MO431A - Tarefa 1- Versão 1

Jacques Wainer

Versão 1

Data limite: Entregue um pdf, via moodle até meia noite de 13/4

Pode ser feito individualmente ou em grupos de até 3 pessoas.

Para este projeto você precisa ter o Python 3.X, numpy, matplotlib e sklearn. O projeto pode ser feito também em Matlab e R.

Provavelmente você deve também usar o Jupyter como modo interativo já que ele imprime as imagens (veja abaixo) na pagina com a interação.

Entregue apenas 1 **pdf** com o programa e resultados e seus comentários. Pode ser o pdf que mistura programas, resultados e suas respostas (por exemplo usando o Jupyter) ou um texto com os resultados e respostas, com o programa ao final.

1 Leia o arquivo dados.npy

npy é um formato do numpy para armazenar matrizes de forma mais compacta que usando, por exemplo, um .csv

o numpy.load lê arquivos npy

dados.npy é um arquivo de 10500 linhas e 784 colunas. Mas cada linha é na verdade uma imagem em tons de cinza de 28 por 28 pixels de dígitos (parte do banco de dados MNIST)

dados.npy esta em
[http://www.ic.unicamp.br/~wainer/cursos/1s2021/dados.npy

Vamos chamar a matrix de dados lidos de X

2 Imprima a imagem do 3 primeiros dígitos

A função imshow do subpacote pyplot do matplotlib imprime uma imagem. Mas cada linha da matriz X precisa ser transformada numa matriz 28x28 para que o imshow funcione (veja o reshape do numpy). Ha também a codificação de cores da imagem; como a imagem é em tons de cinza a codificação é a cm.gray

3 Faça a fatoração svd da matriz X.

A função é svd, do subpacote linalg do numpy faz a fatoração svd.

Nao se esqueça de normalizar os dados para média 0. **NAO** normalize para o desvio padrão 1 (se voce fizer quase tudo fica igual, mas nao a questão 7 abaixo).

Faça a fatoração *full_matrix* e a compacta

Verifique o tamanho das matrizes

4 SVD truncado

Vamos usar a redução para 100 dimensões.

4.1 Compute a matriz projetada. Não é preciso imprimi-la (ela sera uma matriz 10500 por 100). Apenas imprima as dimensões.

4.2 compute a matriz reconstruida. De novo não é preciso imprimi-la (ela sera uma matriz 10500 por 784). Apenas imprima as dimensões.

Nesta tarefa voce pode usar as matrizes computadas na tarefa anterior (o SVD nao truncado) e pegar as submatrizes apropriadas e fazer as multiplicações, ou usar o PCA https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html ou o TruncatedSVD https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html do sklearn

5 Imprima a imagem reconstruída dos 3 primeiros dígitos

Compare com as imagens originais impressas acima

6 Imprima os 3 primeiros eigen-dígitos

7 Decidindo o número de dimensões

7.1 Quantas dimensões manter usando a regra de usar singular values maior que 1

7.2 Quantas dimensões manter para capturar 80% da variância dos dados

7.3 Quantas dimensões manter para capturar 95% da variância dos dados