Bayesian Classifier

Machine Learning (CS 306)

Instructor: Dr. Moumita Roy

Teaching Assistants: Indrajit Kalita, Veronica Naosekpam

Email-ids:

moumita@iiitg.ac.in

veronica.naosekpam@iiitg.ac.in

indrajit.kalita@iiitg.ac.in

Mobile No: +91-8420489325 (only for emergency quires)

Reference: Dr. Debasis Samanta, IIT Kharagpur, Classification: Naïve Bayes' Classifier

Classification

- Email: Spam/not spam
- Land-cover: Water-cover/not water-cover
- Customer Behavior Prediction: Sad/Happy
- Tumor: Malignant/not Malignant

$Y \in \{0,1\}$

- coded as binary dependent variable (two-class problem)

  0 → Negative class

  1 → Positive class

Classification

- Classification is a form of data analysis to extract models describing important data classes.

- Essentially, it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes.

  - The term “mutually exhaustive and exclusive” simply means that each object must be assigned to precisely one class

    - That is, never to more than one and never to no class at all.

Simple example of classification

Example 8.1

- Teacher classify students as A, B, C, D and F based on their marks. The following is one simple classification rule:

$$\begin{array}{lcc} \text{Mark} \geq 90 & : & A \\ 90 > \text{Mark} \geq 80 & : & B \\ 80 > \text{Mark} \geq 70 & : & C \\ 70 > \text{Mark} \geq 60 & : & D \\ 60 > \text{Mark} & : & F \end{array}$$

Note:

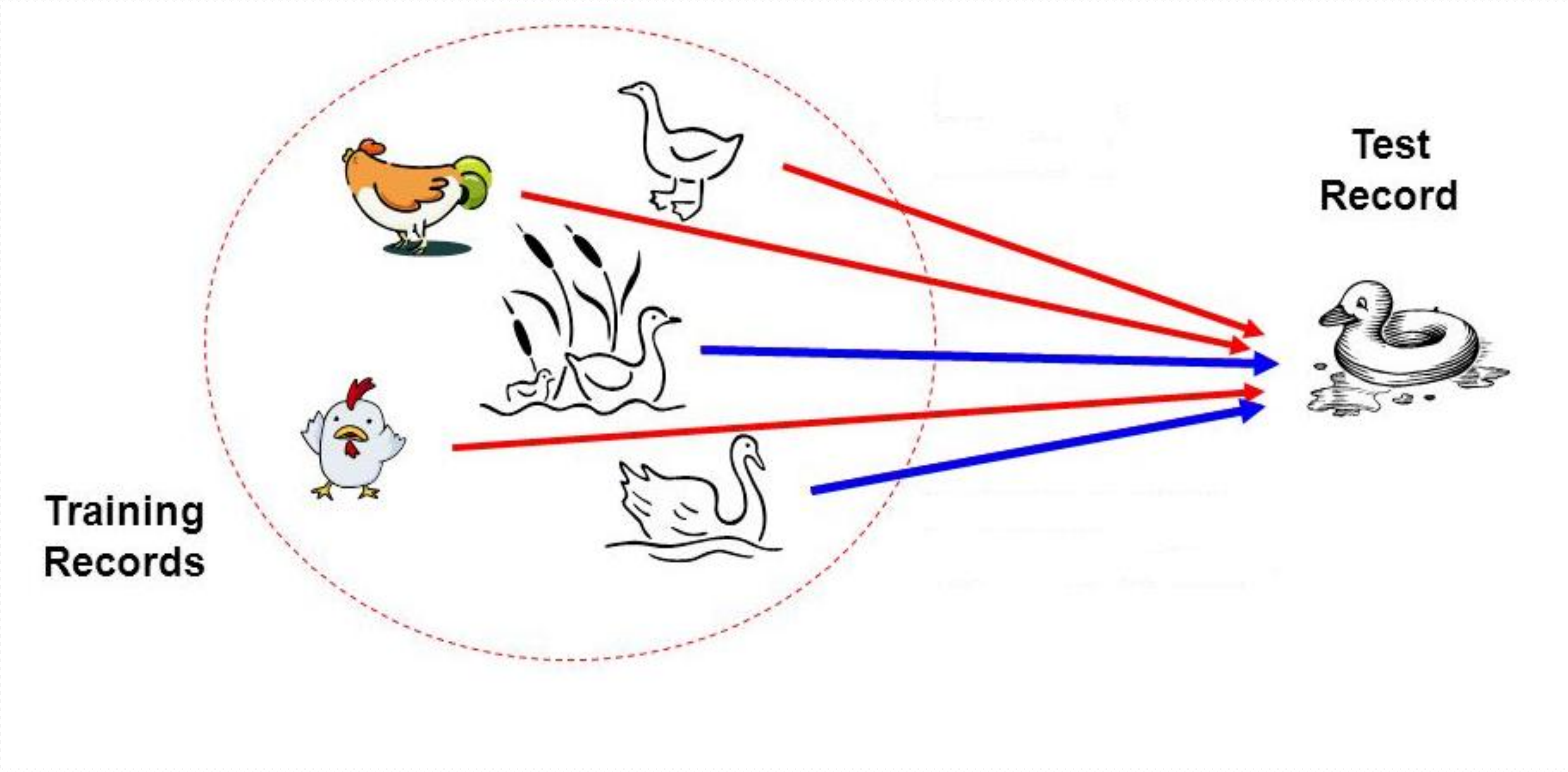
Here, we apply the above rule to a specific data (in this case a table of marks).

Supervised Learning

- Given a collection of data samples (*training set*)
  - Each sample contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes (namely features).
- Goal: Previously unseen samples should be assigned a class as accurately as possible.
    - Satisfy the property of “mutually exclusive and exhaustive”

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/features are independent given the class.
- Called "Naïve" classifier because of these assumptions.

- Before going to discuss the Bayesian classifier, we should have a quick look at Bayes' Theorem.

Bayes' Theorem

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

- P(A|B): conditional probability; the probability of event A occurring given that B is true. It is also called the posterior probability of A given B.

- P(B|A): conditional probability; the probability of event B occurring given that A is true. It can also be interpreted as the likelihood of A given fixed B (i.e. likelihood of an outcome occurring, based on a previous outcome occurring).

- P(A) and P(B): the probabilities of observing A and B respectively without any given conditions; they are known as the marginal probability or prior probability.

Bayes' Theorem

Theorem 8.4: **Bayes' Theorem**

Let $E_1, E_2, \dots \dots E_n$ 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 $\dots \dots E_n$, then

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

Bayesian Classifier

$x \rightarrow$ data sample

$Y \rightarrow$ the class prediction of x

$P(Y|x) \rightarrow$ aim is to find out what is the probability that a pattern x belong to class Y

For this purpose, a data sent / training samples are given:

x	Y
0.2	1
0.6	0
0.1	1
1	0

$\rightarrow$ two-class Problem

From this we need to predict for any unknown sample

$$\left.\begin{array}{l} P(Y=1|x) \\ P(Y=0|x) \end{array}\right\}$$ posterior probability

what we can calculate from data set?

1. prior probability:

$$P(Y=1) = \frac{2}{4} \qquad P(Y=0) = \frac{2}{4}$$

2. Likelihood:

not directly given but we can model.

$$P(x|y)$$

However,

for categorical attributes/feature, we can calculate directly.

3. marginal probability/evidence

$$P(X) = P(x|Y=1)\,P(Y=1) + P(x|Y=0)\,P(Y=0)$$

Rewrite

Bayes' theorem

$$P(Y|x) = \frac{\overbrace{P(x|Y)}^{\text{likelihood}} \; P(Y) \rightarrow \text{prior}}{P(x) \hookrightarrow \text{evidence}}$$

$P(Y|x)$ → posterior probability

For two-class problem

$$P(Y=0|x) = \frac{P(x|Y=0)\,P(Y=0)}{P(x)}$$

$$P(Y=1|x) = \frac{P(x|Y=1)\,P(Y=1)}{P(x)}$$

Assign unknown sample in class where $P(Y|x)$ is maximum

$$P(Y|x) \simeq P(x|Y)\,P(Y)$$

Naive Bayesian classifier

Let, there are K mutually exclusive and exhaustive classes $C_1, C_2, C_3, \ldots C_K$ with prior probability $P(C_1), P(C_2), P(C_3) \ldots, P(C_K)$

$X = [x_1 \; x_2 \; x_3 \; \ldots \; x_n]$ n-feature attribute

Features are conditionally independent to each other

$$p_i = P(Y=C_i \mid [x_1, x_2 \ldots x_n]) \approx \underbrace{P([x_1, x_2 \ldots x_n] \mid Y=C_i)}\, P(Y=C_i)$$

$$P([x_1\; x_2 \ldots x_n] \mid Y=C_i) = \prod_{j=1}^{n} P(x_j \mid Y=C_i)$$

class assignment of $X = C_x = \arg\max_i p_i$

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 database, there are four features: Day, Season, Fog, Rain with 20 data samples.
- The categories of classes are: 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**	**Heavy**	**???**

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

Naïve Bayesian Classifier

- **Solution:** 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

	Attribute	Class: On Time	Late	Very Late	Cancelled
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
Fog	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
Fog	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
Rain	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
Rain	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

$$P\left(Y = \text{OnTime} \mid X = [\text{Day}, \text{Season}, \text{Fog}, \text{Rain}]\right)$$

$$\approx P\left(X = [\text{Day}, \text{Season}, \text{Fog}, \text{Rain}] \mid Y = \text{OnTime}\right)$$

$$* P(Y = \text{OnTime})$$

$$\approx P(\text{Day} = \text{weekday} \mid Y = \text{OnTime}) * P(\text{Season} = \text{winter} \mid Y = \text{OnTime})$$

$$* P(\text{Fog} = \text{High} \mid Y = \text{OnTime}) * P(\text{Rain} = \text{Heavy} \mid Y = \text{OnTime})$$

$$* P(Y = \text{OnTime})$$

unseen data sample : [weekday, winter, High, Heavy]

$$\approx .64 * .14 * .29 * .07 * .70$$

Naïve Bayesian Classifier

Instance:

Week Day	Winter	High	Heavy	???

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

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

Exercise

(Find class assignment for pattern [1 1 1 0])

X1	X2	X3	X4	Y
0	1	1	1	1
1	0	1	1	0
1	1	0	1	1
1	1	1	1	0
0	1	1	0	1

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

- **Approach to overcome the limitations in Naïve Bayesian Classification**
- Estimating the posterior probabilities for continuous attributes
  - In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.
  - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.

1. We can discretize each continuous attributes and then replace the continuous values with its corresponding discrete intervals.
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

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

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

Naïve Bayesian Classifier

- For each class C_i, the posterior probabilities for attribute A_j (it is the numeric attribute) can be calculated following Gaussian normal distribution as follows.

$$P(A_j = a_j|C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(a_j - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Here, the parameter μ_{ij} can be calculated based on the sample mean of attribute value of A_j for the training records that belong to the class C_i.

Similarly, σ_{ij}^2 can be estimated from the calculation of variance of such training records.

Exercise

(Find class assignment for pattern [-4 -3]; Gaussian Distribution for Likelihood)

X1	X2	Y
7	2	1
2	2	1
-2	-3	0
-2	-4	0
2	5	1
-7	-3	0