

Linear algebra review and distance measurements

CSCI 347 - Data Mining

Assignment 1, Quiz 1

- Assignment 1 will be released tonight.
- You should be able to finish this after this class.
- Quiz 1 is available.
 - Only 1 attempt is allowed (50 mins)
 - Quiz does not take 50 mins to complete.

Common Data Formats

- Data can be represented by data matrix.

$$D = \begin{array}{ccccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & F & 5.3 \\ x_6 & 0.4 & 19 & G & 5.4 \\ x_7 & 1.1 & 11 & A & 5.5 \end{array}$$

Common Data Formats

- Data can be represented by data matrix.

The columns commonly represent attributes/properties of the data

The rows commonly represent entities and their observed values for each attribute

$D =$

	X_1	X_2	X_3	X_4
x_1	0.2	23	A	5.7
x_2	0.4	1	B	5.4
x_3	1.8	0.5	D	5.2
x_4	5.6	50	C	5.1
x_5	-0.5	34	F	5.3
x_6	0.4	19	G	5.4
x_7	1.1	11	A	5.5

Probabilistic view of the data

- Data can be represented by data matrix.

Each attribute can be thought of as a random variable as well.

The rows commonly represent entities and their observed values for each attribute

$D =$

	X_1	X_2	X_3	X_4
x_1	0.2	23	A	5.7
x_2	0.4	1	B	5.4
x_3	1.8	0.5	D	5.2
x_4	5.6	50	C	5.1
x_5	-0.5	34	F	5.3
x_6	0.4	19	G	5.4
x_7	1.1	11	A	5.5

Probabilistic view of the data

- Assumes that each numeric attribute X is a random variable.
- The data set becomes an experiment that has observations of these random variable.
- $X: O \rightarrow \mathbb{R}$, where O is the domain of X and \mathbb{R} is the codomain of X .
- If the outcomes are numeric and represent the observed values of the random variable, then X is basically the identity function.
 - $X: O \rightarrow O$, where $X(v) = v$

Review of stats

- Estimated Mean $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$
- Estimated variance $\hat{\sigma}_j^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_{ij} - \hat{\mu}_j)^2$
- Estimated standard deviation $\hat{\sigma}_j = \sqrt{\hat{\sigma}_j^2}$
- Estimated covariance $= \hat{\sigma}_{jk} = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_{ij} - \hat{\mu}_j) \cdot (x_{ik} - \hat{\mu}_k)$
- Covariance matrix $\Sigma = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_3^2 \end{pmatrix}$

More stats...

Normalization is the process of adjusting or transforming data into a standard or common format.

- Mean centering $x'_{ij} = x_{ij} - \hat{\mu}_j$
- Z – score normalization $x'_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$
 - Mean is still 0.
 - SD = 1
 - Variance is 1
 - Can detect outliers based on the standard deviation.
 - Can use when features have different units and scales.
 - Does not change the distribution of the data.
- Range normalization $x'_{ij} = \frac{x_{ij} - \min_i \{x_{ij}\}}{\max_i \{x_{ij}\} - \min_i \{x_{ij}\}}$
 - Good for visualization.
 - Good for distance-based algorithms.

Z-Score normalization example

Z-score or standard score normalization tells us how many standard deviations each entity value is from the attribute mean:

$$x'_{11} = \frac{(0.2 - 1.285)}{2.035} = -0.533$$

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5



	X_1	X_2	X_3
x'_1	-0.5	0.2	1.7
x'_2	-0.4	-1.0	0.1
x'_3	0.3	-1.1	-0.9
x'_4	2.1	1.7	-1.4
x'_5	-0.9	0.8	-0.4
x'_6	-0.4	-0.04	0.1
x'_7	-0.1	-0.5	0.7

Range normalization

$$x'_{ij} = \frac{x_{ij} - \min_i \{x_{ij}\}}{\max_i \{x_{ij}\} - \min_i \{x_{ij}\}}$$

- Ensure comparability across features
 - Some algorithms may disproportionately weigh feature with larger ranges. (Ex: k-means)
- Facilitate Visualization and Interpretation.
- Some algorithms assume normalized inputs.
- Question: what would be the new range of values?

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

$$x'_{11} = \frac{0.2 - (-0.5)}{5.6 - (-0.5)} = 0.11$$

Range normalization

$$x'_{ij} = \frac{x_{ij} - \min_i \{x_{ij}\}}{\max_i \{x_{ij}\} - \min_i \{x_{ij}\}}$$

- Activity: find the range normalized values for attribute X_2 .

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

Range normalization example

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5



	X_1	X_2	X_3
x'_1	0.11	0.46	1.00
x'_2	0.15	0.02	0.50
x'_3	0.38	0.00	0.17
x'_4	1	1	0.00
x'_5	0.00	0.68	0.33
x'_6	0.15	0.38	0.50
x'_7	0.26	0.22	0.67

Geometric view of data

- Projection

In linear algebra and geometry, projection refers to "dropping" one vector onto another (or onto a subspace), capturing the "shadow" or component of one vector in the direction of the other.

- $D =$

	X_1	X_2	X_3	X_4
x_1	0.2	23	A	5.7
x_2	0.4	1	B	5.4
x_3	1.8	0.5	D	5.2
x_4	5.6	50	C	5.1
x_5	-0.5	34	F	5.3
x_6	0.4	19	G	5.4
x_7	1.1	11	A	5.5

Geometric view of data

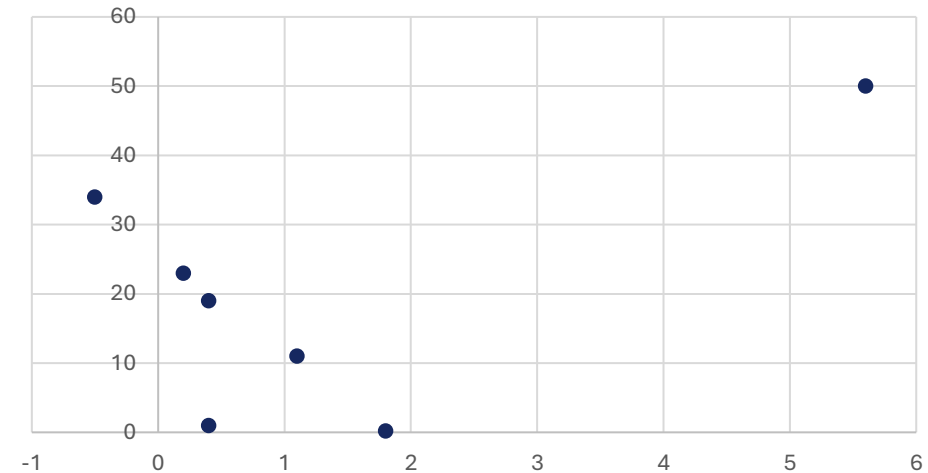
- Projection

$$D = \begin{array}{ccccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & F & 5.3 \\ x_6 & 0.4 & 19 & G & 5.4 \\ x_7 & 1.1 & 11 & A & 5.5 \end{array} \xrightarrow{\pi_{12}} \begin{array}{ccc} & X_1 & X_2 \\ x_1 & 0.2 & 23 \\ x_2 & 0.4 & 1 \\ x_3 & 1.8 & 0.2 \\ x_4 & 5.6 & 50 \\ x_5 & -0.5 & 34 \\ x_6 & 0.4 & 19 \\ x_7 & 1.1 & 11 \end{array}$$

Geometric view of data

- Projection

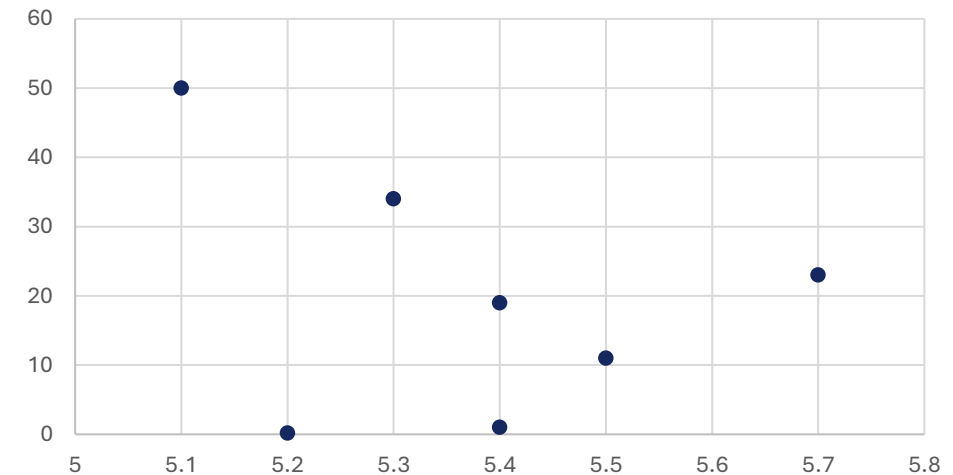
$D =$		X_1	X_2	X_3	X_4	$\xrightarrow{\pi_{12}}$		X_1	X_2
	x_1	0.2	23	A	5.7		x_1	0.2	23
	x_2	0.4	1	B	5.4		x_2	0.4	1
	x_3	1.8	0.5	D	5.2		x_3	1.8	0.2
	x_4	5.6	50	C	5.1		x_4	5.6	50
	x_5	-0.5	34	F	5.3		x_5	-0.5	34
	x_6	0.4	19	G	5.4		x_6	0.4	19
	x_7	1.1	11	A	5.5		x_7	1.1	11



Geometric view of data

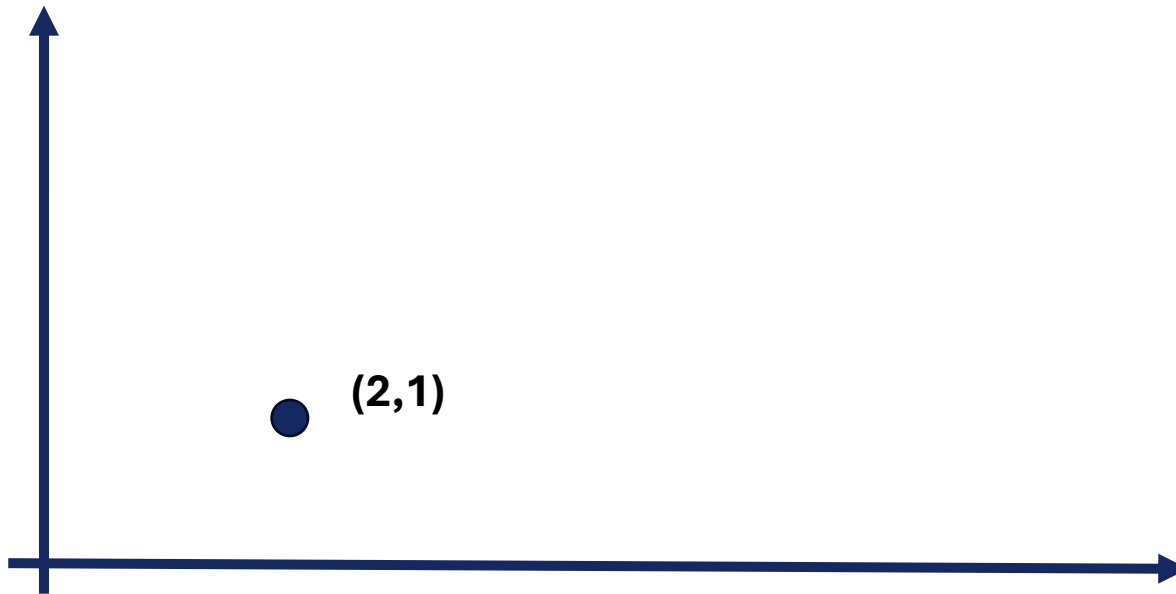
- Projection and re-label

$D =$		X_1	X_2	X_3	X_4	$\xrightarrow{\pi_{42}}$		X'_1	X'_2
	x_1	0.2	23	A	5.7		x_1	5.7	23
	x_2	0.4	1	B	5.4		x_2	5.4	1
	x_3	1.8	0.5	D	5.2		x_3	5.2	0.5
	x_4	5.6	50	C	5.1		x_4	5.1	50
	x_5	-0.5	34	F	5.3		x_5	5.3	34
	x_6	0.4	19	G	5.4		x_6	5.4	19
	x_7	1.1	11	A	5.5		x_7	5.5	11



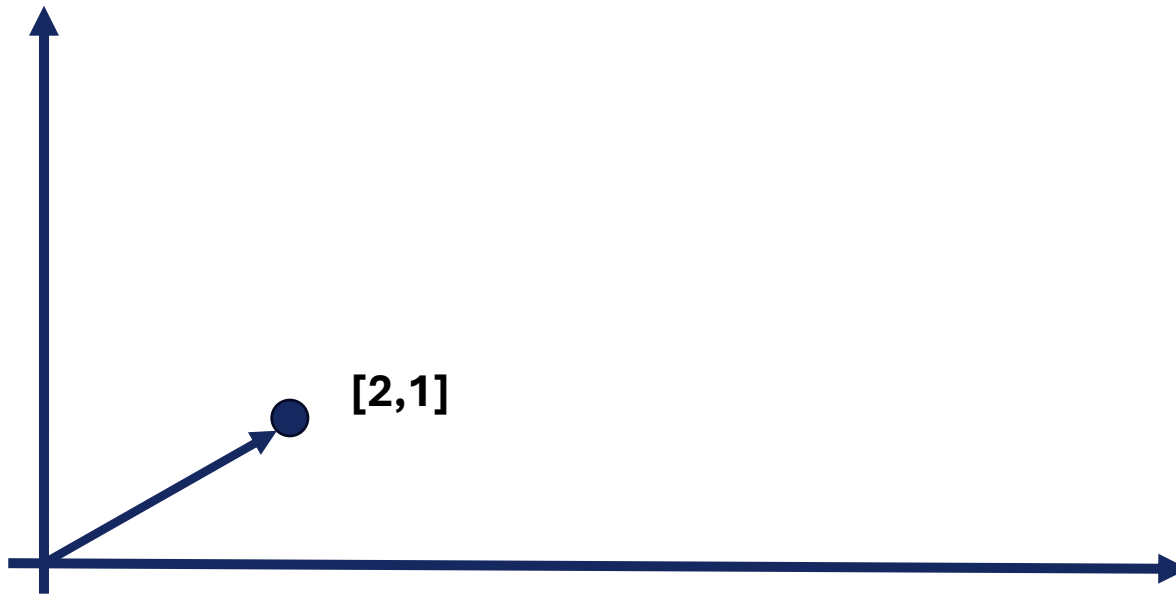
Geometric view of data

- Points
 - Location in a coordinate system
 - No direction or magnitude (simply a position)



Geometric view of data

- Vector
 - A **vector** represents a quantity with both magnitude (size) and direction.



Geometric view of data

- Vector
 - Vector in $n - \text{dimensional}$ space can be considered as:
 - $v = (v_1, v_2, v_3, \dots, v_n) \in \mathbb{R}^n$
 - Each $\forall i \in [1..n] : v_i$ is called a component of the vector.
 - Or equivalently, it can be considered as a $n - \text{dimensional}$ column vector
 - (All vectors are assumed to be column vectors by default)

$$\bullet \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ \vdots \\ v_n \end{pmatrix} = (v_1, v_2, v_3, \dots, v_n)^T \in \mathbb{R}^n$$

Common Data Formats

- Data can be represented by data matrix.
- Each row/entity of a data matrix can be represented as a vector.

- $D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$ $x_4 = (5.6 \ 50 \ 5.1)$

- Essentially, the dataset can be considered as a collection of vectors

Geometric view of data

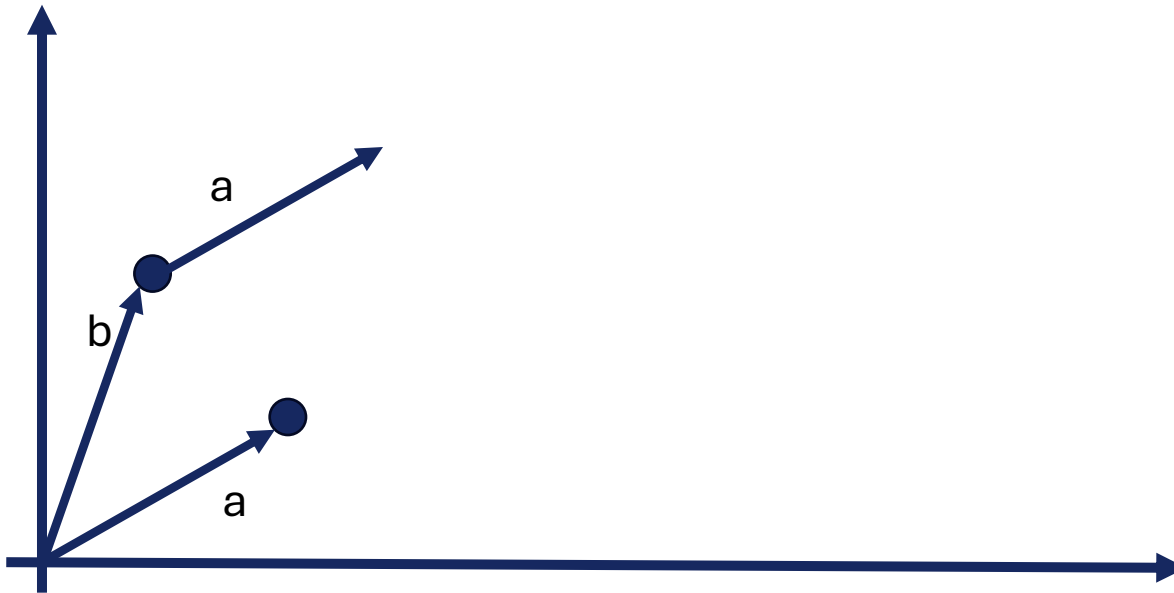
- Vector addition: $a + b = (a_x + b_x \quad a_y + b_y)$



$$a = (a_x \quad a_y)$$
$$b = (b_x \quad b_y)$$

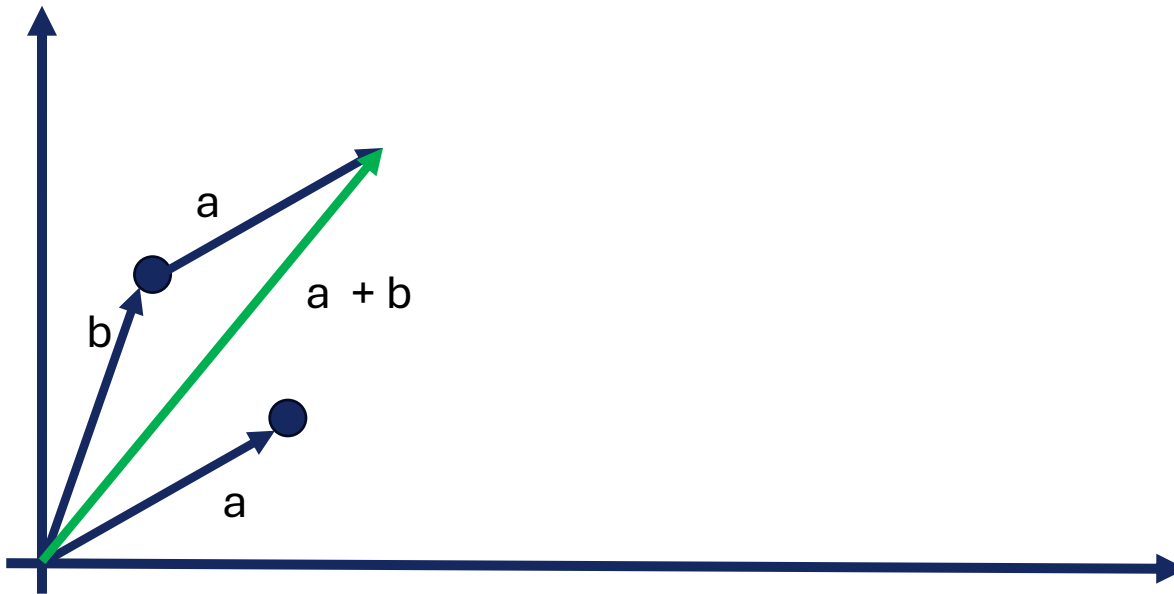
Geometric view of data

- Vector addition $a + b = (a_x + b_x \quad a_y + b_y)$



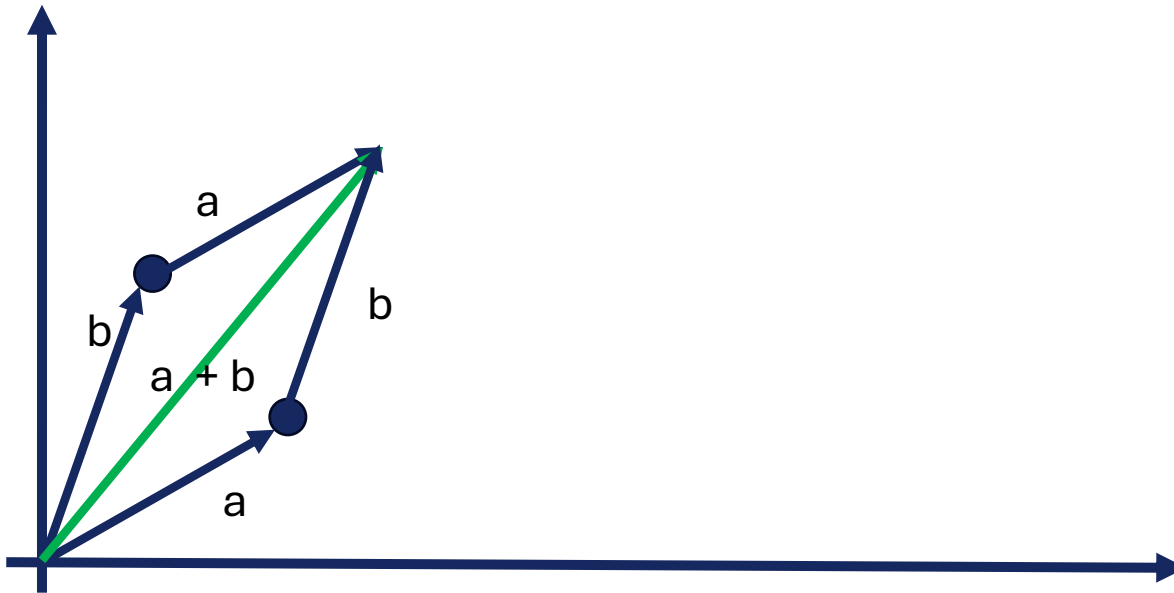
Geometric view of data

- Vector addition $a + b = (a_x + b_x \quad a_y + b_y)$



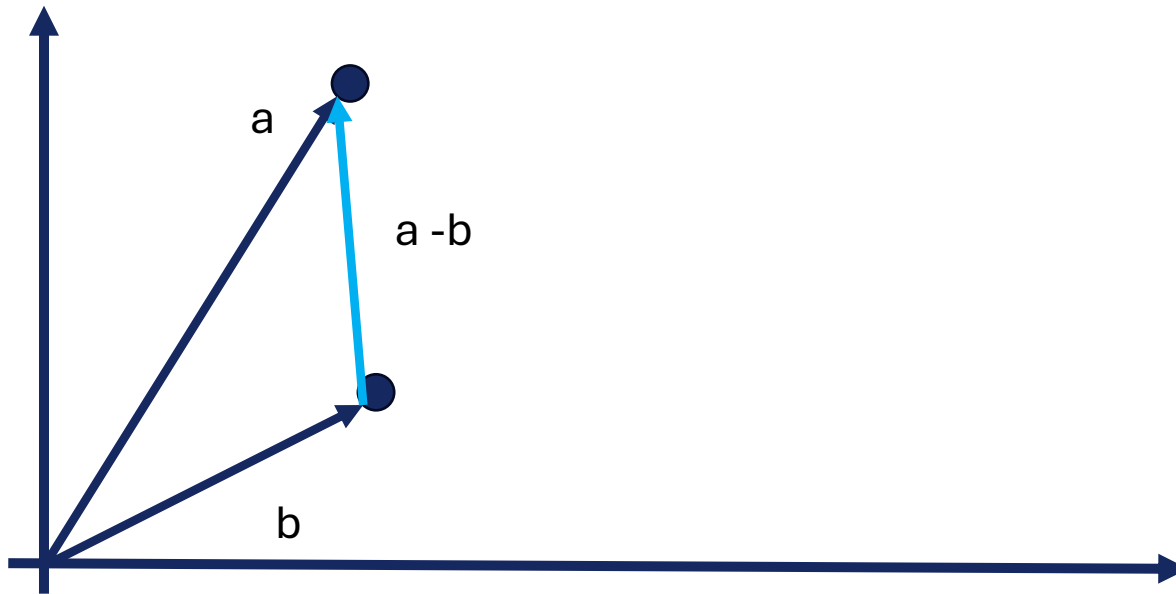
Geometric view of data

- Vector addition $a + b = (a_x + b_x \quad a_y + b_y)$



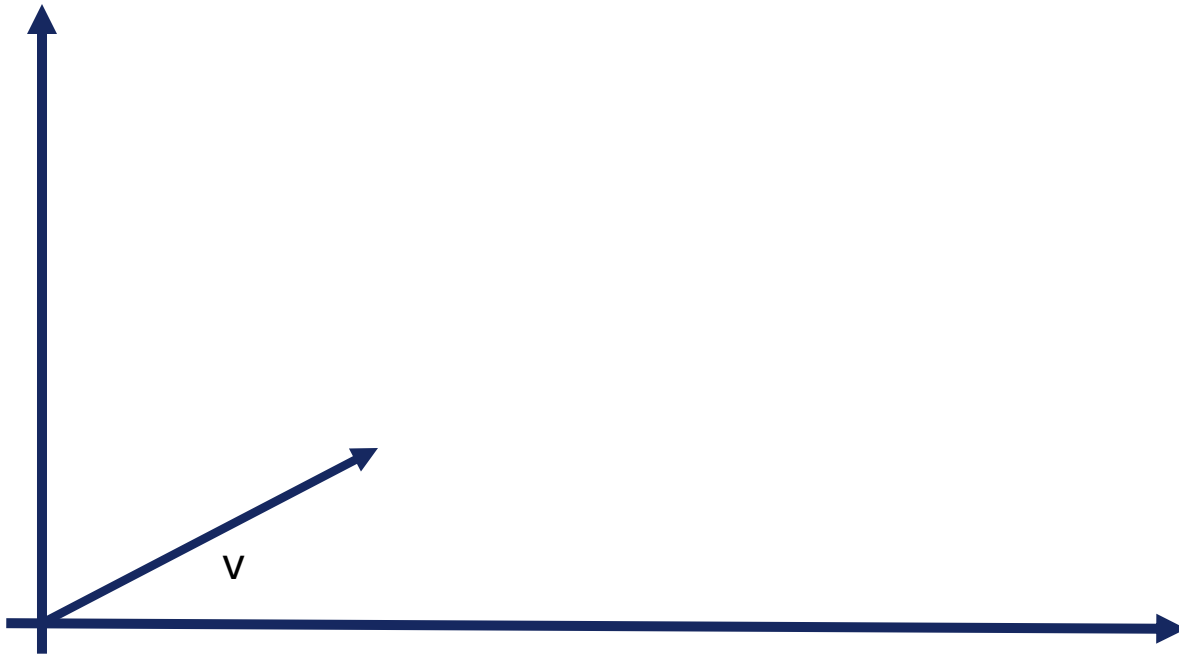
Geometric view of data

- Vector subtraction $a - b = (a_x - b_x \ a_y - b_y)$



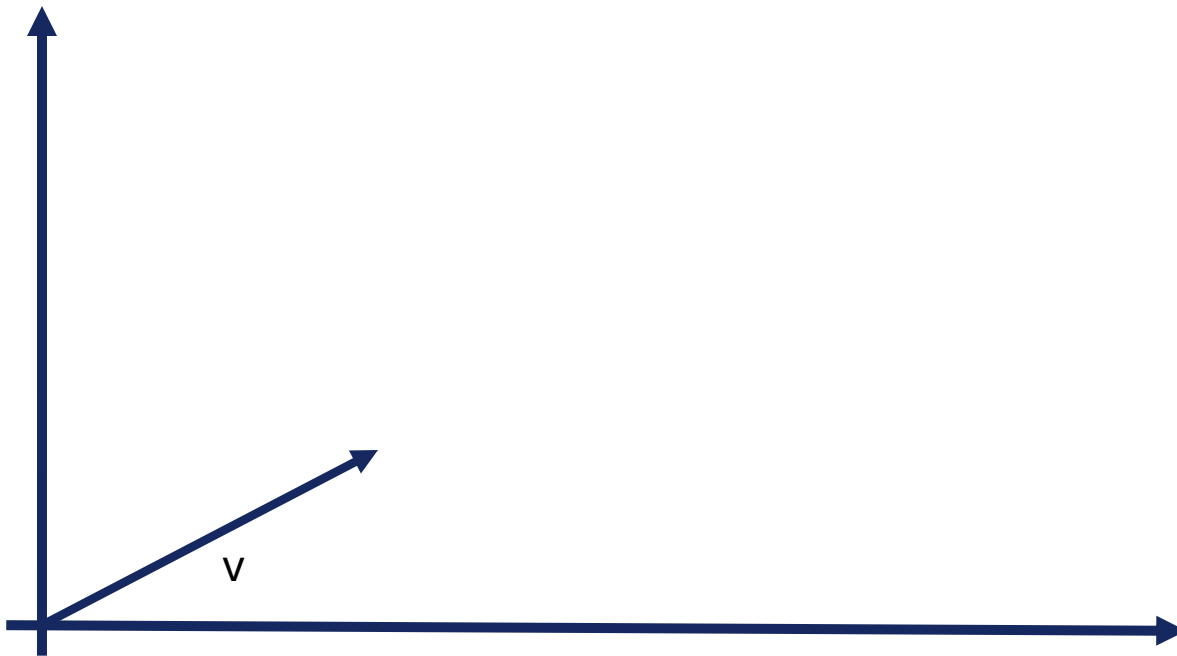
Geometric view of data

- Scaling: For $\alpha \in \mathbb{R}$, $\alpha v = (\alpha v_x \quad \alpha v_y)$



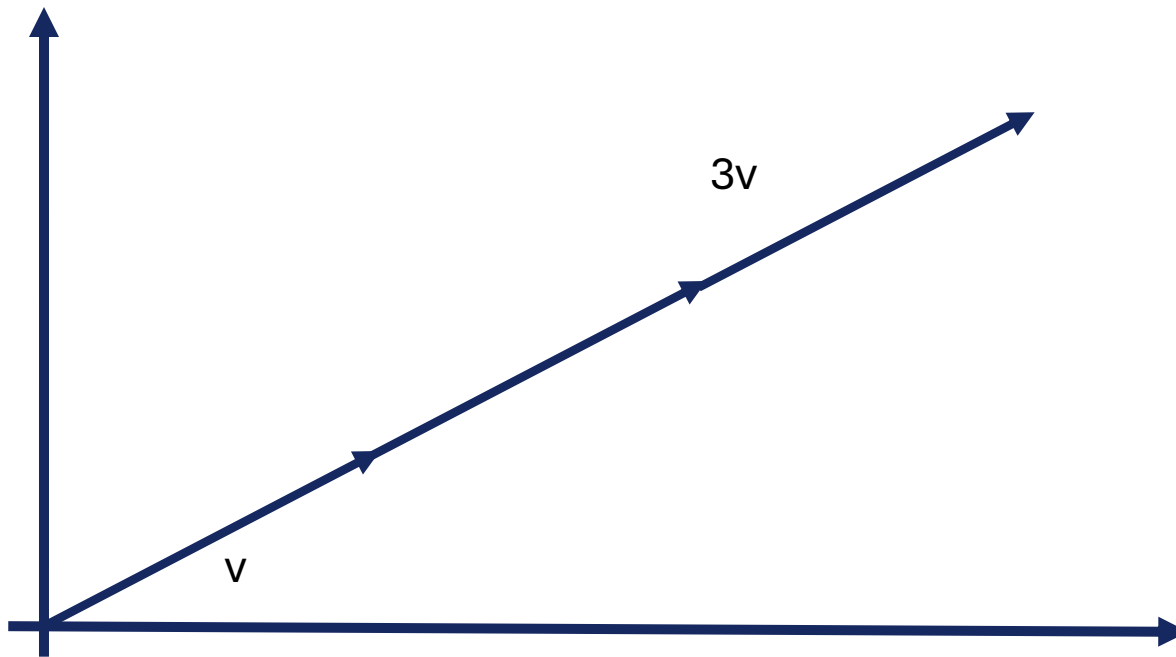
Geometric view of data

- Scaling: For $\alpha \in \mathbb{R}$, $\alpha v = (\alpha v_x \quad \alpha v_y)$
 - Basically, take α copies of v , and add them.



Geometric view of data

- Scaling: For $\alpha \in \mathbb{R}$, $\alpha v = (\alpha v_x \quad \alpha v_y)$
 - Let $\alpha = 3$, then



Distance between vectors

- We are interested in some measure of distance between vectors representing separate entities.
- First, we need to define magnitude of a vector.
- Norm of a vector is a measure of magnitude (non-negative) in the given vector space.
 - At its core, **“norm” means “size” or “length.”**
- There are different types of norms we can define. (different types of distance measurements)
 - $L_1, L_2, L_3, \dots, L_p, \dots, L_\infty$

Distance between vectors

- L_2 norm of the vector x_i with m dimensions (columns/attributes)

$$\|x_i\|_2 = \sqrt{\sum_{k=1}^m x_{ik}^2}$$

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

Distance between vectors

- L_2 is known and Euclidean norm (2-norm)
- L_2 norm of the vector x_i with m dimensions (columns/attributes)

$$\|x_i\|_2 = \sqrt{\sum_{k=1}^m x_{ik}^2}$$

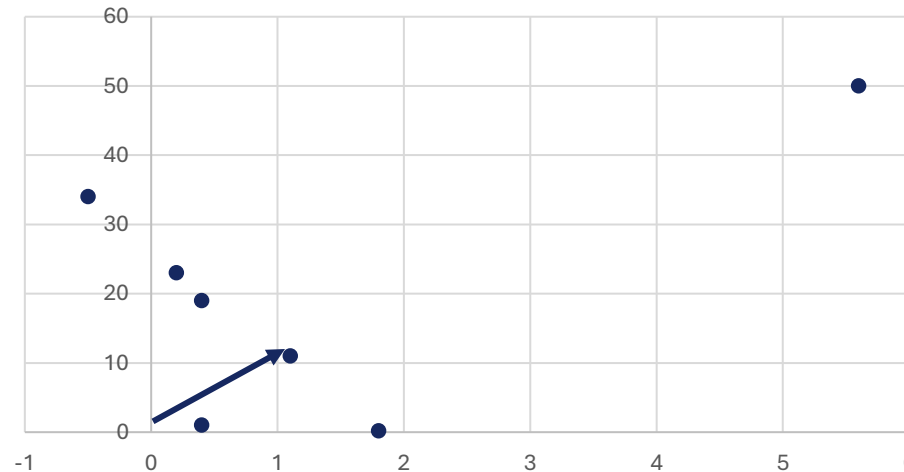
	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

Distance between vectors

- L_2 example

- $\|x_7\|_2 = \sqrt{\sum_{k=1}^m (x_{7k}^2)} = \sqrt{(x_{71}^2 + x_{72}^2)} = \sqrt{(1.1^2 + 11^2)} =$

11.05



	x_1	x_2
x_1	0.2	23
x_2	0.4	1
x_3	1.8	0.2
x_4	5.6	50
x_5	-0.5	34
x_6	0.4	19
x_7	1.1	11

Distance between vectors

- We are interested in some measure of distance between vectors representing separate entities.
- L_2 norm between two vectors where x_i and x_j are m dimensional vectors.

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

	x_1	x_2	x_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

- We are interested in some measure of distance between vectors representing separate entities.
- L_2 norm between two vectors where x_i and x_j are m dimensional vectors.

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^3 (x_{1k} - x_{2k})^2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2} = 22$$

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

$D =$

Distance between vectors

L_1 norm of the vector x with m dimensions (columns/attributes)

$$\|x\|_1 = \sqrt[1]{\sum_{k=1}^m |x_k|^1} = (|x_1|^1 + |x_2|^1 + \cdots + |x_m|^1)^1$$

Basically, L_1 norm of a vector is the sum of the its component values.

Distance between vectors

L_1 norm of the vector x_i and x_j with m dimensions
(columns/attributes)

$$\|x_i - x_j\|_1 = \sqrt[1]{\sum_{k=1}^m |x_{ik} - x_{jk}|^1} = \sum_{k=1}^m |x_{ik} - x_{jk}|$$

Distance between vectors

L_1 norm of the vector x_1 and x_2

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

$$\begin{aligned}\|x_1 - x_2\|_1 &= \sqrt[1]{\sum_{k=1}^3 |x_{1k} - x_{2k}|^1} = \sum_{k=1}^3 |x_{1k} - x_{2k}| = |x_{11} - x_{21}| + |x_{12} - x_{22}| + |x_{13} - x_{23}| \\ &= |0.2 - 0.4| + |23 - 1| + |5.7 - 5.4| = 22.5\end{aligned}$$

What does L_1 norm between two vectors mean?

- L_1 norm is the sum of absolute values of the components of the vector.
- This is also called Manhattan distance/norm or taxicab norm.



L_p norm

- L_p norm of two vectors x with m dimensions is defined as follows:
- $\|x\|_p = \sqrt[p]{\sum_{k=1}^m |x_k|^p} = (|x_1|^p + |x_2|^p + \cdots + |x_m|^p)^{\frac{1}{p}}$
- This is defined for any integer p greater than 0.

L_p norm

- L_p norm of two vectors x_i, x_j with m dimensions is defined as follows:
- $\|x_i - x_j\| = \sqrt[p]{\sum_{k=1}^n |x_{ik} - x_{jk}|^p}$
- Let's do an example where $p = 4$. . (Find the distance between x_1, x_2 ; In class activity)

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

L_p norm

- L_p norm of two vectors x_i, x_j with m dimensions is defined as follows:
- $\|x_i - x_j\| = \sqrt[p]{\sum_{k=1}^n |x_{ik} - x_{jk}|^p}$
- Let's do an example where $p = 4$. (Find the distance between x_1, x_2 ; In class activity)

• $D =$

	X_1	X_2	X_3
x_1	0.2	23	5.7
x_2	0.4	1	5.4
x_3	1.8	0.5	5.2
x_4	5.6	50	5.1
x_5	-0.5	34	5.3
x_6	0.4	19	5.4
x_7	1.1	11	5.5

$$\begin{aligned}
 \|x_1 - x_2\| &= \sqrt[4]{\sum_{k=1}^3 |x_{1k} - x_{2k}|^4} \\
 &= \sqrt[4]{|0.2 - 0.4|^4 + |23 - 1|^4 + |5.7 - 5.4|^4} \\
 &= 22.0000002277 \\
 &\cong 22
 \end{aligned}$$

L_∞ norm

- L_∞ norm of two vectors x_i, x_j with m dimensions is defined as follows:

$$\|x_i - x_j\|_\infty = \lim_{p \rightarrow \infty} \sqrt[p]{\sum_{k=1}^m |x_{ik} - x_{jk}|^p} = \max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \dots, |x_{im} - x_{jm}|)$$

- Basically, the largest component of the difference of the vectors.

Dot product of two vectors

- Dot product: Given two vectors a, b of m dimensions

$$a \cdot b = a^T b = (a_1 \quad a_2 \quad \cdots \quad a_m) \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \sum_{i=1}^m a_i b_i$$

Dot product of two vectors

- Dot product between x_3 and x_5 :

	x_1	x_2	x_3
$D =$	0.2	23	5.7
	0.4	1	5.4
	1.8	0.5	5.2
	5.6	50	5.1
	-0.5	34	5.3
	0.4	19	5.4
	1.1	11	5.5

$$\begin{aligned}x_3 \cdot x_5 &= \sum_{k=1}^3 x_{3k} x_{5k} \\&= x_{31} x_{51} + x_{32} x_{52} + x_{33} x_{53} \\&= (1.8)(-0.5) + (0.5)(34) + (5.2)(5.3) \\&\quad 43.66\end{aligned}$$

Dot product of two vectors (Geometric explanation)

- Dot product: Given two vectors a, b of m dimensions

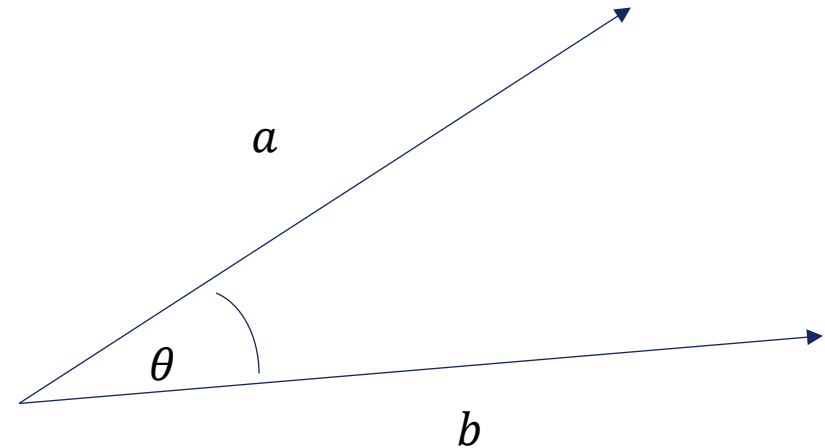
$$a \cdot b = a^T b = (a_1 \quad a_2 \quad \cdots \quad a_m) \times \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \sum_{i=1}^m a_i b_i$$

$$= \|a\|_2 \|b\|_2 \cos \theta$$

θ is the angle between two vectors a, b .

Cosine of the angle θ between two vectors a, b is calculated as follows:

$$\cos \theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$$



- Example: cosine of the angle between x_2, x_3 is:

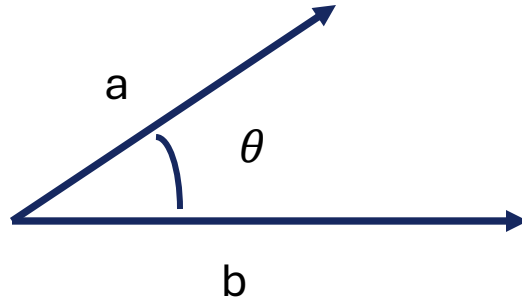
		X_1	X_2	X_3	
	x_1	0.2	23	5.7	
	x_2	0.4	1	5.4	
	x_3	1.8	0.5	5.2	
	x_4	5.6	50	5.1	
	x_5	-0.5	34	5.3	
	x_6	0.4	19	5.4	
	x_7	1.1	11	5.5	

$$D = \frac{x_2 \cdot x_3}{\|x_2\|_2 \|x_3\|_2}$$

$$= \frac{(0.4 \ 1 \ 5.4)(1.8 \ 0.5 \ 5.2)}{\sqrt{0.4^2 + 1^2 + 5.4^2} \sqrt{1.8^2 + 0.5^2 + 5.2^2}}$$

$$= 0.96$$

Distance between vectors



$$\cos \theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$$

$$\cos \theta_1 \approx 1$$



$$\cos \theta_1 \approx 0$$



$$\cos \theta_1 = -1$$



What if we have categorical attributes?

	X_1	X_2	X_3	X_4
x_1	0.2	23	A	5.7
x_2	0.4	1	B	5.4
x_3	1.8	0.5	D	5.2
x_4	5.6	50	C	5.1
x_5	-0.5	34	F	5.3
x_6	0.4	19	G	5.4
x_7	1.1	11	A	5.5

$$\begin{aligned}\|x_1 - x_2\|_2 &= \sqrt{\sum_{k=1}^4 (x_{1k} - x_{2k})^2} \\ &= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2} \\ &= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (A - B)^2 + (5.7 - 5.4)^2}\end{aligned}$$

What can we do about this?

What if we have categorical attributes?

Integer encoding

	X_1	X_2	X_3	X_4
x_1	0.2	23	1	5.7
x_2	0.4	1	2	5.4
x_3	1.8	0.5	4	5.2
x_4	5.6	50	3	5.1
x_5	-0.5	34	6	5.3
x_6	0.4	19	7	5.4
x_7	1.1	11	1	5.5

$$\begin{aligned}\|x_1 - x_2\|_2 &= \sqrt{\sum_{k=1}^4 (x_{1k} - x_{2k})^2} \\ &= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2} \\ &= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (1 - 2)^2 + (5.7 - 5.4)^2} = 22.02\end{aligned}$$

What if we have categorical attributes?

	X_1	X_2	X_3	X_4
x_1	0.2	23	1	5.7
x_2	0.4	1	2	5.4
x_3	1.8	0.5	4	5.2
x_4	5.6	50	3	5.1
x_5	-0.5	34	6	5.3
x_6	0.4	19	7	5.4
x_7	1.1	11	1	5.5

$$\begin{aligned}\|x_1 - x_2\|_2 &= \sqrt{\sum_{k=1}^4 (x_{1k} - x_{2k})^2} \\ &= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2} \\ &= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (1 - 2)^2 + (5.7 - 5.4)^2} = 22.02\end{aligned}$$

- This is appropriate if the categorical attribute is ordinal (there is an order)
- Not a good approach for nominal attributes
- Easy to implement

One-Hot encoding

- Appropriate for nominal data.
- Create a binary column for each possible value in the category.
 - Ex: X_3 is replaced by columns $X_{3A}, X_{3B}, X_{3C}, X_{3D}, X_{3F}, X_{3G}$
 - If category A is present in the row, then the observed value for the new column X_{3A} is 1, and all the other new columns are 0.

	X_1	X_2	X_3	X_4		X_1	X_2	X_{3A}	X_{3B}	X_{3C}	X_{3D}	X_{3F}	X_{3G}	X_4	
x_1	0.2	23	A	5.7		x_1	0.2	23	1	0	0	0	0	5.7	
x_2	0.4	1	B	5.4		x_2	0.4	1	0	1	0	0	0	5.4	
x_3	1.8	0.5	D	5.2		x_3	1.8	0.5	0	0	0	1	0	5.2	
x_4	5.6	50	C	5.1		x_4	5.6	50	0	0	1	0	0	5.1	
x_5	-0.5	34	F	5.3		x_5	-0.5	34	0	0	0	0	1	5.3	
x_6	0.4	19	G	5.4		x_6	0.4	19	0	0	0	0	0	1	5.4
x_7	1.1	11	A	5.5		x_7	1.1	11	1	0	0	0	0	0	5.5

One-Hot encoding

	X_1	X_2	X_{3A}	X_{3B}	X_{3C}	X_{3D}	X_{3E}	X_{3G}	X_4
x_1	0.2	23	1	0	0	0	0	0	5.7
x_2	0.4	1	0	1	0	0	0	0	5.4
x_3	1.8	0.5	0	0	0	1	0	0	5.2
x_4	5.6	50	0	0	1	0	0	0	5.1
x_5	-0.5	34	0	0	0	0	1	0	5.3
x_6	0.4	19	0	0	0	0	0	1	5.4
x_7	1.1	11	1	0	0	0	0	0	5.5

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^9 (x_{1k} - x_{2k})^2}$$

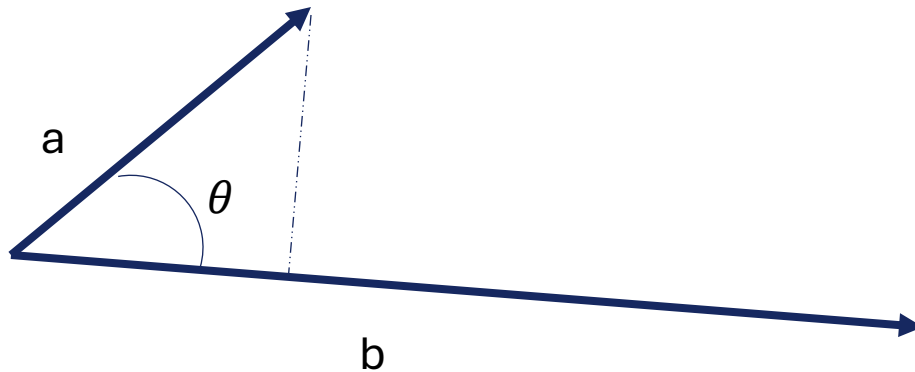
$$=$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (1 - 0)^2 + (0 - 1)^2 + (0 - 0)^2 + (0 - 0)^2 + (0 - 0)^2 + (0 - 0)^2}$$

$$= 22.05$$

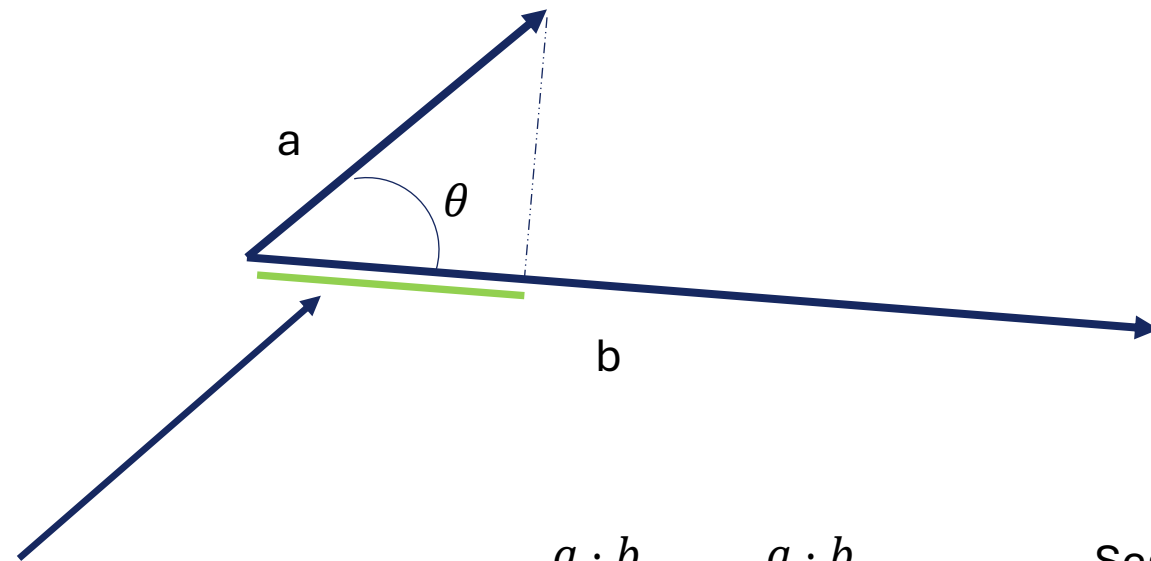
Putting it together

- Application: How far is a from the line through b ?
 - How long is the dotted line?



Putting it together

- Application: How far is a from the line through b ?
 - How long is the dotted line?

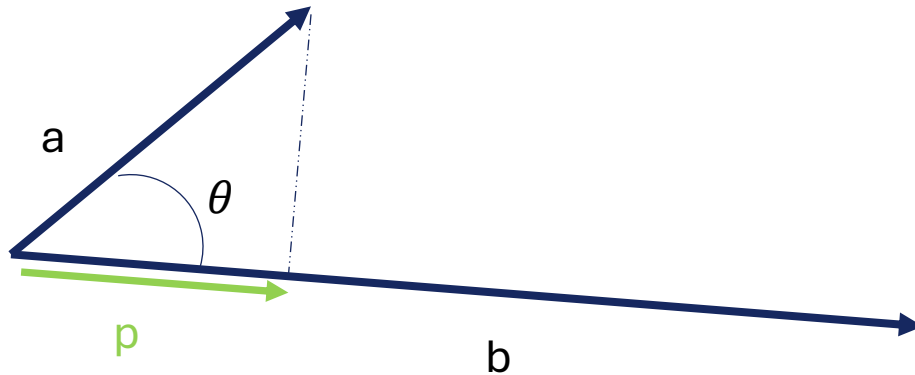


$$a_b = \|a\|_2 \cos \theta = \|a\|_2 \cdot \frac{a \cdot b}{\|a\|_2 \|b\|_2} = \frac{a \cdot b}{\|b\|_2}$$

Scalar projection of a in direction of b

Putting it together

- Application: How far is a from the line through b ?
 - How long is the dotted line?

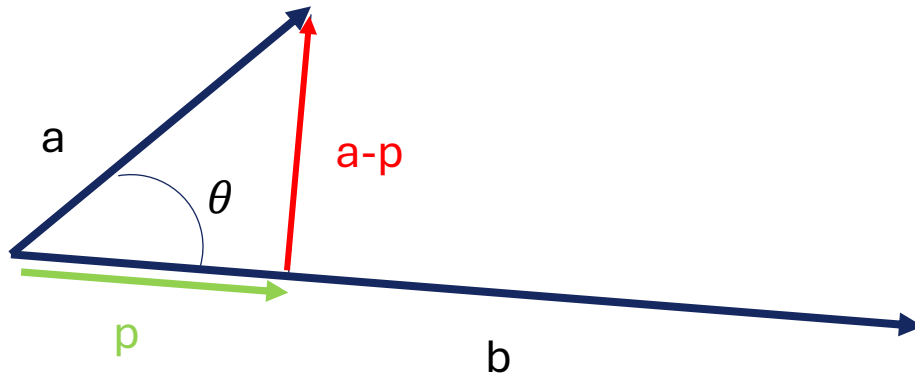


$$\cos \theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$$

$$p = \frac{a \cdot b}{\|b\|^2} b = (\|a\| \cos \theta) \frac{b}{\|b\|} = \left(\frac{a \cdot b}{\|b\|^2} \right) b$$

Putting it together

- Application: How far is a from the line through b ?
 - How long is the dotted line?



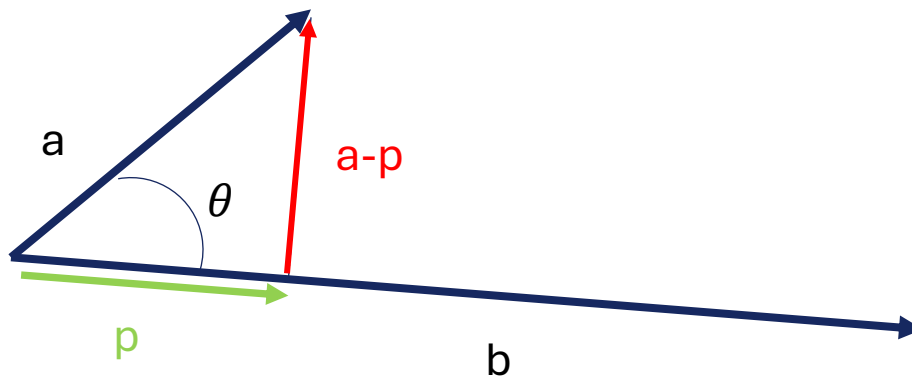
$$\cos \theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$$

$$p = \frac{a_b b}{\|b\|} = (\|a\| \cos \theta) \frac{b}{\|b\|} = \left(\frac{a \cdot b}{\|b\|^2} \right) b$$

Putting it together

- Application: How far is a from the line through b ?
 - How long is the dotted line?

Answer: $\|a - p\|_2$



$$\cos \theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$$

$$p = \frac{a \cdot b}{\|b\|^2} b = (\|a\| \cos \theta) \frac{b}{\|b\|} = \left(\frac{a \cdot b}{\|b\|^2} \right) b$$