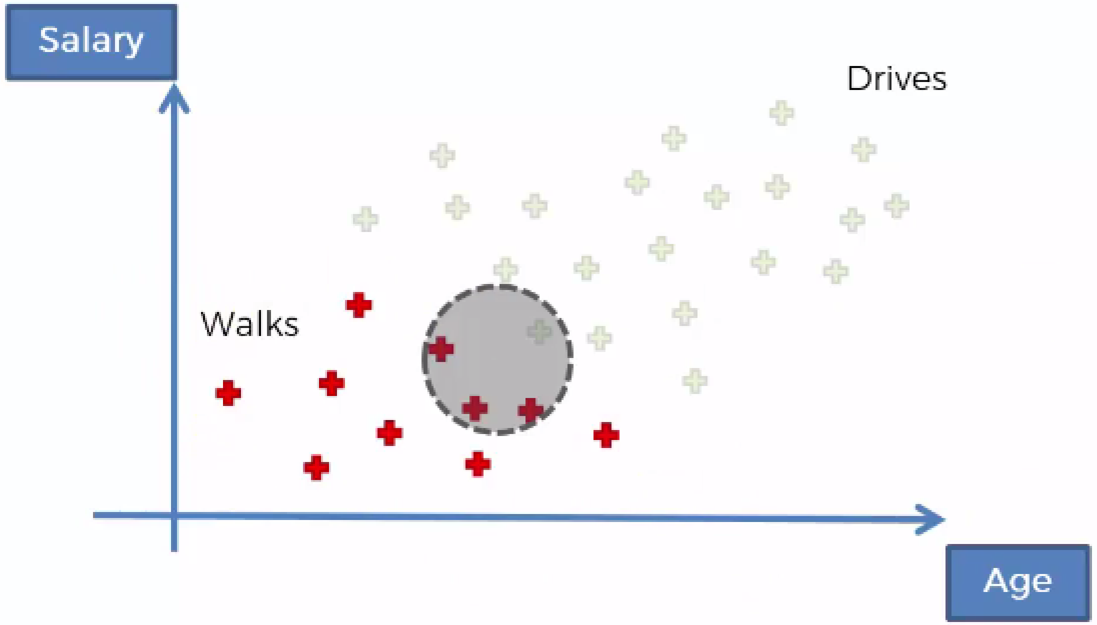


#2. P(X)

$$P(X) = \frac{Number\ of\ Similar\ Observat}{Total\ Observations}$$

$$P(X) = \frac{4}{30}$$

Naïve Bayes: Step 1

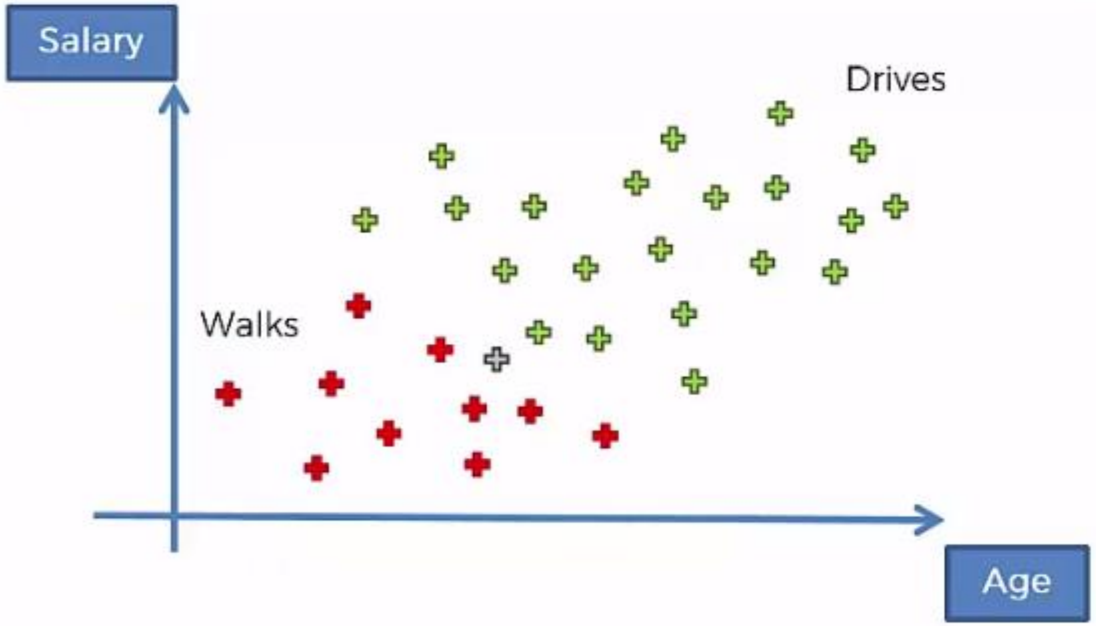


#3. P(X|Walks)

$$P(X|Walks) = \frac{\text{Number of Similar Observations Among those who Walk}}{\text{Total number of Walkers}}$$

$$P(X|Walks) = \frac{3}{10}$$

Naïve Bayes: Step 2

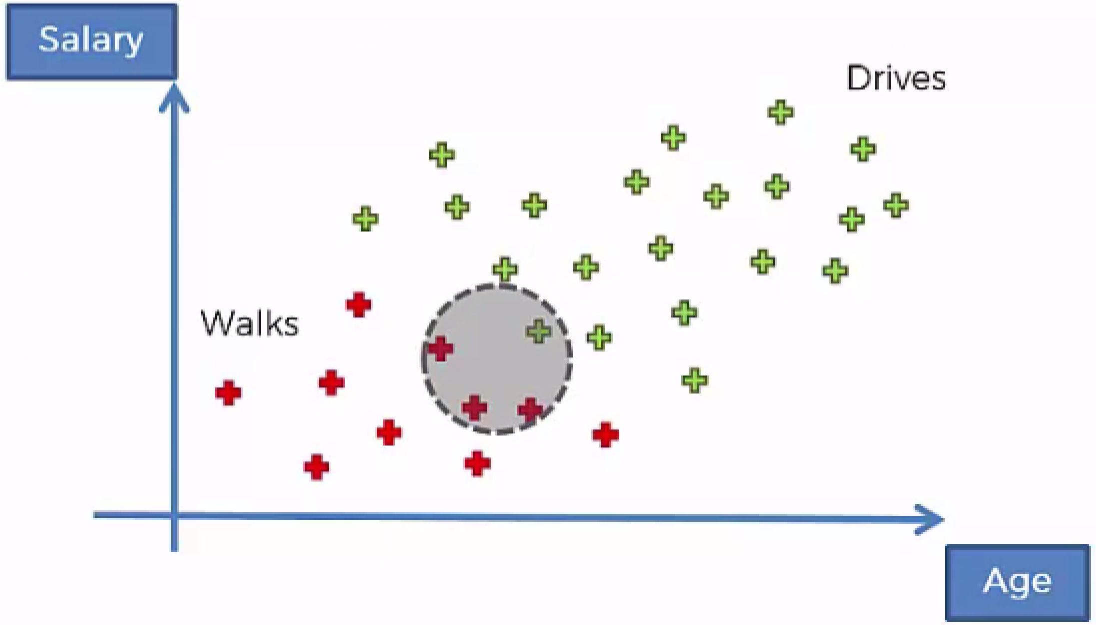


#1. P(Drives)

$$P(Drives) = \frac{Number\ of\ Drivers}{Total\ Observations}$$

$$P(Drives) = \frac{20}{30}$$

Naïve Bayes: Step 2



#2. P(X)

$$P(X) = \frac{Number\ of\ Similar\ Observations}{Total\ Observations}$$

$$P(X) = \frac{4}{30}$$

Step 3

$$P(Walks|X) \ v.s. \ P(Drives|X)$$

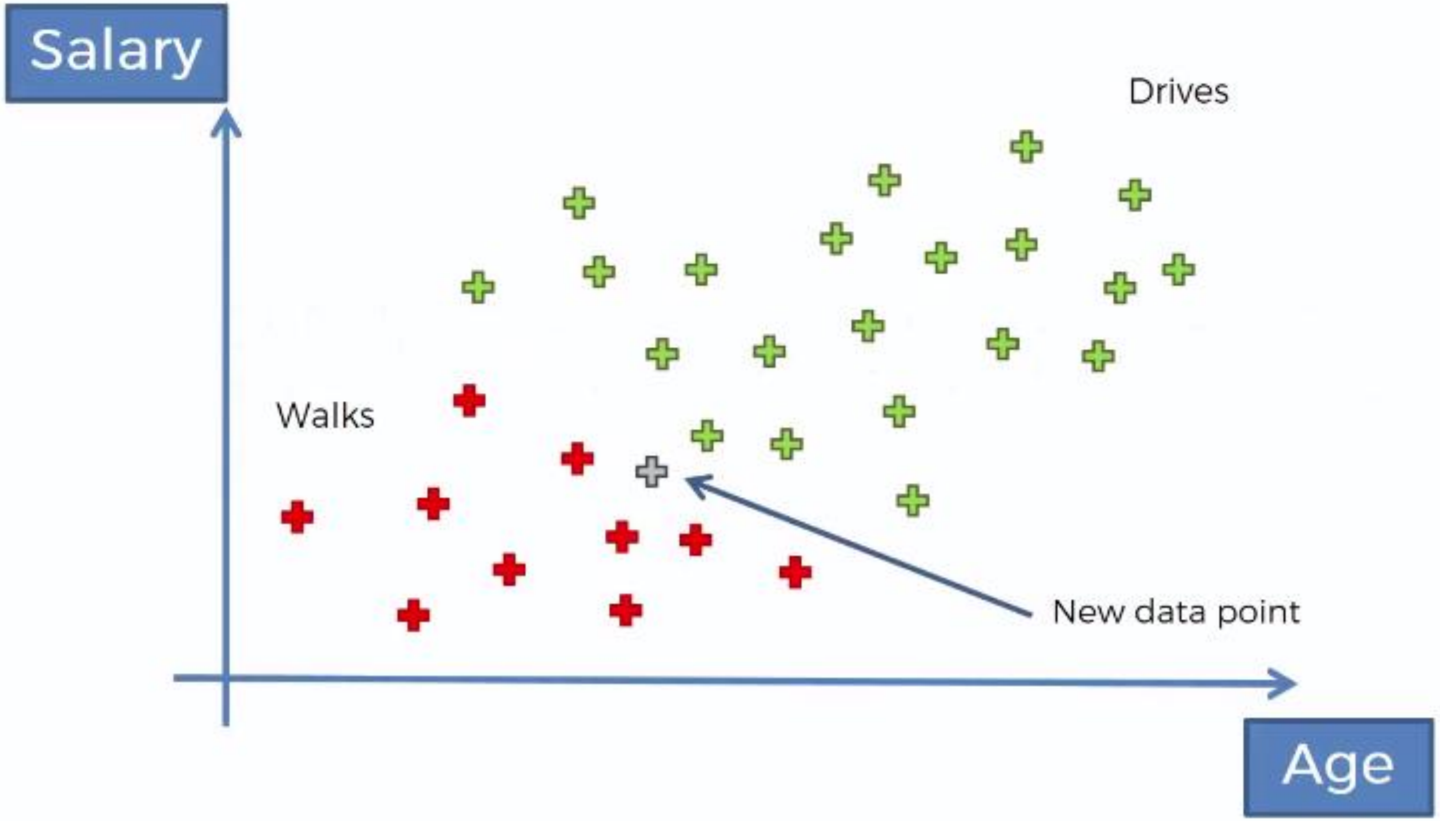
Step 3

$$0.75 > 0.25$$

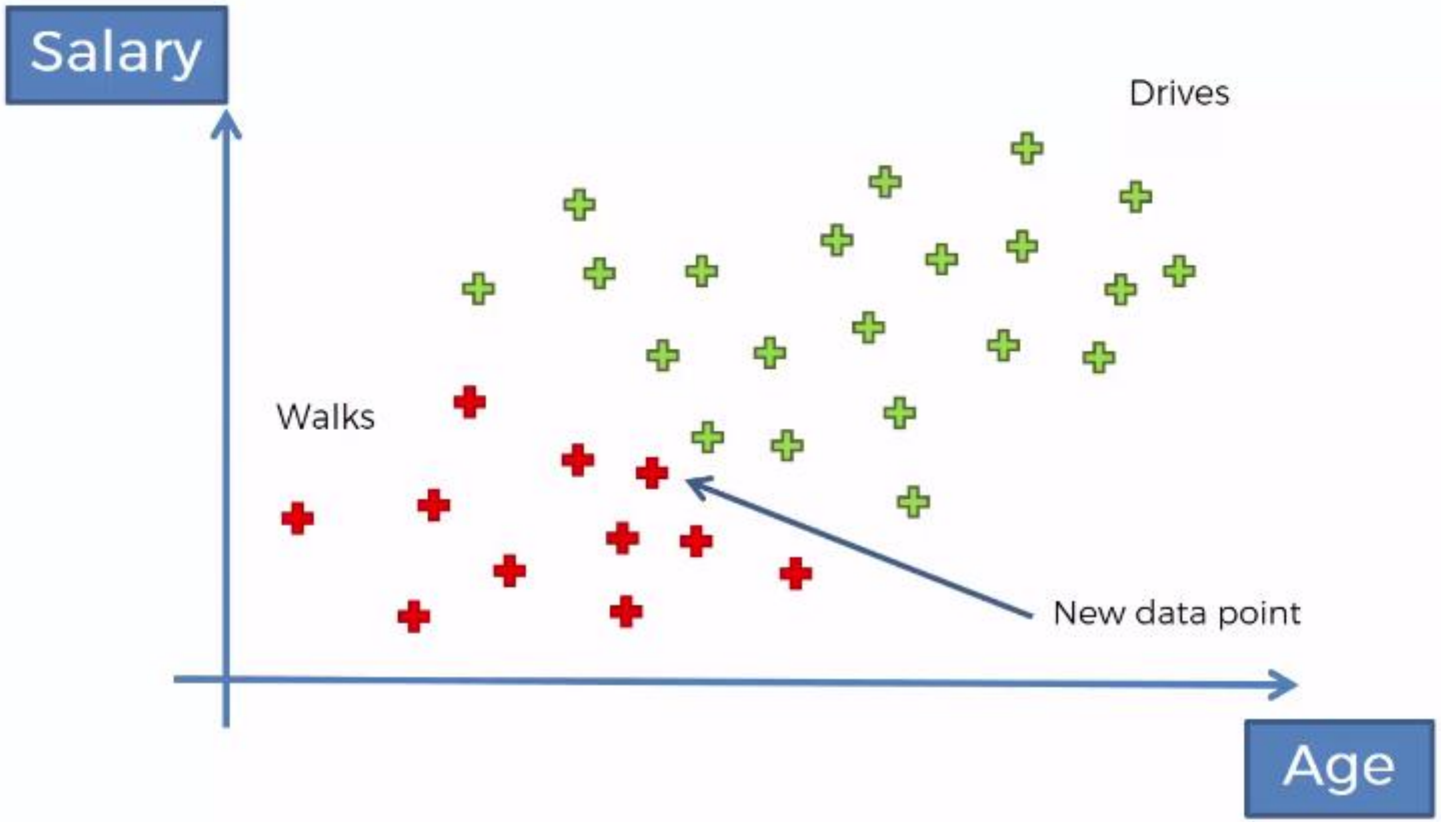
Step 3

$$P(Walks|X) > P(Drives|X)$$

Naïve Bayes



Naïve Bayes



Thank you