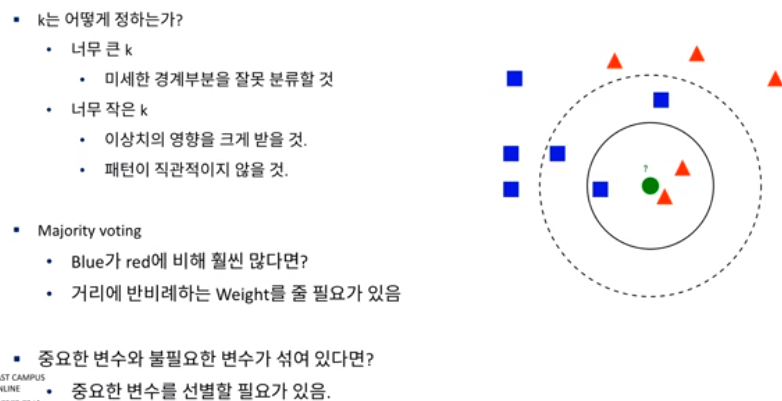
**<k-Nearest neighborhood>**

주변 관측치에 대한 정보를 이용하여 분류



거리는 어떻게 구하나?

연속형 변수 : 유클리드거리, 맨하탄거리

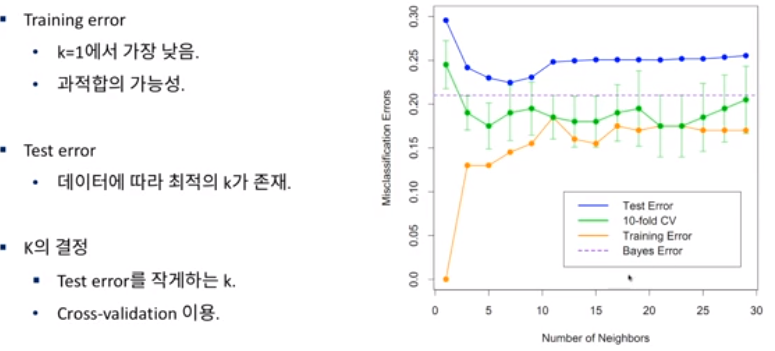
범주형 변수 : 해밍거리

<Cross validation>

* 테스트를 하려면 데이터를 나누어야하므로 데이터의 소실이 발생. 과소추정, 과대추정
* 훈련데이터와 테스트 데이터를 섞어서 테스트. K-fold cross-validation



KNN에서 적절한 k는 데이터마다 다르다. 따라서 k를 적절하게 구할수있어야한다.



Bayes error : true error 여기에 가까우면 모델이 적절하다는 표현



1차원 x변수만 가지고있을때는 knn을 썼을 때 녹색4 빨간1로 녹색으로 분류함.

y변수가 추가되었을 때 원형으로 knn을 보면 녹2, 빨3으로 결과가 달라짐

* 이 변수들을 분류하는데 y변수는 거의 의미가없음(y변수는 분포를 설명하지 못함)

2차원에서도 이런 문제가 발생하는데 어느 변수가 중요한지 모르고 때려 박으면 차원의저주 때문에 결과가 이상하게 나옴 -> 그래서 feature selection을 하거나 PCA를 통해 차원을 축소한후 KNN진행