기계학습

9강 표집법, PCA(1)

장필훈 교수

학습목차

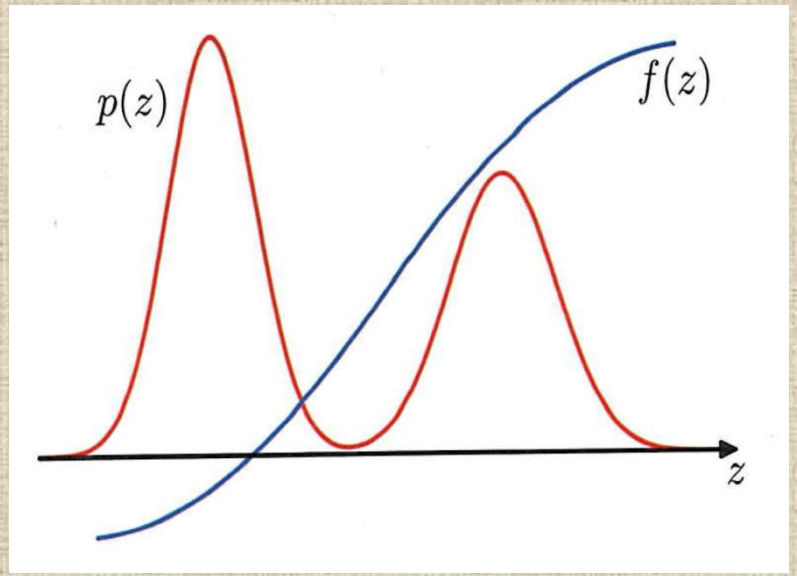
1. 표집법(sampling)
2. PCA(1)

01
표집법

1-1 sampling

- 확률적 모델은 정확한 추론을 시행하기가 까다롭다
  - 정확한 사전 분포를 알 수없기 때문.
  - 대부분 근사한다.
- 주로 어떤 함수의 특정 분포에 대한 기댓값을 구하는 문제

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig11.1

1-1 sampling

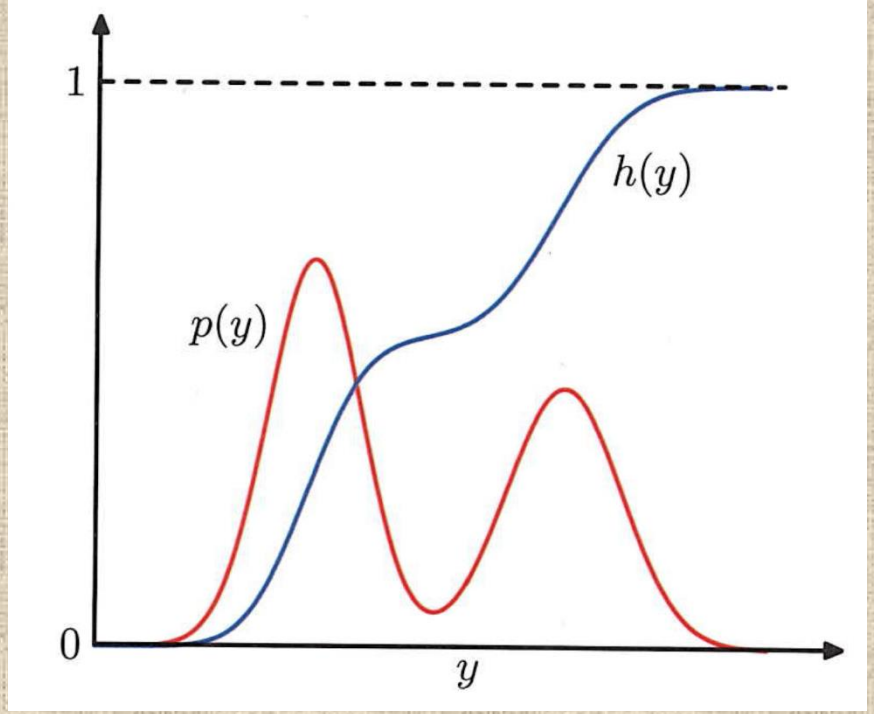
- 실제로는 표본을 샘플링해서 근사한다
- 표본들이 독립적이지 않을 수 있음
  - $p(\mathbf{z})$가 불균등하므로, 기댓값이 왜곡될 수 있다.

1-2 임의분포의 난수 생성

- 0~1 사이의 균등한 분포의 난수가 주어졌다고 가정

$$\mathbf{z} = h(y) \equiv \int_{-\infty}^{y} p(\hat{y})d\hat{y}$$

$$y = h^{-1}(\mathbf{z})$$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig11.2

1-2 임의분포의 난수 생성

- 우리가 쓰려는 분포에 대해 부정적분 + 그 역
  - 해석 불가능한 경우가 대부분
  - 표집법을 쓴다.
    - 단변량에만 사용 가능

1-3 rejection sampling(거부 표집법)

- $p(\mathbf{z})$로부터 직접 샘플링하는것이 어렵다고 가정
- 하지만,

$$p(z) = \frac{1}{Z_p}\hat{p}(z)$$

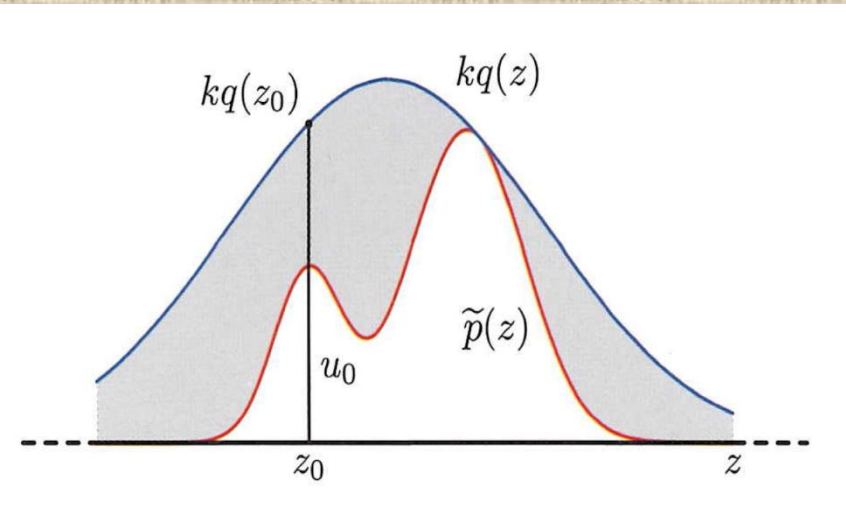
에서, $\hat{p}(z)$는 계산 가능하다고 가정.

(Z_p)를 알기가 어려움.

1-3 rejection sampling(거부 표집법)

- 제안분포(proposal dist.)를 두고, 샘플링을 시행한다.

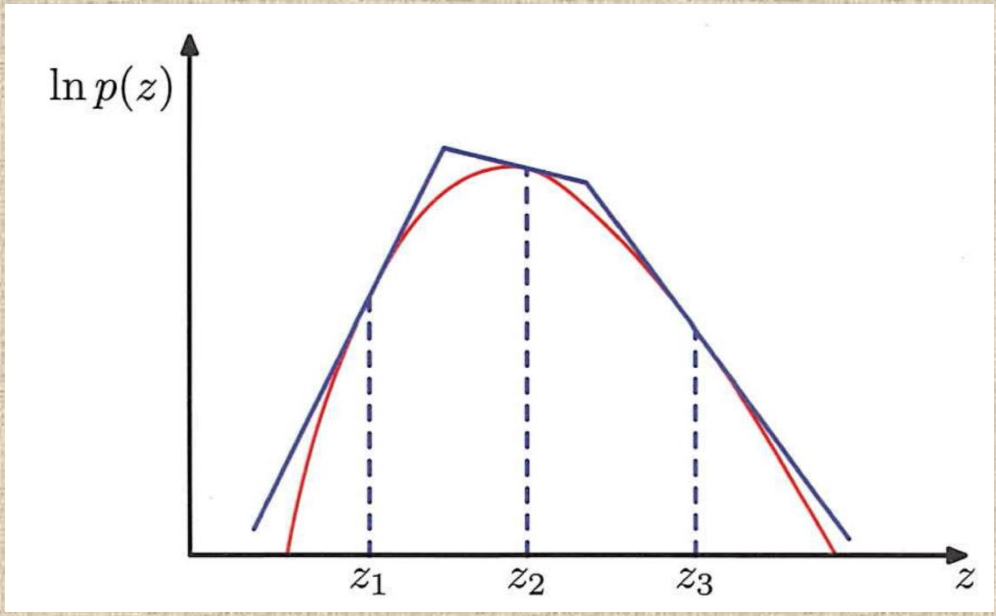
승인확률 $p = \int \frac{\hat{p}(z)}{kq(z)} q(z)dz = \frac{1}{k}\int \hat{p}(z)dz$



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. 11.4

1-3 rejection sampling(거부 표집법)

- 모든 z에 대해 $kq(z) \geq \hat{p}(z)$가 성립하는 k를 찾아야 한다.
- 가능한 한 작은 k를 찾아야 한다.
- 적응적 거부표집
  - 포괄함수를 그때그때 만들어 쓰는 것.



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. 11.6

1-3 rejection sampling(거부 표집법)

- 고차원공간에서는 쓸 수가 없다.
  - 다변량 가우시안(공분산 $\sigma_p^2\mathbf{I}$가정)일때,

    $kq(z) \geq p(z)$인 k가 존재하려면 $\sigma_q^2 \geq \sigma_p^2$ 이고,

    D차원의 경우 $k = \left(\frac{\sigma_q}{\sigma_p}\right)^D$

    따라서 승인률은 D가 커짐에 따라 기하급수적으로 감소.

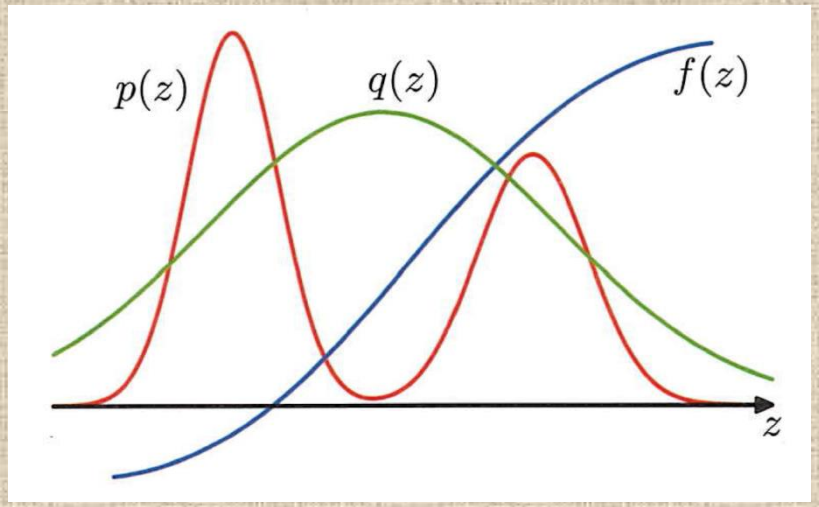
1-4 importance sampling(중요도표집법)

- 구역을 나눠서 기댓값을 구한다는 아이디어.
- $q(z)$로부터 $z^{(l)}$을 추출하고,
  각 항을 적절히 가중하여 합한다.

$$\sum f(z)\frac{p(z)}{q(z)}q(z)dz$$

$$\cong \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})$$

중요도 가중치



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. 11.8

1-4 importance sampling(중요도표집법)

- 표집분포가 원 분포에 얼마나 근접하느냐가 중요
- $p(z)$가 굉장히 변동이 심하면, 소수의 가중치가 큰 값을 가지게 되고 그 결과 유효표본숫자는 L보다 훨씬 더 적게 됨.
  - $p(z)$가 중요한 구간에서는 $q(z)$도 커야 한다

1-4 importance sampling(중요도표집법)

- SIR(sampling importance resampling)
  1. $q(z)$ 로부터 L 개의 표본을 추출
  2. 중요도 표집에 따라 각각의 가중치 결정
  3. 앞에 뽑은 L개로부터 resampling, 이때 가중치는 2단계에서 결정한 가중치 사용.
- $L \to \infty$일때 옳다는 것이 알려져 있음.

1-5 Markov chain Monte Carlo(MCMC)

- 몬테카를로 방법 : 예제 – π값 계산

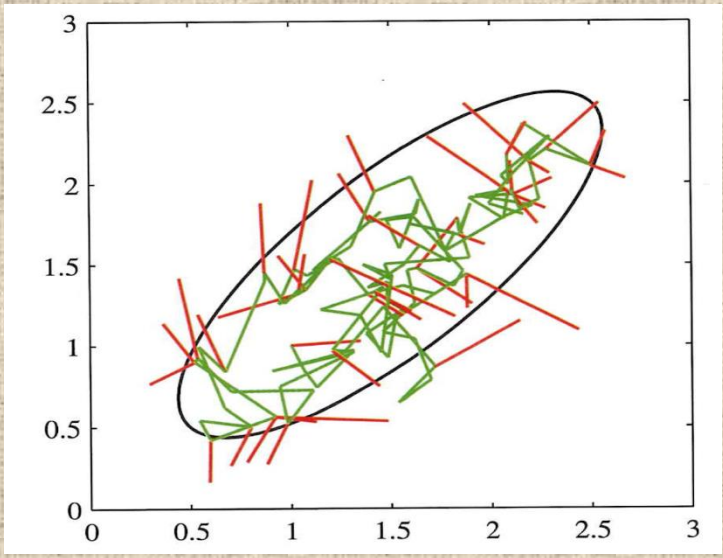
  1. 정사각형 안에 사분원을 그린다
  2. 일정한 개수(다다익선)의 점을 *균일*하게 찍는다
  3. 사분원 내부 점 수를 센다($x^2 + y^2 \leq 1$)
  4. 사분원 내부와 외부 점 수 비율 이용해서 구한다.

1-5 Markov chain Monte Carlo(MCMC)

- 표본공간차원이 고차원일때도 사용 가능
- 기본적인 메트로폴리스 알고리즘
  - 제안분포가 대칭임을 가정
  - 승인확률 $= \frac{\hat{p}(z^*)}{\hat{p}(z^{(\tau)})} > 1$
  - 표본이 승인되면 $\mathbf{z}^{(\tau+1)} = \mathbf{z}^*$, 거부되면 $\mathbf{z}^{(\tau)}$계속.

1-5 Markov chain Monte Carlo(MCMC)

- 메트로폴리스 알고리즘과 거부표집법 차이
  - 메트로폴리스에서는, 거부되었을 경우 버리지 않고 이전 샘플이 다시 포함됨 = 샘플이 중복됨
- 메트로폴리스 알고리즘 예시
  - 가우시안분포



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. 11.9

1-5 Markov chain Monte Carlo(MCMC)

- 마르코프 연쇄
  - 다음 조건부 독립성이 $m \in \{1, \dots, M-1\}$에 대해 성립

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

  - 전이확률

$$T_m\left(z^{(m)}, z^{(m+1)}\right) \equiv p(z^{(m+1)} | z^{(m)})$$

  - 모든 m에 대해 전이확률이 동일하면 동질적 homogeneous

1-5 Markov chain Monte Carlo(MCMC)

- 마르코프 연쇄(계속)
  - 특정 변수에 대한 주변확률을 다음과 같이 표현 가능

marginal probability

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

  - 연쇄에서 각 단계가 분포를 변화시키지 않으면,

    그 분포를 마르코프연쇄에 대해 불변, 정류.

    invariant, *stationary*

1-5 Markov chain Monte Carlo(MCMC)

- 마르코프 연쇄를 사용해서 주어진 분포로부터 표집하려면
  - 해당 분포가 연쇄에 대해 불변이어야 한다.
  - $m \to \infty$인 경우 어떤 초기분포를 택해도 $p(z^{(m)})$이 해당 불변분포 $p^*(z)$로 수렴해야 한다
    - 이 성질을 '에르고딕성' ergodicity
    - 이 불변분포를 '평형분포' equilibrium dist.

1-5 Markov chain Monte Carlo(MCMC)

- 메트로폴리스 헤이스팅스
  - 제안분포가 대칭일 필요가 없다.
  - 승인확률 $\frac{\tilde{p}(z^*)q_k(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q_k(z^*|z^{(\tau)})}$
- 제안분포는 주로 가우시안을 쓴다.
  - 분산이 너무 크면 거부율이 높고,
  - 너무 작으면 원 분포를 다 재현해 내는 데 오래 걸린다.

1-6 기브스 표집법 (Gibbs sampling)

- 메트로폴리스 헤이스팅스의 특수케이스
- 변수들 중 하나의 값을 나머지 변수들에 대한 해당변수의 조건부 분포에서 추출한 값으로 바꿈.
  - 예) 세 개의 변수들에 대한 분포 $p(z_1, z_2, z_3)$
    먼저 z_1을 $p(z_1|z_2, z_3)$으로부터 샘플링한 값으로 변경
    z_2, z_3의 경우도 마찬가지.

1-6 기브스 표집법 (Gibbs sampling)

- M-H의 특수케이스 증명을 위해,

  동질적 마르코프 연쇄의 경우 $p^*(z)$는 다음의 경우에 불변.

$$p^*(z) = \sum_{z'} T(z', z) p^*(z')$$

- 세부균형(detailed balance)을 만족하면 불변이다.

$$p^*(z) T(z, z') = p^*(z') T(z', z)$$

1-6 기브스 표집법 (Gibbs sampling)

- 증명

$$\sum_{z'} p^*(z')T(z',z) = \sum_{z'} p^*(z)T(z,z') = p^*(z)\sum_{z'} p(z'|z) = p^*(z)$$

- 깁스 샘플링은 세부균형을 만족한다.

$q(z'|z) = p(z'_k, z_{-k}|z_{-k}) = p(z'_k|z_{-k})$이므로,

$$p(z)q(z'|z) = p(z_k, z_{-k})p(z'_k|z_{-k}) = p(z_k|z_{-k})p(z_{-k})p(z'_k|z_{-k})$$

$$= p(z_k|z_{-k})p(z_{-k}, z'_k) = q(z|z')p(z') \because z'_{-k} = z_{-k}$$

1-6 기브스 표집법 (Gibbs sampling)

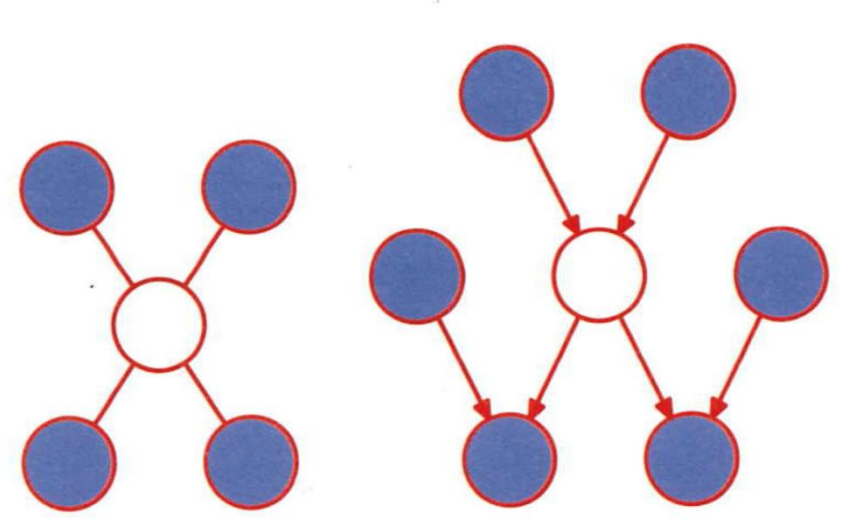
- 따라서 깁스샘플링은 불변, 동질이다.
- 깁스 샘플링에서, M-H의 승인확률은 언제나 1이다.

$$\frac{p(z')q(z|z')}{p(z)q(z'|z)} = \frac{p(z'_k|z'_{-k})p(z'_{-k})p(z_k|z'_{-k})}{p(z_k|z_{-k})p(z_{-k})p(z'_k|z_{-k})} = 1$$

- 따라서 깁스 샘플링은, M-H의 특수한 형태.

1-6 기브스 표집법 (Gibbs sampling)

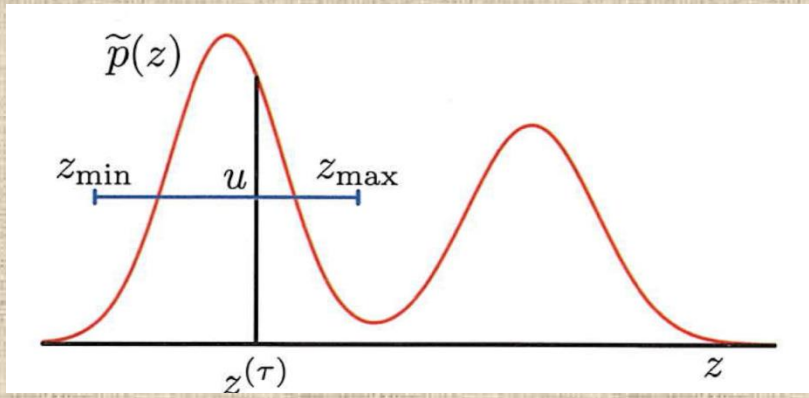
- 한 변수를 제외한 변수 모두의 조건부 분포는,
  그래프 관점에서 보면 마르코프 블랭킷에 존재하는
  노드들에만 의존한다는 뜻. : 깁스샘플링하기 쉽다.



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. 11.12

1-7 조각 표집법(slice sampling)

- 메트로폴리스 알고리즘의 난점
  - 각 단계 크기가 크면 높은 거부율로 비효율적
  - 크기가 작으면 임의보행행동으로 비상관화가 느림
- 그래서 분포의 성질에 대해 단계 크기를 맞추는 방식.



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. 11.13(b)

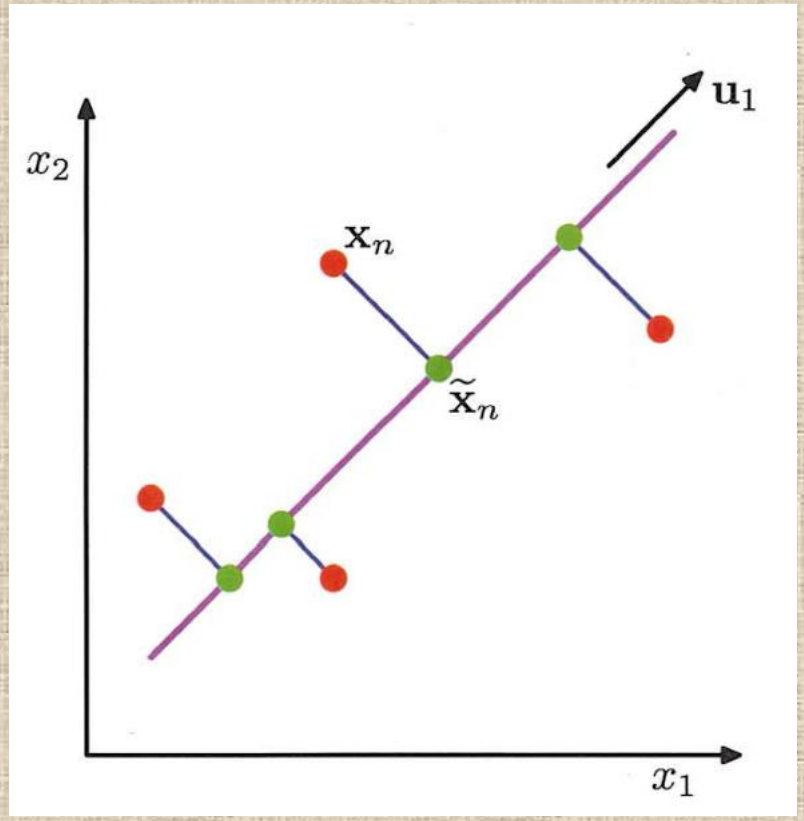
02
PCA(1)

2-1 Principle Component Analysis

- 차원수 감소, 데이터 압축, 특징추출, 데이터 시각화
- 데이터를 principle subspace에 투영
  - 보통 더 낮은 차원
  - 선형변환
  - 데이터의 분산을 최대화하는 방향으로 이루어짐
  - 평균투영비용을 최소화하는 방향도 동일
    - 비용 = 데이터포인트와 투영체간 평균제곱거리

2-1 Principle Component Analysis

- 예



Bishop. Fig 12.1

2-1 Principle Component Analysis

- 최대분산
  - 공간의 방향을 D차원벡터 $\mathbf{u}_1$으로 정의
  - $\mathbf{u}_1$은 단위벡터라고 가정(절대값 무관)
  - 각각의 데이터포인트는 $\mathbf{u}_1^T\mathbf{x}_n$에 투영됨.
  - 투영된 데이터의 평균은 $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$일 때 $\mathbf{u}_1^{\mathrm{T}}\bar{\mathbf{x}}$

2-1 Principle Component Analysis

- 최대분산(계속)
  - 투영된 데이터의 분산은, 공분산행렬$\mathbf{S}$가

    $\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^{\mathrm{T}}$ 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 주어짐.

  - $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$을 최대화 하는 것이 목표.

2-1 Principle Component Analysis

- 최대분산($\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$최대화)
  - $\mathbf{u}_1$이 단위벡터이므로,$\mathbf{u}_1^T\mathbf{u}_1 = 1$
  - 제약조건을 얻었으므로 라그랑주 승수 도입

$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1)$$

  - $\mathbf{u}_1$에 대한 미분을 0으로 두면,

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

2-1 Principle Component Analysis

- 최대분산(계속)
  - $\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$이라는 것은, $\mathbf{u}_1$이 $\mathbf{S}$의 고유벡터라는 뜻.
  - 양변에 $\mathbf{u}_1^{\mathrm{T}}$곱하면,

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1$$

  - $\mathbf{u}_1$을 가장 큰 eigenvalue λ_1을 가지는 eigenvector로 설정하면, 최대의 분산을 가지게 된다는 뜻.

2-1 Principle Component Analysis

- eigenvector, eigenvalue (고유벡터, 고윳값)
  - 체(field) K에 대한 벡터공간 V위의 선형변환 V 가 주어지면,
  - 어떤 $v \in V$와 $\lambda \in K$가
  - $Tv = \lambda v,\ v \neq 0$
- λ가 실수일 때, 정방행렬 A와 $Ax = \lambda x$성립하고,

  x는 nonzero vector.

2-1 Principle Component Analysis

- 가장 큰 eigenvalue를 가지는 방향이 '제1주성분'
- 찾은 eigenvector와 직교하는 모든 가능한 방향 중, 다시 가장 큰 분산을 가지는 방향을 찾을 수 있다.
  - 반복하면서 계속 추가 주성분을 구할 수 있다.
  - 모두 찾으면, 공분산행렬$\mathbf{S}$의 가장 큰 eigenvalue $\lambda_1, \ldots, \lambda_M$에 해당하는 고유벡터들 $\mathbf{u}_1, \ldots, \mathbf{u}_M$

2-1 Principle Component Analysis

- 요약
  - 데이터집합의 평균과 공분산행렬을 계산하고,
  - 공분산행렬의 가장 큰 M개의 고윳값들에 해당하는 M개의 고유벡터들을 찾는 과정.

2-1 Principle Component Analysis

- 최소오류공식화
  - 완전히 정규직교하는 D차원 기저벡터들 $\mathbf{u}_i$

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \delta_{ij}, \qquad (\delta\text{: Kronecker delta})$$

  - 따라서 모든 데이터 포인트는 이 기저벡터들의 선형결합.

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni}\mathbf{u}_i$$

2-1 Principle Component Analysis

- 최소오류공식화
  - 선형결합은 다음과 같이 적을 수 있다.($\mathbf{u}_i$는 정규직교)

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni}\mathbf{u}_i = \sum_{i=1}^{D} (\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i$$

  - 목표는 이것을 M<D인 M개의 변수로 근사해 내는것.
  - (다음시간에 계속)

다음시간

10강

- PCA(2)