Insper

Machine Learning

Aula 13 – Redução de dimensionalidade

2020 – Engenharia
Fábio Ayres <fabioja@insper.edu.br>

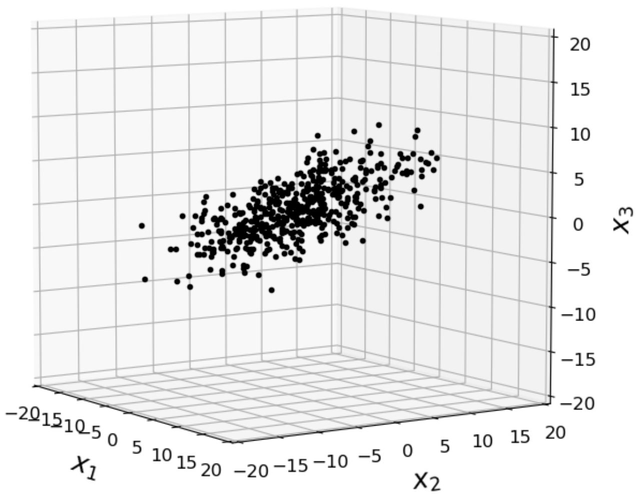
Redução de dimensionalidade

- Como visualizar um dataset de dimensão $n = 200$?

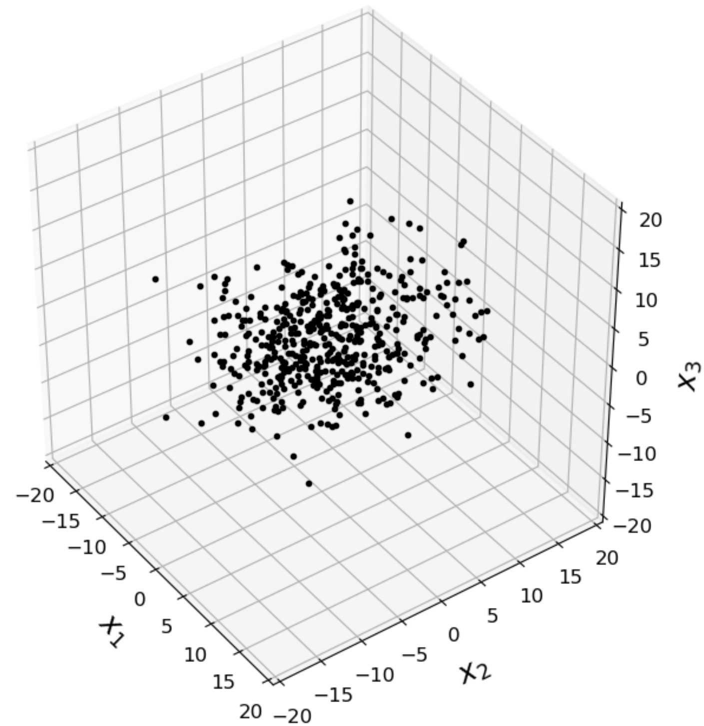
- Será que em um dataset de dimensão $n = 200$ realmente temos tudo isso de informação independente?

- Como "enxugar" um dataset?

Exemplo



20
15
10
5
0
−5
−10
−15
−20
x_3
−20
−15
−10
−5
0
5
10
15
20
x_1
−20
−15
−10
−5
0
5
10
15
20
x_2



20
15
10
5
0
−5
−10
−15
−20
x_3
−20
−15
−10
−5
0
5
10
15
20
x_1
−20
−15
−10
−5
0
5
10
15
20
x_2

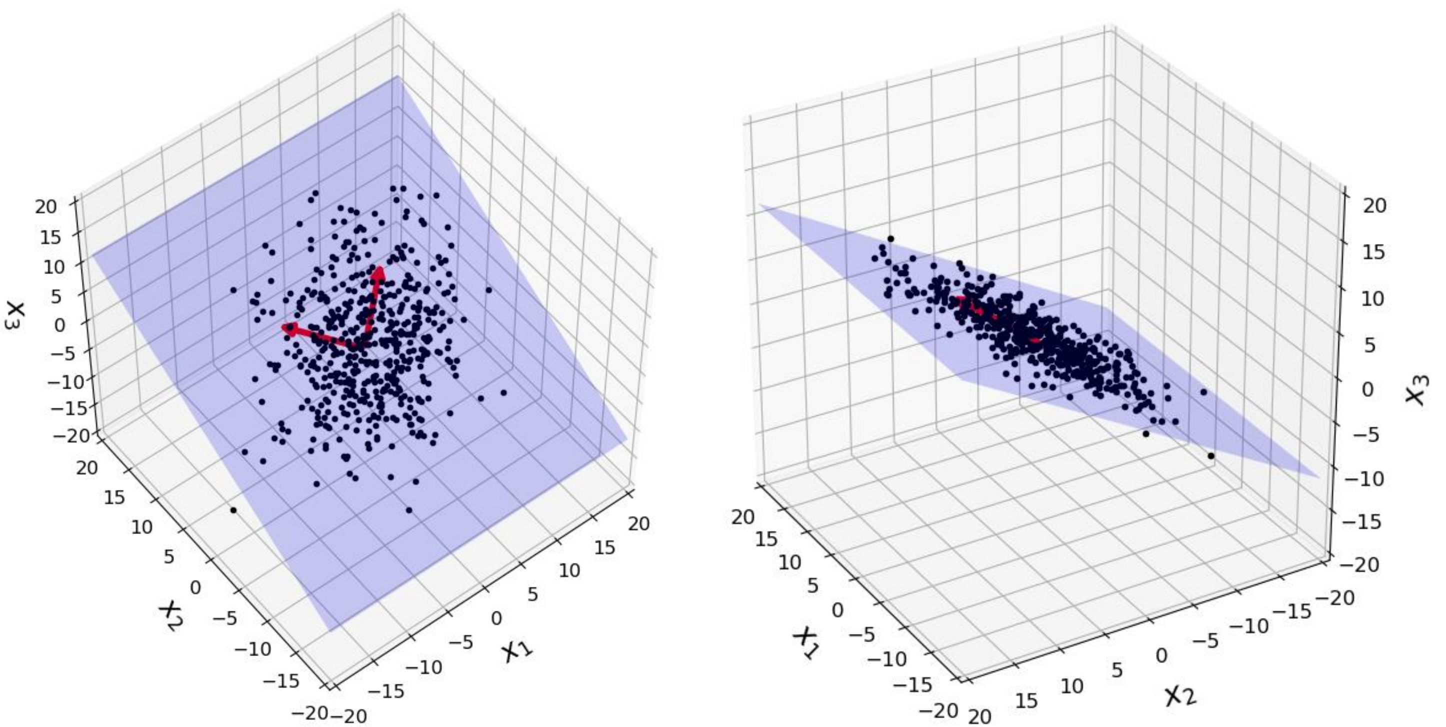
Ideia: projeção

Podemos projetar esses dados em um hiperplano de menor dimensão!

Assim só precisamos guardar:

- A origem do plano
- Os vetores determinando o hiperplano
- As coordenadas de projeção por ponto: $d \ll n$

Exemplo



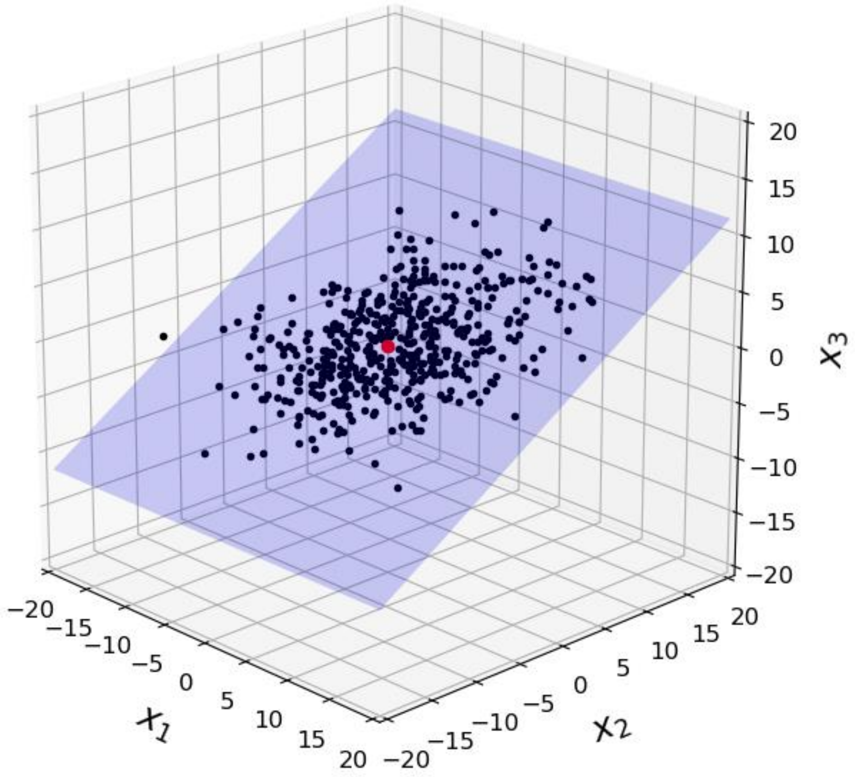
Aproximações sucessivas

- $d = 0$: não guardo nenhuma coordenada por ponto!
  - Guardo apenas o centroide da nuvem!

- $d > 0$: começo a gerar componentes de projeção por ponto.

$$\hat{x}_i = p + \alpha_{i1} w_1 + \alpha_{i2} w_2 + \cdots + \alpha_{id} w_d$$

Zero-ésima aproximação



Zero-ésima aproximação

- Vamos chamar de $\hat{x}_i$ a aproximação do ponto x_i

- Como temos uma aproximação de um ponto só, $\hat{x}_i = p$ para algum ponto p. Que ponto é esse?

Zero-ésima aproximação

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\|x_i - p\|^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - p)^T(x_i - p)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(x_i^T x_i - 2p^T x_i + p^T p\right)$$

$$\frac{\partial}{\partial p}MSE = \frac{1}{m}\sum_{i=1}^{m} -2x_i + 2p = 0$$

$$\Rightarrow \boxed{p = \frac{1}{m}\sum_{i=1}^{m} x_i}$$

O ponto ótimo é o centroide!

Próximas aproximações

- Agora vamos descobrir a próxima aproximação: $d = 1$

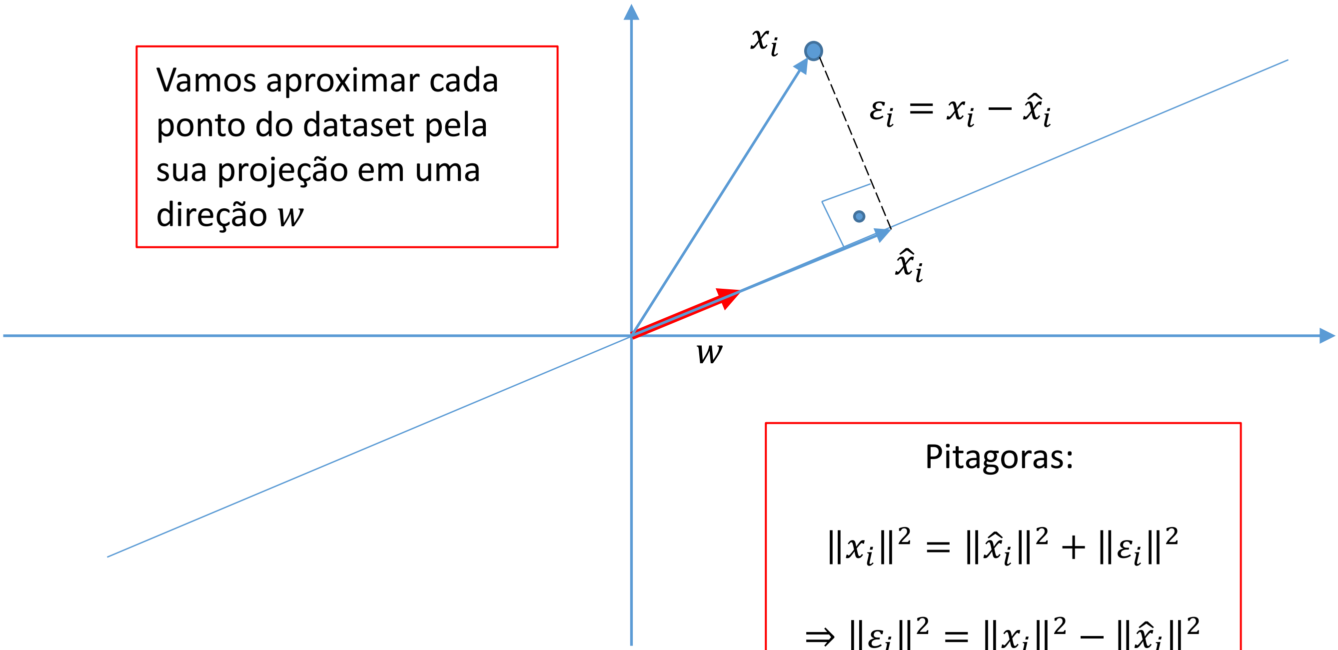
- Vamos remover a zero-ésima aproximação do dataset:

$$x_i \leftarrow (x_i - p)$$

- Nosso dataset (remanescente) tem uma nuvem de pontos cujo centróide é a origem

Próximas aproximações

Vamos aproximar cada ponto do dataset pela sua projeção em uma direção w



Pitagoras:

$$\|x_i\|^2 = \|\hat{x}_i\|^2 + \|\varepsilon_i\|^2$$

$$\Rightarrow \|\varepsilon_i\|^2 = \|x_i\|^2 - \|\hat{x}_i\|^2$$

Próximas aproximações

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\|\varepsilon_i\|^2 = \frac{1}{m}\sum_{i=1}^{m}\left(\|x_i\|^2 - \|\hat{x}_i\|^2\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\|x_i\|^2 - \frac{1}{m}\sum_{i=1}^{m}\|\hat{x}_i\|^2$$

Portanto,

$$w_1 = \arg\min_{w} MSE = \arg\max_{w} \frac{1}{m}\sum_{i=1}^{m}\|\hat{x}_i\|^2$$

Maximizando as projeções

$$\hat{x}_i = \text{projeção de } x_i \text{ na direção } w$$

$$\Rightarrow \hat{x}_i = \left(\frac{w^T x_i}{w^T w}\right) w$$

$$M(w) = \frac{1}{m}\sum_{i=1}^{m} \|\hat{x}_i\|^2 = \frac{1}{m}\sum_{i=1}^{m} \hat{x}_i^T \hat{x}_i$$

$$= \frac{1}{m}\sum_{i=1}^{m} \left(\frac{w^T x_i}{w^T w}\right)^2 w^T w = \frac{1}{m}\sum_{i=1}^{m} \frac{(w^T x_i)^2}{w^T w}$$

Maximizando as projeções

$$M(w) = \frac{1}{m}\sum_{i=1}^{m}\frac{(w^T x_i)^2}{w^T w} = \frac{1}{m}\sum_{i=1}^{m}\frac{w^T x_i x_i^T w}{w^T w}$$

$$\Rightarrow M(w) = \frac{w^T C w}{w^T w}$$

onde

$$C = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T$$

Matriz de covariância

linhas

colunas

$$X = \begin{bmatrix} \boxed{x_1^T} \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} \qquad X^T = \boxed{[x_1} \quad x_2 \quad \cdots \quad x_m]$$

$$\frac{1}{m} X^T X = \frac{1}{m} [x_1 \quad x_2 \quad \cdots \quad x_m] \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T$$

Logo: $C = \frac{1}{m} X^T X$, que é a matriz de covariância dos dados!

Maximizando as projeções

Portanto, queremos achar w que maximiza

$$M(w) = \frac{w^T C w}{w^T w}$$

E agora?

Autovalores e autovetores

- Imagine o que acontece se w for um autovetor (dentre vários) da matriz C: $Cw = \lambda w$
- Portanto

$$M(w) = \frac{w^T C w}{w^T w} = \lambda \frac{w^T w}{w^T w} = \lambda$$

- Agora temos uma estratégia para maximizar $M(w)$:

$$w = \text{autovetor de } C \text{ correspondente ao maior autovalor}$$

Próximas aproximações?

Repita o processo:

- Remova a aproximação feita até o momento

$$X_k = X - \hat{X}_{k-1}$$

- Calcule a matriz de covariância atual

$$C_k = Cov[X_k] = \frac{1}{m} X_k^T X_k$$

- Calcule o autovetor do maior autovalor da matriz C_k

Principal Component Analysis

Ufa! Se você chegou até aqui, parabéns! Você acabou de derivar o método das componentes principais, um dos algoritmos mais importantes da estatística!

Principal Component Analysis

Calcule os parâmetros da PCA:

1. A média p das amostras

2. A matriz de covariância C

3. Os d autovetores w_k correspondentes aos maiores autovalores de C

Principal Component Analysis

Agora calcule as componentes principais de cada ponto de dados:

1. Remova a média p de cada amostra x_i

2. Projete os pontos resultantes nas direções principais

Principal Component Analysis

- Medida de “desempenho”: soma dos autovalores
  - Representa o quanto da variabilidade original do dataset foi capturada no dataset reduzido

- Aplicações
  - Análise das direções: podem indicar os aspectos mais importantes do dataset
  - Compressão: representar aproximadamente o dataset com menos dados
  - Visualização: projetar em 2D ou 3D para observar como os dados se espalham
  - Redução de computação: treinar modelos mais rapidamente sem perder muita qualidade

Principal Component Analysis

Implementações:

- Decomposição por Valores Singulares (SVD: Singular Value Decomposition)

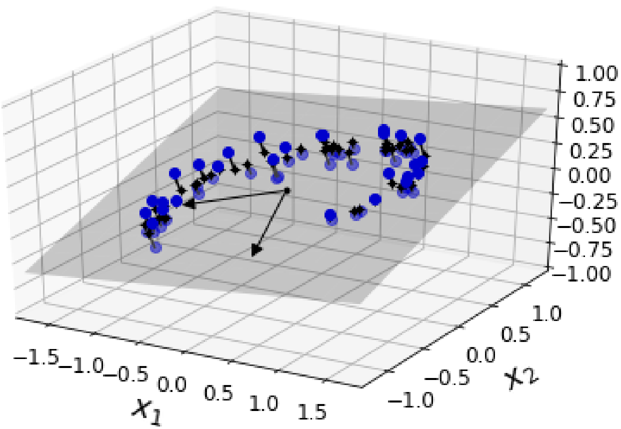
- PCA incremental

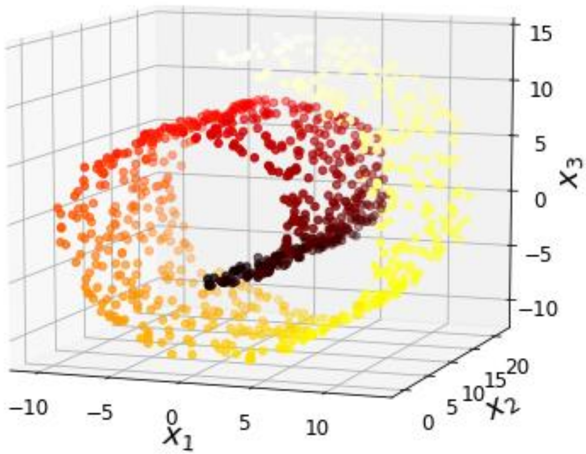
- PCA aleatorizada

Estude mais sobre essas técnicas em seu livro texto

Limitações

- Só projeta os pontos em hiperplanos
- E se os pontos estão em outras superfícies?





Outras técnicas

Existem técnicas mais sofisticadas para redução de dimensionalidade, que lidam com a não-linearidade:

- Kernel PCA
- Local Linear Embedding
- t-Distributed Stochastic Neighbor Embedding (t-SNE)