Distributed word representations: vector comparison

Chris Potts
Stanford Linguistics

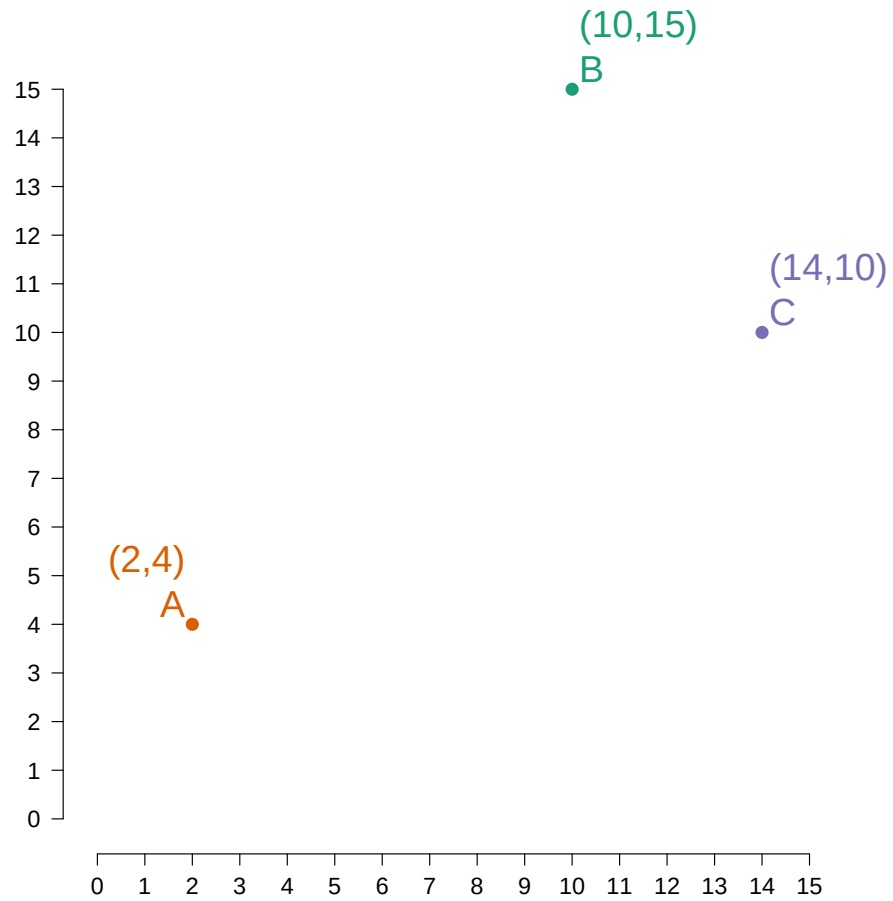
CS 244U: Natural language understanding

Running example

	d_x	d_y
A	2	4
B	10	15
C	14	10

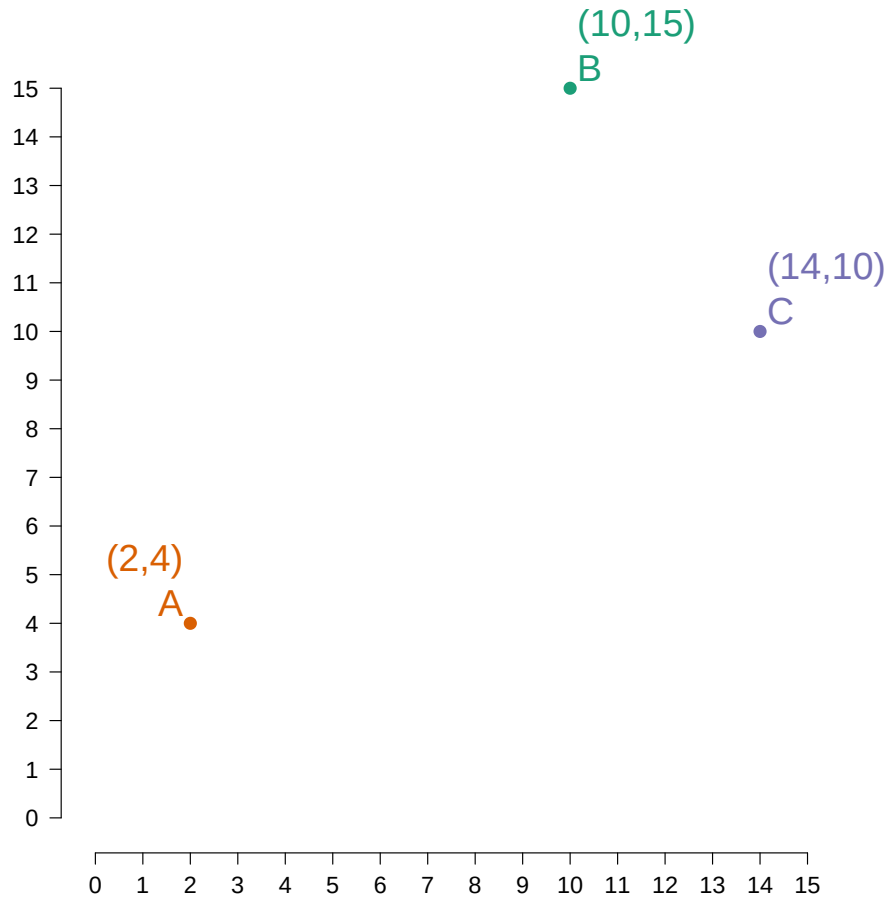
Running example

	d_x	d_y
A	2	4
B	10	15
C	14	10



Running example

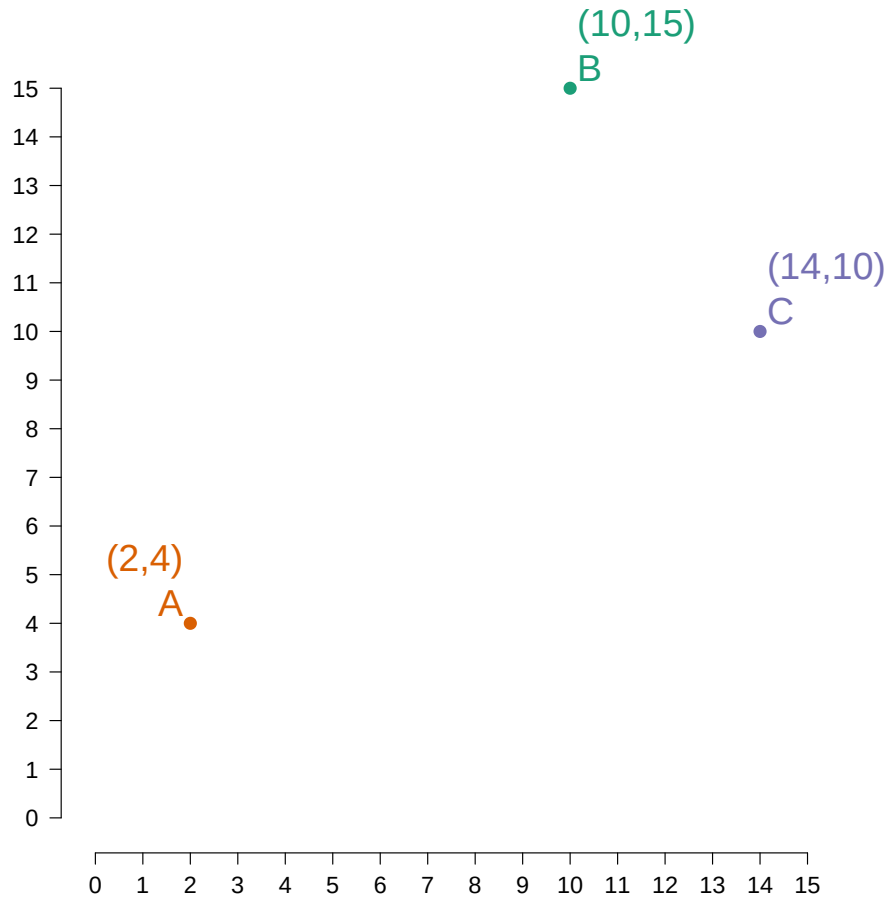
	d_x	d_y
A	2	4
B	10	15
C	14	10



- Focus on distance measures

Running example

	d_x	d_y
A	2	4
B	10	15
C	14	10



- Focus on distance measures
- Illustrations with row vectors

Euclidean distance

Definition

Between vectors u and v of dimension n:

$$\sqrt{\sum_{i=1}^{n} |u_i - v_i|^2}$$

Euclidean distance

Definition

Between vectors u and v of dimension n:

$$\sqrt{\sum_{i=1}^{n} |u_i - v_i|^2}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10

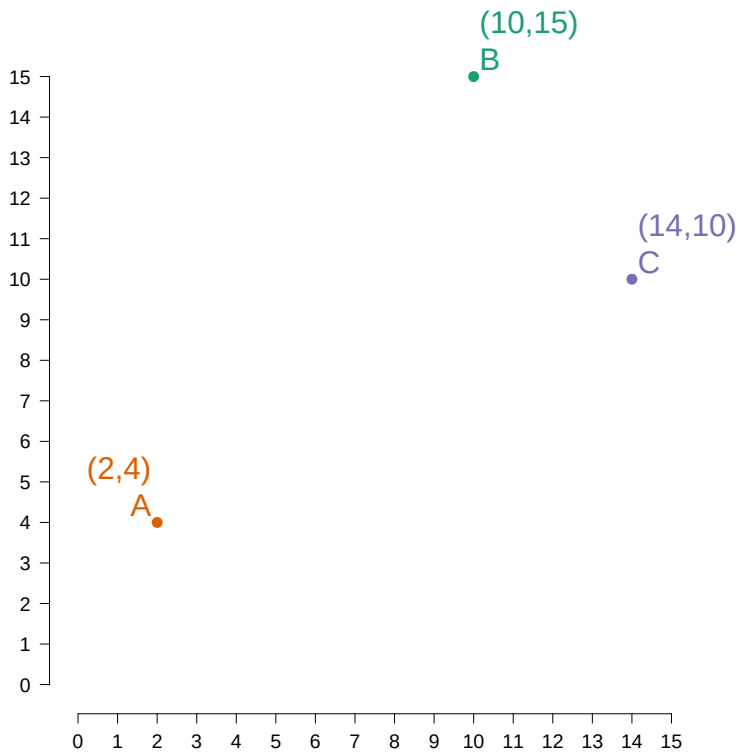
Euclidean distance

Definition

Between vectors u and v of dimension n:

$$\sqrt{\sum_{i=1}^{n} |u_i - v_i|^2}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



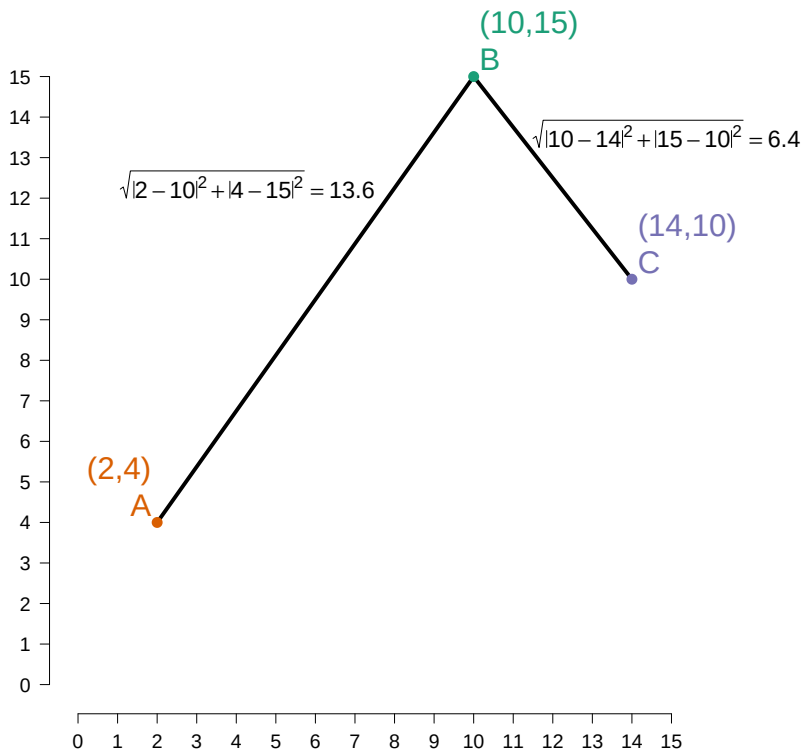
Euclidean distance

Definition

Between vectors u and v of dimension n:

$$\sqrt{\sum_{i=1}^{n} |u_i - v_i|^2}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



Vector L2 (length) normalization

Definition

Given a vector u of dimension n, the normalization of u is a vector $\hat{u}$ of dimension n obtained by dividing each element of u by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

Vector L2 (length) normalization

Definition

Given a vector u of dimension n, the normalization of u is a vector $\hat{u}$ of dimension n obtained by dividing each element of u by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

	d_x	d_y
A	2	4
B	10	15
C	14	10

Vector L2 (length) normalization

Definition

Given a vector u of dimension n, the normalization of u is a vector $\hat{u}$ of dimension n obtained by dividing each element of u by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

	d_x	d_y
A	2	4
B	10	15
C	14	10

L2 norm the rows
$\Rightarrow$

	d_x	d_y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58

Vector L2 (length) normalization

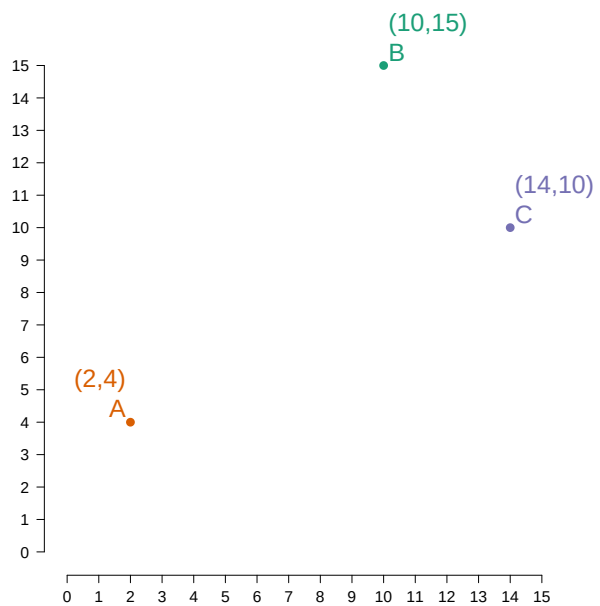
Definition

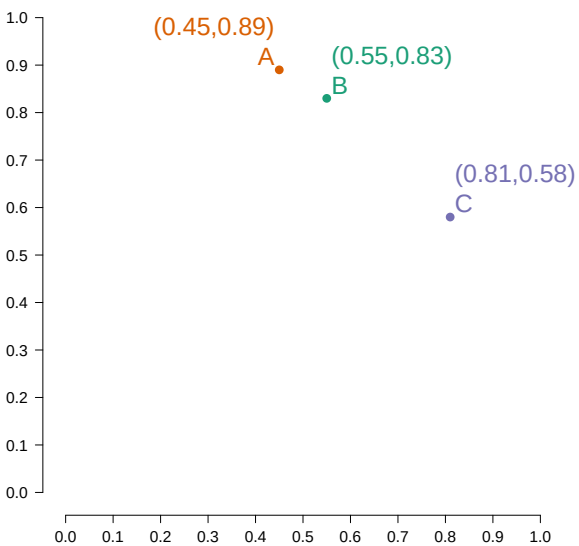
Given a vector u of dimension n, the normalization of u is a vector $\hat{u}$ of dimension n obtained by dividing each element of u by $\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2}$.

	d_x	d_y
A	2	4
B	10	15
C	14	10

L2 norm the rows $\Rightarrow$

	d_x	d_y
A	0.45	0.89
B	0.55	0.83
C	0.81	0.58





Cosine distance

Definition (Cosine distance)

Between vectors u and v of dimension n:

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

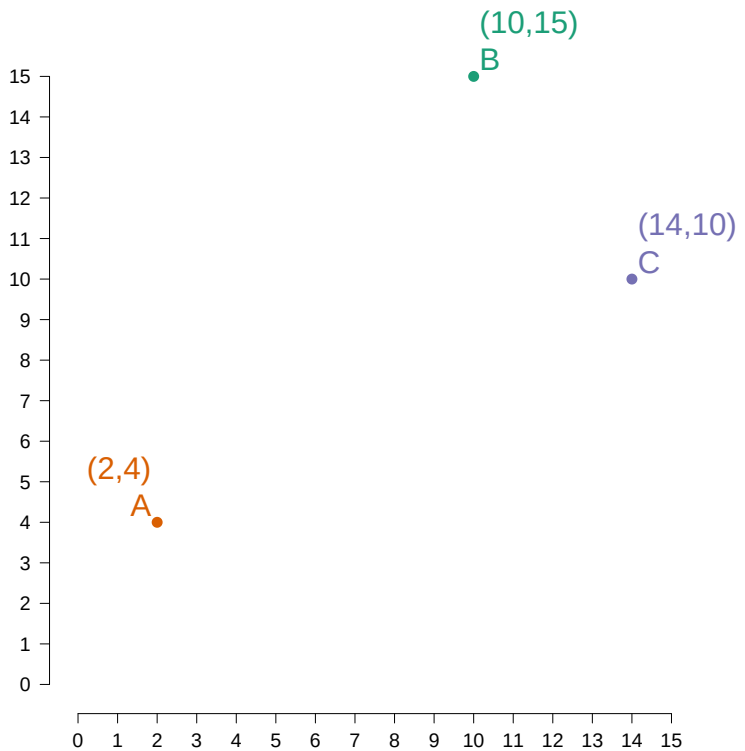
Cosine distance

Definition (Cosine distance)

Between vectors u and v of dimension n:

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



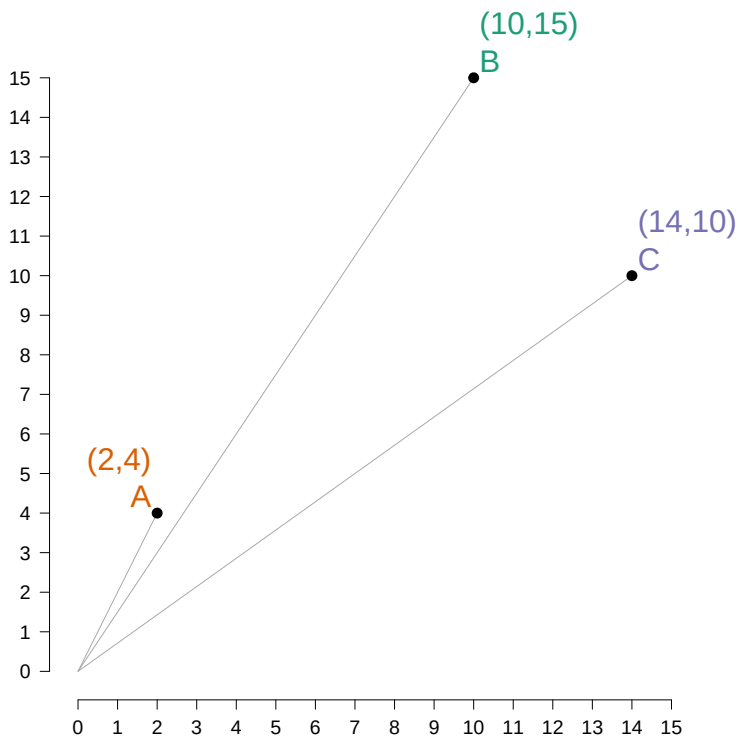
Cosine distance

Definition (Cosine distance)

Between vectors u and v of dimension n:

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



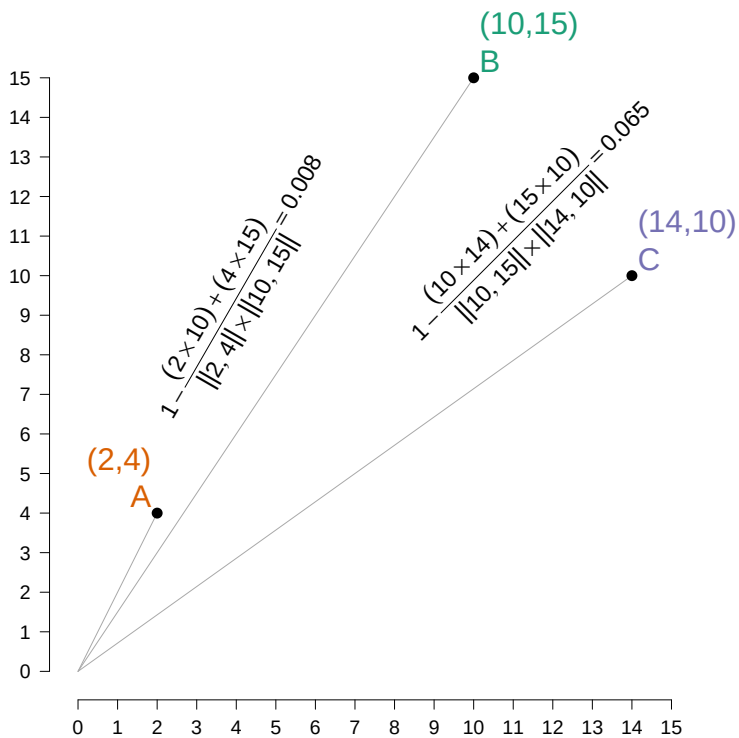
Cosine distance

Definition (Cosine distance)

Between vectors u and v of dimension n:

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



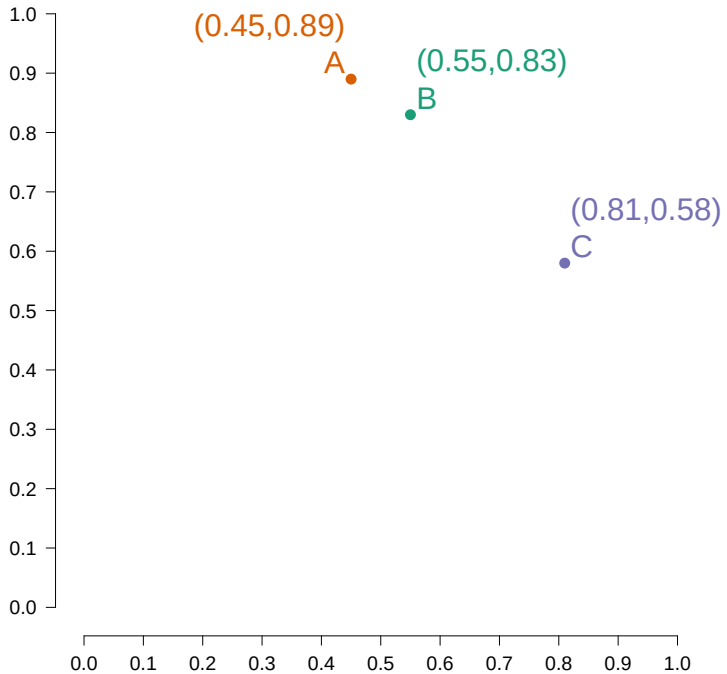
Cosine distance

Definition (Cosine distance)

Between vectors u and v of dimension n:

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



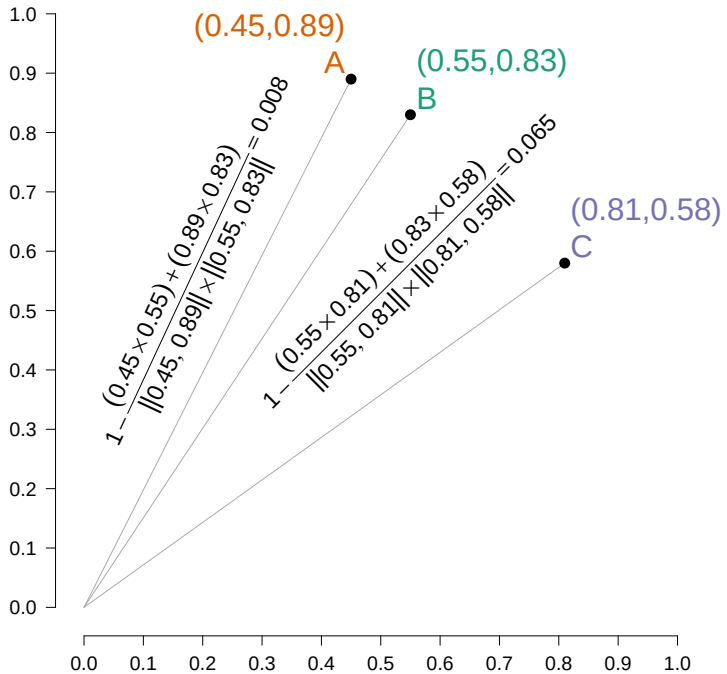
Cosine distance

Definition (Cosine distance)

Between vectors u and v of dimension n:

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

	d_x	d_y
A	2	4
B	10	15
C	14	10



Deciding which comparison method to use

Deciding which comparison method to use

	d_x	d_y
A	2	4
B	10	15
C	14	10

Deciding which comparison method to use

	d_x	d_y
A	2	4
B	10	15
C	14	10

$$\|A\| = 4.47$$
$$\|B\| = 18.03$$
$$\|C\| = 17.20$$

Deciding which comparison method to use

	d_x	d_y
A	2	4
B	10	15
C	14	10

$$\|A\| = 4.47$$
$$\|B\| = 18.03$$
$$\|C\| = 17.20$$

	A and B closer than B and C?
Euclidean distance	No
Cosine distance	Yes

Other comparison methods

- Manhattan distance
- KL divergence
- Symmetric KL divergence
- KL divergence with skew
- Jensen–Shannon distance
- Matching coefficient
- Dice coefficient
- Jaccard coefficient