Contents

Projection

- 운동 경기장에 카메라를 움직일 rail을 설치한다고 하자. 운동선수들의 움직임을 가장 잘 살펴볼 수 있는 위치는 어디일까?
- 선수들의 움직임이 가장 큰 방향에 직각이 되는 위치에서 촬영하면 움직임을 가장 잘 잡아낼 수 있다.
  - noise: SNR(signal-to-noise ratio)를 가장 작게
  - reduncancy: 일부 축의 정보로 다른 축의 위치를 알기 어렵게

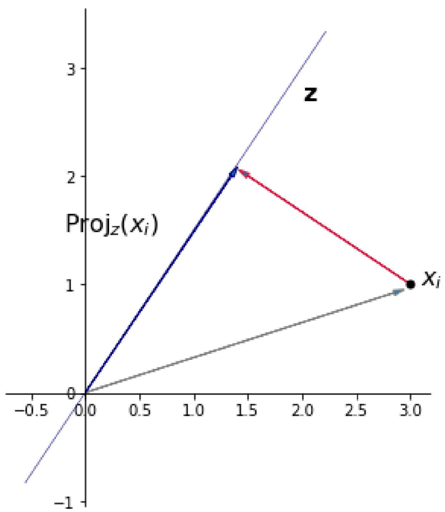
- 최근 word embedding space에서 필요한 해석이 가능한 축을 생성할 수 있도록 하는 embedding 을 시도하고 있다.
- Projection 방법은 다른 대상을 '직접' 비교할 수 있는 아주 유용한 방법이다.

- 선형변환, 혹은 matrix multiplication은 machine learning에서 광범위하게 사용하며, vector space의 기본 연산방법이다.
- Feature들의 column space에 투사한 후 projection을 비선형으로 변환하여 확률이나 예측 class를 대응시키는 것으로 해석하면 연속변수를 다루는 여러 통계모형들을 projection의 응용으로 해석할 수 있다.
- 한편 projection을 벡터의 선형변환으로 생각해보면 차원이 높은 vector space의 연산의 의미를 이해하는데 유용하다.

- 우리가 이해하고 있는 세상에 분석대상을 비추어 해석한다.
- 시야가 좁을수록 사물의 핵심을 제대로 파악하기 어렵지만, 지나치게 넓어도 문제가 있다.

- Vector space V의 vector $y \in V$의 특성을 가장 잘 반영하는 subspace $S \subset V$의 vector $z \in S$를 찾는 문제를 생각해보자.

Figure 1. Projection



Contents

1 용어

Othorgonal projection theorem 주어진 $y \in V$와 subspace $S \subset V$에 대해 다음 극소화문제는 유일한 해를 갖는다.

$$\hat{y} = \arg\min_{z \in S} \|y - z\|^2$$

여기서 $\hat{y} \in S$이며 $(y - \hat{y}) \perp S$이다.

- Orthogonal은 vector와 space의 최단거리는 둘이 직각으로 만나는 점에서 결정되기 때문에 붙여진 이름이다.
- 위 문제의 극소값은 y에 선형변환 P를 적용하여 계산할 수 있다. (orthogonal proejction mapping onto S)

$$\hat{y} = Py$$

Theorem X의 column들이 S의 basis라고 하면,

$$P = X(X^\intercal X)^{-1}X^\intercal \text{ such that } Py \in S \text{ and } (y - Py) \perp S$$

Definitions

- A **projection** on a vector space V is a linear operator $P : V \to V$ such that $P^2 = P$.
- It is **orthogonal** if $P^2 = P = P^\intercal$. Otherwise it is an oblique projection.

2 Scalar projection 과 vector projection

- $\mathbf{x} \in \mathbb{R}^n$ 를 $\mathbf{z} \in \mathbb{R}^n$ 에 project 하면, $\mathbf{z}$에 대하 $\mathbf{x}$의 projection $\mathrm{Proj}_{\mathbf{z}}(\mathbf{x})$는 다음과 같이 계산할 수 있다.

$$\mathrm{Proj}_{\mathbf{z}}(\mathbf{x}) = \underbrace{w}_{\text{scalar projection}} \times \mathbf{z}$$

2.1 Scalar projection

- 위 식에서 $w \in \mathbb{R}$는 scalar로 scalar projection of x onto z로 부르며 다음과 같이 계산한다.

$$\mathbf{z} \cdot (\mathbf{x} - w\mathbf{z}) = 0, \quad w = \frac{\mathbf{x} \cdot \mathbf{z}}{\mathbf{z} \cdot \mathbf{z}} = \underbrace{\frac{\mathbf{x} \cdot \mathbf{z}}{\|\mathbf{x}\| \cdot \|\mathbf{z}\|}}_{\cos\theta} \times \frac{\|\mathbf{x}\|}{\|\mathbf{z}\|}$$

- $\mathbf{z} = (1, \cdots, 1)$ 이면 $\mathbf{x}$의 $\mathbf{z}$에 대한 scalar projection은 산술평균이다.

- 연산에 사용하는 자료를 모두 단위 vector로 변환하여 사용한다면 다음과 같이 쓸 수 있다.

$$w = \hat{\mathbf{x}} \cdot \hat{\mathbf{z}}$$

- Machine learning algorithm에선 scalar projection을 주로 사용한다.
- 각 feature들의 label을 바꾸는 것은 예측결과에 아무런 영향도 미치지 않으므로 해당 축 위에서 상대적인 위치만 중요하다.

Contents

2.2 Dot product 와 similarity

- Unit vector들의 dot product는 두 vector 사이의 사잇각의 cosine 값이다.
- 이 방법은 difference와 함께 vector들 사이의 유사성을 측정하는데 많이 사용한다.

$$\cos\theta = \hat{\mathbf{x}} \cdot \hat{\mathbf{z}}$$

- 따라서 scalar projection이 1에 가까우면 projection vector $\hat{\mathbf{z}}$와 유사성이 높으며, 0에 가까우면 연관성이 없어진다.
- 실제 응용에선 $\hat{\mathbf{z}}$의 관계는 생략하고 vector들 사이의 scalar projection의 차이로 유사성을 판단한다.

2.3 Vector projection

- 다음과 같이 정리할 수 있다.

$$\mathrm{Proj}_{\mathbf{z}}(\mathbf{x}) = \frac{\mathbf{x} \cdot \mathbf{z}}{\|\mathbf{z}\|} \times \frac{\mathbf{z}}{\|\mathbf{z}\|} = (\hat{\mathbf{z}} \cdot \mathbf{x})\,\hat{\mathbf{z}}$$

- Projection vector $\mathbf{z}$에 대한 scalar projection을 새로운 feature로 사용하기도 한다.
- Classification으로 사용하기도 하지만 기본용도는 dimension reduction이므로 이론모형 검증 등의 목적이 아니라면 projection vector의 방향을 분석에 사용하지 않는다.

Contents

3 Descriptive statistics

3.1 Mean

- $o = (1, 1, \cdots, 1)$ 를 n-component column vector, X는 (n, K) matrix이다

$$\bar{X} = \hat{\mu} = \frac{o^\intercal X}{o^\intercal o}$$

3.2 Deviation, variance

- Deviation 은 평균이 0인 projection vector로 project 하는 것으로 생각할 수 있다.
- Projection operator P는 각 원소에서 평균을 빼는 역할을 하므로 $P^2 = P$이고 $P^\intercal = P$ 또한 만족한다.

$$X - \bar{X} = X - \frac{o^\intercal X}{o^\intercal o} o = \left(I - \frac{o\, o^\intercal}{o^\intercal o} \right) X \equiv PX$$

- 분산의 불편추정량은 deviation의 제곱합을 자유도로 나눈 것이므로 다음과 같이 구할 수 있다.

$$\mathrm{Var}(x_k) = \frac{\|Px_k\|^2}{n-1} = \frac{(Px_k)^\intercal (Px_k)}{n-1} = \frac{x_k^\intercal P x_k}{n-1}$$

3.3 Covariance and Correlation

$$\mathrm{Cov}(X) = \frac{(X - \bar{X})^\intercal (X - \bar{X})}{n-1} = \frac{(PX)^\intercal (PX)}{n-1} = \frac{X^\intercal P X}{n-1}$$

$$\mathrm{corr}(x_h, x_k) = \frac{\mathrm{Cov}(x_h, x_k)}{\sqrt{\mathrm{Var}(x_h) \times \mathrm{Var}(x_k)}} = \frac{(Px_h)^\intercal (Px_k)}{\|Px_h\| \|Px_k\|} = \frac{x_h^\intercal P x_k}{\|Px_h\| \|Px_k\|}$$

4 Overdetermined linear equation system & OLS

$$Ax = b$$

- b를 A의 columns의 선형결합으로 표시할 수 없다면 해가 존재않고, 이는 $b \notin \mathrm{span}(A)$을 의미한다.
  - $\mathrm{rank}(A|b) = \mathrm{rank}(A)$ 라면 해가 존재하지만 $\mathrm{rank}(A|b) > \mathrm{rank}(A)$ 이라면 해가 존재하지 않는다.
- b를 A의 columns의 선형결합으로 설명하기 위해서는 근사값을 사용해야한다.

$$Ax + \varepsilon = b$$

- 대부분 b와 $\mathrm{span}(A)$의 거리를 가장 작게하는 '해' 혹은 weight를 구한다.

$$\min_x \|b - Ax\|^2 = \min_x (b - Ax)^\intercal (b - Ax)$$