Naive Bayes Classifier

Tushar B. Kute,
http://tusharkute.com

Joint Probability

- The joint probability is the probability of two (or more) simultaneous events, often described in terms of events A and B from two dependent random variables, e.g. X and Y.
- The joint probability is often summarized as just the outcomes, e.g. A and B.
  - Joint Probability: Probability of two (or more) simultaneous events, e.g. P(A and B) or P(A, B).

Conditional Probability

- The conditional probability is the probability of one event given the occurrence of another event, often described in terms of events A and B from two dependent random variables e.g. X and Y.
  - Conditional Probability: Probability of one (or more) event given the occurrence of another event, e.g. P(A given B) or P(A | B).

Summary

- The joint probability can be calculated using the conditional probability; for example:
  - P(A, B) = P(A | B) * P(B)
- This is called the product rule. Importantly, the joint probability is symmetrical, meaning that:
  - P(A, B) = P(B, A)
- The conditional probability can be calculated using the joint probability; for example:
  - P(A | B) = P(A, B) / P(B)
- The conditional probability is not symmetrical; for example:
  - P(A | B) != P(B | A)

Alternate way for conditional prob

- Specifically, one conditional probability can be calculated using the other conditional probability; for example:
  - $P(A|B) = P(B|A) * P(A) / P(B)$
- The reverse is also true; for example:
  - $P(B|A) = P(A|B) * P(B) / P(A)$
- This alternate approach of calculating the conditional probability is useful either when the joint probability is challenging to calculate (which is most of the time), or when the reverse conditional probability is available or easy to calculate.

Bayes Theorem

- Bayes Theorem: Principled way of calculating a conditional probability without the joint probability. It is often the case that we do not have access to the denominator directly, e.g. P(B).
- We can calculate it an alternative way; for example:
  - P(B) = P(B|A) * P(A) + P(B|not A) * P(not A)
- This gives a formulation of Bayes Theorem that we can use that uses the alternate calculation of P(B), described below:
  - P(A|B) = P(B|A) * P(A) / P(B|A) * P(A) + P(B|not A) * P(not A)

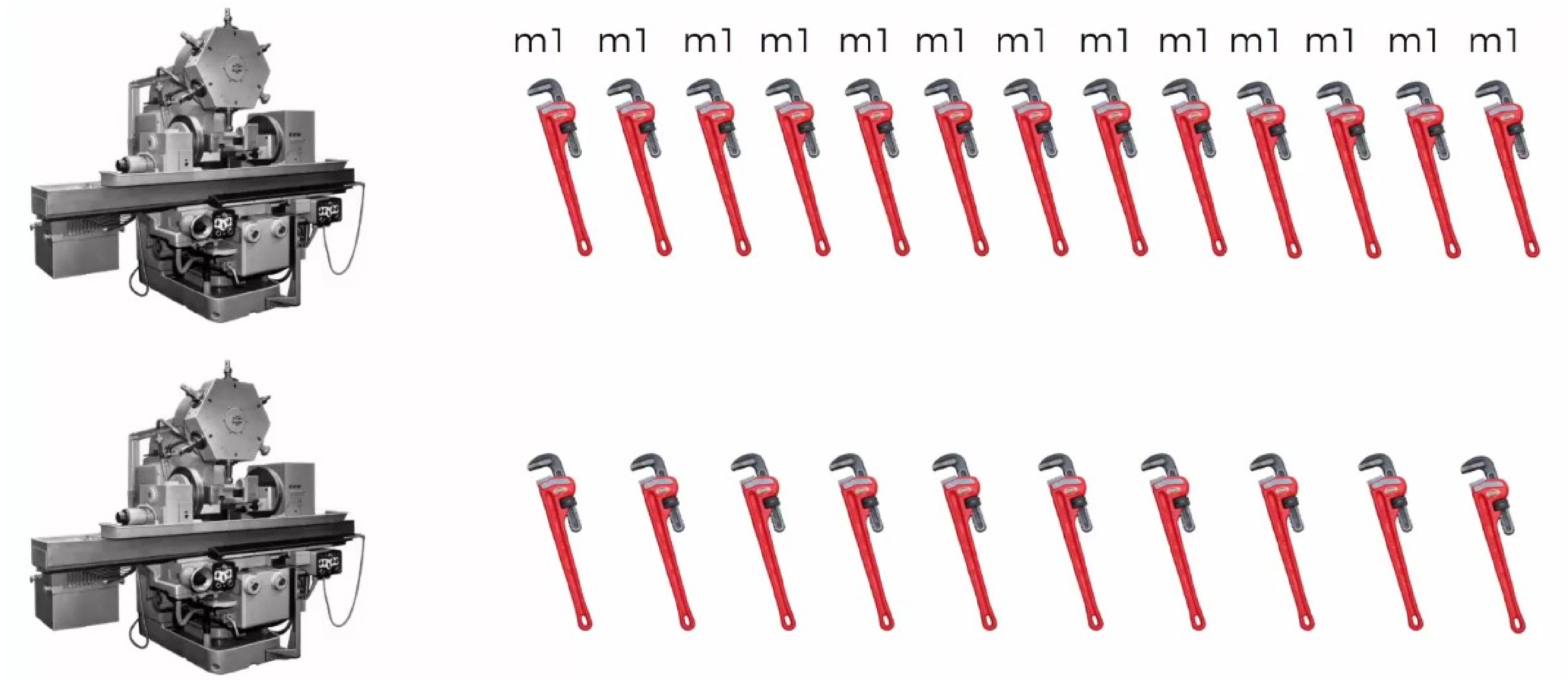
Bayes Theorem

- Firstly, in general, the result P(A|B) is referred to as the posterior probability and P(A) is referred to as the prior probability.
  - P(A|B): Posterior probability.
  - P(A): Prior probability.
- Sometimes P(B|A) is referred to as the likelihood and P(B) is referred to as the evidence.
  - P(B|A): Likelihood.
  - P(B): Evidence.
- This allows Bayes Theorem to be restated as:
  - Posterior = Likelihood * Prior / Evidence

Naive Bayes Classifier

- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.
- It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes Theorem

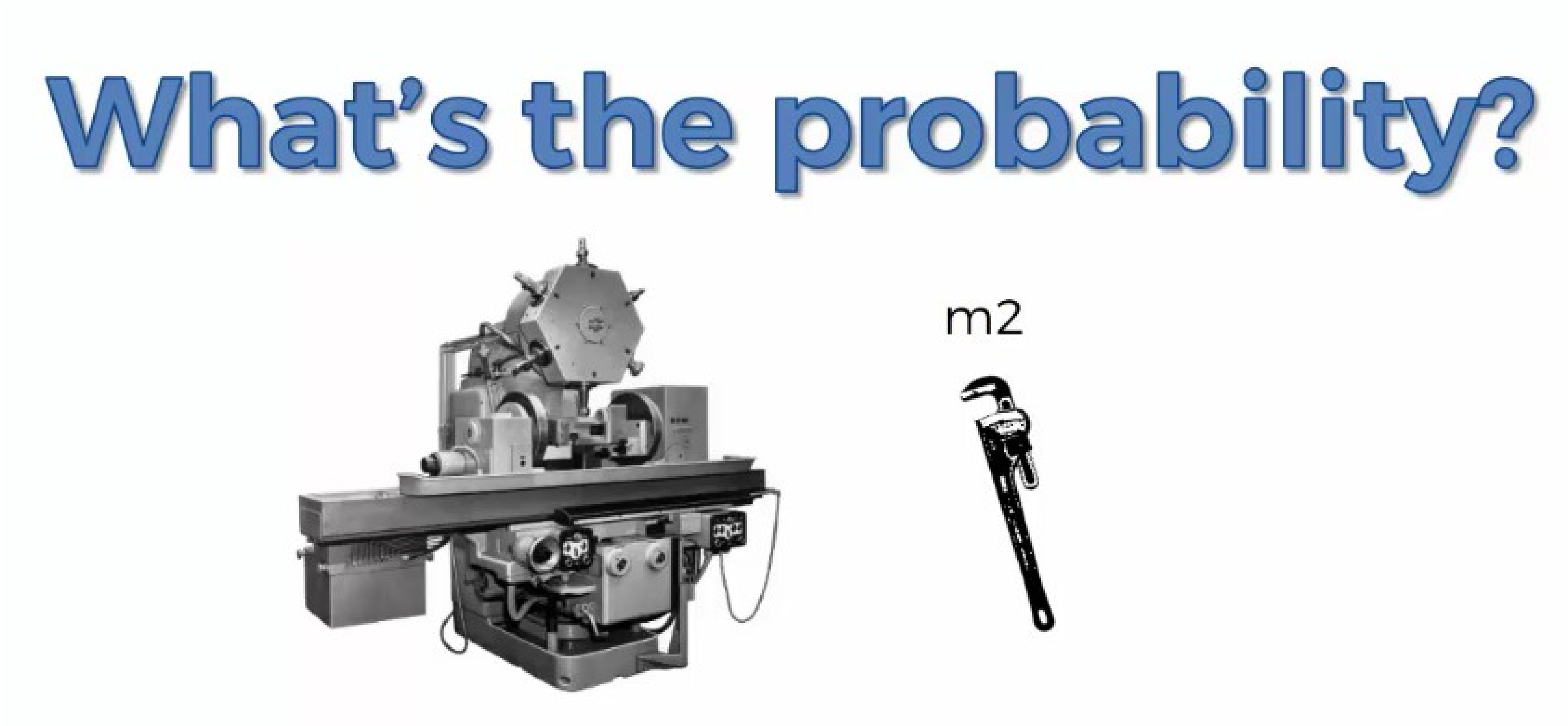


Example Reference: Super Data Science

Bayes Theorem

Defective Spanners

Bayes Theorem



Bayes Theorem

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

Bayes Theorem

Mach1: 30 wrenches / hr -> P(Mach1) = 30/50 = 0.6
Mach2: 20 wrenches / hr -> 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 -> P(Mach1 | Defect) = 50%
And 50% came from mach2 -> P(Mach2 | Defect) = 50%

Question:
What is the probability that a part produced by mach2 is defective = ? -> P(Defect | Mach2) = ?

Bayes Theorem

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

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

That’s intuitive

$$P(Defect \mid Mach2) = \frac{P(Mach2 \mid Defect) * P(Defect)}{P(Mach2)} = 1.25\%$$

<u>Let’s look at an example:</u>

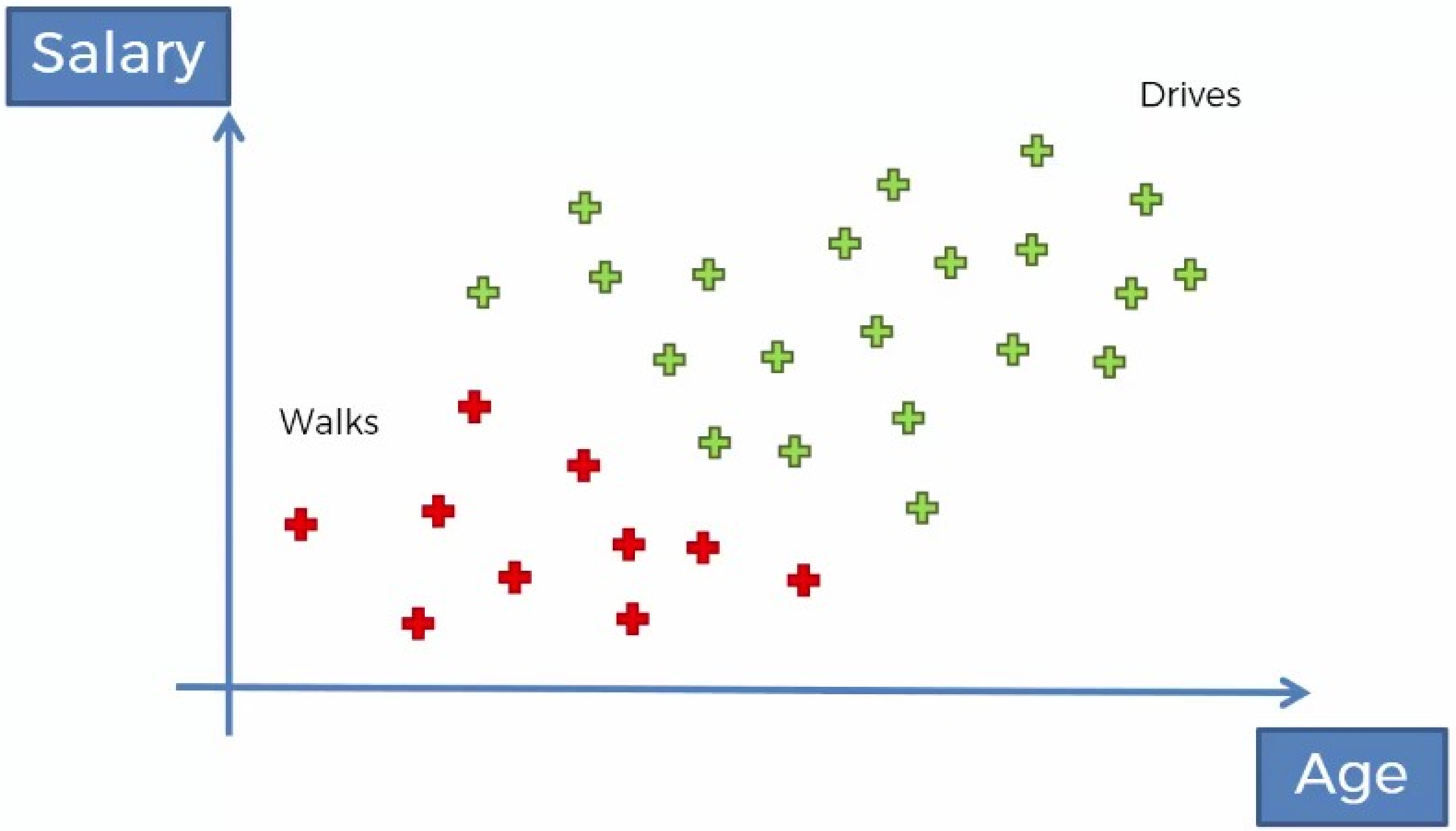
- 1000 wrenches
- 400 came from Mach2
- 1% have a defect = 10
- of them 50% came from Mach2 = 5
- % defective parts from Mach2 = 5/400 = 1.25%

Exercise

Quick exercise:

P(Defect | Mach1) = ?

Example:

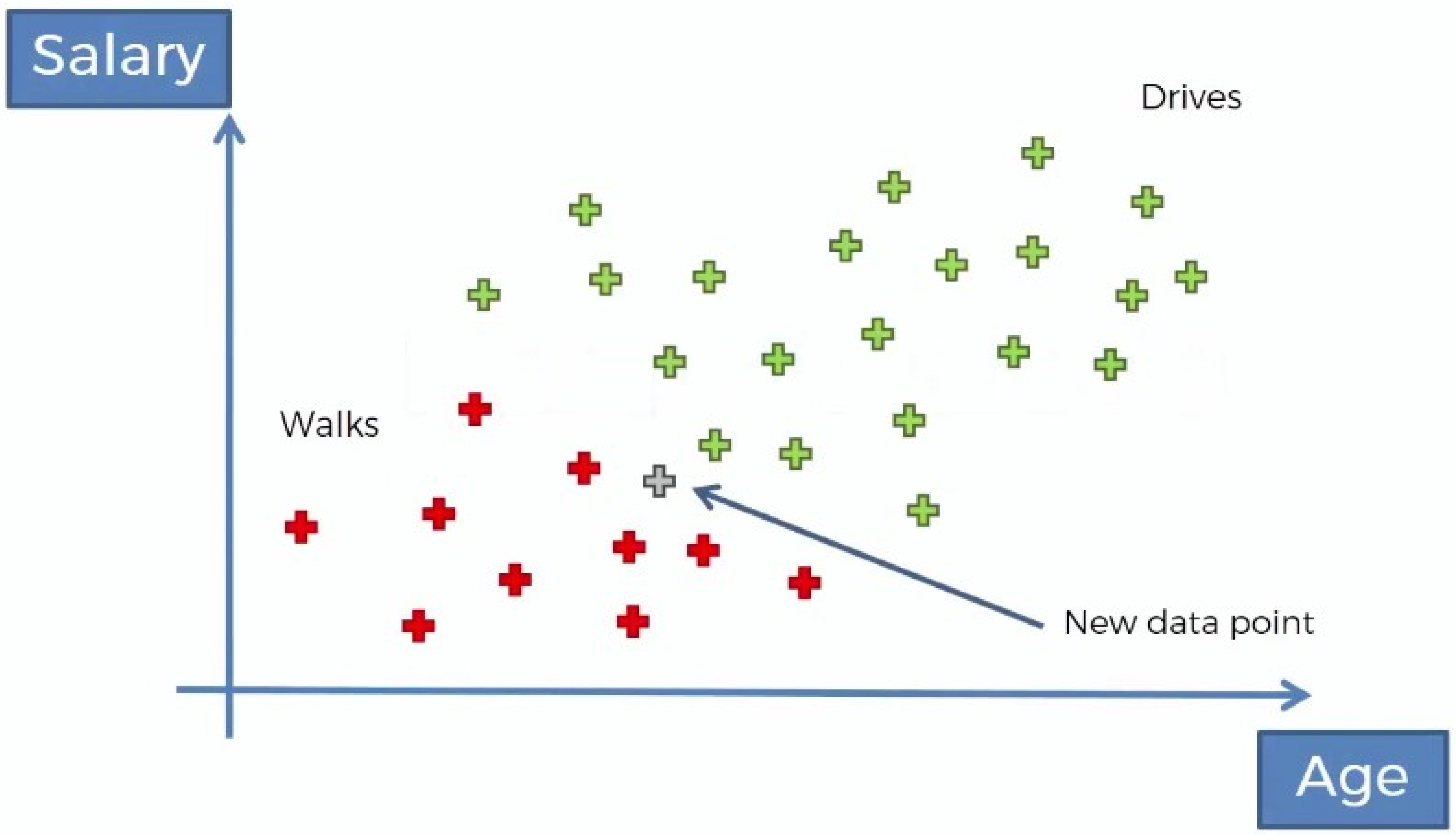


Step-1

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

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

Step-1



Step-1

#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

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

#2 Marginal Likelihood

Step-2

#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

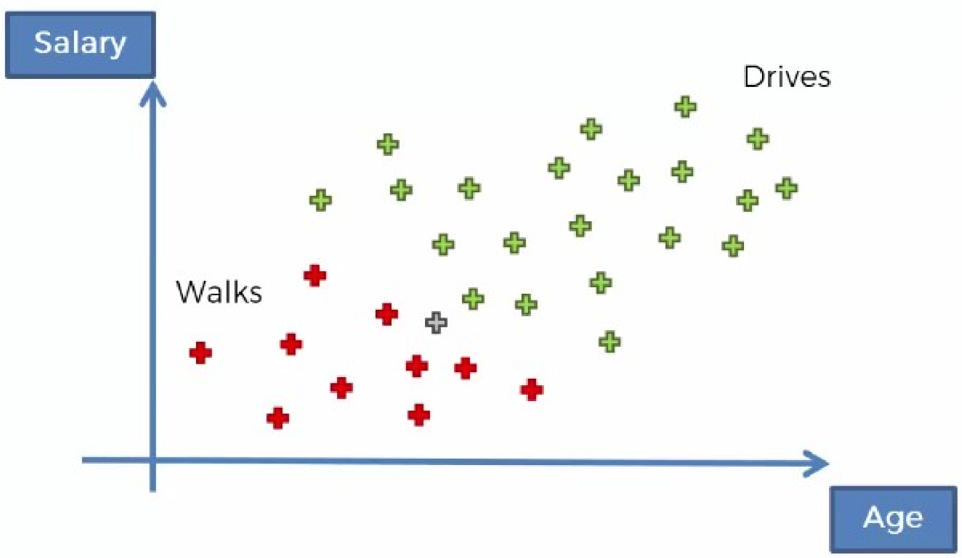
$$P(Drives|X) = \frac{P(X|Drives) * P(Drives)}{P(X)}$$

#2 Marginal Likelihood

Step-3

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

Naive Bayes – Step-1

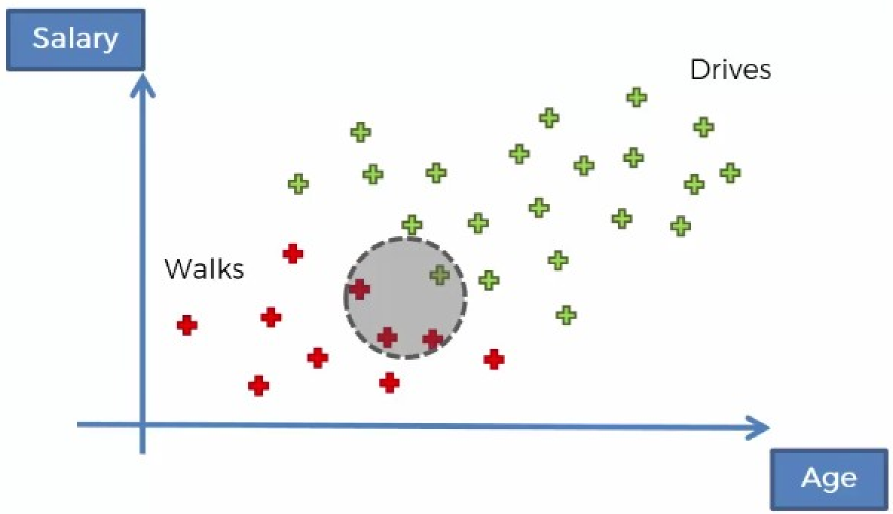


#1. P(Walks)

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

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

Naive Bayes – Step-2

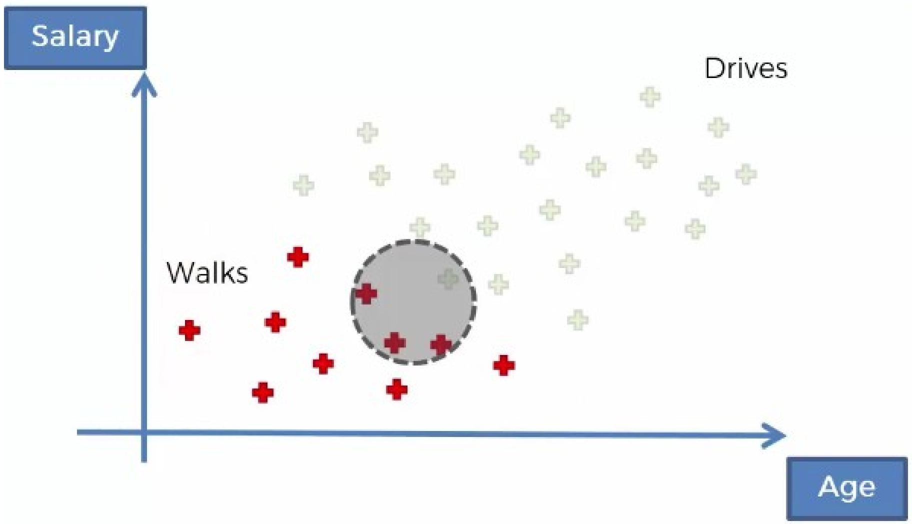


#2. P(X)

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

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

Naive Bayes – Step-3



#3. P(X|Walks)

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

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

Combining altogether

#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

$$P(Walks|X) = \frac{\frac{3}{10} * \frac{10}{30}}{\frac{4}{30}} = 0.75$$

#2 Marginal Likelihood

Naive Bayes – Step-4

#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

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

#2 Marginal Likelihood

Naive Bayes – Step-5

#4 Posterior Probability

#3 Likelihood

#1 Prior Probability

$$P(Drives|X) = \frac{\frac{1}{20} * \frac{20}{30}}{\frac{4}{30}} = 0.25$$

#2 Marginal Likelihood

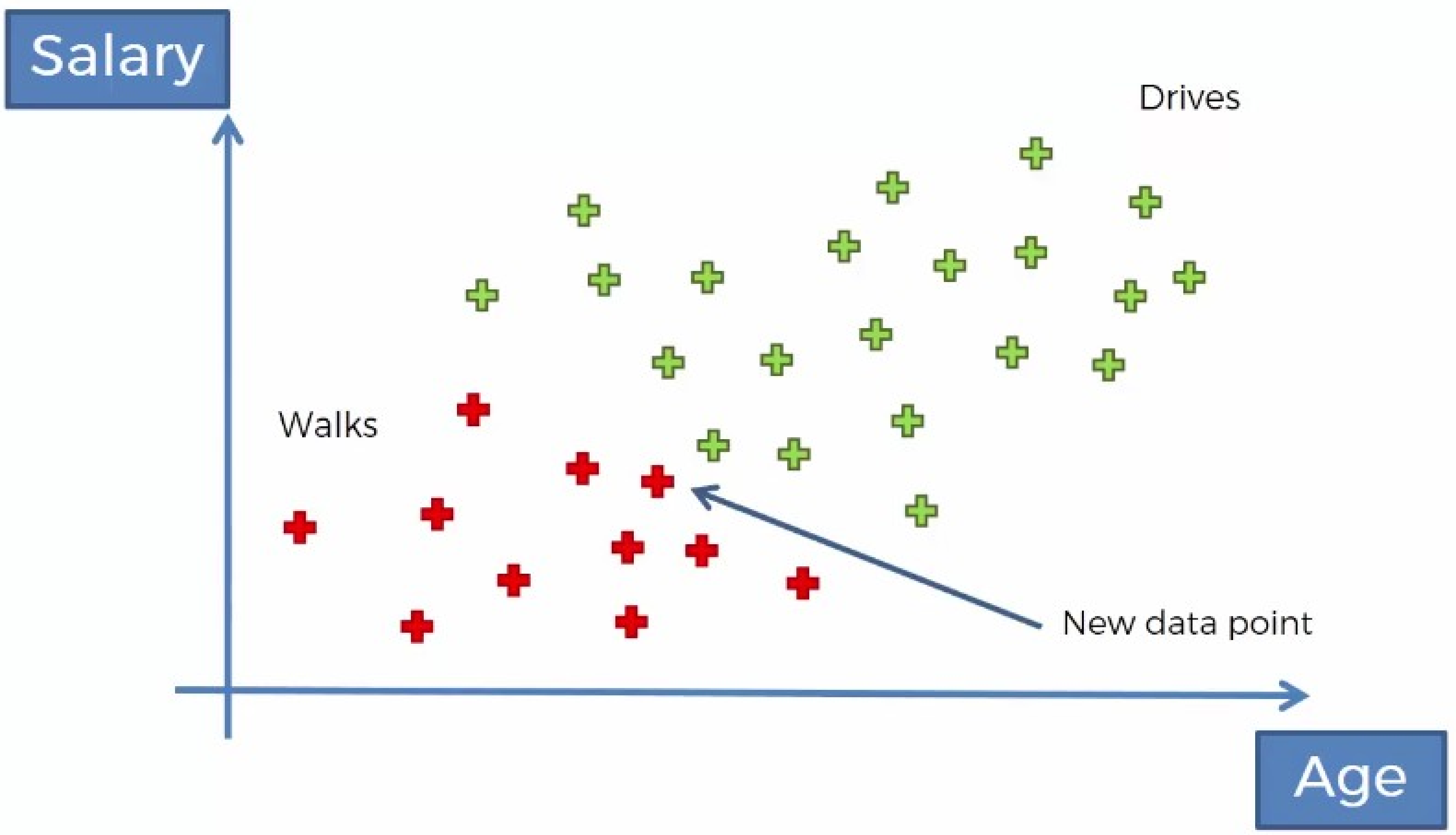
tusharkute.com

Types of model

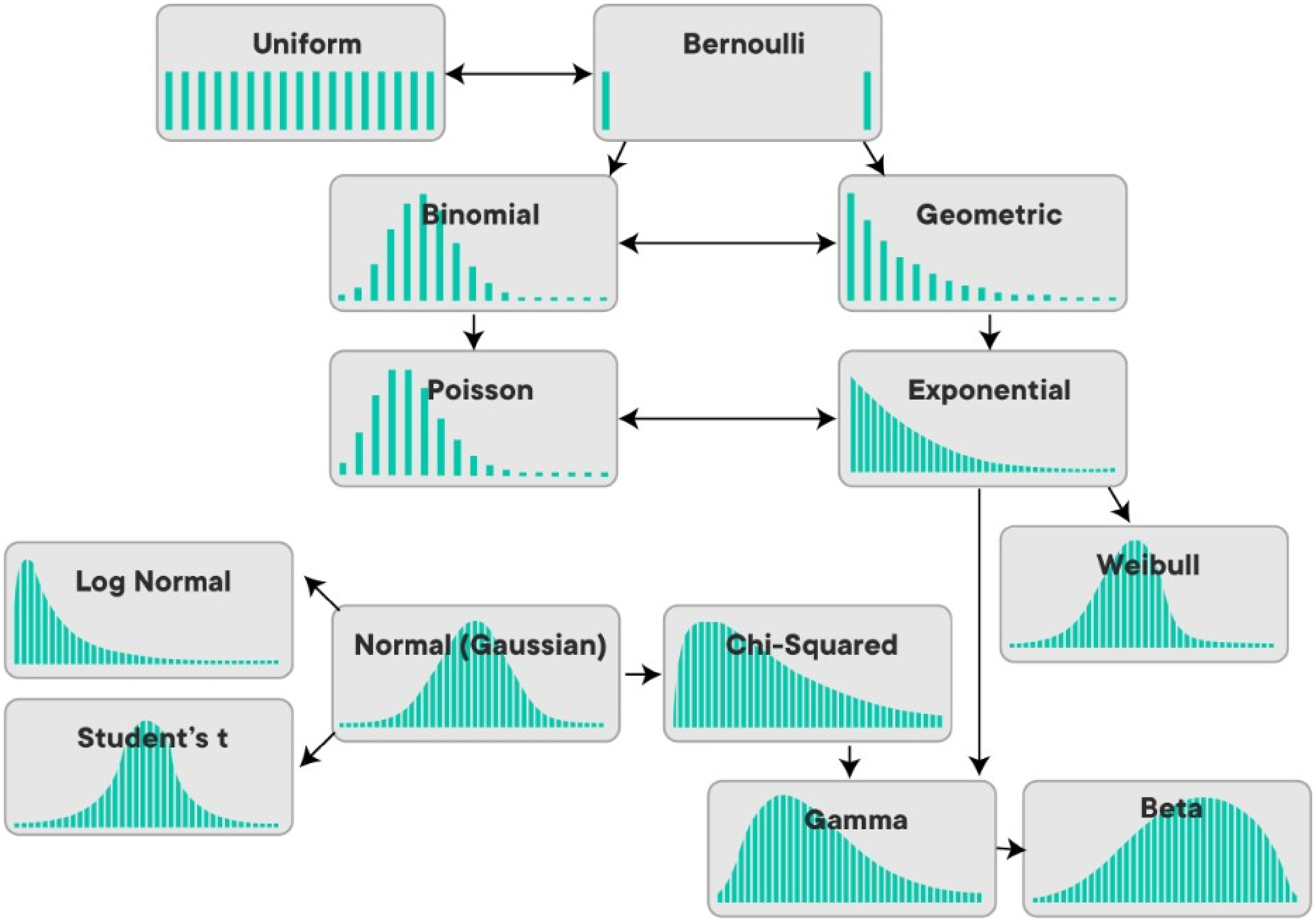
$$P(Walks|X)\ v.s.\ P(Drives|X)$$

$$0.75\ v.s.\ 0.25$$

Final Classification



Probability Distribution



Types of Naive Bayes Classifier

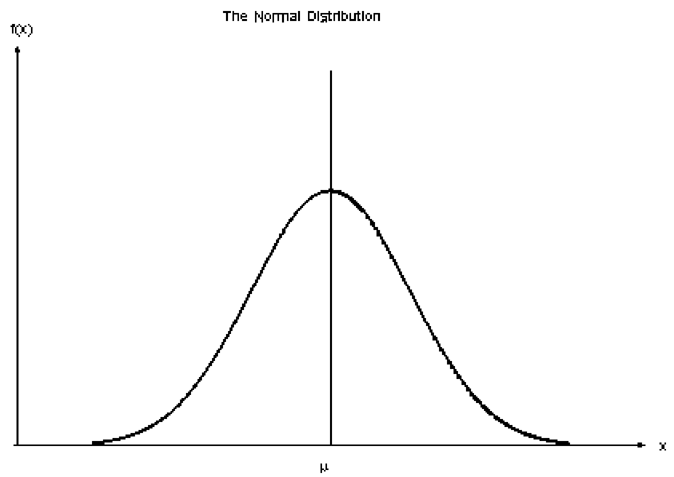
- Multinomial Naive Bayes:
  - This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc.
  - The features/predictors used by the classifier are the frequency of the words present in the document.

Types of Naive Bayes Classifier

- Bernoulli Naive Bayes:
  - This is similar to the multinomial naive bayes but the predictors are boolean variables.
  - The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

Types of Naive Bayes Classifier

- Gaussian Naive Bayes:
  - When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.



Advantages

- When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.
- Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.
- Naive Bayes is also easy to implement.

Disadvantages

- Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.
- If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com