워크북

교과목명 : 머신 러닝

차시명: 9차시

◈ 담당교수: **장 필 훈**

⊙ 세부목차

- 표집법
  - 거부표집법
  - 중요도표집법
  - MCMC
  - 기브스 표집법
  - 조각표집법
- PCA

학습에 앞서

■ 학습개요

실제 사용하는 확률모델들은 정확한 추론을 직접 시행하기가 까다롭다. 그래서 샘플링을 한 후 분포를 근사하는 때가 많고, 대표적으로 몬테카를로 테크닉에 대해 다룬다. 대부분의 경우 근본적인문제는 어떤 함수f의 확률분포p에 대한 기댓값을 구하는 것이다. 대표적인 샘플링 기법으로 리젝션 샘플링, 중요도 표지법(importance sampling), 깁스 샘플링을 본다.

다음으로 PCA에 대해 배운다. 주성분 분석은 실제로도 매우 자주 쓰이므로 수식의 전개부터 의미의 이해까지 모두 다룬다.

■ 학습목표

1	여러종류의 샘플링방법을 이해하고 각각의 장단점을 파악한다.
2	몬테카를로 방법, 마르코프 체인등 자주 쓰이는 개념을 숙지한다.
3	주성분분석이 필요한 이유, 응용사례등을 배운다.
4	주성분분석의 수식전개를 이해한다.

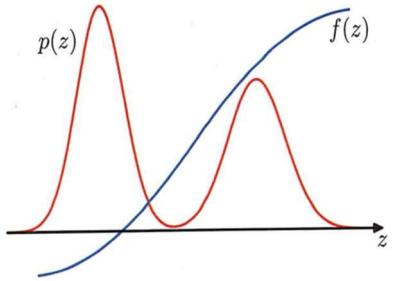
■ 주요용어

용어	해설
몬테카를로 방법	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.
메트로폴리스 알고리즘	메트로폴리스-헤이스팅스 알고리즘. 직접 표본을 얻기 어려운 확률분포로부터 표본의 수열을 생성해내는 데 사용하는 알고리즘. 표집법의 하나.
고윳값, 고유벡터	행렬 A, 상수 λ, 벡터 v가 $Av=\lambda v$관계를 만족하면, λ를 고윳값, v를 고유벡터라 한다.

학습하기

오늘은 표집법(샘플링)에 대해 알아보겠습니다. 확률적 모델은 모집단의 정확한 분포를 알수 없는 경우가 대부분이므로 샘플링해서 근사합니다. 샘플링을 무한히 할 수는 없으므로 이 방법으로는 '매우 정확한' 분포를 알 수 없겠지만, 우리가 원하는 만큼의 정확도로 근사할 수만 있으면 되겠지요. 모집

단에서 랜덤하게 아무렇게나 여러개 추출해도 대충의 분포 모양을 짐작할 수 있습니다. 하지만 어느정도 샘플링을 해야 '어느정도 정확하게' 할 수 있을까요. 얼마나 정확하게를 말하는 것일까요. 좀 더 구체적으로 이야기 하자면, 샘플링의 목표는 어떤 함수의 특정 분포에 대한 기댓값을 구하는 문제입니다.



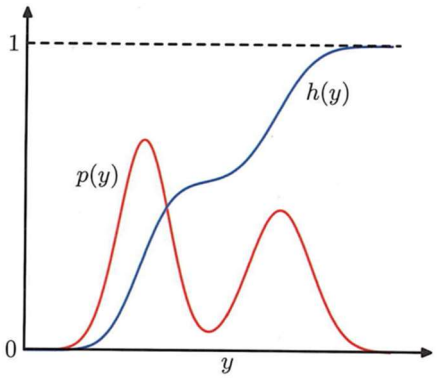
Bishop Fig 11.1

위 그림을 보시면, 확률분포 p(z)에 대한 f(z)의 기댓값($\int f(z)p(z)dz$)를 구하는 것이 목표입니다. 그런데 f(z)의 값이 큰 쪽에서 p(z)가 값이 작고, f(z)가 작은 쪽에서 p(z)가 큰 영역이 존재함을 알 수 있습니다. 그러면 f(z)p(z)를 적분한 값이 p(z)가 작은 지역에 의해 더 좌지우지될 가능성이 크다고 볼 수도 있습니다(p(z)가 큰 지역에서는 f(z)가 작기 때문). 그러면 f(z)p(z)를 정확하게 근사할 수 없겠지요. 더 많은, 혹은 더 정확한 샘플링이 필요할 것입니다.

그리고 이와 반대로, 샘플링은 임의의 형태의 분포를 생성하는 데도 사용될 수 있습니다. 좀 더 정확히 말하면 우리가 생성하고자 하는 분포를 '충분히 대표해낼 수 있는' 집단의 생성입니다. 이때 데이터포인트를 생성해 내는것도 샘플링이라고합니다. 본질적으로 이 두 활동은 동일한데, 아예 모르는 분포로부터 '대표성 있는' 작은 집단을 추출하는 것이나, 우리가 이미 알고있는 분포로부터 집단을 생성해 내는 것이나 결국 샘플링 과정은 동일하기 때문입니다.

그러면, uniform distribution으로부터 우리가 원하는 분포의 샘플집단을 골라내는 방법에 대해 이야기해보겠습니다.(이것은 위에 말한 샘플링의 두가지 과정 중 두번째에 가깝습니다)

일단, 0~1 사이의 균등한 분포의 난수가 주어졌다고 가정하고, 이때 위의 p(z)와 같은 분포를 얻어내도록 하겠습니다. 일단 누적분포 h(y)를 다음 그림과 같이 그려보겠습니다.



Bishop. Fig. 11.2

$\int_{-\infty}^{y} p(y)dy = h(y)$이고 $h(y)$를 z라 하면, $h^{-1}(z) = y$로 쓸 수 있습니다.

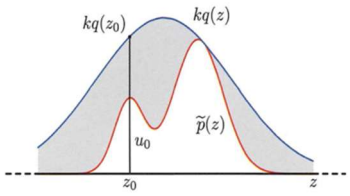
즉 우리가 쓰려는 분포에 대해 부정적분을 취하고 그 역함수를 구하면, 우리가 원하는 샘플 y를 얻을 수 있습니다. 하지만 이렇게 p(z)를 부정적분해서 h(z)를 얻을 수 있는 경우가 극히 드묾으로 주로 표집법을 씁니다.

<거부표집법>

p(z)로부터 직접 샘플링하는것은 어렵지만 정규화계수를 모르는 상태의 $\hat{p}(z)$는 계산 가능하다고 가정합시다. 종종 이것은 사실입니다. 즉, $p(z) = \frac{1}{Z_p}\hat{p}(z)$에서 Z_p를 알기 어렵다는 이야기입니다.

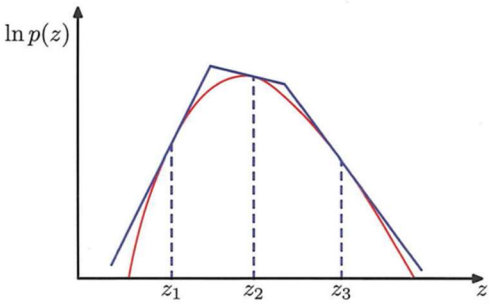
일단 제안분포(우리가 가정하는 분포)를 두고 샘플링을 시행합니다. 그러면 승인(acceptance) 확률은 다음과 같습니다.

$$p = \int \frac{\hat{p}(z)}{kq(z)} q(z)dz = \frac{1}{k}\int \hat{p}(z)dz$$



Bishop. Fig. 11.4

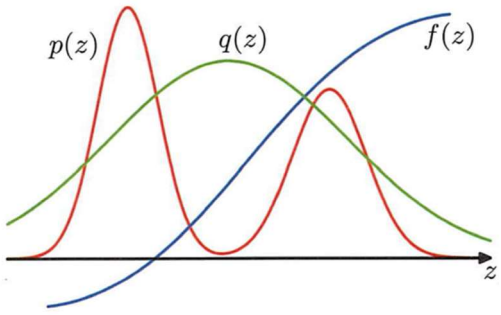
이때 모든 z에 대해 kq(z) >= $\hat{p}$(z)가 성립하는, 가능한 한 작은 k를 찾아야 합니다. 일단 여기까지만 해도 성공이긴 하지만, 제안분포가 위의 그림과 같이 우리가 얻어내고자 하는 분포와 모양이 많이 다르면 어떻게 할까요. 그때 쓸 수 있는 방법중 하나로 적응적 거부표집이 있습니다. 포괄함수를 그때그때(구간별로) 만들어 쓰는 것을 뜻합니다. 오른쪽 그림을 참고하세요.(Bishop. Fig. 11.6)



거부표집법은 고차원공간에서 쓸 수 없는 단점이 있습니다. 차원수가 증가함에 따라 승인률이 기하급수적으로 감소하기 때문입니다.

<중요도 표집법>

중요도 표집법은 위에 우리가 그림 11.1에서 본것과 같은 문제를, 구간을 나누어 해결하겠다는 아이디어 입니다. 즉, 아래 그림에서 q(z)로부터 적당히 데이터 포인트를 추출하고, 그 항을 적절히 가중해서 합하겠다는 아이디어 입니다.



$$\sum f(z)\frac{p(z)}{q(z)}q(z)dz \approx \frac{1}{L}\sum \frac{p(z^{(l)})}{q(z^{(l)})} f(z^{(l)})$$

표집분포가 원 분포에 얼마나 근접하느냐가 중요하고, p(z)가 변동이 너무 심하면 소수의 가중치가 너무 큰 값을 가지게 되고 결과적으로 우리가 뽑은 샘플중에 유효한 샘플의 수가 너무 적게 됩니다.

따라서 p(z)가 중요한 구간에서는 q(z)도 커야 합니다.

이 변형으로 SIR이라는 것이 있습니다. sampling importance resampling인데 과정은 다음과 같습니다.

1. q(z)로부터 L개의 표본을 추출
2. 중요도 표집에 따라 각각의 가중치 결정
3. 앞의 L개로부터 resampling. 이때 가중치는 2에서 결정한 가중치를 사용

<MCMC(Markov chain Monte Carlo)>

우선 몬테카를로방법에 대해 알아보겠습니다. 좋은 예제로 원주율 계산이 있습니다.

일단 정사각형 안에 사분원을 그립니다. 그리고 그 안에 격자점을 찍습니다. 그 다음 원의 내부에 찍힌 점의 수와 밖에 찍힌 점의 수를 세어서 그 비율로 원주율을 계산합니다. 이 방법은 표본공간차원이 고차원일때도 사용이 가능합니다.

다음으로 마르코프연쇄에 대해 알아보겠습니다. 마르코프연쇄는 쉽게 말하면 '다음상태가 바로 이전 상태에만 의존하는 것'을 말합니다. 식으로 나타내면 다음과 같습니다.

$$p\left(z^{(m+1)}\middle|z^{(1)},\dots,z^{(m)}\right)=p\left(z^{(m+1)}\middle|z^{(m)}\right)$$

이 확률을 전이확률이라고 합니다. $T_m\left(z^{(m)},z^{(m+1)}\right)\equiv p\left(z^{(m+1)}\middle|z^{(m)}\right)$

모든 m에 대해 전이확률이 동일하면 동질적(homogeneous)이라고 합니다.

연쇄에서 각 단계가 분포를 변화시키지 않으면 그 분포를 마르코프 연쇄에 대해 불변이라고 합니다.(invariant. stationary하다고도 합니다)

마르코프 연쇄를 사용해서 주어진 분포로부터 샘플링을 하려면 해당분포가 연쇄에 대해 불변이어야 하고, 무한히 반복해도 어떤 확률분포로 수렴해야 합니다. 이 수렴하는 분포를 평형분포라고 하고, 시작분포가 어떤 분포이든 평형분포에 도달하는 성질을 에르고딕성이라고 합니다.

이 마르코프 연쇄를 사용해서 샘플링하는 방법중 대표적인 것이 메트로폴리스 헤이스팅스 알고리즘입니다. 메트로폴리스와 메트로폴리스 헤이스팅스에 관해 녹화강의에 더 자세히 설명했으므로 참고 바랍니다.

<기브스 표집법>

메트로폴리스 헤이스팅스의 특수케이스입니다. 변수들 중 하나의 값을 나머지 변수들에 대한 해당변수의 조건부분포에서 추출한 값으로 바꾼 것입니다. 예를 들어 세개의 변수들에 대한 분포 p(z1,z2,z3)로부터 z1을 빼고, 남은 z2, z3만 조건부로 넣고 z1을 새로 샘플링해서 replace하는 것입니다. 이것을 z2, z3에 대해서도 반복하고 다시 z1에 대해서도 반복할 수 있습니다.

기브스 표집법이 세부균형을 만족한다는 점을 강의시간에 증명했으므로 참고 바랍니다.

<조각표집법>

메트로폴리스 알고리즘의 경우 각 단계 크기가 크면 높은 거부율때문에 비효율적입니다. 반대로 크기가 작으면 임의보행행동으로 비상관화가 느립니다. 그래서 분포의 성질을 보고 단계 크기를 맞추는 방식이 가능한데 그것이 조각표집법입니다.

<PCA>

차원수 감소의 가장 대표적인 선형방법이 주성분분석(Principal Component Analysis)입니다. 데이터를 subspace에 투영해서 차원감소를 시도하는 아이디어 입니다. 이때 데이터의 분산을 최대화 하는 방향으로 투영이 이루어집니다. 평균투영비용(데이터포인트와 투영체간 평균제곱거리)을 최소화 하는 방향으로 투영할수도 있는데 이 두가지는 사실 같은 것입니다. 이 둘이 같음을 보이겠습니다.

일단 최대분산방향을 구해보겠습니다.

1차원에 투영한다고 할때, 투영되는 1차원 공간의 방향을 D차원 벡터 u_1으로 정의하겠습니다. 방향만 중요하므로 u_1은 단위벡터이고, 각각의 데이터포인트는 $u_1^T x_n$으로 투영됩니다.(내적)

투영된 데이터의 분산은 공분산행렬 S가 $\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^{\mathrm{T}}$일 때, $\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\bar{\mathbf{x}}\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$으로 주어집니다. 이 값을 최대화하는 것이 목표입니다. 라그랑주 승수법으로 해결하며, 구체적인 과정을 녹화강의에 담았습니다. 참고 바랍니다.

설명 도중에 고유벡터와 고윳값에 대한 설명도 포함되어 있습니다. 기초적인 개념이고 중요한 것이므로 강의가 부족하면 다른 것들이라도 찾아보면서 알아두시기 바랍니다.

요약하면, 데이터집합의 평균과 공분산행렬을 계산하고, 공분산행렬의 가장 큰 M개의 고윳값에 해당하는 M개의 고유벡터를 찾는 과정입니다.

다음시간에는 최소오류공식화방법으로 해보고 이것이 최대분산방향과 같음을 보이겠습니다.

	연습문제

1. 샘플링 할 때 각 단계의 스텝 크기에 따른 트레이드오프가 있다
    a. O
    b. 각 단계의 크기가 크면 높은 거부율 때문에 비효율적이고, 크기가 작으면 임의 보행 행동으로 비 상관화가 느리다
2. 0~1사이의 균등한 난수가 주어졌다고 가정하고, 우리가 쓰려는 분포에 대해 부정적분이 해석적으로 가능하면 임의의 분포에 대해 난수 생성이 가능하다.
    a. O
    b. 가능하다. $h(y) = \int_{-\infty}^{y} p(\hat{y})d\hat{y}$으로 정의하면 $y = h^{-1}(z)$이므로 가능하다. 본문참조.
3. 제안분포를 쉽게 알기 어려우므로 포괄함수르 그때그때 만들어 쓰는 방법도 연구 되어 있다.

a. O

b. 적응적 거부표집이라고 한다

4. sampling importance resampling을 할 때 무한히 반복하면 원하는 정확도로 근사해낼 수 있다.

a. O

b. $L \to \infty$일때 옳음이 알려져 있다.

5. 표본공간차원이 고차원이면, 어떤 샘플링을 해도 적절한 수준의 횟수만 반보하면 원하는 만큼 원분포를 근사해낼수 있다.

a. X

b. 차원이 높아질수록 거부율이 기하급수적으로 커진다. 따라서 MCMC를 주로 쓴다.

6. 메트로폴리스 알고리즘과 거부 표집법은 승인 확률에서 조금 차이가 있다.

a. X

b. 샘플이 거부되었을 때 버리느냐 혹은 다시 포함 되느냐의 차이가 있고 승인 확률은 같다.

7. 마르코프 연쇄에서 각 단계가 분포를 변화시켜지 않으면 동질적(homogeneous)이라고 한다.

a. X

b. 불변, 정류적 이라고 한다.(invariant, stationary)

8. 모든 표집법은 제안분포가 대칭이라는 것을 가정 한다.

a. X

b. 메트로폴리스 헤이스팅스는 제안분포가 대칭일 필요가 없다. 메트로폴리스 헤이스팅스의 특수케이스인 깁스샘플링도 대칭일 필요가 없다.

9. 고유벡터와 고윳값을 구했을 때, 가장 큰 고윳값에 대응하는 고유벡터는 제1주성분이 된다.

a. O

b. PCA의 정의를 공분산최대화로 두나, 최소오류공식화로 두나 모두 같은 식을 유도해내고, 제1주성분은 위의 설명과 같다.

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

참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

다음 차시 예고

- 최소오류공식화
- 적용예
- 피셔선형판별과 비교
- PPCA
- 커널PCA
- autoencoder
- 순차데이터 집합
- 마르코프 모델
- 은닉마르코프모델