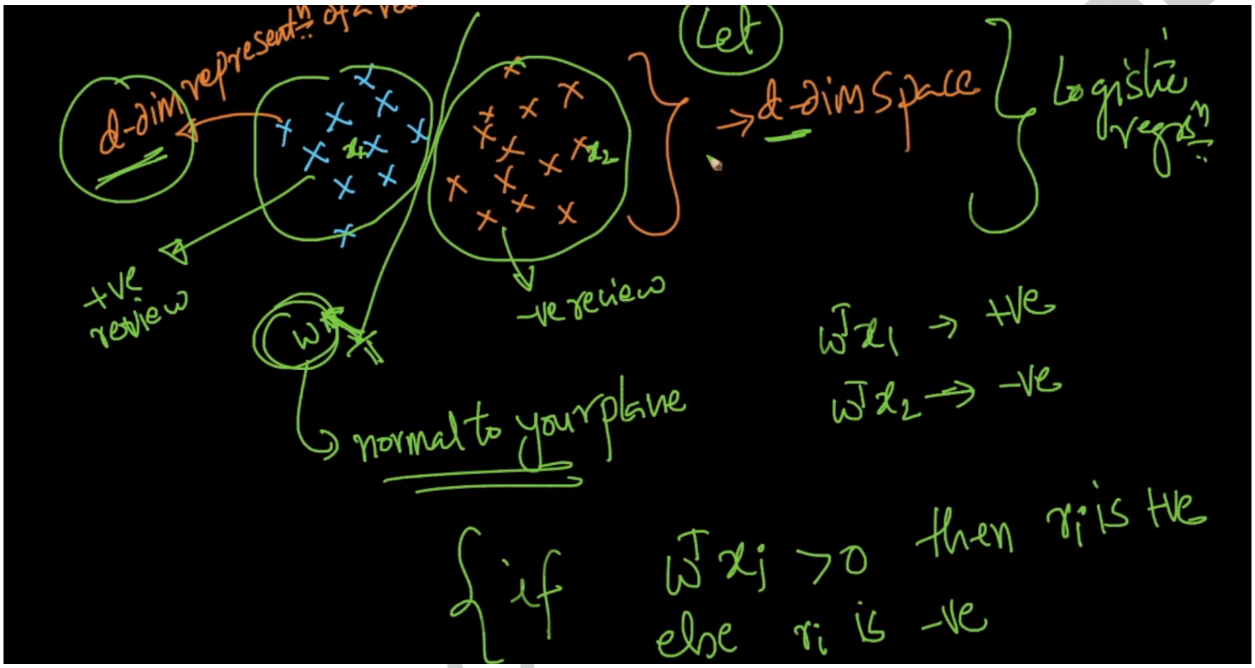
28.3 Why convert a text to a vector?

Given any problem, if we are able to convert the data into vector form, we can leverage the whole power of Linear Algebra.

When we are given the text data, if we could convert it into d-dimensional vector format, and plot those vectors/points in d-dimensional coordinate space, we can find a hyperplane that could separate the points/vectors belonging to different classes.

Below is the representation that was explained starting from the timestamp 3:25.



This is a d-dimensional representation of each vector. Each vector represents a d-dimensional review. Here the vector ‘w’ is normal to the hyperplane ‘π’.

If $w^T.x_1>0$, then we can say that the vector ‘w’ is in the same direction as that of ‘x_1’ and.

If $w^T.x_2<0$, then we can say that the vector ‘w’ is in the direction opposite to that of ‘x_2’.

For a given query point ‘x_i’,

If $w^T.x_i>0$, then the point ‘x_i’ is classified as positive.

If $w^T.x_i<0$, then the point ‘x_i’ is classified as negative.

Properties required to convert a text into a d-dimensional vector

If we have 3 vectors ‘r_1’, ‘r_2’ and ‘r_3’ which are semantically similar, then

If similarity(r_1, r_2) > similarity(r_1, r_3), then distance(v_1, v_2) < distance(v_1, v_3)

$v_1 \rightarrow$ Vector form of the review ‘r_1’.

$v_2 \rightarrow$ Vector form of the review ‘r_2’.

It means, if the reviews ‘r_1’ and ‘r_2’ are similar, then the vectors ‘v_1’ and ‘v_2’ must be close.

If similarity(v_1, v_2) > similarity(v_1, v_3), then length(v_1-v_2)<length(v_1-v_3).