2-09-2024

IR — Asst 2 Syllabus

- TF, DF, IDF, Score of doc for query
- cosine similarity
- Probabilistic Ranking
- Okapi Retrieval model / method
- Language Modelling — unigram - n-gram
- Performance Measure (A, P, R, f1)

# CLASSIFICATION

divide into categories.

i/p - document

classification — class A / class B

- Techniques by classifier

input : $X = x_1, x_2, x_3, \ldots x_n$

class : $C = c_1, c_2, c_3, \ldots, c_m$

$m \ll n$

02-09-24

## Supervise Classfn:

$X_1$ — $C_2$
$X_2$ — $C_2$
$X_3$ — $C_1$
$X_4$ — $C_2$
$X_5$ — $C_1$
$X_6$ — $C_1$
$X_7$ — $C_2$
$X_8$ — $C_2$
$X_9$ — $C_2$
$X_{10}$ — $C_1$

## Training Phase:

- features } will be studied by
- behaviour } classifiers.

## Testing Phase:

New Data:

$X_{50} \longrightarrow C$

3-09-24

## Text Classification problem

A description $d \in X$ is given, $X$ is the document space and a fixed set of class $C = \{c_1, c_2, \ldots c_j\}$. Class are also called category or level.

$\langle d, c \rangle \Rightarrow$ document & class pair $\langle d, c \rangle \in X \times C$

description ↓ a sentence

< the main contributor of microprocessor & chips, china ∉ c

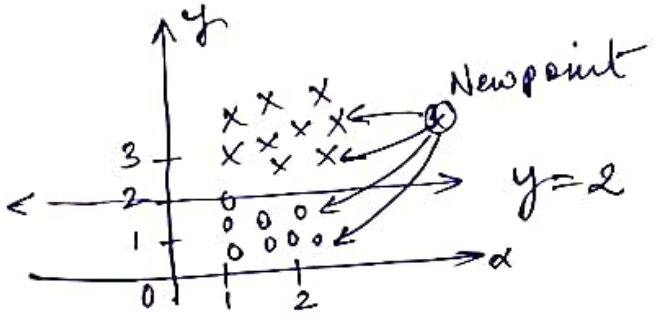
using a learning method, we want to learn a classifier or classification function $\gamma$ that map the documents to class

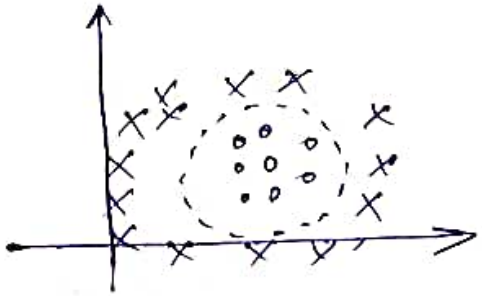
$\gamma : X \longrightarrow C$.

This type of learning is called supervised learning.

# Classifiers.

X → class A
O → class B.



K-Nearest neighbour

- calc distance ⇒ then check to which class does it belong.

09.09.2024

# Naive Bayes Classification

↳ Probability based algorithm

The probability of a doc d being in class C is:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

$P(c)$ – prior probability of document occuring & being in class C.

$P(t_k|c)$ – conditional prob of term $t_k$ occuring in class C.

$n_d$ – no. of terms in the documents

↳ Sum of the log of the probability/probabilities

$$P(c|d) = \log_e P(c) + \sum_{1 \le k \le n_d} \log_e [P(t_k|c)]$$

↳ $P(c) = \frac{N_c}{\text{Total of doc.}}$

Q1 09.24

| DocID | Text | Class |
|---|---|---|
| 1 | Dear friend, lunch money. | N |
| 2 | Friend money money | N |
| 3 | Dear money money | S |
| Test 4 | Money dear money money | ? |

$$P(\_|N) = \frac{x + x_i}{y + y_i}$$ (+1 in numerator; $y_i$ = no. of vocab.)

$$P(N|4) = P(N) \prod_{1 \le k \le 3} P(t_k|N)$$

$$= P(N)\left[P(\text{Money}|N) * P(\text{dear}|N)\right] * P(\text{money}|N) * P(\text{money}|N).$$

$P(N) = \frac{2}{3}$

$P(S) \frac{1}{3}$

$P(\text{money}|N) = \frac{3}{7}$

$P(\text{dear}|N) = \frac{1}{7}$

$P(\text{money}|S) = \frac{2}{3}$

$P(\text{dear}|S) = \frac{1}{3}$

$$= \frac{2}{3} \times \frac{3}{7} \times \frac{1}{7} \times \frac{3}{7} \times \frac{3}{7} = \frac{6 \times 3}{343 \times 7} = 0.0175 \times \frac{3}{7}$$

$$= \underline{0.00749}$$

$$P(S|4) = P(S) * P(\text{money}|S) * P(\text{dear}|S) * P(\text{money}|S) * P(\text{money}|S)$$

$$= \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3}$$

$$= \frac{4 \times 2}{81 \times 3} = 0.0494 \times \frac{2}{3}$$

$$= \underline{0.0329}.$$

∴ doc 4 will be in class S (spam).

09 09.09.2024

# 8.

| | DocID | Words in doc | class |
|---|---|---|---|
| TRAINING | 1 | Chinese Beijing chinese | C |
| | 2 | Chinese chinese shangai | C |
| | 3 | chinese Macau | C |
| | 4 | Tokyo Japan chinese | NC |
| Test: | 5 | chinese chinese chinese Tokyo Japan | ? |

$$P(\text{Tokyo}|C) = \frac{1}{8+6} = \frac{1}{14}. \quad \because t_{\text{Tokyo}|C} = 0$$

$$P(\text{Japan}|C) = \frac{1}{8+6} = \frac{1}{14}$$

$$P(\text{Chinese}|C) = \frac{6}{14} = \frac{3}{7}$$

$$P(C) = 3/4$$

$$P(C|5) = P(C) * P(\text{Chin}|C)^3 * P(\text{Tokyo}|C) * P(\text{Jap}|C)$$

$$= \frac{3}{4} \times \frac{3}{7} \times \frac{3}{7} \times \frac{3}{7} \times \frac{1}{14} \times \frac{1}{14}.$$

$$= 0.0003012$$

$$P(NC) = \frac{1}{4}$$

$$P(\text{Tokyo}|NC) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Japan}|NC) = \frac{2}{9}$$

$$P(\text{Chin}|NC) = \frac{2}{9}$$

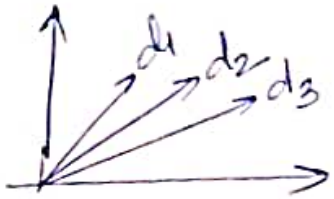
$P(NC|S) = \frac{1}{4} \times \frac{2}{7} \times \frac{2}{7} \times \frac{2}{7} \times \frac{2}{7} \times \frac{2}{7}$

$= 0.000135$

∴ Doc 5 will be in C

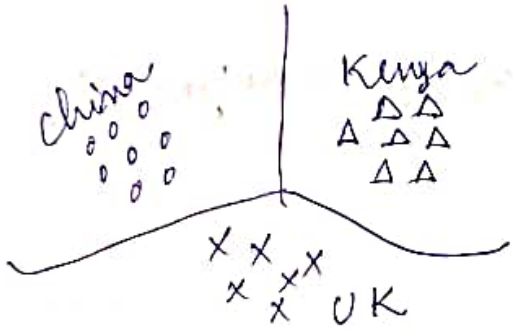
13-09-2024

## # Vector Space Collection




Assumption : contiguity hypothesis

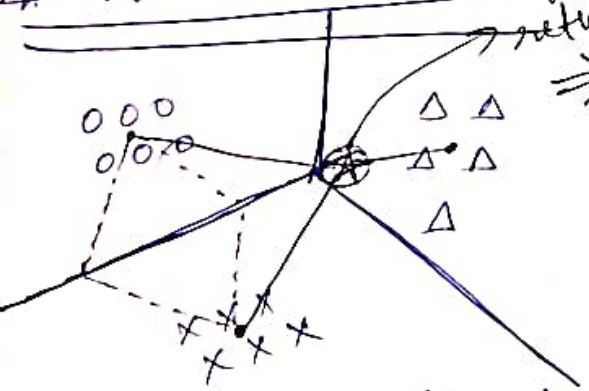
Doc of same class form a contiguous region & region of diff. classes do not overlap.




Based on this
i) Rocchio
ii) KNN.

## # ROCCHIO classifier

return $\min^m$ distance from centroid
⇒ class of new data point



Algorithm :
- calc the centroid / centre of mass.
- hyperplane to seperate the classes.
  → set of points i.e., equidistant from the centroid.

Advantage : easy to calc; efficient

Disadvantage : if pts belonging to same class are scattered, then centroid may not be proper.

09. 13.09.2024

# KNN classifier

Algorithm:

- Assign the value of k
- Calc the dist. from new data pt. to all the existing data pt.
- Arrange them in ascending order & consider the top K values
- Perform majority voting & assign the class accordingly.

Data pts.

$x_1$ $y_1$ — $d_1$

$x_2$ $y_2$ — $d_2$

⋮

$x_n$ $y_n$ — $d_n$

New Data pt.

$x_k, y_k$

$(x_k y_k - x_1 y_1) - d_1$

$(x_k y_k - x_2 y_2) - d_2$

⋮

$(x_k y_k - x_n y_n) - d_n$

i) Euclidean
ii) Manhattan
iii) others

Eg: Let $K = 3$ Top 3.

$d_1 < d_2 < d_3 < d_4 \cdots < d_n$

$c_1$ $c_1$ $c_2$

majority = $c_1$

∴ return $c_1$

13.09.2024

| x | y | Quadrant |
|---|---|---|
| 1 | 2 | I |
| 2 | 3 | I |
| 4 | 4 | I |
| -2 | 3 | II |
| -3 | 1 | II |
| -1 | +5 | ? |

| II. | I |
|---|---|
| | |

Calc using Euclidean dis

$k = 3$

$x_k = -1 \quad y = 5$

$d_1 = \sqrt{4+9} = \sqrt{13}$

$d_2 = \sqrt{9+4} = \sqrt{13}$

$d_3 = \sqrt{25+1} = \sqrt{26}$

$d_4 = \sqrt{1+4} = \sqrt{5}$

$d_5 = \sqrt{4+16} = \sqrt{20}$

$d_4 < d_1 = d_2 < d_5 \; | \; < d_3$

II I I II | I

I

majority = II.

(-1, 5) — II.

04 11.04.2024

# Q.

| docID | Words in doc | in C = china |
|---|---|---|
| 1 | chinese Beijing chinese | Yes |
| 2 | chinese chinese Shanghai | Yes |
| 3. | chinese Macau | Yes |
| 4 | Tokyo Japan chinese | No |
| 5 | chinese chinese Tokyo Japan chinese | |

Using Rocchio algorithm.

Step-1 (Centroid)

term weight tabble (tf-idf) $= (1+\log_{10} tf_{t,d})\left(\log_{10} \frac{N}{df_t}\right)$ $N=4$

| Vector | chinese | Japan | Tokyo | Macau | Beijing | Shangai |
|---|---|---|---|---|---|---|
| $d_1$ | 0 | 0 | 0 | 0 | 0.602 | 0 |
| $d_2$ | 0 | 0 | 0 | 0 | 0 | 0.602 |
| $d_3$ | 0 | 0 | 0 | 0.602 | 0 | 0 |
| $d_4$ | 0 | 0.602 | 0.602 | 0 | 0 | 0 |
| $d_5$ | 0 | 0.602 | 0.602 | 0 | 0 | 0 |
| $\mu_c$ (centroid for C-Yes) | 0 | 0 | 0 | 0.2 | 0.2 | 0.2 |
| $\mu_{\bar{c}}$ (centroid for C-No) | 0 | 0.602 | 0.602 | 0 | 0 | 0 |

$\mu_c = \frac{1}{3}(d_1+d_2+d_3)$ } Basically avg.

$\mu_{\bar{c}} = \frac{1}{1}(d_4)$.

$|\mu_c - d_5|$ } minm distance

$|\mu_{\bar{c}} - d_5|$

⇓

Manhattan distance.

17.09.24

min^n distance : (using Eucledian distance).

$d_5$ & $\mu_C$

$= \sqrt{0^2+0^2+0^2+0^2+0^2+0^2} = \sqrt{0} = 0.$

$d_5$ & $\mu_C$

$= \sqrt{0^2+0.602^2+0.602^2+0.2^2+0.2^2+0.2^2}$

$= 0.918.$

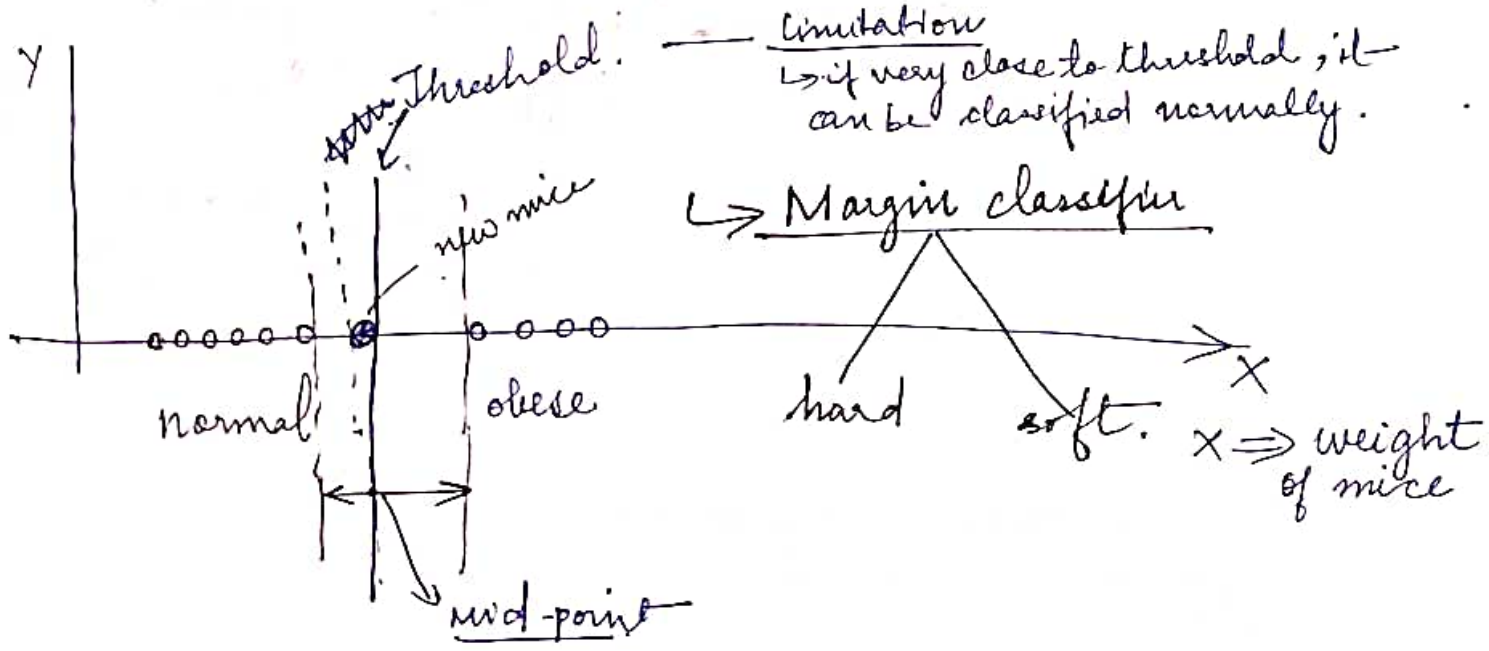
$\therefore \min^m(0, 0.918)$

min^m is $\mu_C$.

$\therefore d_5$ belongs to $\bar{C}$ i.e No.

24.09.24.

Support Vector Machine (SVM).

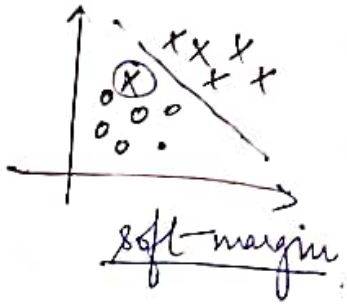



11

09 24.09.24

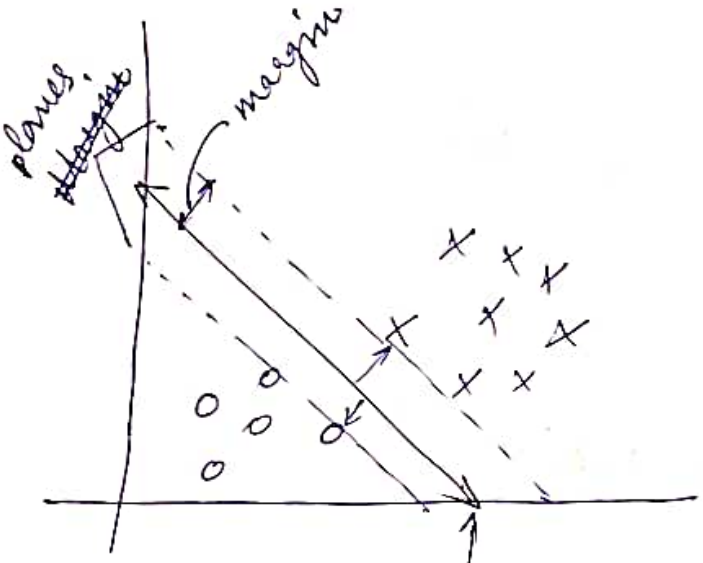
# Hard Margin → No outliar is allowed.

Soft Margin → Allow certain misclassification



soft margin

## # SVM → Maxm margin classifier



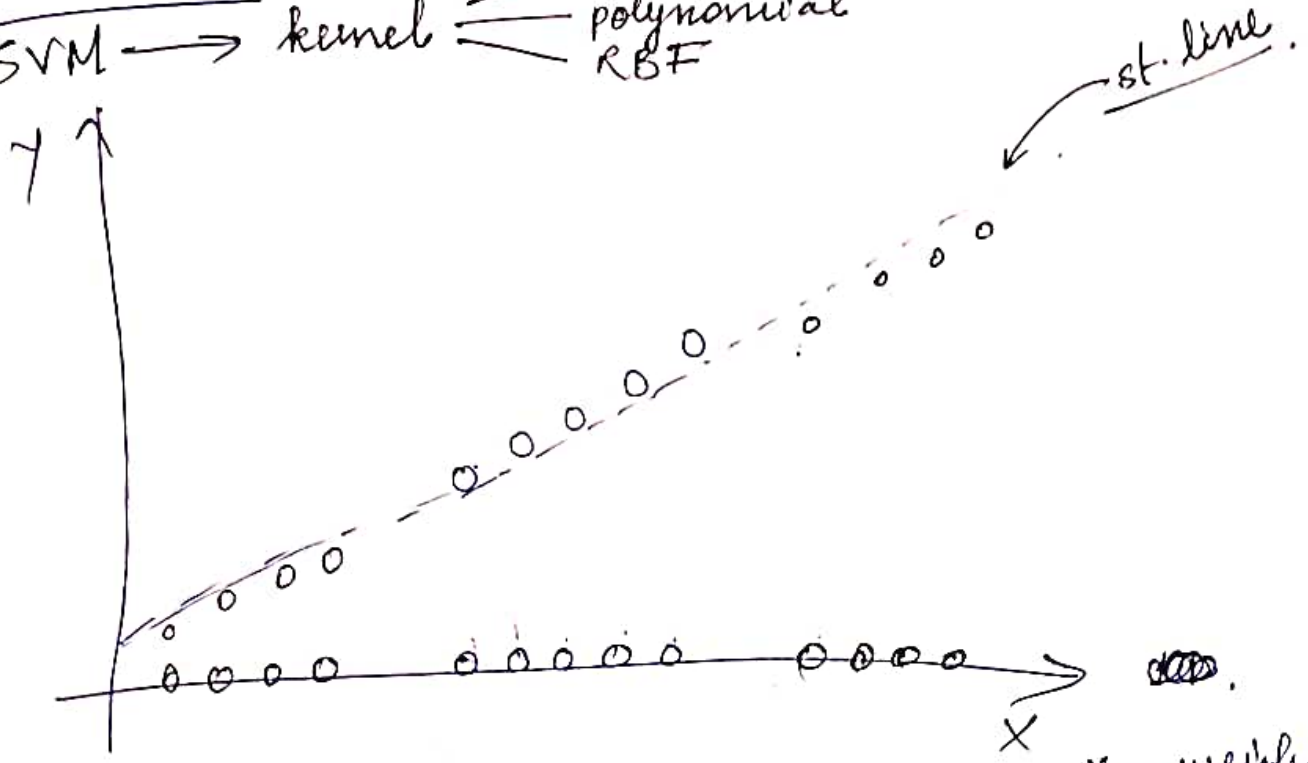

hyperplane → ( maxm dist. from the support vector )

↓

data points which helps to create hyperplane

24.09.24

# SVM → kernel → linear, polynomial, RBF




- Transform the data points to diff dimension using kernel so as it is easy to classify the scattered datapoints.

x = weight

y = (weight)²

plot (w, w²).