9 강 표집법, PCA(1)

◈ 담당교수: 장필훈

▣ 주요용어

용어	해설
몬테카를로방법	빈도주의에 바탕한 방법. 특정 값의 근사치를 구하기 위해 난수를 이용해 확률적으로 구한다. 일종의 시뮬레이션으로 이해할 수도 있다. 계산하려는 목표가 해석불가능한 함수거나 구하기가 극히 어려울 때 사용한다.
마르코프체인	정확히는 '마르코프 성질을 가진 이산 확률과정', 줄여서 '이산시간 확률과정'. 시간에 따른 계의 상태변화를 나타내는데 관찰 시간이 이산적이어서 이산시간이다. 미래$(\tau+1)$의 상태는 현재(τ)에만 의존하고 과거$(\tau-1)$에는 의존하지 않아야 한다. 조건부 확률이 '과거상태와 독립'이라고 표현하기도 한다.
깁스샘플링	결합확률분포로부터 일련의 표본을 생성하는 알고리즘. 메트로폴리스 헤이스팅스의 특별한 예.
제안분포	원 분포를 근사해내기 위해 샘플링 과정에서 필요한 분포. 원 분포를 최대한 타이트하게 포함하도록 설정되며, 다루기 쉬운 분포(예-가우시안)를 고른다. 샘플링 방법에 따라 사용하는 승인률이 다르고 해당 승인률에 따라 원 분포를 추정해낸다.
메트로폴리스알고리즘	메트로폴리스-헤이스팅스 알고리즘. 직접 표본을 얻기 어려운 확률분포로부터 표본의 수열을 생성해내는 데 사용하는 알고리즘. 표집법의 하나.
고유값, 고유벡터	행렬 A, 상수 λ , 벡터 v가 $Av=\lambda v$관계를 만족하면, λ를 고윳값, v를 고유벡터라 한다.

▣ 정리하기

1. 대부분의 경우 정확한 사전 분포를 알 수 없기 때문에 확률적 모델은 정확한 추론을 시행하기가 까다롭다.
2. 그래서 실제로는 표본을 샘플링해서 근사 한다. 표본들이 독립적이지 않을 수 있으므로 기대값이 왜곡될 수있다.
3. 대부분의 경우 정규화상수를 알기가 어렵지만, 확률에 비례하는 값을 얻어 내는 것은 쉽다.
   다시말해, $p(z)=\frac{1}{Z_p}\hat{p}(z)$에서 $\hat{p}(z)$를 얻어내는 것은 쉽지만 Z_p를 알아내는 것은 어렵다.
4. 거부표집법은 제안분포를 두고 샘플링을 시행하되 승인 확률을 따른다.
   a. 이때 제안분포가 원분포를 모두 포함 해야 하고, 타이트하게 포함할수록 좋다.
   b. 고차원일때는 승인율이 기하급수적으로 감소하므로 쓸 수 없다
5. 중요도 표집법은 기대값을 바로 구하겠다는 아이디어를 바탕으로 한다.
   a. 샘플링 된 데이터를 적절하게 가중하여 합하는데 이때 가중치를 중요도 가중치라고 한다.
   b. 중요도 표집법을 한번 거친 데이터를 대상으로 리샘플링 하는 방법도 있다
6. 몬테카를로 방법은 거의 균일하게 분포하는 점의 개수를 세는 식으로 원하는 값을 근사해 내는 방법을 통칭한다.
7. 다음 상태가 이전의 모든 상태에 의존 하는 것이 아니라 바로 전 상태에만 의존 할 때, 마르코프 체인이라고 한다.

8. 기본적인 메트로폴리스 알고리즘은 제안 분포가 대칭임을 가정 한다.
9. 마르코프 연쇄를 사용하여 주어진 분포로부터 표집하려면 어떤 초기 분포를 택해도 결국 해당 불변분포로 수렴해야 하는데 이 성질을 에르고딕성이라고 하고, 이 분포를 평형분포라고 한다.
10. 메트로폴리스 헤이스팅스 알고리즘은 제안분포가 대칭일 필요가 없다.
11. 깁스 샘플링은 메트로폴리스 헤이스팅스의 특수 케이스다.
12. 주성분 분석은 차원감소, 데이터 압축, 특징 추출, 데이터 시각화 등에 응용된다.
13. 데이터의 최대 분산을 찾는 것과 최소 오류 공식화는 같은 결과를 안는다