LECTURE FIVE

TEXT CLASSIFICATION

5.1 Standing queries

The path from IR to text classification. You have an information need to monitor, say: *"Unrest in the Niger delta region*".

You want to rerun an appropriate query periodically to find new news items on this topic. You will be sent new documents that are found.

Such queries are called standing queries. Long used by "information professionals". A modern mass instantiation is Google Alerts. Standing queries are (hand-written) text classifiers

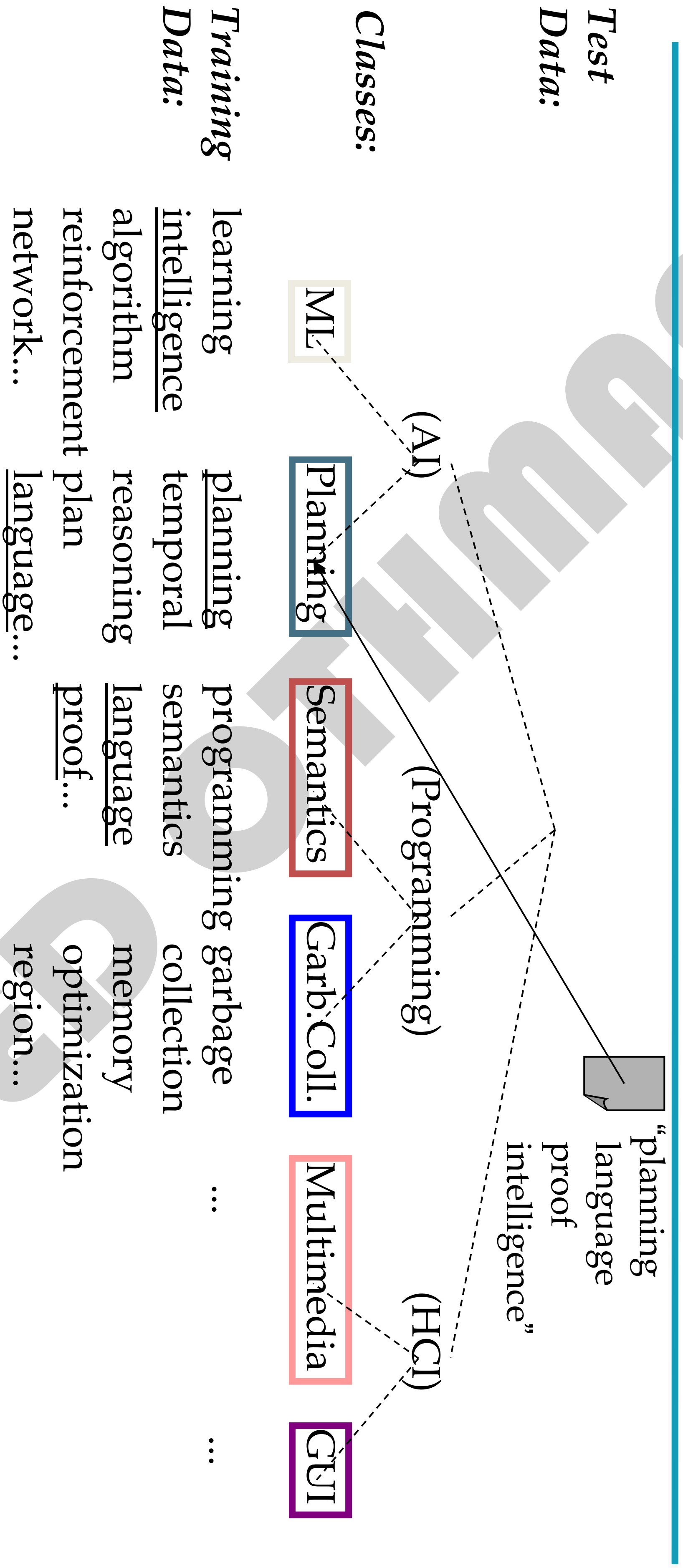

Test Data:
"planning language proof intelligence"
Classes:
(AI)
(Programming)
(HCI)
ML
Planning
Semantics
Garb.Coll.
Multimedia
GUI
Training Data:
learning intelligence algorithm reinforcement network...
planning temporal reasoning plan language...
programming semantics language proof...
garbage collection memory optimization region...
...
...

Figure 5.1: Document Classification

5.2 Classification: Basic Concepts

Classification: A form of data analysis that extracts model describing important data classes.

Supervised learning (classification): Class label of each training tuple is provided. New data is classified based on the training set.

Unsupervised learning (clustering): The class labels of training data are unknown. Number of classes to be learned may not be known in advance.

Data classification: Is a two-step process, consisting of a

- Learning step: Where a classification model is constructed.
- Classification step: Where the model is used to predict class labels for given data.

5.3 Linear Classifiers

All feature vectors from the available classes can be classified correctly using a **linear classifier**, and some techniques can be used for the computation of the corresponding linear functions.

Design of linear classifiers described by linear discriminant functions **(hyperplanes) g(x).**

The major advantage of linear classifiers are:

- Their simplicity.
- Computational attractiveness.

Example 1

Boolean functions, AND and OR, are linearly separable. The corresponding truth tables for the AND and OR operations are given in table and the respective class positions in the two-dimensional space are shown in figure.

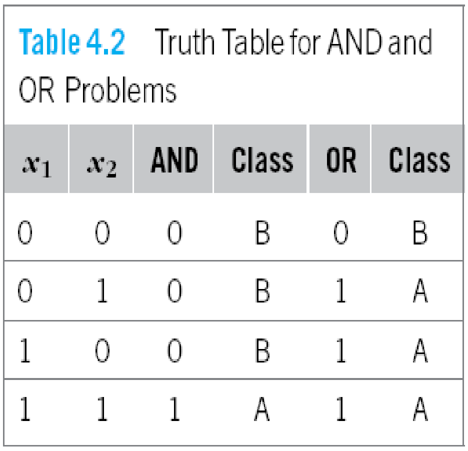
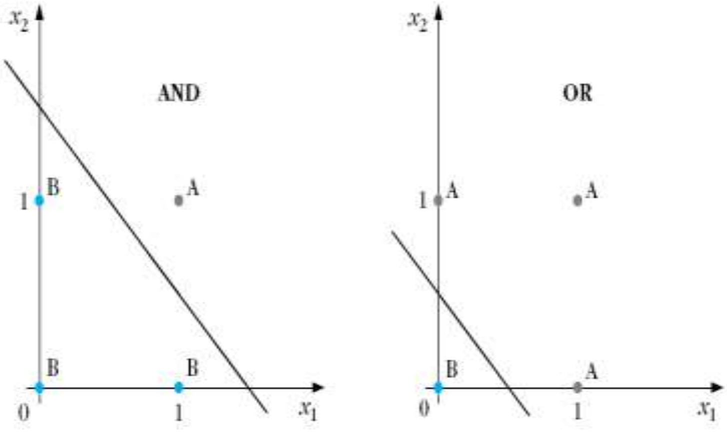


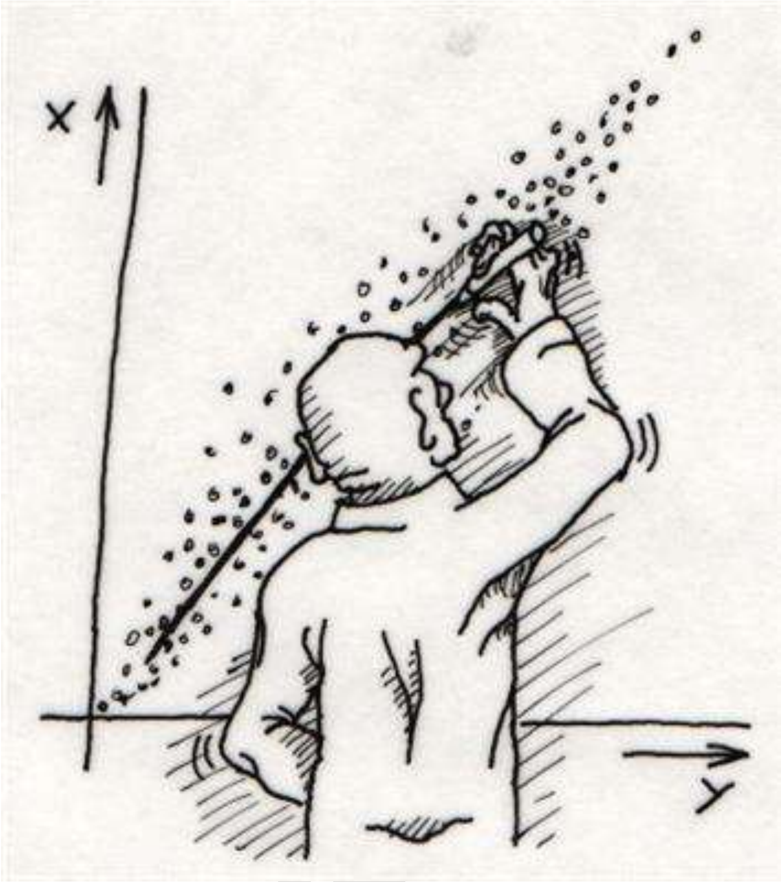
Table 4.2 Truth Table for AND and OR Problems

x_1	x_2	AND	Class	OR	Class
0	0	0	B	0	B
0	1	0	B	1	A
1	0	0	B	1	A
1	1	1	A	1	A



5.4 Linear Least Squares

Is the line of best fit for a group of points. It seeks to minimize the sum of all data points of the square differences between the function value and data value. It is the earliest form of linear regression.



Example 2

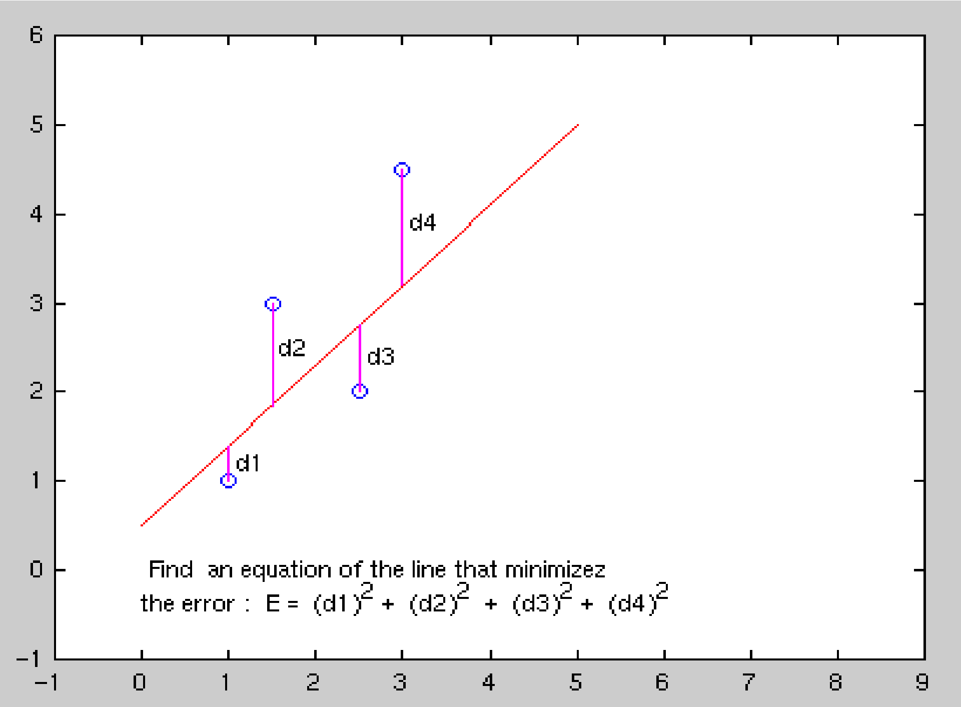
Consider the points (1 , 2.1), (2 , 2.9), (5 , 6.1), and (7 , 8.3) with the best fit line f(x) = 0.9x + 1.4

The squared errors are:

$x_1=1$	$f(1)=2.3$	$y_1=2.1$	$e_1= (2.3 – 2.1)^2$ = 0.04
$x_2=2$	$f(2)=3.2$	$y_2=2.9$	$e_2= (3.2 – 2.9)^2$ =.0 09
$x_3=5$	$f(5)=5.9$	$y_3=6.1$	$e_3= (5.9 – 6.1)^2$ = .004
$x_4=7$	$f(7)=7.7$	$y_4=8.3$	$e_4= (7.7 – 8.3)^2$ = 0.36

So the total squared error is 0.04 + 0.09 + 0.04 + 0.36 = .53

By finding better coefficients of the best fit line, we can make this error smaller…



We want to minimize the vertical distance between the point and the line.

- $E = (d_1)^2 + (d_2)^2 + (d_3)^2 + \ldots + (d_n)^2$ for n data points
- $E = [f(x_1) - y_1]^2 + [f(x_2) - y_2]^2 + \ldots + [f(x_n) - y_n]^2$
- $E = [mx_1 + b - y_1]^2 + [mx_2 + b - y_2]^2 + \ldots + [mx_n + b - y_n]^2$
- $E = \sum(mx_i + b - y_i)^2$

E must be MINIMIZED! How do we do this?

$$E = \sum(mx_i + b - y_i)^2$$

Treat x and y as constants, since we are trying to find m and b. So…PARTIALS!

$$\partial E/\partial m = 0 \text{ and } \partial E/\partial b = 0$$

But how do we know if this will yield maximums, minimums, or saddle points?

E = $\sum(mx_i + b - y_i)^2$ is minimized when the partial derivatives with respect to each of the variables is zero.

$$\partial E/\partial w = 0 \quad \textbf{and} \quad \partial E/\partial b = 0$$

$$m = \frac{nSxy - SySx}{nSxx - SxSx}$$

$$b = \frac{SxxSy - SxySx}{nSxx - SxSx}$$

Example: Find the linear least squares approximation to the data: (1,1), (2,4), (3,8)

$Sx = 1+2+3= 6$

$Sxx = 1^2+2^2+3^2 = 14$

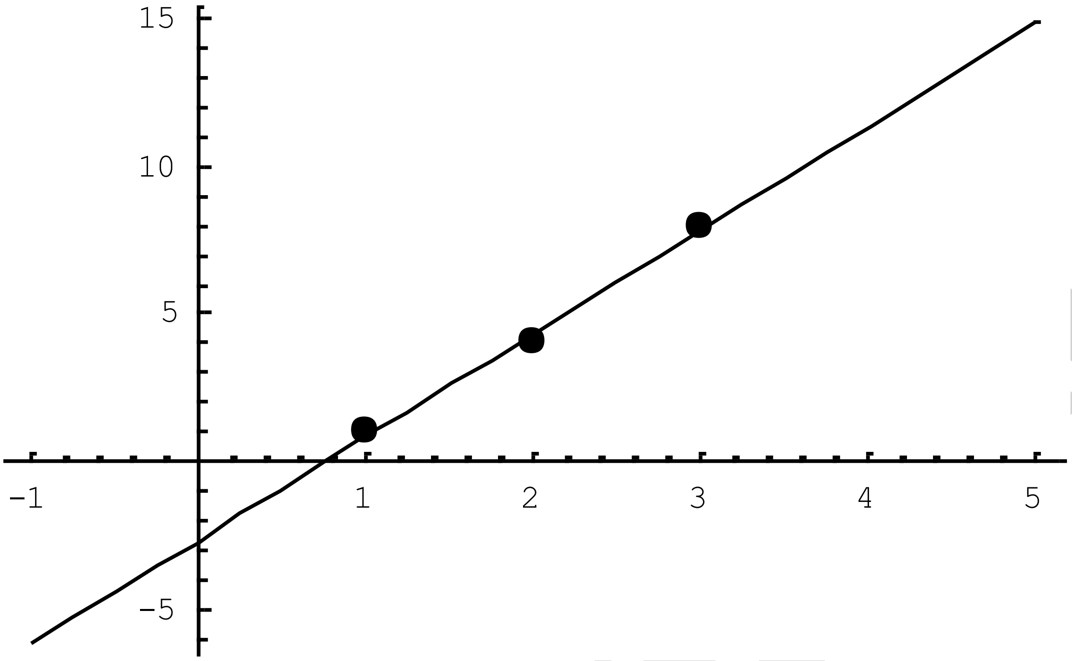
$Sy = 1+4+8 = 13$

$Sxy = 1(1)+2(4)+3(8) = 33$

n = number of points = 3

$$m = \frac{3(33) - 6(13)}{3(14) - 6(6)} = \frac{21}{6} = 3.5$$

$$b = \frac{14(13) - 33(6)}{3(14) - 6(6)} = \frac{-16}{6} = -2.667$$



5.5 Lazy Learners (or Learning from Your Neighbors)

Lazy learning (e.g., instance-based learning): Simply stores training data (or only minor processing) and waits until it is given a test tuple.

Eager learning (the above discussed methods): Given a set of training tuples, constructs a classification model before receiving new (e.g., test) data to classify.

5.6 k-Nearest-Neighbor Classifiers

- The training tuples are described by n attributes.
- Each tuple represents a point in an n-dimensional space.
- all the training tuples are stored in an n-dimensional pattern space.
- When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple.
- The Euclidean distance between two points or tuples, say,
    - X1 = (x11, x12, … , x1n) and X2 = (x21, x22, … , x2n)

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Min-max normalization can be used to transform a value v of a numeric attribute A to v 0 in the range [0, 1] by computing

Min–max normalization:

$$X^* = \frac{X - \min(X)}{\text{range(X)}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Z-score standardization:

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

“But how can distance be computed for attributes that are not numeric, but nominal (or categorical) such as color?”

For nominal attributes, If the two are identical (e.g., tuples X1 and X2 both have the color blue), then the difference between the two is taken as 0. If the two are different (e.g., tuple X1 is blue but tuple X2 is red), then the difference is considered to be 1.

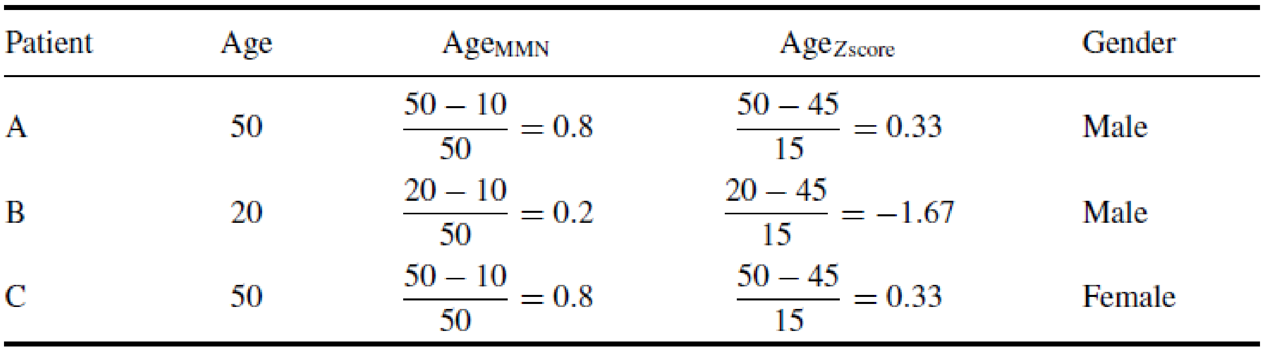
Example: - suppose that we have a new patient who is a 50-year-old male. Which patient is more similar, a 20-year-old male or a 50-year-old female?

Patint	Age	Gender
A	50	Male
B	20	Male
C	50	Female

Suppose that for the age variable, the range is 50, the minimum is 10, the mean is 45, and the standard deviation is 15.

Let patient A be our 50-year-old male, patient B the 20-yearold male, and patient C the 50-year-old female.

The original variable values, along with the min–max normalization (age_{MMN}) and Z-score standardization ($\text{age}_{\text{Zscore}}$), are listed in Table.



Patient	Age	Age_{MMN}	$\text{Age}_{\text{Zscore}}$	Gender
A	50	$\frac{50-10}{50}=0.8$	$\frac{50-45}{15}=0.33$	Male
B	20	$\frac{20-10}{50}=0.2$	$\frac{20-45}{15}=-1.67$	Male
C	50	$\frac{50-10}{50}=0.8$	$\frac{50-45}{15}=0.33$	Female

- The distance between patients A and B is d(A,B)= $\sqrt{(50-20)^2+\ 0^2}=30$

- The distance between patients A and C is d(A,C) =$\sqrt{(20-20\,)^2+\ 1^2}=1$
- We would thus conclude that the 20-yearold male is 30 times more "distant" from the 50-year-old male than the 50-year-old female is.
- In other words, the 50-year-old female is 30 times more "similar" to the 50-year-old male than the 20-year-old male is.

- We use the min–max normalization
- The distance between patients A and B is d_{MMN} (A,B)= $\sqrt{(0.8-0.2\,)^2+\ 0^2}=0.6$
- The distance between patients A and C is d_{MMN} (A,C) =$\sqrt{(0.8-0.8\,)^2+\ 1^2}=1.0$
- which means that patient B is now considered to be more similar to patient A.

- we use the Z-score standardization values
- The distance between patients A and B is d_{Zscore} (A,B)= $\sqrt{(0.33-(-1.67)\,)^2+\ 0^2}=2.0$
- The distance between patients A and C is d_{Zscore} (A,C)= $\sqrt{(0.33-0.33\,)^2+\ 1^2}=1.0$