1 Vectors

1.1 Definitions

Dot product or inner product

$$\vec{v} \cdot \vec{w} = (v_1 w_1 + \ldots + v_n w_n) = \sum_{i=1}^{n} v_i w_i$$

Example We have three goods to buy and sell, their prices are (p_1, p_2, p_3) (price vector $\vec{p}$). The quantities we buy or sell are (q_1, q_2, q_3) (quantity vector $\vec{q}$, their values are positive when we sell and negative when we buy.) Selling the quantity q_1 at price p_1 brings in $q_1 p_1$. The total income is the *dot product*:

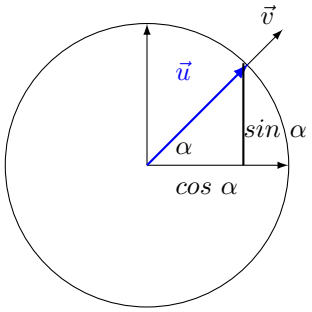
$$\vec{q} \cdot \vec{p} = (q_1, q_2, q_3) \cdot (p_1, p_2, p_3) = q_1 p_1 + q_2 p_2 + q_3 p_3$$

Length $||\vec{v}|| = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{\sum_{i=1}^{n} v_i^2}$

Unit vector is a vector whose length equals one.

$$\vec{u} = \frac{\vec{v}}{||\vec{v}||}$$

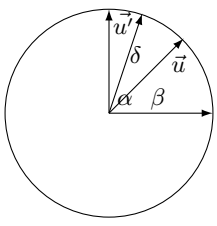
is a unit vector in the same direction as $\vec{v}$. (normalized vector)



$\vec{u} = \frac{\vec{v}}{||\vec{v}||} = (cos\ \alpha, sin\ \alpha)$

Cosine formula Given δ the angle formed by the two unit vectors $\vec{u}$ and $\vec{u'}$, s.t. $\vec{u} = (cos\ \beta, sin\ \beta)$ and $\vec{u'} = (cos\ \alpha, sin\ \alpha)$

$$\vec{u} \cdot \vec{u'} = (cos\ \beta)(cos\ \alpha) + (sin\ \beta)(sin\ \alpha) = cos(\beta - \alpha) = cos\ \delta$$



Given two arbitrary vectors v and w:

$$cos\ \delta = \frac{\vec{v}}{||\vec{v}||} \cdot \frac{\vec{w}}{||\vec{w}||}$$

The bigger the angle δ, the smaller is $cos\ \delta$; $cos\ \delta$ is never bigger than 1 (since we used unit vectors) and never less than -1. It's 0 when the angle is 90^o

Cosine Similarity

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x}}{||\vec{x}||} \cdot \frac{\vec{y}}{||\vec{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\ \sqrt{\sum_{i=1}^{n} y_i^2}}$$

$||\vec{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$

1.2 Exercises

a) Given the vector $\vec{v} = (1, 2, 3)$ and $\vec{w} = (4, 1, 5)$, compute their dot product: $\vec{v} \cdot \vec{w}$.

b) Given the vector $\vec{v}$ in (a) compute its length.

c) Find the unit vector, $\vec{u}$, in the same direction of $\vec{v}$ given in (a).

d) Given the vectors $\vec{v} = (1, 2)$, $\vec{w} = (4, 1)$ and $\vec{z} = (0, 4)$, find the cosine similarity between $\vec{v}$ and $\vec{w}$, $\vec{v}$ and $\vec{z}$. Which vector is more "similar" to $\vec{v}$?

1.3 Solutions

a) $\vec{v} \cdot \vec{w} = 4x1 + 2x1 + 3x5 = 4 + 2 + 15 = 21$

b) $||\vec{v}|| = \sqrt{1 + 4 + 9} = \sqrt{14}$

c) $\vec{u} = \frac{\vec{v}}{||\vec{v}||} = (\frac{1}{\sqrt{14}}, \frac{2}{\sqrt{14}}, \frac{3}{\sqrt{14}})$

d) $\vec{v}$ is more similar to $\vec{z}$ than to $\vec{w}$.

$cos(\vec{v}, \vec{w}) = \frac{\vec{v}}{||\vec{v}||} \cdot \frac{\vec{w}}{||\vec{w}||} = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}) \cdot (\frac{4}{\sqrt{17}}, \frac{1}{\sqrt{17}}) = 0.63$

$cos(\vec{v}, \vec{z}) = \frac{\vec{v}}{||\vec{v}||} \cdot \frac{\vec{z}}{||\vec{z}||} = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}) \cdot (0, 1) = 0.89$

2 Evaluation Measures: Accuracy, Precision, Recall, F-measure

Accuracy Percentage of documents correctly classified by the system.

Error Rate Inverse of accuracy. Percentage of documents wrongly classified by the system

Precision percentage of relevant documents correctly retrieved by the system (TP) with respect to all documents *retrieved by the system* (TP + FP). (how many of the retrieved books are relevant?)

Recall percentage of relevant documents correctly retrieved by the system (TP) with respect to all documents *relevant for the human* (TP + FN). (how many of the relevant books have been retrieved?)

F-Measure Combine in a single measure Precision (P) and Recall (R) giving a *global estimation of the performance* of an IR system

	Relevant	Not Relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not retrieved	False Negative (FN)	True Negative (TN)

Accuracy $\frac{\mathbf{TP + TN}}{TP+TN+FP+FN}$

Error Rate $\frac{\mathbf{FP+FN}}{TP+TN+FP+FN}$

Precision $\frac{TP}{TP+\mathbf{FP}}$

Recall $\frac{TP}{TP+\mathbf{FN}}$

F $\frac{2PR}{R+P}$

2.1 Exercise

a) In a collaction of 100 documents, 40 documents are relavant for a given search. Two IR systems (System I on the left and System II on the right) behave as following w.r.t. the given search and collection. Calculate the above measures.

	Relevant	Not Relevant
Retrieved	30	0
Not retrieved	10	60

	Relevant	Not Relevant
Retrieved	40	50
Not retrieved	0	10

2.2 Solutions

	Acc	ER	P	R	F
System I	0.90	0.1	1	0.44	0.85
System II	0.90	0.5	0.75	1	0.6

3 Purity of clusters

(From Wikipedia)

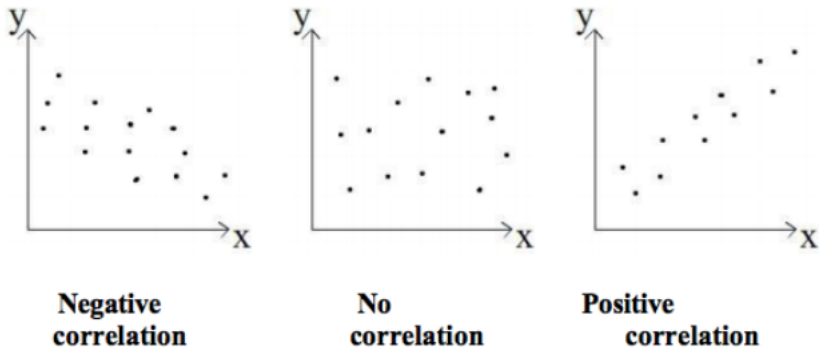
Purity is a measure of the extent to which clusters contain a single class. Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster. Now take the sum over all clusters and divide by the total number of data points. Formally, given some set of clusters M and some set of classes D, both partitioning N data points, purity can be defined as:

$\frac{1}{N} \sum_{m \in M} max_{d \in D} |m \cap d|$

4 Correlation coefficients

If we have studied the behaviour of some data w.r.t. more task/phenomena (variables), we can take a pair of such variables and see if they are correlated. In particular, we can check if one variable increases what happens to the other variable:

- the other variable has a tendency to decrease, then there is a negative correlation.
- the other variable does not tend to either increase or decrease, then there is no correlation.
- the other variable has a tendency to also increase, then there is a positive correlation



Taken from: `http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf` and `http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf`

To decide whether two variables are correlated, we can compute standard correlation coefficients.

Pearson's correlation coefficient is a statistical measure of the strength of a *linear* relationship between paired data, the values are:

$$-1 \leq r \leq 1$$

- positive values denote positive linear correlation
- negative values denote negative linear correlation
- a value of 0 denotes no linear correlation

- the close the value is to 1 or -1, the stronger the linear correlation.

The data must meet the following assumptions:

- interval or ratio level
- lineary related
- bivariante normally distributed.

If the data does not meet the above assumptions, then we should you Spearman's rank correlation:

Sperman Correlation coefficient is a statistical measure of the strenght of a *monotonic*[1] relationship between paired data.

- interval or ratio level or ordinal
- montonically related

[1] A monotonic function is a either never increase or never decreases as its indepdent variable increase.