PRML Chapter.1

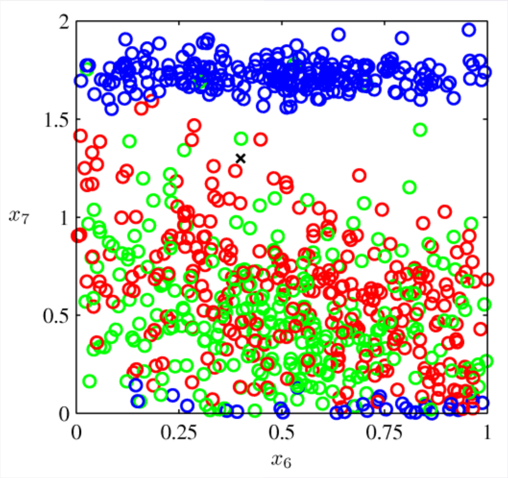
Introduction

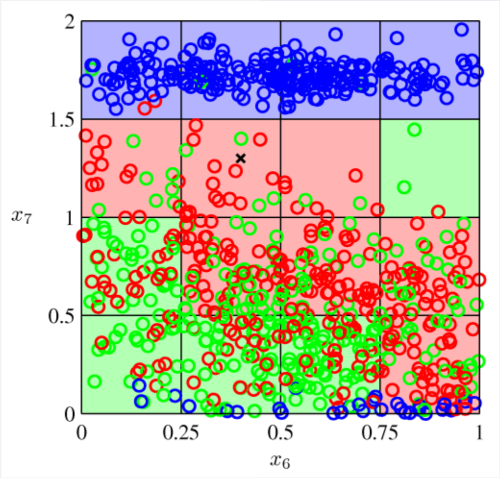
2017010698
수학과 오서영

목차

4. The Curse of Dimensionality

12-Dimensional input vector
3 color -> **3** classes
Solution of Classification?



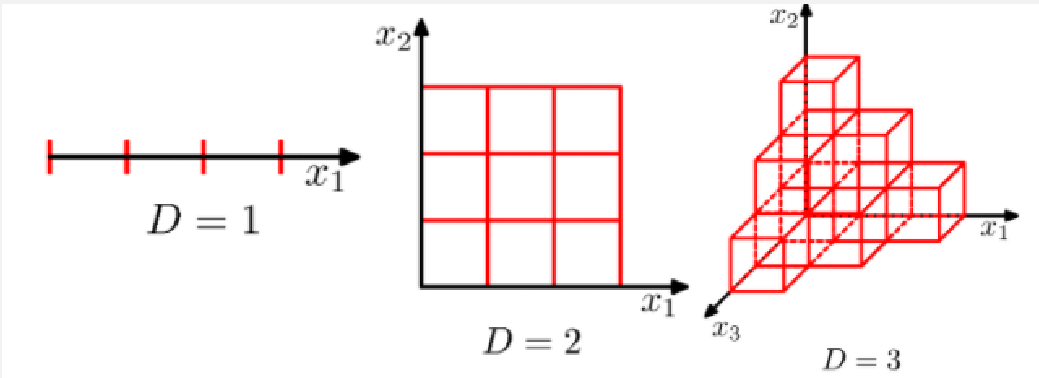


Input space를 작은 단위의 **cell**로 나누기

test point를 적절한 클래스로 분류를 하는 방법이다.

4. The Curse of Dimensionality

입력 데이터의 차원이 증가
-> 이런 방식을 적용하기가 어려워짐
Cell grows **exponentially** with dimensionality D



판별 함수
D : Input variables, general polynomial with order 3
-> 구해야 할 차원은 D^3 까지 증가

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

4. The Curse of Dimensionality

(**EX**) D차원에서 r=1인 구
이 때의 구의 부피는? (입력 차원에 독립적인 일반식)

$$V_D(r) = K_D r^D$$

- r=1인 경우의 구의 부피에서
r = 1−e 인 구의 부피를 빼는 것을 상상

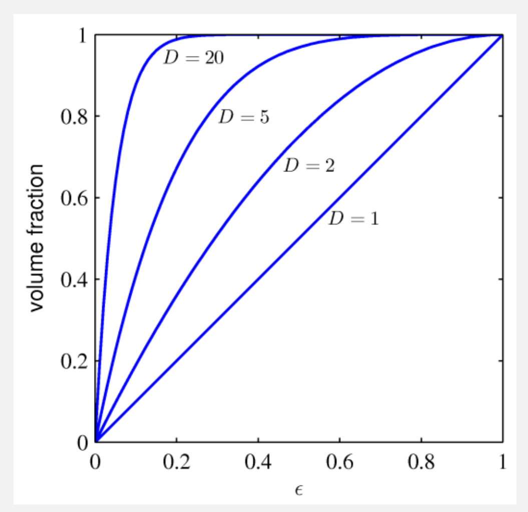
-> e 는 매우 작은 값
-> 다차원 구의 부피는 표면 층의 부피가 된다.
-> e 값을 조절하면서 이 구간의 부피비를 생각하기

$$\frac{V_D(1) - V_D(1-e)}{V_D(1)} = 1 - (1-e)^D$$

원래 구의 부피를 분모로
e 로 인해 결정되는 겉 껍질의 부피를 분자로 놓게 된다.
-> 원래의 부피와 겉면의 부피의 비를 확인

4. The Curse of Dimensionality

e 값이 변화할 때 원래 부피와의 비율



-> 차원이 증가할수록(D가 커질수록)
e 값이 작더라도 원래 volume 크기와 근접하게 됨
-> 차원이 증가할수록 전체 volume 크기의 대부분은
표면에 위치하게 된다는 것

4. The Curse of Dimensionality

다차원 공간에서의 가우시안 분포

- D차원을 가진 샘플 x.
- 이 샘플은 원점으로부터 임의의 거리 r 만큼 떨어져 있다.

이 데이터 x 를 하나의 차원으로 축소

-> | r | : 원점으로부터 떨어진 거리를 의미하는 변수 (양수)

-> D 차원의 정보가 하나의 차원으로 축약된다.

가우시안 분포를 따라 랜덤샘플 생성

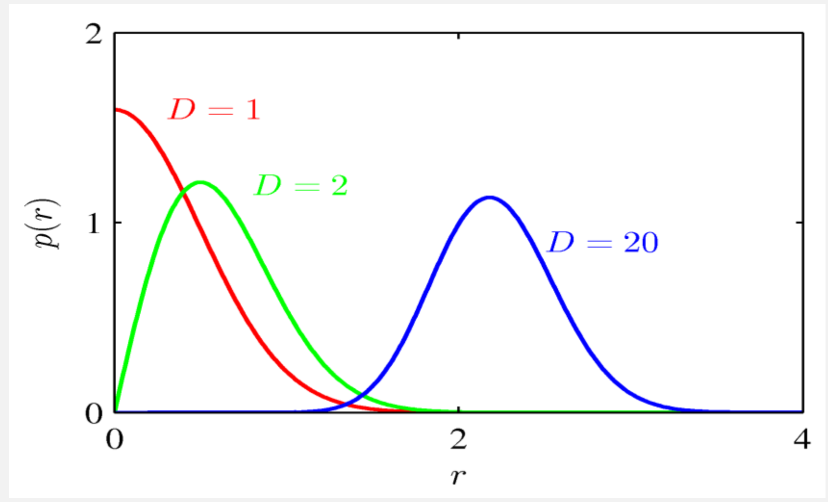
- x 가 원래 가지고 있던 차원을 증가하면서 랜덤하게 생성

-> 실제 데이터가 어느 거리에 많이 존재하는지를 확인

차원이 증가하는 경우 반지름 r 의 위치에
데이터의 분포가 집중
즉, 차원이 증가할수록 전체 부피 중
표면 쪽의 부피 비율이 증가하기 때문에
실제 샘플이 등장할 비율도 표면에 가까워지도록 변화할 것이다.

.

4. The Curse of Dimensionality



차원의 저주 – 정리

데이터의 차원이 증가
-> 그것을 표현하는 데이터의 volume은 exponential하게 증가
-> feature 공간의 희소성이 증가
-> 데이터의 밀도가 낮아지게 된다.
-> 오버피팅

5. The Decision Theory

Decision Theory
: 확률 이론을 바탕으로 불확실성이
관여된 상황에서의 최적의 결정 과정

목표 : 입력 X 와 이에 대한 타겟 T 를 이용하여
새로운 변수 Xnew 에 대응하는 타겟 값 Tnew 를 예측할 수 있다.

1) 오분류 최소화 (Minimizing the misclassification rate)
- 목적 : 잘못 분류될 가능성을 최대한 줄이는 것
-> 모든 x 에 대해서 특정 클래스로 할당시키는 규칙이 필요
-> 입력 공간을 각 클래스 별로 나누게 되는 효과

나누어진 구역 : decision region (Rk)
각 구역의 경계면 : decision boundaries or decision surface

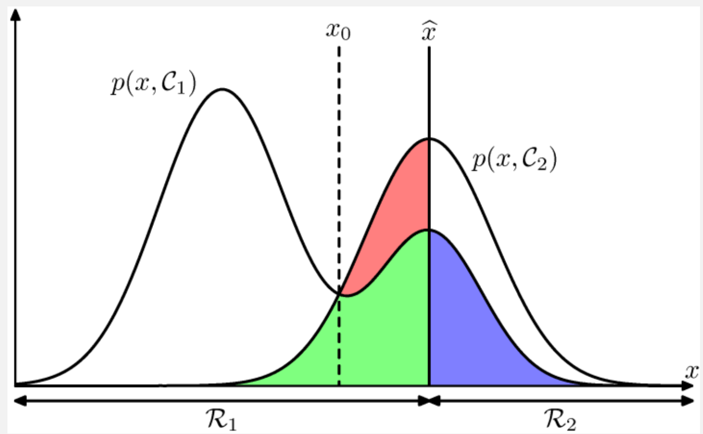
5. The Decision Theory

잘못 분류될 가능성 (오분류될 확률 값을 모두 합한 확률)

$$p(mistake) = p(x \in R_1, C_2) + p(x \in R_2, C_1) = \int_{R_1} p(\mathbf{x}, C_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x}, C_1) d\mathbf{x}$$

-> 이를 최소화 하는 방향으로 모델 설계

5. The Decision Theory



현재 클래스의 구분선을 x^hat 으로 결정
-> x ≥ x^hat 인 영역에서는 해당 클래스가 C2 로 결정
(반대인 경우 C1 으로 할당됨)
-> Error : blue , green , red
-> 이를 최소화하는 영역으로 기준 선이 변경

만약 x^hat 를 왼쪽으로 이동 : blue + green 유지, red 변화
-> 면적을 최소화 : x^hat = x0 인 지점
(제대로 분류될 확률 값을 최대화하는 방향으로 해도 문제 X)

$$p(correct) = \sum_{k=1}^{K} p(x \in R_k, C_k) = \sum_{k=1}^{K} \int_{R_k} p(x, C_K) d\mathbf{x}$$