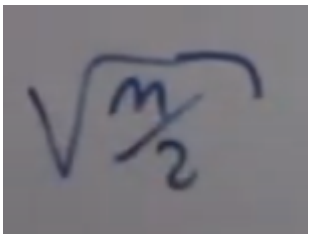
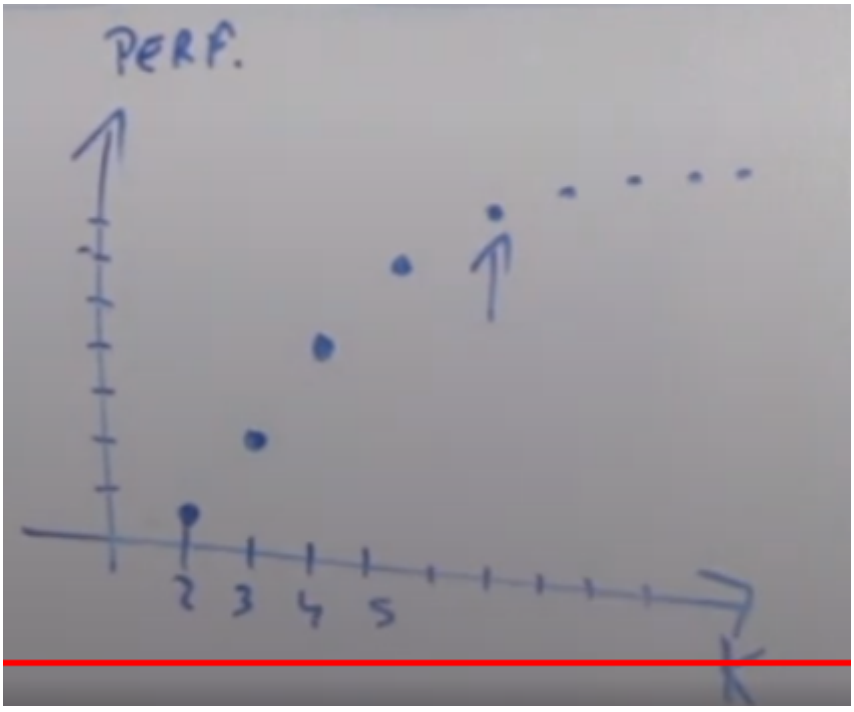
Encontrando o número ideal de clusters - K-means

Existe uma fórmula empírica para se escolher o melhor número de K (número de clusters). Como não podemos testar vários K's diferentes e verificar qual a melhor distância, pois ao aumentar K cada vez a distância em relação aos centroides irão diminuir, então devemos usar a seguinte fórmula:



Sendo N o número de amostras.

Outra forma será utilizar a função de custo (distância dos centroides) como performance em relação ao número de clusters:



Ao fazer isso, podemos escolher um ponto onde a performance já não é tão relevante ao aumentar o número de clusters, ou seja, encontrar o cotovelo da curva (elbow).