Naive Bayes

Classification Techniques

- A number of classification techniques are known, which can be broadly classified into the following categories:

1. Statistical-Based Methods
   - Regression
   - Bayesian Classifier
   -
2. Distance-Based Classification
   - K-Nearest Neighbours
3. Decision Tree-Based Classification
   - ID3, C 4.5, CART
5. Classification using Machine Learning (SVM)
6. Classification using Neural Network (ANN)

Naive Bayes

- Naive Bayes is a popular probabilistic machine learning algorithm for classification tasks.
- It's based on Bayes' theorem, which relates the probability of a hypothesis given evidence to the probability of the evidence given the hypothesis.
- Naive Bayes assumes that the features are conditionally independent given the class label, allowing probabilities to be multiplied together.
- Several variants of Naive Bayes exist, such as Gaussian Naive Bayes and Multinomial Naive Bayes, which model different probability distributions for feature likelihoods.

Naive Bayes

- Naive Bayes is known for its simplicity, efficiency, and ability to handle high-dimensional data, making it useful for tasks like text classification, spam filtering, and sentiment analysis.

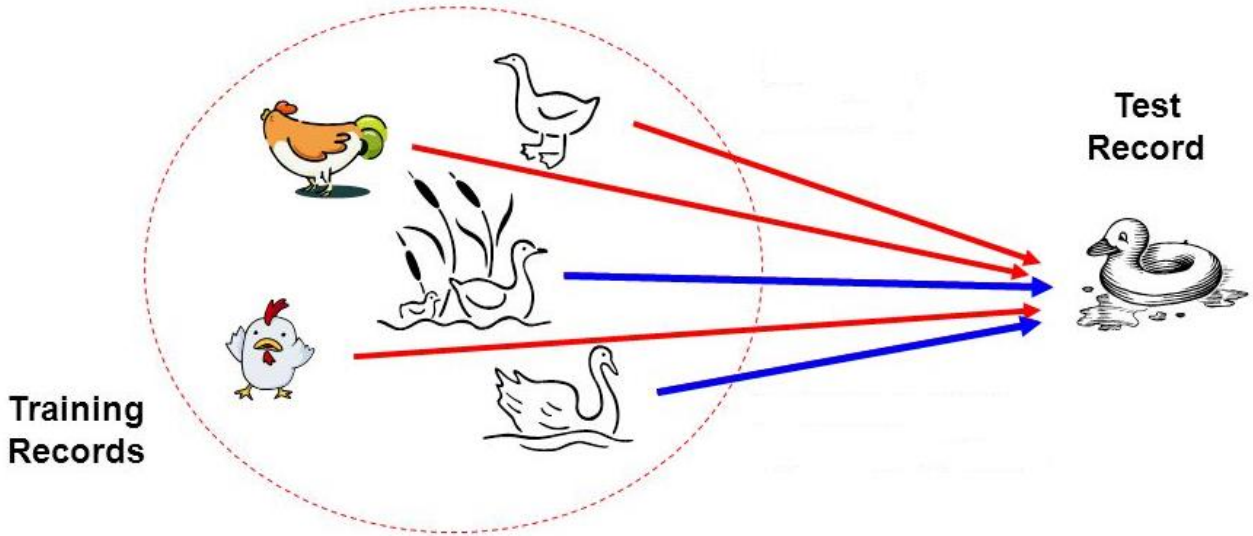
Naive Bayes

- Naive Bayes is a type of classification algorithm that takes in data with multiple features and predicts which class that data belongs to.
- It calculates the probability of each class given the data's features by multiplying the probability of each feature given that class.
- To do this, Naive Bayes assumes that each feature is independent of the others given the class, which is called the naive assumption.
- Naive Bayes calculates the probability of each feature given each class in advance, which is called the prior probability".
- Once the probabilities are calculated, Naive Bayes selects the class with the highest probability as the predicted class for the input data.

Bayesian Classifier

Bayesian Classifier

- Principle
  - If it walks like a duck, quacks like a duck, then it is probably a duck

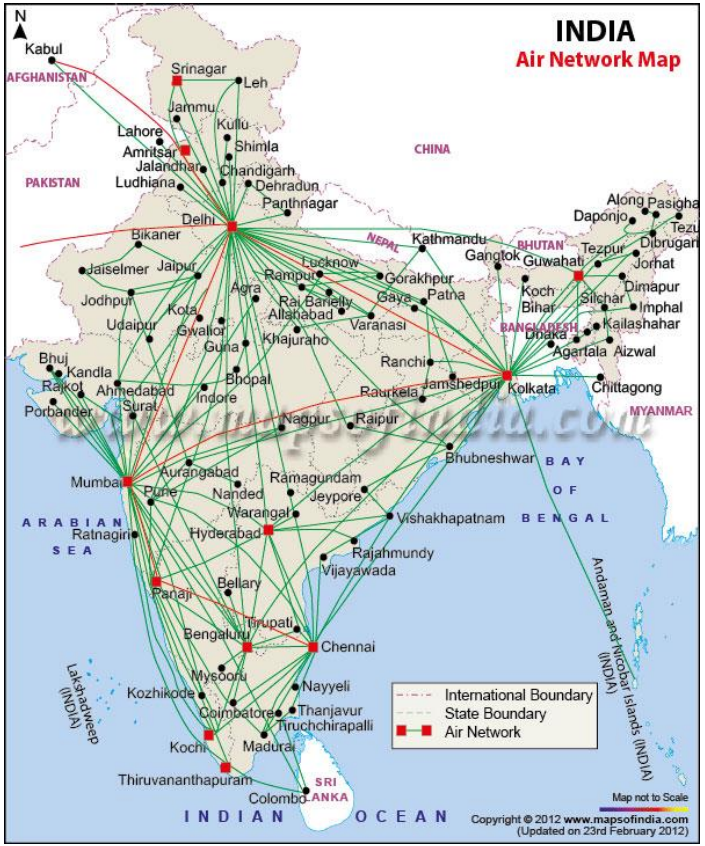


Bayesian Classifier

- A statistical classifier
  - Performs *probabilistic prediction, i.e.,* predicts class membership probabilities
- Foundation
  - Based on Bayes' Theorem.
- Assumptions
  1. The classes are mutually exclusive and exhaustive.
  2. The attributes are independent given the class.
- Called "Naïve" classifier because of these assumptions
  - Empirically proven to be useful.
  - Scales very well.

Example: Bayesian Classification

- **Example 7.2:** Air Traffic Data
  - Let us consider a set observation recorded in a database
    - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	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

Cond. to next slide…

Air-Traffic Data

Cond. from previous slide…

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

- In this dataset, there are four attributes

  A = [Day, Season, Fog, Rain]

  with 20 tuples.
- The categories of classes are:

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

- Given this is the knowledge of data and classes, we are 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.

Bayesian Classifier

- In many applications, the relationship between the attributes set and the class variable is non-deterministic.
  - In other words, a test cannot be classified to a class label with certainty.
  - In such a situation, the classification can be achieved probabilistically.
- The Bayesian classifier is an approach for modelling probabilistic relationships between the attribute set and the class variable.
- More precisely, Bayesian classifier use Bayes' Theorem of Probability for classification.
- Before going to discuss the Bayesian classifier, we should have a quick look at the Theory of Probability and then Bayes' Theorem.

Bayes’ Theorem of Probability

Simple Probability

Definition 7.2: Simple Probability

If there are n elementary events associated with a random experiment and m of n of them are favorable to an event A, then the probability of happening or occurrence of A is

$$P(A) = \frac{m}{n}$$

Simple Probability

- Suppose, A and B are any two events and *P(A)*, *P(B)* denote the probabilities that the events *A* and *B* will occur, respectively.

- **Mutually Exclusive Events:**
  - Two events are mutually exclusive, if the occurrence of one precludes the occurrence of the other.

Example: Tossing a coin (two events)

Tossing a ludo cube (Six events)

Can you give an example, so that two events are not mutually exclusive?

Hint: Tossing two identical coins, Weather (sunny, foggy, warm)

Simple Probability

- **Independent events:** Two events are independent if occurrences of one does not alter the occurrence of other.

Example: Tossing both coin and ludo cube together.

(How many events are here?)

Joint Probability

Definition 7.3: Joint Probability

If *P(A)* and *P(B)* are the probability of two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, then $P(A \cap B) = 0$
If A and B are independent events, then $P(A \cap B) = P(A).P(B)$

Thus, for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

Conditional Probability

Definition 7.2: **Conditional Probability**

If events are dependent, then their probability is expressed by conditional probability. The probability that A occurs given that B is denoted by $P(A|B)$.

Suppose, A and B are two events associated with a random experiment. The probability of A under the condition that B has already occurred and $P(B) \neq 0$ is given by

$$P(A|B) = \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B}$$

$$= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B}$$

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

Conditional Probability

Corollary 7.1: **Conditional Probability**

$$P(A \cap B) = P(A).P(B|A), \quad if\ P(A) \neq 0$$

or $$P(A \cap B) = P(B).P(A|B), \quad if\ P(B) \neq 0$$

For three events *A*, *B* and *C*

$$P(A \cap B \cap C) = P(A).P(B).P(C|A \cap B)$$

For *n* events $A_1, A_2, \ldots, A_n$ and if all events are mutually independent to each other

$$P(A_1 \cap A_2 \cap \ldots\ldots\ldots\ldots \cap A_n) = P(A_1).P(A_2) \ldots\ldots\ldots\ldots P(A_n)$$

Note:

$P(A|B) = 0$ if events are **mutually exclusive**

$P(A|B) = P(A)$ if *A* and *B* are **independent**

$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ otherwise,

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

Conditional Probability

- Generalization of Conditional Probability:

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

$$= \frac{P(B|A) \cdot P(A)}{P(B)} \qquad \because P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$$

By the law of total probability : P(B) = $P[(B \cap A) \cup (B \cap \bar{A})]$, where $\bar{A}$ denotes the compliment of event A. Thus,

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P[(B \cap A) \cup (B \cap \bar{A})]}$$

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

Conditional Probability

In general,

$$P(A|D) = \frac{P(A) \cdot P(D|A)}{P(A) \cdot P(D|A) + P(B) \cdot P(D|B) + P(C) \cdot P(D \mid C)}$$

Total Probability

Definition 7.3: Total Probability

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 *or* E_2 *or* $\ldots \ldots E_n$, then

$$P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2) + \cdots \ldots \ldots \ldots . + P(E_n).P(A|E_n)$$

Bayes' Theorem

Theorem 7.4: Bayes' Theorem

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 *or* E_2 *or* $\ldots \ldots E_n$, then

$$P(E_i|A) = \frac{P(E_i).P(A|E_i)}{\sum_{i=1}^{n} P(E_i).P(A|E_i)}$$

Naïve Bayesian Classifier

- Suppose, *Y* is a class variable and *X* = $\{X_1, X_2, \ldots\ldots, X_n\}$ is a set of attributes, with instance of *Y*.

INPUT (X)	CLASS(Y)
… … …	
… … …	…
$x_1, x_2, \ldots, x_n$	y_i
… … …	…

- The classification problem, then can be expressed as the class-conditional probability

$$P\big(Y = y_i | (X_1 = x_1) \text{ AND } (X_2 = x_2) \text{ AND } \ldots\ldots (X_n = x_n)\big)$$

Naïve Bayesian Classifier

- Naïve Bayesian classifier calculate this posterior probability using Bayes' theorem, which is as follows.

- From Bayes' theorem on conditional probability, we have

$$P(Y|X) = \frac{P(X|Y)\cdot P(Y)}{P(X)}$$

$$= \frac{P(X|Y)\cdot P(Y)}{P(X|Y=y_1)\cdot P(Y=y_1) + \cdots + P(X|Y=y_k)\cdot P(Y=y_k)}$$

where,

$$P(X) = \sum_{i=1}^{k} P(X|Y=y_i)\cdot P(Y=y_i)$$

Note:

- $P(X)$ is called the evidence (also the total probability) and it is a constant.
- The probability *P(Y|X)* (also called class conditional probability) is therefore proportional to *P(X|Y)*· $P(Y)$.
- Thus, *P(Y|X)* can be taken as a measure of *Y* given that *X*.

$$P(Y|X) \approx P(X|Y)\cdot P(Y)$$

Naïve Bayesian Classifier

- Suppose, for a given instance of *X* (say *x* = ($X_1 = x_1$) and ….. ($X_n = x_n$)).
- There are any two class conditional probabilities namely *P(Y=* y_i *|X=x)* and *P(Y=* y_j *| X=x)*.
- If *P(Y=* y_i *| X=x) > P(Y=* y_j *| X=x)*, then we say that y_i is more stronger than y_j for the instance *X = x*.
- The strongest y_i is the classification for the instance *X = x*.

Naïve Bayesian Classifier

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	½ = 0.5	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	0/1 = 0
	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

Naïve Bayesian Classifier

		Class			
Attribute		On Time	Late	Very Late	Cancelled
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	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	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

Naïve Bayesian Classifier

Instance:

Week Day	Winter	High	Heavy	???

Case1: Class = On Time : $0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013$

Case2: Class = Late : $0.10 \times 0.50 \times 1.0 \times 0.50 \times 0.50 = 0.0125$

Case3: Class = Very Late : $0.15 \times 1.0 \times 0.67 \times 0.33 \times 0.67 = 0.0222$

Case4: Class = Cancelled : $0.05 \times 0.0 \times 0.0 \times 1.0 \times 1.0 = 0.0000$

Case3 is the strongest; Hence correct classification is **Very Late**

Naïve Bayesian Classifier

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
  - It relies on all attributes being categorical.
  - If the data is less, then it estimates poorly.

Naïve Bayesian Classifier

- Estimating the posterior probabilities for continuous attributes

Approach 2:

We can assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. A Gaussian distribution is usually chosen to represent the posterior probabilities for continuous attributes. A general form of Gaussian distribution will look like

$$P(x: \mu, \sigma^2) = \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

where, μ and σ^2 denote mean and variance, respectively.

Reference

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

DSamanta, IIT Kharagpur, Data Analytics resource