Md. Sahil
001710501029

1.

a) False,

Naive bayes classifier assumes that each feature x_i is conditionally independent of each other feature x_j for $j \neq i$

b) ~~false~~, True

The success of pattern classification scheme using decission function depends on two factors :-

i) The form of the decision function $d(x)$

ii) ~~The~~ One's ability to determine the coefficients.

The first factor is governed by the geometrical properties of the classes under consideration.

c) ~~False~~, ~~True~~, False

~~The~~ In syntactic pattern recognition the set of ~~input~~ pattern primitives & ~~gra~~ the grammar is required for the classification.

Training phase is not an essential step.

d) False, The points that are ~~also~~ within the margen of the hyperplane can only be the support vectors.

e) True,

Density-based clustering algorithms groups points together under one cluster when they occur in a high density region.

Outliers generally occur in low ~~de~~ density regions, hence are not classified.

f). True,

Naive Bayes classifier assumes that the classes are conditionally independent.

For example consider a naive-bayes classifier to classify spam & important emails. The classifier will only consider the occurance of certain key words in the mails inorder to classify the mail. The ordering of the ~~keywords is not~~ words ~~poss~~ in not ~~is~~ paid importance, which is not true in real ~~f~~ life.

g) False,

~~Sta~~ Syntactic Pattern Recognition attempts to classify ~~problems~~ patterns based on a set of extracted features called pattern primitives and their geometrical model ~~repres~~ represented through the grammar. ~~to~~

h) False,

Hierarchial clustering methods help in exploring data at different levels of ~~gra~~ granularity.

i) False,

A Hopfield net is mainly used for optimization.

j) False,

For a supervised pattern classification proplem having M-classes ~~wh~~; where the classes are pairwise seperable, the classifier needs to compute $[M(M-1)]/2$ number of decission surfaces.

k). False,

Data processing is required to ensure
- accuracy, consistency, completeness, timeliness, believability &
- interpretability.

2.

M - patterns.

prototypes : $Z_1, Z_2, \dots Z_M$. classes : $\omega_1, \omega_2, \omega_3, \dots, \omega_M$

~~We dan determine the~~

we use euclidean distance of a point from any given prototype as a measure of similarity of that point to ~~a partial~~ the particular class belonging to the prototype.

$$D_i = \|X - Z_i\| = \sqrt{(X - Z_i)'(X - Z_i)}$$

Minimum - distance classification is used here, i.e. a given point X is assigned to the class that has the minimum euclidean distance from it.

X is assigned to class ω_i if $D_i < D_j \ \forall j \neq i$.

(Ties are resolved arbitrarily)

$$D_i^2 = \|X - Z_i\|^2 = (X - Z_i)'(X - Z_i)$$

$$= X'X - X'Z_i - Z_i'X + Z_i'Z$$

$$= X'X - 2(X'Z_i - \tfrac{1}{2} Z_i'Z)$$

We need to choose i such that D_i is minimum

$\Rightarrow$ D_i^2 is minimum (as D_i is always a +ve value)

$\Rightarrow$ choosing i such that $(X'Z_i - \frac{1}{2} Z_i'Z)$ is maximum.

as $X'X$ is independent of i.

$\therefore$

$\therefore$ we can define decision function as follows;

$$d_i(x) = X'Z_i - \frac{1}{2} Z_i' Z \quad , i = 1, 2, \ldots M$$

when X is assigned to the class ω_i

when $d_i(x) > d_j(x) \quad \forall\, i \neq j$

we can see that $d_i(x)$ is a linear decission function.

3.

Baye's theorem describes the probaility of an event, based on prior knowledge of conditions that might be related to the event.

Mathematically :-

let X be a data sample

let H be a hypothesis that X belongs to a class C.

$P(H) \rightarrow$ initial provability.

$P(X) \rightarrow$ (evidence) probability that sample data is observed

$P(X|H) \rightarrow$ (likelyhood) the probability of observing X given that the hypothesis holds.

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

Let D be a set of N tuples $\{X_1, X_2, \ldots X_N\}$

when $X_i \in \mathbb{R}^n$.

$\rightarrow$ ~~class~~ K - classes $C_1, C_2, \ldots C_K$.

$\therefore$ From bayes theorem

~~$P(C_k|X) = \frac{P(X|C_k).P(C_k)}{P(X)}$~~ $\quad P(C_i|X) = \frac{P(X|C_i)\, P(C_i)}{P(X)}$

Since the denominator is not dependent on ~~C_k~~ C_i and all values of the ~~feature~~ features are known, ∴ The denominator is effectively constant.

∴ Only numerator is significant, which is equivalent to the joint probability model $P(C_k, x_1, \ldots x_n)$

$$P(C_i) P(X|C_i) = P(C_k, X)$$

or $P(C_i) P(x_1, x_2, \ldots x_n | C_k) = P(C_k, x_1, x_2 \ldots x_n)$

with $[X = (x_1, x_2, x_3 \ldots x_n)]$

using chain rule for repeated application of definition of conditional probability:

$$P(C_i, x_1, x_2 \ldots x_n) = P(x_1, x_2 \ldots, x_n, C_i)$$

$$= P(x_1 | x_2, \ldots x_n, C_i) P(x_2, x_3 \ldots x_n, C_i)$$

$\vdots$

$$= P(x_1 | x_2 \ldots x_n, C_i) P(x_2 | x_3, \ldots x_n, C_i) \ldots P(x_{j-1} | x_j, \ldots x_n, C_i)$$

$$\ldots P(x_{n-1} | x_n, C_k) P(x_n | C_k) P(C_k)$$

Naive Bayes assumes that each feature x_i is conditionally independent of every other feature x_j for $i \neq j$.

$$\Rightarrow P(x_j | x_{j+1}, \ldots x_n, C_i) = P(x_j | C_i)$$

Thus the joint model can be expressed as :-

$\cancel{P(C_i | x_1, \ldots x_n) \propto P(C_k, x_1, \ldots x_n)}$

$= \cancel{P(C_k) P(x_1 | C_k) P(x_2 | C_k) P(x_3 | C_k)}$

$$P(C_i, x_1, x_2 \ldots x_n) = P(C_i) . P(x_1 | C_i) P(x_2 | C_i) \ldots P(x_n | C_i)$$

$$= P(C_i) \prod_{j=1}^{n} P(x_j | C_i)$$

In order to classify X, we pick the hypothesis ~~that is~~ most probable, i.e.

∴ The corresponding classifier, a Bayes Classifier, is the function that assigns a class label $\hat{y} = C_i$ for some i as follows:

$$\hat{y} = \underset{i \in \{1, \dots k\}}{\operatorname{argmax}} P(C_i) \prod_{j=1}^{n} P(x_j | C_i).$$

Let A_j denote the jth feature of a given data sample X.

Now A_j can be either categorical or continuous valued.

$P(x_j | C_i)$ has to be computed ~~differential~~ differently for the above mentiond two cases.

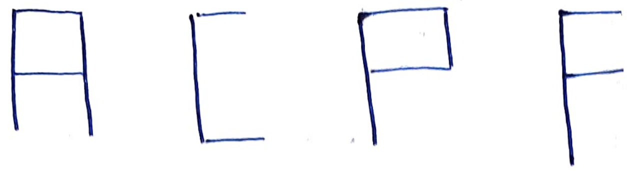
If A_j is categorical, $P(x_j | C_i)$ is the number of tuples in C_i having value x_j of A_j divided by $|C_{i,D}|$, number of tuples of C_i in D.

If A_j is continous-valued, $P(x_j | C_i)$ is usually computed based on Gaussian distribution with a mean μ and SD σ.

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

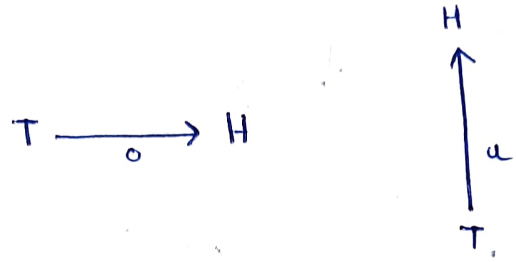
so, $P(x_j | C_i) = g(x_j, \mu_{C_i}, \sigma_{C_i})$

5.



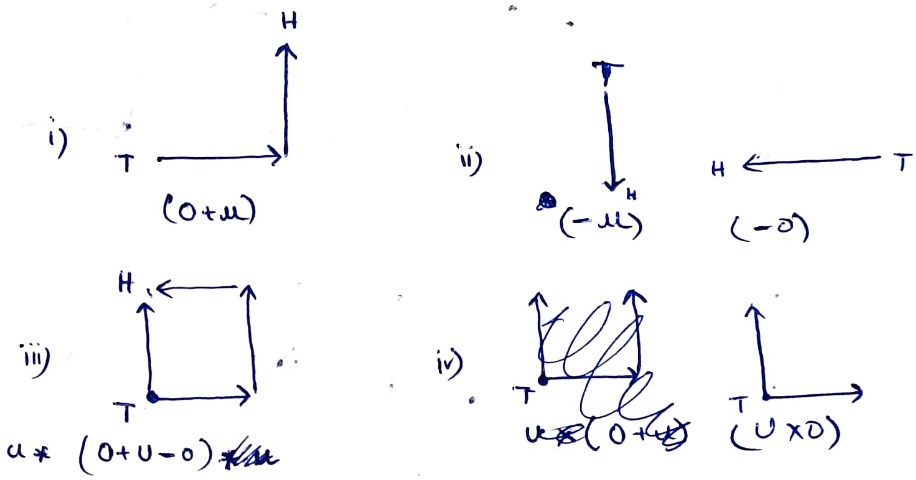
Primitives required :-

{o, u}



Operation allowed on primitives :-

{ *, -, +, x }.



* ⟶ represents head-head & tail-tail attachment.

+ ⟶ represents head to tail concatenation

- ⟶ represents head-tail reversal

x ⟶ represents tail-tail attachment.

The grammar consists of :

1. Start symbol S
2. Set of terminals { u, o, *, -, +, x, |, (,) }

   (where "|" is or-operator, & "(" & ")" are opening & closing parentheses).
3. Set of non-terminals, { S, A, C, P, F }.
4. Set of production rules P = { S → A | C | P | F,

   A → u + ((o u + o + -u) * o) + -u,

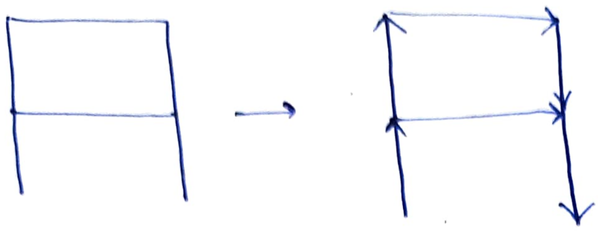
   C → -o + u + u + o,

   P → u + ((u + o + -u) * o),

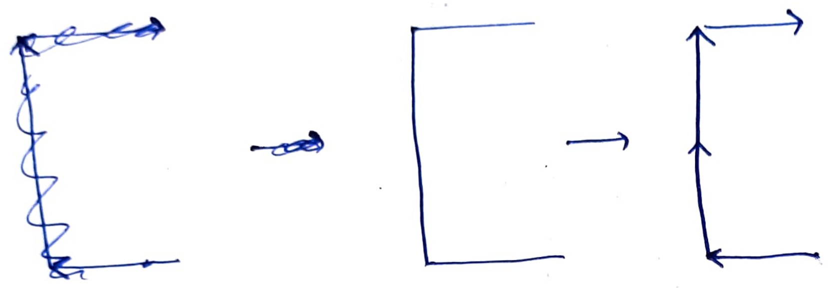
   F → u + (o x u) + o

   }

Premitive representation & interconnections :-

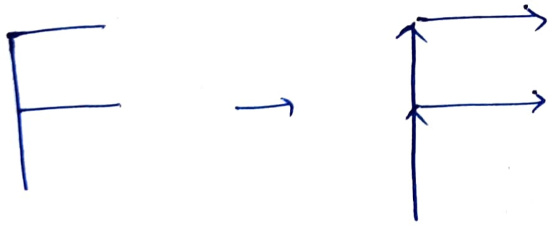


$A = u + ((u + o + -u) * o) + -u$



$C = -o + u + u + o$

$P = u + ((u + o + -u) * o)$



$F = u + (o \times u) + o$

6. Pattern recognition is the automated recognition of patterns and relularities in data.

It can be divided into two use cases

i) Recognizing concrete items.

Ex- pictures.

ii) Recognizing abstract items.

Ex- voical audio,

Most conventional approachs of pattern recognition are based on direct computation through machines which are math-related techniques. These conventional approachs include :- feature extraction, classification, clustering etc.

We can also use application of biological concets of neurons ~~inside~~ in humans for computing. This lead to the development of neural networks.

ANN (Artifical Neural Networks):-

An Artificial Neural Network is a paralleled distributed information processing structure in the form of a directed graph.

It consists of a larger number of simple processing units (perceptrons) ~~with a high degree of dist~~ arranged into ~~layers~~ one or more layers with a high degree of interconnection between each layer of units.

The processing units works parallelly and in coordination with each other.

The design & function of neural networks simulate some functionality of biological brains and neural system.

Pattern recognition can be done using both conventional computers and neural networks.

→ Neural network require simple processors as opposed to general computers which use few complex processes.

→ Neural networks use fewer processing steps.

→ Neural networks use the concept of distributed processing making them faster.

→ Neural nets are trained by example helping them to achieve far better results even for unknown data.

→ Neural nets are tolerable to noisy patterns.

Due to the adaptive - learning, self organizing and fault tolerance capabilities of neural nets, ANNs are used for various pattern recognition applications.

<u>Neural networks as classifiers :-</u>

<u>Neural Net Classifiers</u>

- Binary input
  - Supervised learning
    - Hopfield net
    - Hamming net
      - optimum classifiers
  - Unsupervised learning
    - carpenter classifier
      - leader clustering.
- continuous valued input
  - Supervised learning
    - Perceptron
      - gaussian classifier
    - Multi-layer perceptron
      - ~~K-neighb~~ K nearest neighbours
  - unsupervised learning
    - Kohonen Self-organising feature map
      - K-means ~~sets~~ clustering.

We can see the importance of ANN in Pattern Recognition by looking at the diversity of applications that ANN's have In Pattern Recognition problems.

Algorithms	Type	usage
Hopfield	recursive	optimization
Multi-layer perceptron	feed forward	classification
Kohonan	self-organising	data-coding
Temporal differences	predictive	forecasting.

∴ Neural networks provide the following advantages:-

- Can work with incomplete information once trained
- Fault tolerance (robust to outliers)
- Distributed
- Parallel
- Can learn non-linear and complex relationships
- ~~be~~ trained by example
- Generalizability ~~&~~ ~~&~~ i.e. can infer unseen relationships one trained.