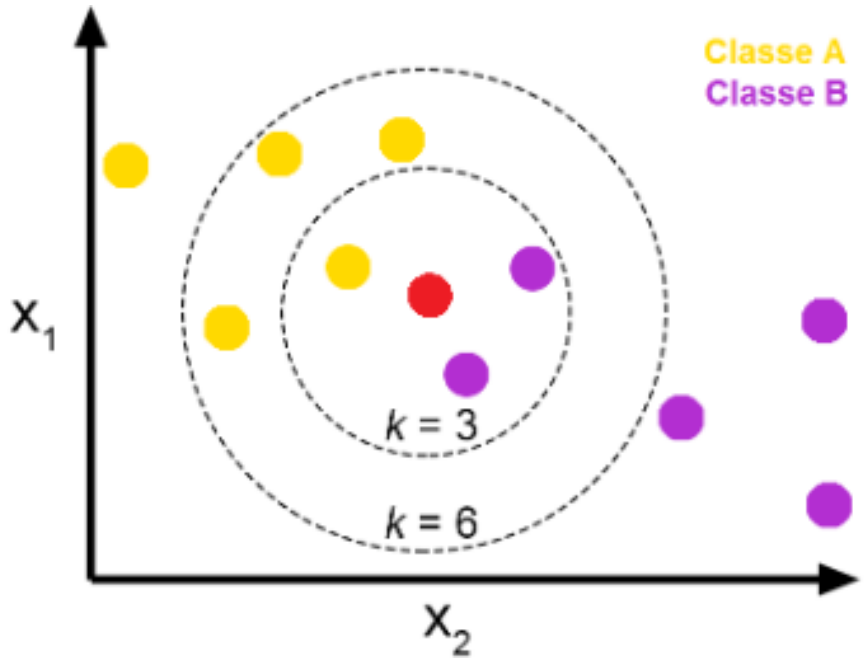
K-NN(K-Nearest Neighbors, 최근접 이웃)

K-NN이란?

데이터의 주변 K개의 데이터를 살펴본 후 이를 기반으로 분류하는 방식이며 여기서 K는 비교해볼 데이터의 개수이다.



위 그림을 보면 K=3일 때는 Class B로 분류 되지만 K=6일 때는 Class A로 분류되는 것 을 알 수 있다.

KNN은 K를 어떻게 정하냐에 따라 결과값이 달라질 수 있다.

K는 너무 크지도 작지도 않은 값을 사용해야 하며 노이즈가 많을수록 k가 클수록 좋고 반대로 노이즈가 적을수록 k가 작을수록 좋다.

또, 두 클래스의 값이 동점 되는 경우를 대비하기 위해 홀수로 지정하는 것이 좋다.

이렇게 결정한 K개의 데이터를 이용하는 두가지 방법이 있다.

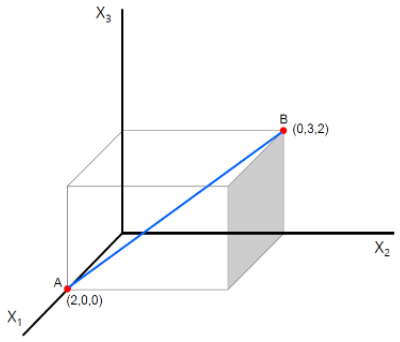
1. 개수만으로 분류하는 방법(연속형)
2. 거리에 반비례하는(1/d) 가중치를 부여하는 방법(범주형)

위 두가지 방법이 있으며 그에따른 거리 측정 방법은 3가지가 있다

거리 측정 방법

1.유클리드 거리(Euclidean Distance)

-연속형에 쓰는 방법이며, 점과 점사이의 직선의 거리를 구하는 방법이다.



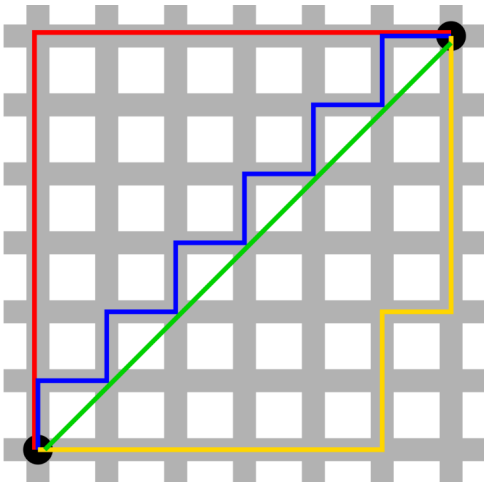
$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

다음과 같은 수식으로 데이터와 데이터 사이의 직선 거리를 계산한다.

2.맨해튼 거리(Manhattan Distance)

-연속형에 쓰는 방법이며, 각 차원을 따라서 간 거리를 구하는 방법이다.



$$d\left(A,B\right) = \sum_{i=1}^{n} |a_i - b_i|$$

다음과 같은 수식으로 데이터와 데이터 사이의 각차원의 절댓값의 합으로 거리를 계산한다.

3.해밍거리(Hamming Distance)

-범주형에 쓰는 방법이며, 같은 위치에 있는 값들을 비교해 거리를 구하는 방법이다.

- '1011101'과 '1001001'사이의 해밍 거리는 2이다. (1011101, 1001001)
- '2143896'과 '2233796'사이의 해밍 거리는 3이다. (2143896, 2233796)
- "toned"와 "roses"사이의 해밍 거리는 3이다. (toned, roses)

거리기반 모델의 경우에는 치명적인 단점이 있다. 각데이터의 중요도 보다는 값의 크기에 따라 결정되는 것이다.
이를 해결하기 위해 값의 범위를 재조정 해야하는데 이에는 2가지 방법이 있다.

1. 최소-최대 정규화

모든 데이터에 대해 각데이터의 최소값은 0 최대값은 1 으로 설정하는 것이다.

$$Z = (X - min(X))/(max(X) - min(X))$$

이 방식에는 치명적인 단점이 있다. 바로 이상치에 많은 영향을 받는다.
예를들어 10개의 데이터중 9개는 1~5에 머무르지만 하나의 데이터가 100이라는 값을 가진다고 가정해보자. 그럼 9개의 데이터는 0부터 0.05사이의 값으로만 변환이 된다.

2.Z-점수 표준화

데이터를 표준 정규분포에 해당하도록 값을 바꾸는 것이다.

$$Z = (X - 평균)/표준편차$$

이 방식은 이상치의 영향을 피할 수 있다.

장점

1. 구현이 쉽다.
2. 결과가 일관성 있다.
3. 정확도가 높다.
4. 데이터에 대한 가정이 없다.
5. 오류 데이터는 비교대상에서 제외

단점

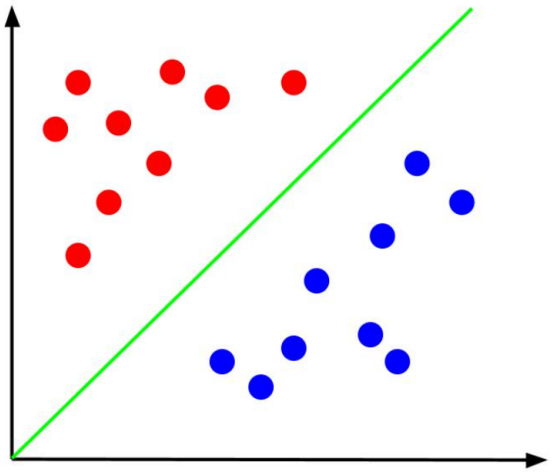
1. 데이터가 많아지면 분류가 느리다.
2. 적절한 K를 결정해야 한다.
3. 많은 데이터를 활용하기 때문에 고사양의 하드웨어가 필요함.

SVM(Support Vector Machine)

SVM란?

결정경계(Decision Boundary) 분류를 위한 기준 선을 정하는 모델

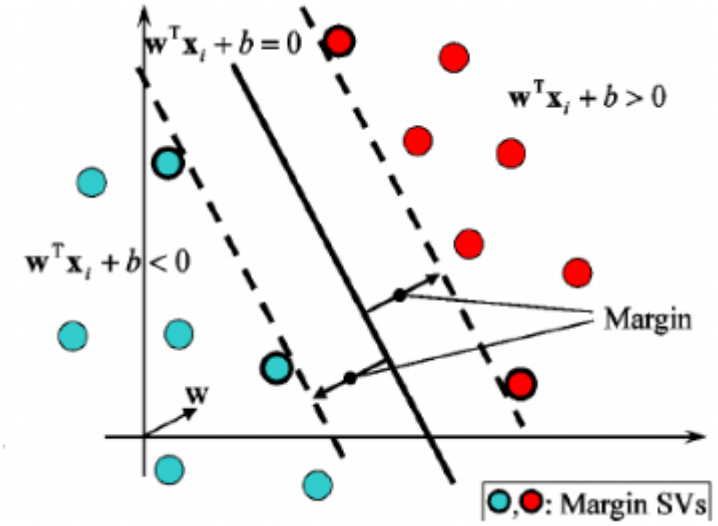
데이터를 기준선의 어느쪽에 속하는지 확인후 분류하는 방식이다.



여기서 중요한것은 데이터의 분류하는 결정경계이다. 최적의 결정경계를 결정하기 위해서 알아야할 여러가지가 있다.

1. Margin

Margin이란 선과 가장 가까운 양 옆 데이터와의 거리다.

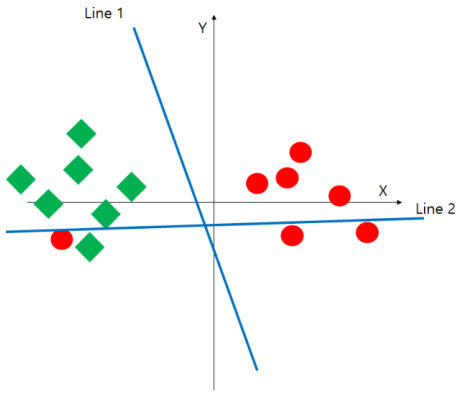


위 세가지 선중 가운데 선이 가장 Margin이 최대화 된 선이다.

Margin을 최대화 하는 방향으로 결정경계를 잡아야한다.

2. Outlier

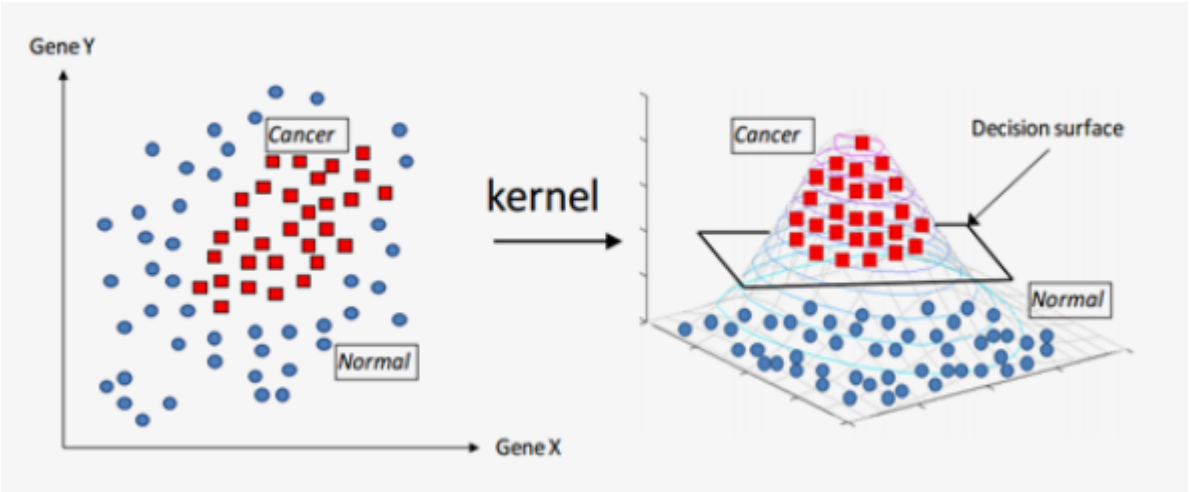
어느정도 데이터를 무시하는걸 말한다.



다음과같은 경우는 2번보다는 1번이 더좋다. 이런경우 Outlier처리를 하는게 더 최적의 결정경계에 가깝다.

3. Kernel Trick

차원을 높여 매핑하는 방법이다.



2차원일때 곡선이여야 했던 결정경계가 3차원으로 바뀌면서 직선에 가까워 졌다.

4. C

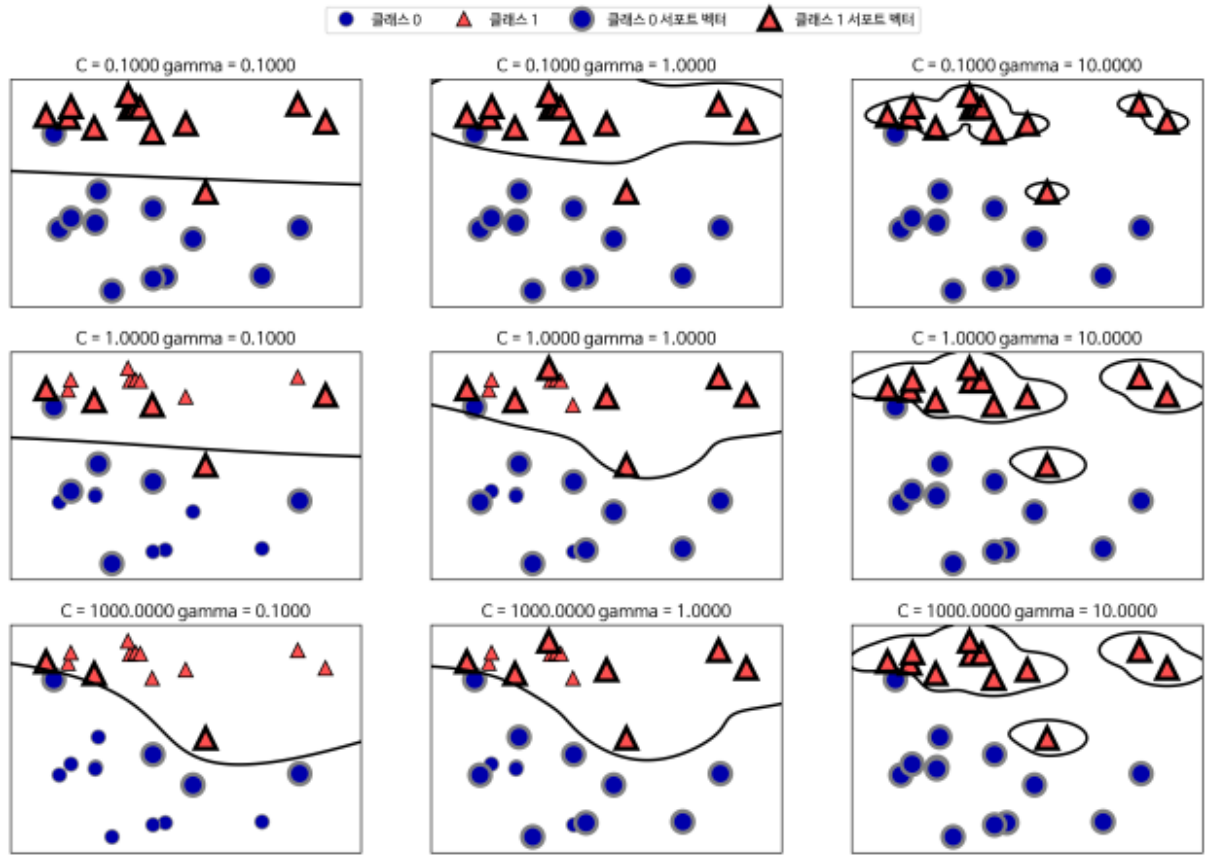
오류를 얼마나 허용하느냐, 얼마나 정확하게 구분할건지를 의미한다.

C값이 증가함에 따라 결정경계가 직선이 아닌 곡선에 가까워 지며 훈련데이터분류에 더 적합해진다.

5. Gamma

결정경계의 곡률을 결정한다.

Gamma값도 증가함에 따라 결정경계가 곡선에 가까워지며 훈련데이터분류에 더 적합해진다.



위 그림을 보면 c와 gamma값의 증가함에 따른 결정경계의 변화를 볼 수 있다.

하지만, 이럴경우 오버피팅이 일어났다고 한다.

6. Overfitting(과적합)

훈련 데이터를 지나치게 학습하는것을 의미한다.

결정경계가 곡선에 가까워질수록 훈련 데이터에는 정확해지나 이는 최적의
결정경계가 아니며 직선으로된 최적경계가 더 적합하다.
그러므로 C와 Gamma의 값을 적당하게 설정하는것이 중요하다

장점

1. 범주나 수치 예측 문제에 사용 가능하다.
2. 오류데이터의 영향을 적게 받는다.
3. 과적합되는 경우가 적다.

단점

1. 학습속도가 느리다.
2. 최적의 결정경계를 위한 커널과 모델피라미터에서 다양한 테스트가 필요하다.