Machine Learning, Artificial Intelligence, and Big Data Analytics (IL, 4th Semester)

Lecture 6

FH KREMS
UNIVERSITY OF APPLIED SCIENCES / AUSTRIA

Agenda

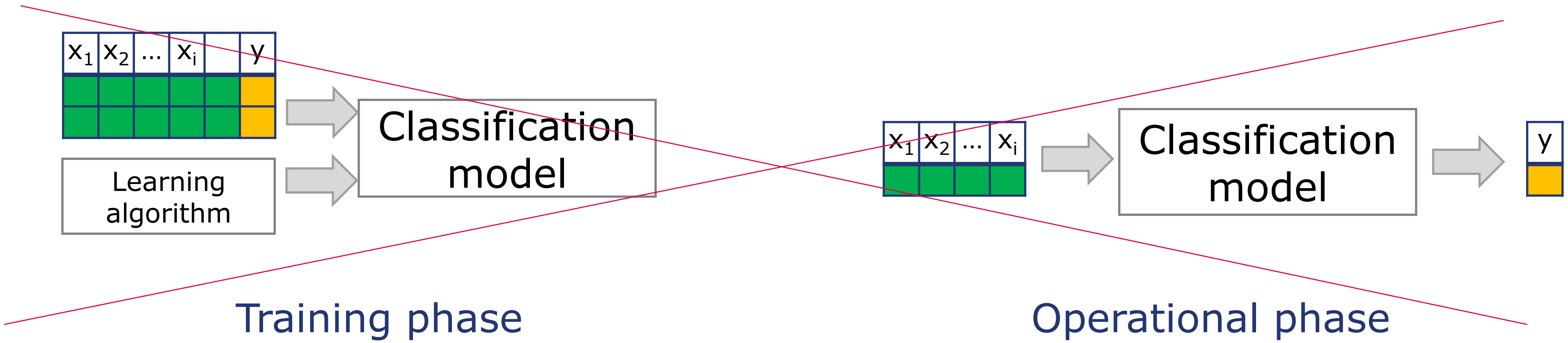
- K-Nearest Neighbors

- Distance metrics

kNN

- A simple classification technique.
- It does not even need training…

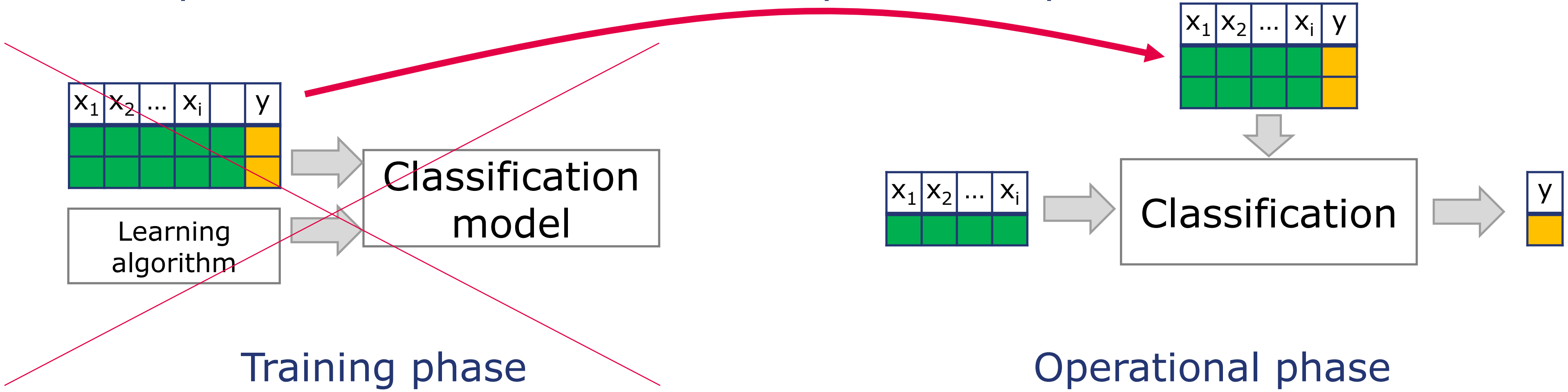
  … but it requires a bit more effort in the operational phase.



kNN

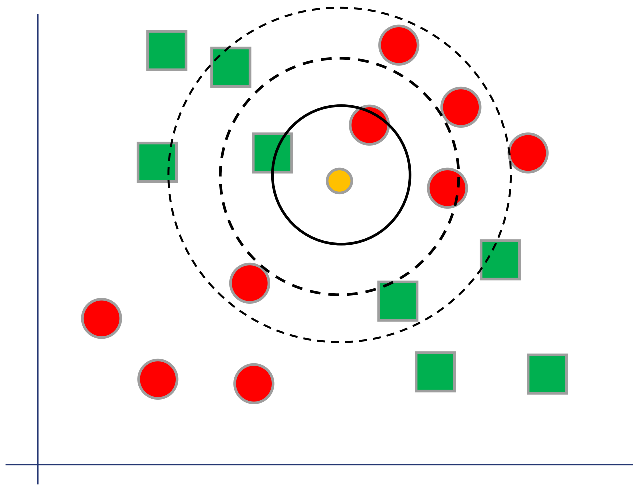
- A simple classification technique.
- It does not even need training…

  … but it requires a bit more effort in the operational phase.



kNN

Basic concept

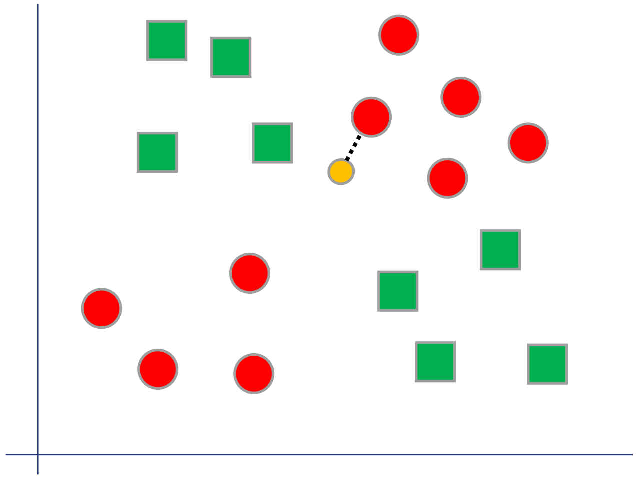


- Classify a sample based on its neighbors
- When we want to determine the label of a sample, we look at the label of its closest neighbors.
- ***k*** is the number of nearest neighbors to consider

kNN

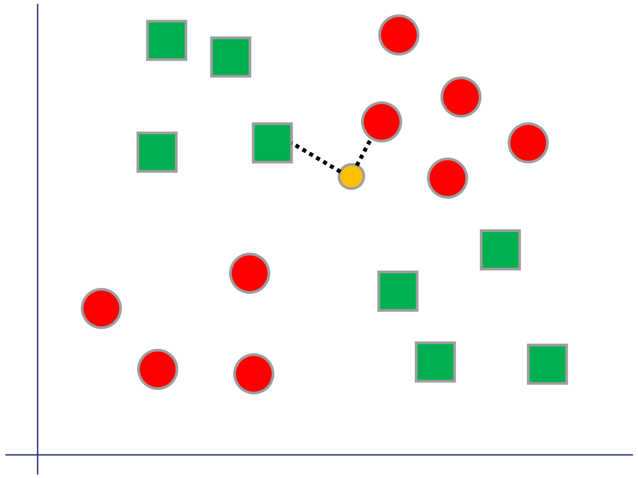
Basic concept

K=1



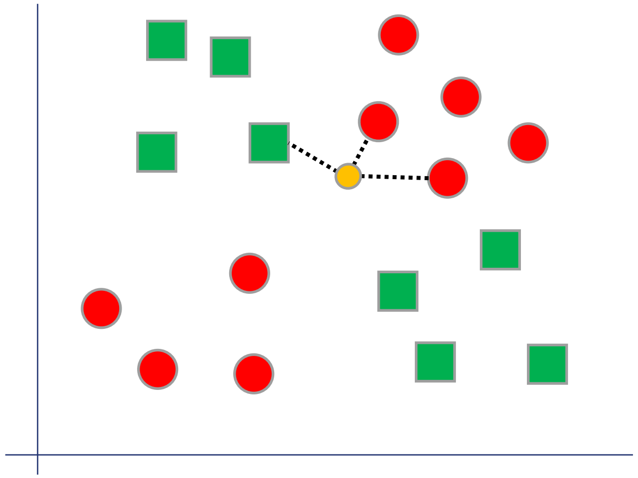
Red circle

K=2



Red - (Tie) Random or the closest

K=3

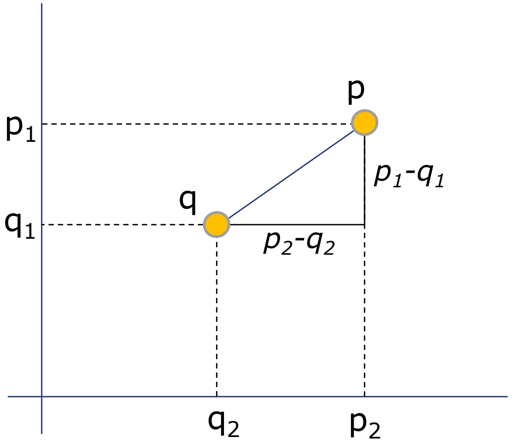


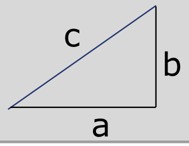
Red circle – (Majority vote)

Distance and similarity measures

How do we measure the distance between two samples?

- The Euclidean distance: The first choice in case of numerical features





$c = \sqrt{a^2 + b^2}$ Pythagoras (mid-school math)

2D $d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$

3D $d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$

$\vdots$

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

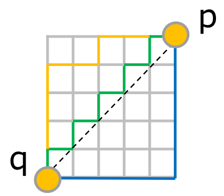
Distance and similarity measures

How do we measure the distance between two samples?

- Minkowski. A generalized distance metric

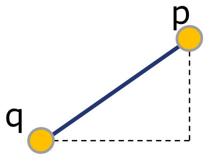
$$d(p,q) = \sqrt[r]{\sum_{i=1}^{n} |p_i - q_i|^r}$$

r = 1 ⇒ $d(p,q) = \sum_{i=1}^{n} (p_i - q_i)$



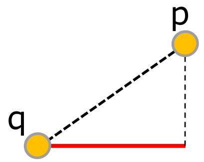
Manhattan distance

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



Euclidean distance

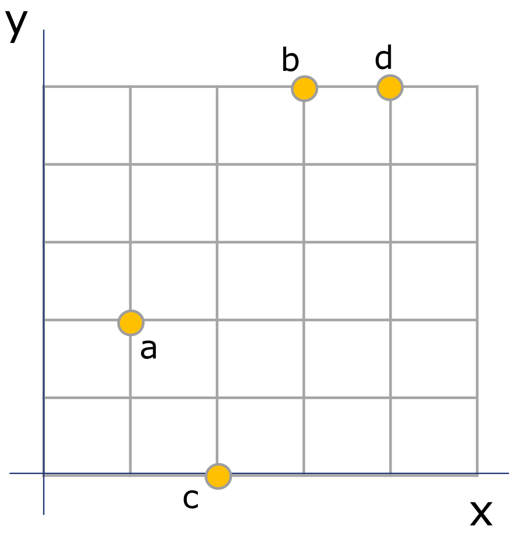
r = ∞ ⇒ $d(p,q) = \max_{\mathrm{i}} (p_i - q_i)$



Chebychev distance

Distance and similarity measures

L-norm distance example



	x	y
a	1	2
b	3	5
c	2	0
d	4	5

data

	a	**b**	**c**	**d**
a	0			
b	5	0		
c	3	6	0	
d	6	1	7	0

Manhattan (L1)

	a	**b**	**c**	**d**
a	0			
b	3.61	0		
c	2.24	5.1	0	
d	4.24	1	5.39	0

Euclidean (L2)

	a	**b**	**c**	**d**
a	0			
b	3	0		
c	2	5	0	
d	3	1	5	0

Chebyshev ($L\infty$)

Distance and similarity measures

Limitations of Euclidean distance → Different scales

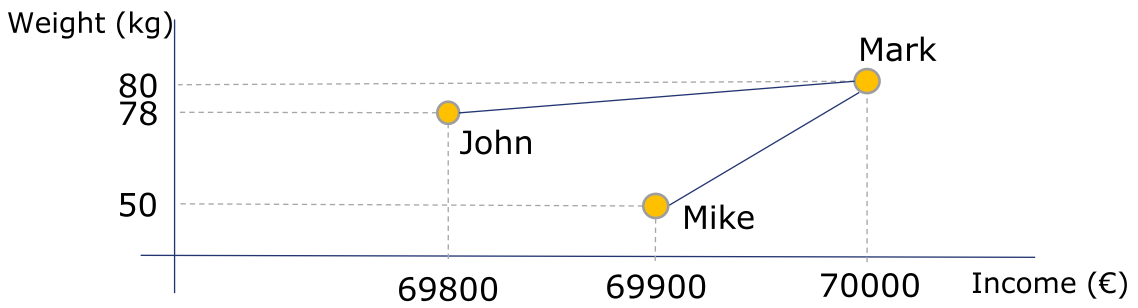
	Weight	Height	Diabetes
A	85	175	Yes
B	65	170	No
C	70	180	No

$$d(A,B) = \sqrt{(175-170)^2 + (85-65)^2} = 20.6$$
$$d(A,C) = \sqrt{(175-180)^2 + (85-70)^2} = 15.8$$
$$d(B,C) = \sqrt{(170-180)^2 + (65-70)^2} = 11.2$$

- Euclidean distance works good with numeric variables but let’s look at this example…

	Weight	Salary	Diabetes
Mark	80	70k	Yes
Mike	50	69.9k	No
John	78	69.8k	No

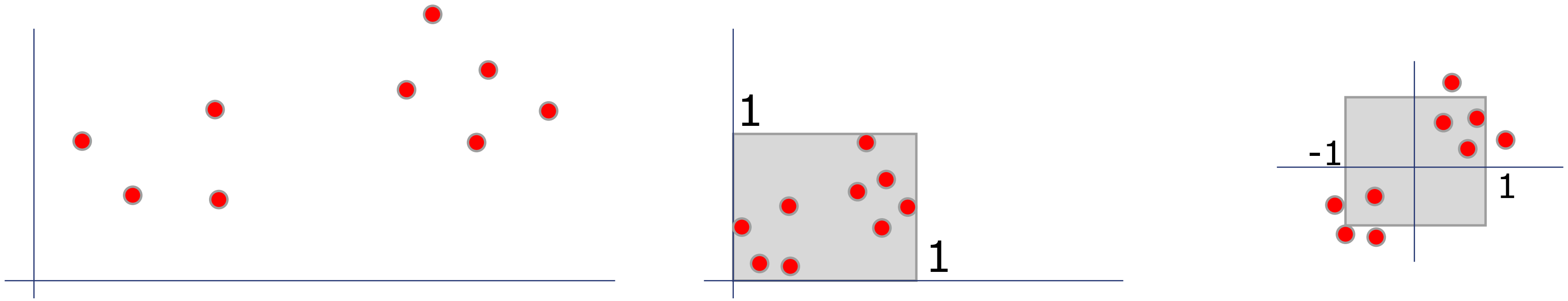
Who is more similar to Mark? Mike or John?



Distance and similarity measures

Rescale

- Variables on larger scale dominate the distance! **Solution: Rescale all variables**
- Two common ways:
  - Normalization to [0,1] → $x' = (x - \mathit{min}(x)) / (\mathit{max}(x) - \mathit{min}(x))$
  - Standardization: zero mean and unit std dev → $x' = (x - \mathit{mean}(x)) / \mathit{sd}(x)$



Original dataset

After normalization

After standardization

Distance and similarity measures

Limitations of Euclidean distance → Categorical variables

	Weight	Height	Diabetes
A	85	175	Yes
B	65	170	No
C	70	180	No

$$d(A,B) = \sqrt{(175-170)^2 + (85-65)^2} = 20.6$$

$$d(A,C) = \sqrt{(175-180)^2 + (85-70)^2} = 15.8$$

$$d(B,C) = \sqrt{(170-180)^2 + (65-70)^2} = 11.2$$

	Weight	Height	Gender	Country	Diet	Income	Diabetes
A	85	175	M	AT	Vegan	70K	Yes
B	65	170	F	IT	Vegetarian	59K	No
C	70	180	M	FR	Omnivorous	50K	No

?

Distance and similarity measures

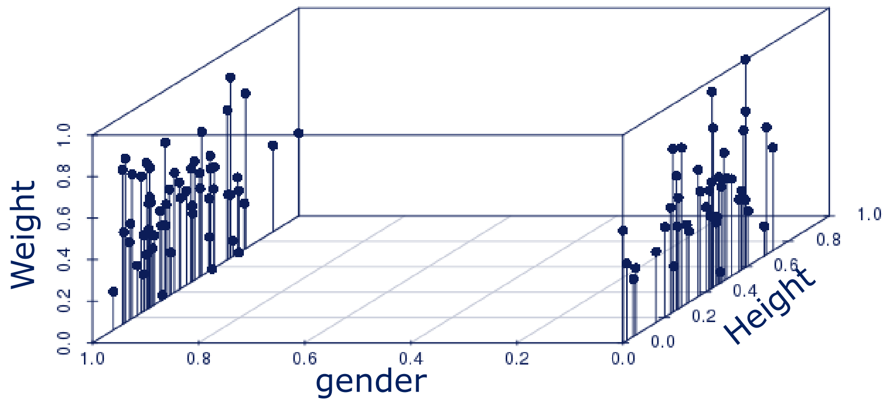
What about categorical variables?

	Weight	Height	Gender	Income
A	85	175	M	70K
B	65	170	F	59K
C	70	180	M	50K

	Weight	Height	isMale	Income
A	85	175	1	70K
B	65	170	0	59K
C	70	180	1	50K

	Weight	Height	isMale	Income
A	1	0.5	1	1
B	0	0	0	0.45
C	0.25	1	1	0

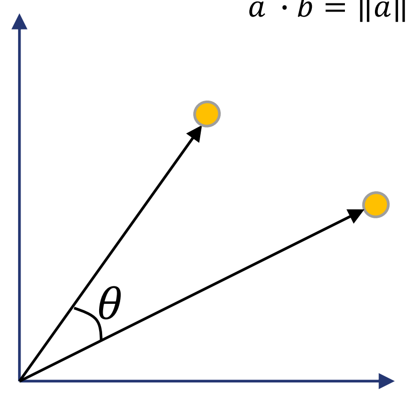
- One could one hot encode or dummify categorical variables. However, **the distance will be dominated by the one-hot-encoded variables**.
- Solution: Weight each variable or look at other distance metrics.



Distance and similarity measures

Cosine similarity

$$\vec{a} \cdot \vec{b} = \|\vec{a}\|\|\vec{b}\| \cos\theta$$



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

$$\cos 0° = 1$$

$$\cos 90° = 0$$

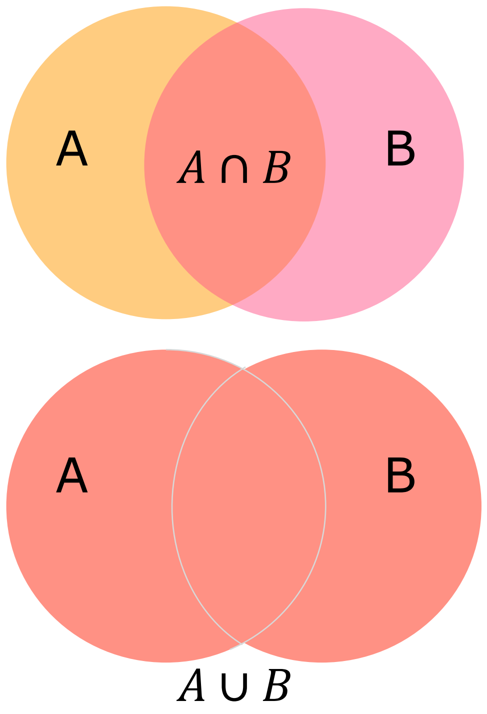
- It measures only different orientations but not different magnitude of a vector.
- 1 - cosine_sim = Cosine Dissimilarity
- It is not a proper distance metrics, but it can be very useful in kNN.

Distance and similarity measures

Jaccard index (or Tanimoto coefficient)

- Measure of similarity
- Only for binary features!

$$\text{Jaccard index} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{A + B - |A \cap B|}$$



Distance and similarity measures

Example – A binary Customer/Product matrix (very common problem)

Products

customers

	p_1	p_2	p_3	p_4	p_5	…	p_i	…	p_n
c_1	.	1	1	.	1	1	1	.	.
c_2	.	.	1	1	1	1	1	.	1
c_3	1	.	.	1	.	1	1	.	.
…									
c_m									

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
c_1	.	1	1	.	1	1	1	.	.

Jaccard index $= \dfrac{|A \cap B|}{|A \cup B|} = \dfrac{|A \cap B|}{A + B - |A \cap B|} = \dfrac{4}{7} = 0.57$

Number of features where both observations are ones (i.e., p_3, p_5, p_6, p_7)

Number of features where at least one observation is one (i.e., $p_2, p_3, p_4, p_5, p_6, p_7, p_9$ → 7)
or equivalently:
Number of 1s in C_1 + num of 1s in C_2 - features where both observations are ones (i.e., 5+6-4=7)

Cosine simil. $= \dfrac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \dfrac{4}{\sqrt{5} * \sqrt{6}} = \dfrac{4}{5.48} = 0.73$

Number of features where both observations are ones (i.e., p_3, p_5, p_6, p_7)

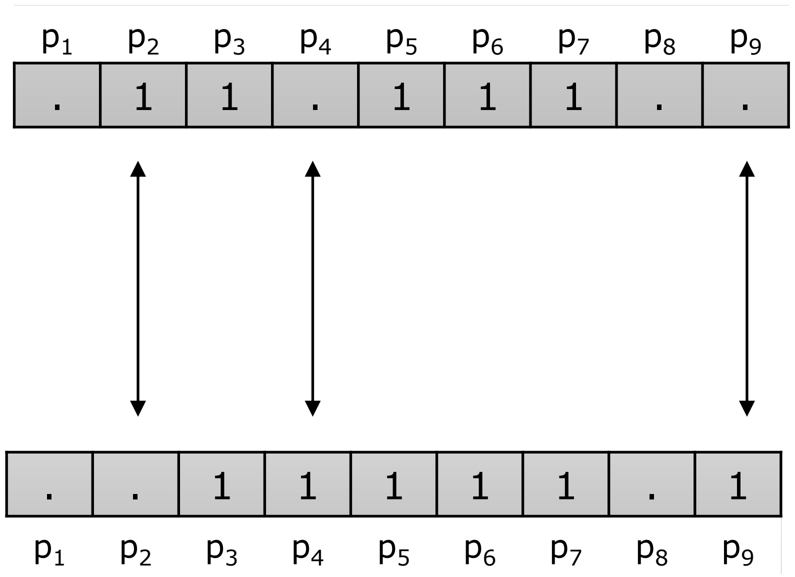
Magnitude of C_1 * magnitude of C_2

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
c_2	.	.	1	1	1	1	1	.	1

Distance and similarity measures

Hamming distance

- Straightforward: the number of symbols that differ between two vectors



Hamming distance = 3

Distance and similarity measures

Summary

		Binary
Manhattan	$d(A,B) = \sum_{i=1}^{n}(A_i - B_i)$	$d(A,B) = a + b - 2c$
Euclidean	$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$	$d(A,B) = \sqrt{a + b - 2c}$
Cosine	$s(A,B) = \frac{\vec{a} \cdot \vec{b}}{\lVert\vec{a}\rVert \lVert\vec{b}\rVert}$	$s(A,B) = \frac{c}{\sqrt{ab}}$
Tanimoto	---------	$s(A,B) = \frac{c}{a+b-c}$
Hamming	---------	*d(A,B)=length(xor(A,B))*
		a=num. of 1s in A. b=num. of 1s in B. c=num. of common 1s.