Computing for Data Sciences

Lecture 1

Vectors

Vectors are indexed set of elements. They are an array of numbers, with each number representing its component in each dimension.

It can thus be represented as:

Vector, $X = (X_1, X_2, X_3, \ldots X_N)$,

Where N represents the number of dimensions. N is usually an integer between $[2, \infty)$.

The physical implication of assigning values to components of vector would be its degree of freedom. Taking an example, $3\ X_1 + 4\ X_2$ would represent a vector in two dimensions (considering X_1 and X_2 as independent non-zero vectors) and altering the values of X_1 and X_2 would cover the entire 2-dimensional plane. Since, both X_1 and X_2 can be specified independently of each other, its degree of freedom would be 2. However, if the expression is equated to a constant such as $3\ X_1 + 4\ X_2 = 7$, this would put a constraint on the values of X_1 and X_2 and thus reduce exactly one degree of freedom. On changing the values of X_1, the value of X_2 would be automatically defined and the vector could no longer encompass the entire plane as earlier. Now only one of the values can be specified independently and the degree of freedom is thus reduced to one. If on the other hand, the vector is equated to a variable such as in the expression $3\ X_1 + 4\ X_2 = Y$ (assuming no constraints on Y), the vector would retain its degree of freedom and both variables could once again be assigned values independently.

Thus to sum it up,

Degree of freedom, $X = (X_1, X_2, X_3, \ldots X_N) = \{$ *Y (variable), then degree=N*

K (constant), then degree=N-1 $\}$

Distance between vectors

Distance is an abstract concept. Attempts have been regularly made to define such abstract entities. For example, 1 was attempted to be defined as common property of all sets containing just a single entity. And Peano's axioms could be helpful in the operations domain.

To calculate distance, we first need to define a notion of distance. The function for distance can be different in each dimension, such as the common notion of distance between vectors $X(X_1)$and $Y(Y_1)$ in 1 Dimension is $| X_1 - Y_1 |$ whereas the distance between vectors $X(X_1, X_2)$ and $Y(Y_1, Y_2)$ is calculated by Euclidean formula as $\sqrt{\{(X_1 - Y_1)^2 + (X_2 - Y_2)^2\}}$.

The notion of distance can be defined differently in different cases. However, it needs to satisfy certain predefined rules which would make it consistent across dimensions. Since Distance function should also have domain and a range, it also needs a specification of this 'space' where its definition would be valid.

The distance defined over the space of all real numbers, $\mathbb{R}$ should satisfy the following rules:

Dist(X, X) → $\mathbb{R}$ where Dist(X, X) indicates distance between vector X and vector Y

1) Dist(X, Y) ≥ 0
2) Dist(X, X) = 0
3) Dist(X, Y) = Dist(Y, X)
4) Dist(X, Y) + Dist(Y, Z) ≥ Dist(X, Z) (Also known as the triangle inequality)

Triangle Inequality

Triangle inequality is a rule based on distances between points in a triangle. It explains that if the points in a triangle can be represented by points X, Y and Z then sum of the distances between points X, Y and Y, Z would be greater than the distance between the points X, Z and this should be followed in the common notions of distance. The sum of the distances between points X, Y and Y, Z would be equal to the distance between points X, Z only when X, Y and Z lie in a straight line in order.

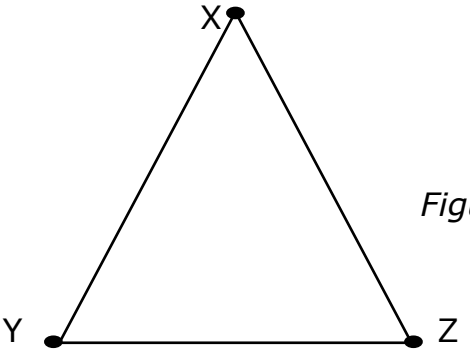

Figure a: Dist(X, Y) + Dist(Y, Z) > Dist(X, Z)

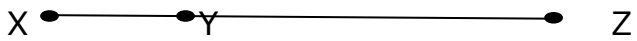

Figure b: Dist(X, Y) + Dist(Y, Z) = Dist(X, Z)

Common terms related to vectors:

1. **Metric:** Any function which obeys the rules predefined over the space is known as a metric. In this case, any notion of distance which follows the above defined rules would be a distance metric.
2. **Metric space:** The space over which the metric is defined is known as metric space. Here the metric space is the set of all real numbers, $\mathbb{R}$
3. **Norm:** Every metric space has a norm which indicates that all values are calculated with respect to a reference point. For a vector X, it is denoted as ||X||. For distance function the norm would be denoted as ||X|| = Dist(X, Ref) where Ref is the reference point(also known as the origin)

We would now define a few notion of distance which follows the above rules

1. $\text{Dist}(X, Y) = \sum |X-Y|$
2. $\text{Dist}(X, Y) = \sum ((X_1 - Y_1)^6)^{1/6}$
3. A more generalized form of the above, $\text{Dist}(X, Y) = \sum ((X_1 - Y_1)^P)^{1/P}$

We can check that all of the three notions of distance defined above follow the rules specified earlier*. Hence, they all can be used as distance metrics and would indicated different values of distance between the same vectors in the same dimensional plane. This implies that the notion of distance is just an abstract concept and is used to have a general idea of distance between vectors and not an absolute indicator. Hence any such notion of distance can be used as long as it is a metric.

*Note: For generalized case 3, the notion of distance is true only if P>1