Lecture 2.2

- Nearest neighbor
- K Nearest Neighbor

Parametric and Non-parametric Models

- A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a **Parametric model**
  - No matter how much data you throw at a parametric model, it won’t change its mind about how many parameters it needs
- **Nonparametric methods** seek to best fit the training data in constructing the mapping function, whilst maintaining some ability to generalize to unseen data
  - As such, they are able to fit a large number of functional forms.

K-Nearest Neighbours

- The **k-nearest neighbors (KNN)** algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point
- **Nearest neighbor search (NNS)**, as a form of proximity search, is the optimization problem of finding the point in a given set that is closest (or most similar) to a given point
  - **Nearest neighbor** refers to the single data point closest to a given point, while k nearest neighbors refers to the group of "k" data points that are closest to a given point

k-NN Algorithm

- In the **training phase** the kNN algorithm stores the dataset
- In the **prediction/testing phase** for a given test point, kNN calculates the distances to all points in the training dataset, selects the k nearest neighbors, and determines the output based on their labels or values
- Advantages
  - Easy to implement and understand, no explicit training phase, adaptable to classification and regression problems
- Disadvantages
  - Computationally expensive, sensitive to irrelevant or redundant features, performance can degrade if data is not normalized or scaled

Distance Metrics

- **Euclidean Distance**:

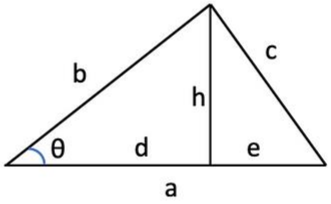
$$d_{Euc}(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- **Cosine Similarity**:

$$S_{cos}(\boldsymbol{A},\boldsymbol{B}) = \cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

$$Cosine\ Distance = 1 - S_{cos}(\boldsymbol{A},\boldsymbol{B})$$

Cosine Similarity in Detail



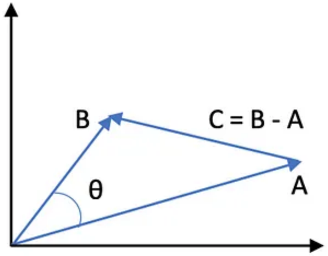
$c^2 = b^2 + a^2 - 2ab\cos(\theta)$

$\cos(\theta) = d/b$
$d = b\cos(\theta)$

$\sin(\theta) = h/b$
$h = b\sin(\theta)$

$c^2 = h^2 + e^2$
$c^2 = (b\sin(\theta))^2 + (a-d)^2$
$c^2 = (b\sin(\theta))^2 + (a - b\cos(\theta))^2$
$c^2 = b^2\sin^2(\theta) + a^2 - 2ab\cos(\theta) + b^2\cos^2(\theta)$
$c^2 = b^2(\sin^2(\theta) + \cos^2(\theta)) + a^2 - 2ab\cos(\theta)$

$c^2 = b^2 + a^2 - 2ab\cos(\theta)$



$\vec{A} = \begin{matrix} a1 \\ a2 \end{matrix}$ $\quad \vec{B} = \begin{matrix} b1 \\ b2 \end{matrix}$ $\quad \vec{B} - \vec{A} = \begin{matrix} b1 - a1 \\ b2 - a2 \end{matrix}$

$\vec{A}.\vec{B} = a1.b1 + a2.b2$

Length of vector $\vec{A} = ||\vec{A}|| = \sqrt[2]{a1^2 + a2^2}$

$$||\vec{B} - \vec{A}||^2 = ||\vec{B}||^2 + ||\vec{A}||^2 - 2||\vec{A}||.||\vec{B}||.\cos(\theta)$$

$$2||\vec{A}||.||\vec{B}||.\cos(\theta) = ||\vec{B}||^2 + ||\vec{A}||^2 - ||\vec{B} - \vec{A}||^2$$
$$2||\vec{A}||.||\vec{B}||.\cos(\theta) = b1^2 + b2^2 + a1^2 + a2^2 - ((b1 - a1)^2 + (b2 - a2)^2)$$
$$2||\vec{A}||.||\vec{B}||.\cos(\theta) = b1^2 + b2^2 + a1^2 + a2^2 - ((b1^2 - 2b1.a1 + a1^2) + (b2^2 - 2b2.a2 + a2^2))$$
$$2||\vec{A}||.||\vec{B}||.\cos(\theta) = 2(a1.b1 + a2.b2)$$

$$\cos(\theta) = \frac{a1.b1 + a2.b2}{||\vec{A}||.||\vec{B}||}$$

$$\cos(\theta) = \frac{\vec{A}.\vec{B}}{||\vec{A}||.||\vec{B}||}$$

k-NN algorithm using Euclidean Distance Problem 1

SL	Length	Width	Class
1	5.1	3.5	Setosa
2	4.9	3.0	Setosa
3	6.3	3.3	Versicolor
4	6.1	2.9	Versicolor

Test point: (5.9, 3.2)
K=3

k-NN algorithm using Euclidean Distance Solution 1

SL	Length	Width	Class	Distance
1	5.1	3.5	Setosa	$\sqrt{(5.9-5.1)^2+(3.2-3.5)^2}=0.85$
2	4.9	3.0	Setosa	$\sqrt{(5.9-4.9)^2+(3.2-3.0)^2}=1.02$
3	6.3	3.3	Versicolor	$\sqrt{(5.9-6.3)^2+(3.2-3.3)^2}=0.41$
4	6.1	2.9	Versicolor	$\sqrt{(5.9-6.1)^2+(3.2-2.9)^2}=0.36$

K	Length	Width	Class	Distance
1	6.1	2.9	Versicolor	$\sqrt{(5.9-6.1)^2+(3.2-2.9)^2}=0.36$
2	6.3	3.3	Versicolor	$\sqrt{(5.9-6.3)^2+(3.2-3.3)^2}=0.41$
3	5.1	3.5	Setosa	$\sqrt{(5.9-5.1)^2+(3.2-3.5)^2}=0.85$

Class
Versicolor

k-NN algorithm using Cosine Distance Problem 2

SL	Length	Width	Class
1	5.1	3.5	Setosa
2	4.9	3.0	Setosa
3	6.3	3.3	Versicolor
4	6.1	2.9	Versicolor

Test point: (5.9, 3.2)
K=3

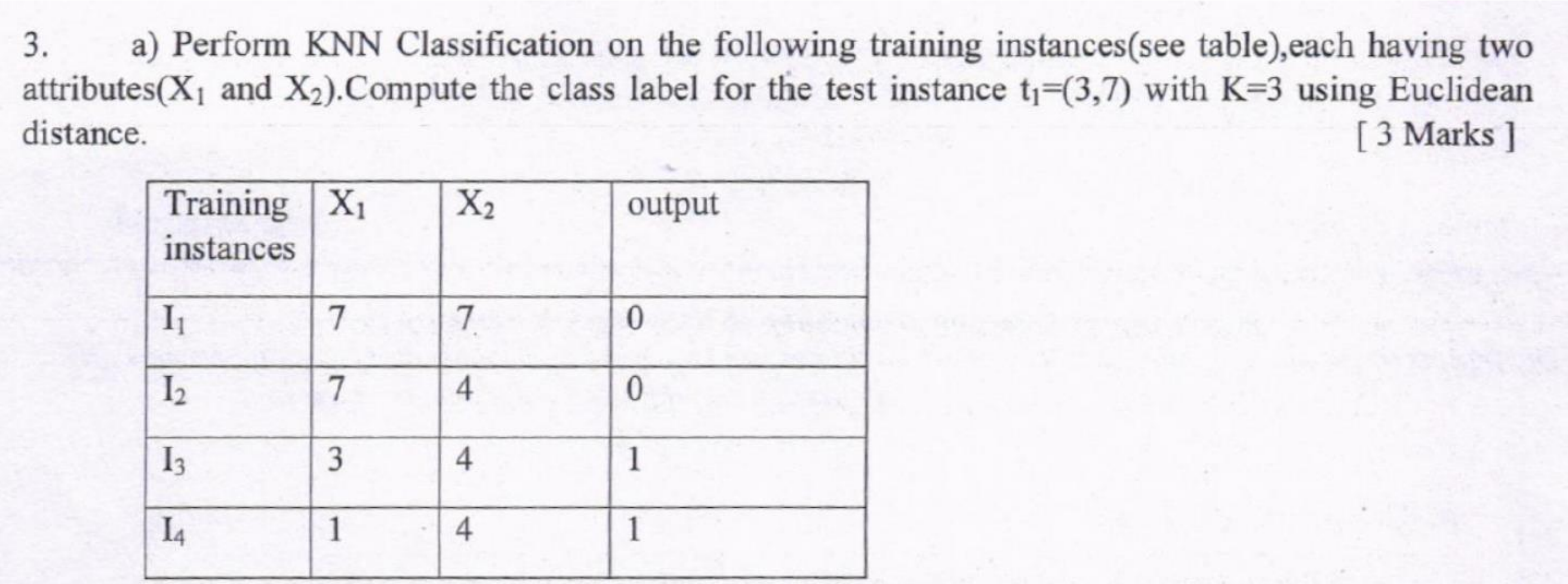
k-NN algorithm using Cosine Distance Solution 2

SL	Length	Width	Class	Distance
1	5.1	3.5	Setosa	$1-\frac{\sum_{i=1}^{n} A_iB_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} = 1-\frac{5.1\times5.9+3.5\times3.2}{\sqrt{5.1^2+3.5^2}\sqrt{5.9^2+3.2^2}} = 0.005$
2	4.9	3.0	Setosa	$1-\frac{\sum_{i=1}^{n} A_iB_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} = 1-\frac{4.9\times5.9+3.0\times3.2}{\sqrt{4.9^2+3.0^2}\sqrt{5.9^2+3.2^2}} = 0.0013$
3	6.3	3.3	Versicolor	$1-\frac{\sum_{i=1}^{n} A_iB_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} = 1-\frac{6.3\times5.9+3.3\times3.2}{\sqrt{6.3^2+3.3^2}\sqrt{5.9^2+3.2^2}} = 0.0001$
4	6.1	2.9	Versicolor	$1-\frac{\sum_{i=1}^{n} A_iB_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} = 1-\frac{6.1\times5.9+2.9\times3.2}{\sqrt{6.1^2+2.9^2}\sqrt{5.9^2+3.2^2}} = 0.0014$

K	Length	Width	Class	Distance
1	6.3	3.3	Versicolor	0.0001
2	4.9	3.0	Setosa	0.0013
3	6.1	2.9	Versicolor	0.0014

Class
Versicolor

Classwork



3. a) Perform KNN Classification on the following training instances(see table),each having two attributes(X_1 and X_2).Compute the class label for the test instance t_1=(3,7) with K=3 using Euclidean distance. [3 Marks]

Training instances	X_1	X_2	output
I_1	7	7	0
I_2	7	4	0
I_3	3	4	1
I_4	1	4	1