INF 493/792
Tópicos especiais III
Introdução à mineração de dados

Aula VI: Redução de dimensionalidade

Objetivo de hoje/sexta

Apresentação do problema

Conceitos básicos (revisão)

PCA (construção algébrica) – Parte I

Objeto básico de estudo

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

Possíveis problemas

Quando o valor de **d** (número de atributos) é muito grande:

- Maldição da dimensionalidade (aula passada)
- Dados podem ser “redundantes”

Conceitos e pressupostos básicos

Vamos assumir que as colunas de **D** tem média zero

$$\mathbf{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

Vetores do espação m-dimensional são matrizes com m linhas

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{id} \end{pmatrix}^T \in \mathbb{R}^d$$

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Versão matricial do produto interno

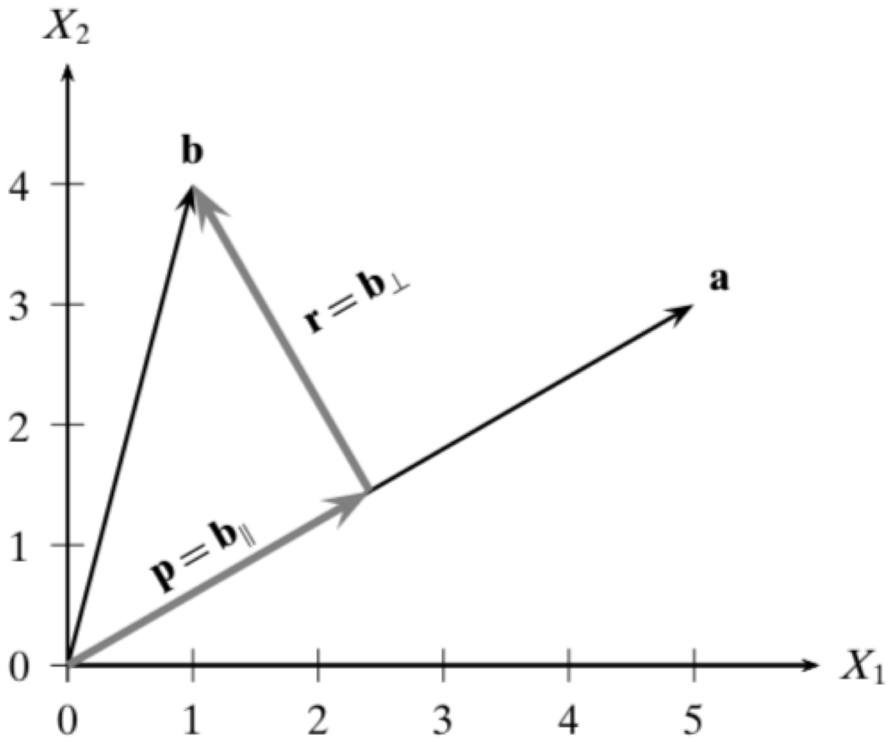
$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y}$$

Autovalores e Autovetores

Seja **T** uma transformação linear

$$\mathbf{Tx} = \lambda \mathbf{x}$$

Projeção ortogonal



$$\mathbf{p} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}\mathbf{a}$$

Matriz de covariância

Variabilidade conjunta de colunas de **D**

$$\mathbf{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

Rank de uma matriz (vs. *rank* efetivo)

Número de linhas/colunas linearmente independentes

Teorema espectral

Seja **A** (**m** x **m**) uma matriz simétrica. Então

Teorema espectral

Seja **A** (**m** x **m**) uma matriz simétrica. Então

1) Autovalores são reais, assim como seus autovetores correspondentes

Teorema espectral

Seja **A** (**m** x **m**) uma matriz simétrica. Então

1) Autovalores são reais, assim como seus autovetores correspondentes
2) Autovetores associados a autovalores distintos são ortogonais
3) $\mathbf{A} = \mathbf{ULU}^T$

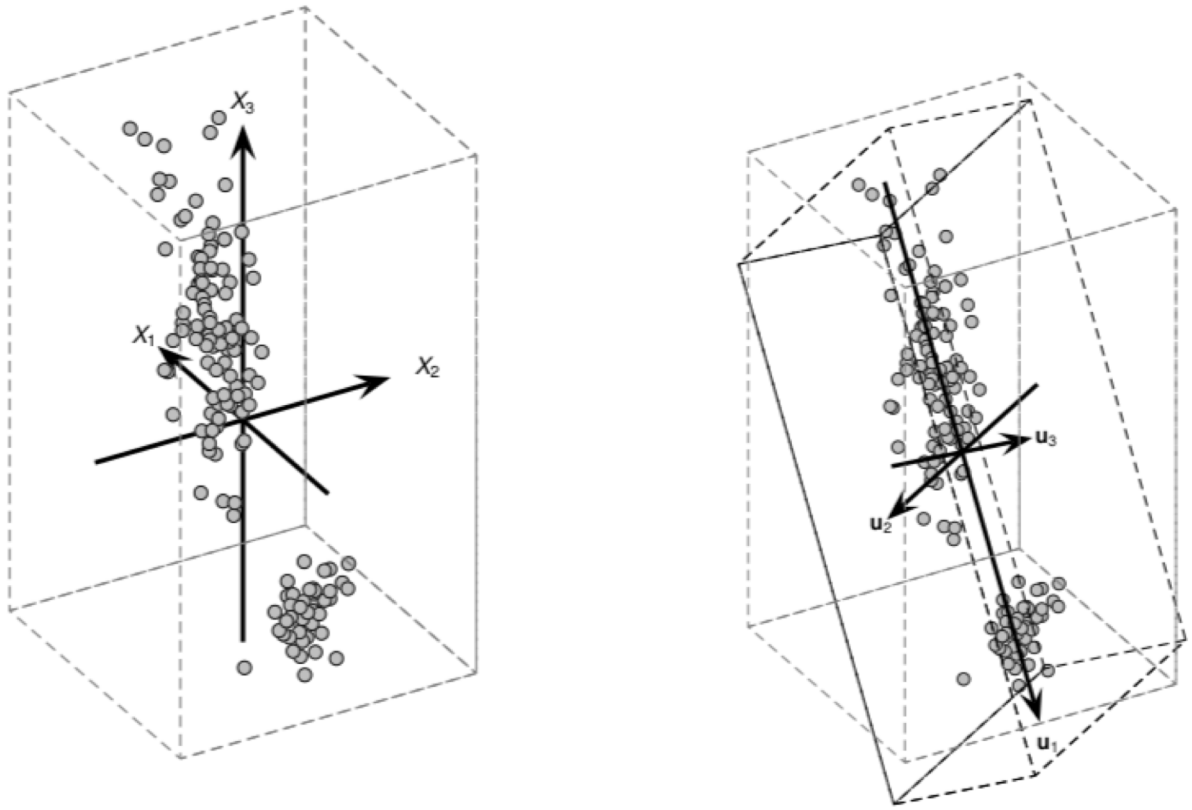
L é uma matriz diagonal (com autovalores de **A**)
U é uma matriz ortogonal (colunas são autovetores de **A**, os quais possuem norma 1)

Mudança de base

Como representar um ponto em uma outra base?

Objetivo de hoje

Encontrar uma base mais adequada para representar os dados



Há uma infinidade de bases

Gostaríamos de encontrar uma que:

- Seja ótima (o que é ótimo?)
- Nos permita, de uma forma natural, reduzir a dimensionalidade de **D** sem perder muita informação

Esses dois requisitos estão relacionados

PCA

Análise de Componentes principais (*Principal Component Analysis*)

Busca uma base de dimensão **r** que melhor captura a variância dos dados

Primeiro caso

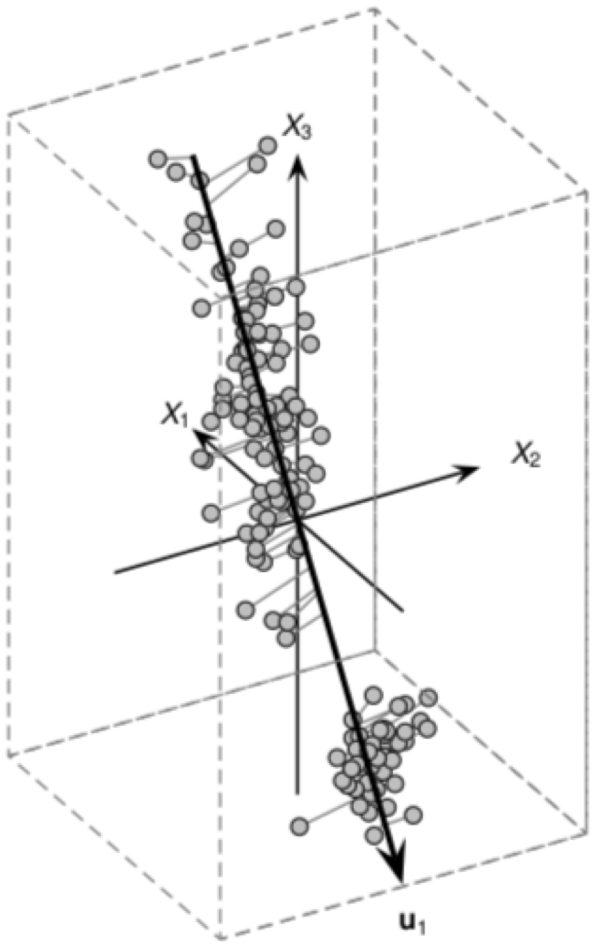
Dada uma matriz **D** (**n** x **d**), qual a melhor aproximação, em uma única dimensão para **D**?

Onde **melhor** significa capturar mais variância dos dados

Fato interessante

A direção da maior variância é também a direção que minimiza o erro quadrático médio.

Visualização



Sobre cálculo matricial

https://atmos.washington.edu/~dennis/MatrixCalculus.pdf

Leitura recomendada

http://www.dataminingbook.info/pmwiki.php

Capítulo 7