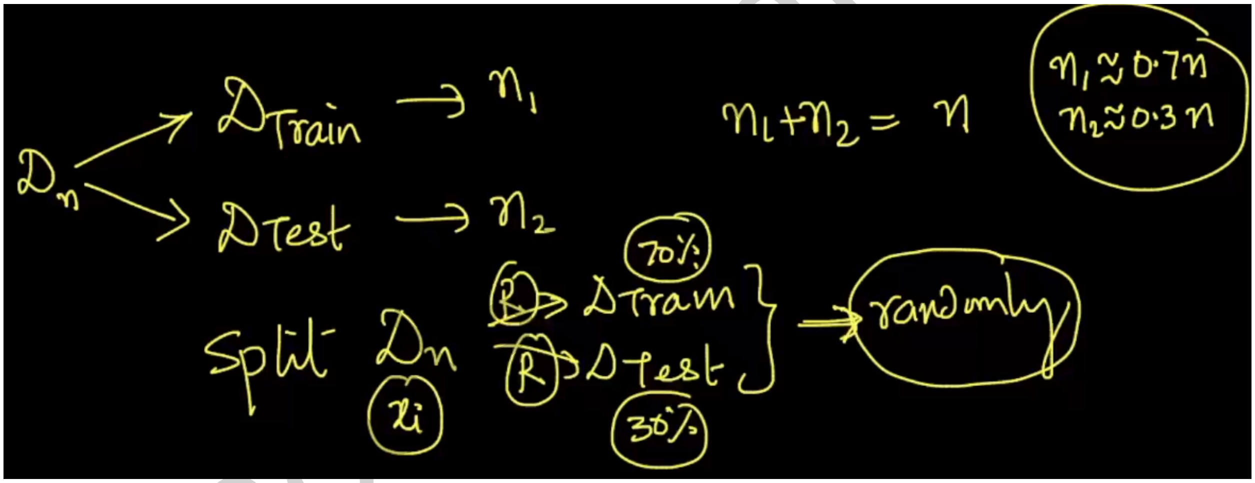
29.8 How to measure the effectiveness of K-NN?

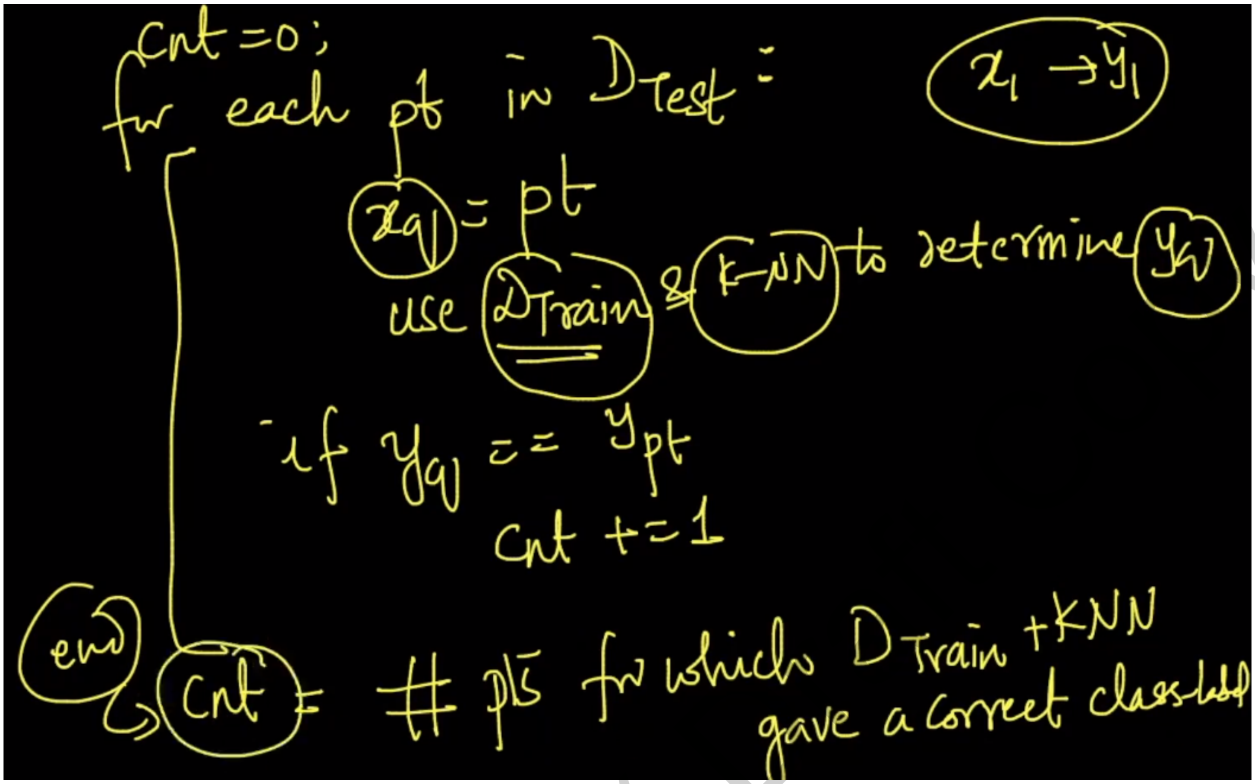
Let us consider the Amazon Fine Food Reviews Dataset which has got 364K reviews(after deduplication). For a given query point 'x_q', we have to predict the class label 'y_q'. Each data point is represented in the form of a numerical vector, and each data point has its own class label.

Procedure to measure the effectiveness of K-NN

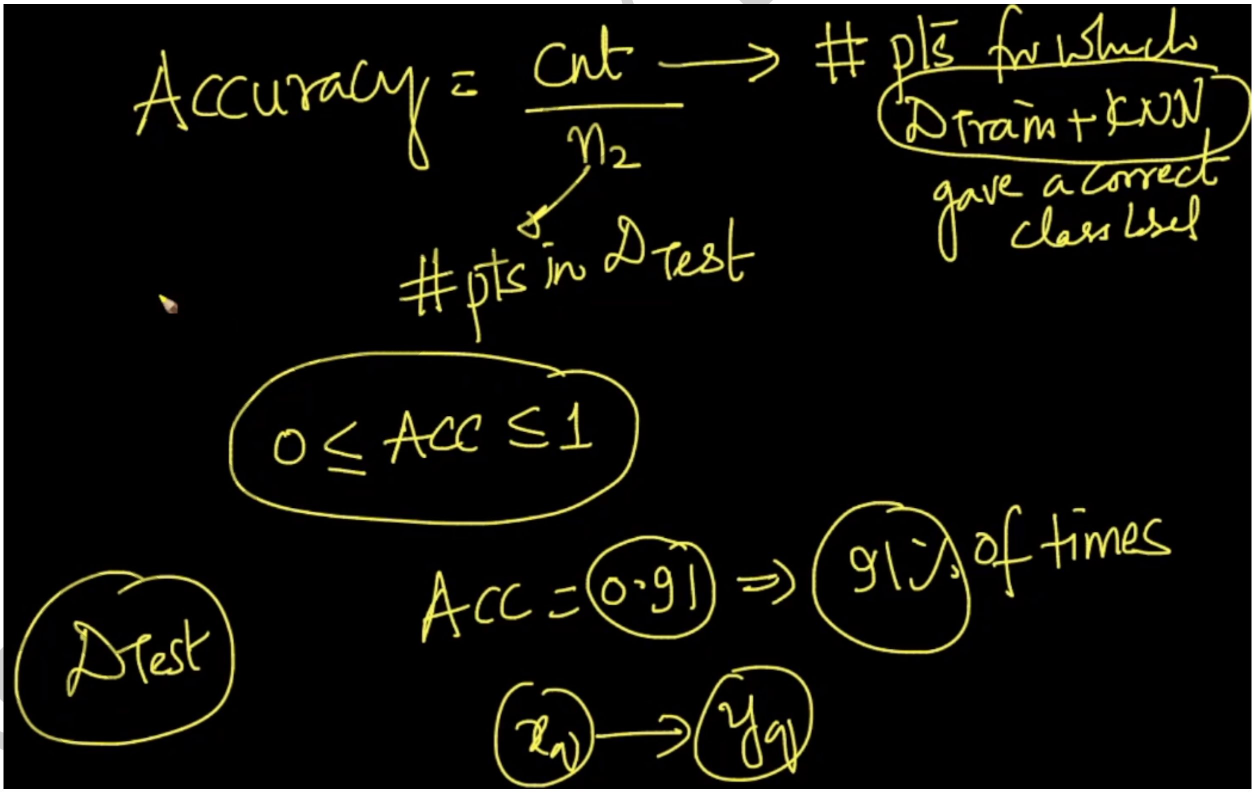
1) Let us assume we are given a dataset $\{D_n\}$ and our inputs are $\{x_i\}_{i=1}^{n}$ and the outputs are $\{y_i\}_{i=1}^{n}$
2) Divide the dataset $\{D_n\}$ into the training set $\{D_{Train}\}$ and the test set $\{D_{Test}\}$. Let 'n_1' be the number of points in $\{D_{Train}\}$ and 'n_2' be the number of points in $\{D_{Test}\}$.
   $(n_1 + n_2 = n)$



3) Now we have to fit the KNN model on 'D_{Train}', so that the entire 'D_{Train}' gets stored. Then for each point 'x_q' in 'D_{Test}', we have to make predictions using the same KNN model and predict the value of y_q'.
4) Let us initialize a variable 'count = 0' and for every data point 'x_q' in 'D_{Test}', if **y_q == y_q'**, then increment the 'count' value by 1.
5) Finally we have to compute the accuracy using the formula
   Accuracy = count/(number of data points in 'D_{Test}') = count/n_2
   Accuracy value typically lies in between 0 and 1.



```
cnt = 0;
for each pt in D_Test :                $x_1 \rightarrow y_1$
    $x_q$ = pt
    use D_Train & K-NN to determine $y_q$
    if $y_q == y_{pt}$
        cnt += 1
end
cnt = # pts for which D_Train + KNN gave a correct class label
```



$$Accuracy = \frac{cnt}{n_2}$$

cnt → # pts for which D_Train + KNN gave a correct class label

n_2 → # pts in D_Test

$$0 \le ACC \le 1$$

ACC = 0.91 ⇒ 91% of times

D_Test

$x_q \rightarrow y_q$

Note: If accuracy = 0.92, it means in 92% of the cases, using the fit on ‘D_{Train}’, the model predicts the output labels accurately.