<Kernel-density Estimation>

:Mixture of Gaussian 의 가정이 완화된 것.

:특정한 분포 가진다는 가정 없이(non-parametric) 주어진 데이터로부터 주변부의 밀도를 추정하는 방식

데이터 샘플링

kernel=’gaussian’ : smoothing

kernel=‘tophat’

kernel=‘epanechnikov’

확률밀도함수 p(X)로부터 생성된 벡터X가 주어진 영역 R 내부에 속할 확률

N 개의 벡터 {x1,x2, xn} 가 동일한 분포로부터 생성되었을 때(independent)

그 중에서 k개가 영역 R내부에 속할 확률.

* : N combination k
* : 성공 확률
* 이항분포 B(N,P) 성질에 의해 (N: 시행 개수)

|  |
| --- |
| E(k) = NP  var = NP(1-P) |

E=,

var= <- N이 무한대로 접근하면 분산은 0에 가까워지고 기댓값은 과 가까워짐

영역 R의 크기는 P(X)의 변화가 없을정도로 매우 작다고 가정을 하면

V= 영역 R의 Volume

N이 무한대로 접근하고 영역 R의 크기는 P(X)의 변화가 없을정도로 매우 작다고 하면

=

이고

즉 =

V: 영역 Volumn

N: 전체 개체 수

K: 원하는 영역에 들어와있는 데이터 포인트 수

확률밀도의 추정은 N이 클수록, 영역 볼륨 V가 작을수록(너무 작게는 안됨) 정확해짐

* 적절한 V 찾기

커널 밀도 추정의 목적: V를 고정시키고 k를 찾기

<Parzen Window Density Estimation>

k개의 객체를 포함하는,

영역 x가 중심으로 하는,

각 면의 길이가 h인

hypercube 의 볼륨은 (d는 차원 수)

Kernel function

hypercube 안에 들어오면 1, 아니면 0

k =

p(x) = =

의 단점

* 불연속적인 추정, hypercube 내에 있는 객체들에 대해 동등한 가중치
* -> 완화하는 방법
* 거리가 같으면 확률 값이 같아지는
* 과 uni kernel인 가우시안 커널
* 가우시안 커널을 사용한 커널 density Estimation : Parzen Window
* 데이터 하나하나가 가우시안 모달 개수
* k = 에서 가우시안 커널 경우 h가

시그마가 크면 멀리 있는 거 영향 줄 수 있지만

시그마가 작으면(커널 width) 수행 못함. -> 싱글 가우시안

Parzen Window Density Estimation의 단점: 오버 피팅 생길 수 있음