Trabalho de Reconhecimento de Padrões

Estudo do uso das técnicas de redução de dimensionalidade para uma imagem binaria

Por: Kevin De Mello Santamaría

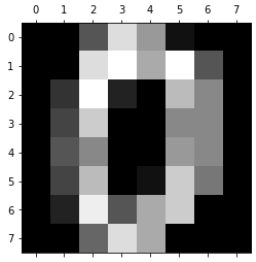
Introdução

Este trabalho em o objetivo de medir o comportamento de cada método de redução de dimensionalidade para cada método de classificação, comparando os scores de cada método. Diga-se de passagem, que o uso de cada método muda dependendo do tipo de dados que você introduzir no sistema. Para esse sistema que representa uma imagem que é um conjunto de dados que representa uma escala de grises, a conservação da dimensionalidade é importante para não perder a informação dos pixeis.

Materiais e métodos

A biblioteca utilizada para realizar este trabalho foi a digits, que é um conjunto de imagens que representam números de uma forma binaria.

Figura 0.1. Imagem de um número.



Na realização de este trabalho se usou a plataforma anaconda que já tem instalado as ferramentas utilizadas no Python, como:

- Matplolib: biblioteca usada para desenhar gráficos e mostrar de forma visual os dados.
- Numpy: biblioteca usada para manipular vetores, matrizes e dados numéricos no geral.
- SciPy: biblioteca que traz ferramentas cientificas para manipulação de dados.

Se usou os seguintes métodos para redução de dimensionalidade na fermenta Sci-Kit Learn:

- PCA: processo usado para transformação ortogonal dos dados fazendo uma mudança de dimensionalidade, convertendo um conjunto de dados de variáveis possivelmente correlacionadas a um conjunto menor. Onde a projeção no primeiro eixo fica os fados de maior variância e no segundo de variância menor e assim por diante.
- LDA: método utiliza informações das categorias associadas a cada padrão para extrair linearmente as características, mas discriminantes, onde a separação e enfatizada através da substituição da matriz covariância total do PCA por uma medida de separabilidade como o critério de Fisher.
- Kernel PCA (COSSINE): é uma extensão da análise PCA usando técnicas de kernel, para lidar com a não linearidade da informação.

- ISOMAP: este método encontra uma relação que preserva a geometria não lineal global preservando a distância geodésica entre os pontos, se dividindo em duas partes em que aproxima as distancias geodésicas e acha uma projeção que preserva essas distancias.
- LLE: este método determina para cada ponto um conjunto de pesos W que melhor reconstrói a partir de seus K-vizinhança mais próximos, com os pesos fixos se determina um conjunto de dados de menor dimensão que preserva as relações de vizinhança desses pesos.
- Laplacian Eigenmaps:
- T-SNE: método criado para dados com várias dimensões, que minimiza a divergência Kullback-Leibler entre medidas de similaridade em D e d d<<D dimensões.

Experimentos e Resultados

A finalidade do projeto é comparar os diferentes métodos de redução de dimensionalidade com o a acurácia dos métodos de aprendizado de máquina, sendo estes o perceptron, regressão logística, N vizinhança, centroide mais perto, SVM, K médios e métodos gaussianos. Se reduziu a dimensionalidade dos dados de 64 para 2 dimensões, para cada um dos métodos e posteriormente aplicando uma função que trazia como resultado uma lista com a acurácia do método e dois gráficos com a clusterização de cada método.

Assim, como observado na tabela 1.1 temos no geral o método que melhor funcionou para a redução da dimensionalidade foi o TSNE, obtendo uma de 86,23% superando a todos os outros métodos em média, só sendo superado pelo KernelPCA (cosseno) no PERCEPTRON, o que quer dizer para uma rede neural simples este método se mostra melhor para técnicas de clusterização.

Tabela 1.1. Resultados do score por cada método de classificação por a redução de dimensionalidade.

	Perceptron	Logistic Regression	MLPClassifier	KNeighborsClassifier	NearestCentroid	QuadraticDiscriminantAnalysis	SVM
PCA	24.24	60.95	64.73	63.73	61.84	65.18	65.29
LDA	43.15	68.29	70.30	67.74	69.63	69.85	71.96
Isomap	24.47	67.74	67.29	80.08	66.85	72.96	77.86
KernelPCA	44.27	59.51	64.40	61.40	61.73	67.63	65.40
LLE	12.56	10.78	20.02	92.65	73.41	74.97	50.16
SpectralEmbedding	21.02	9.12	9.12	77.86	68.74	71.07	76.41
TSNE	32.70	91.10	98.22	98.10	93.88	92.99	96.66

Entre os métodos usados observa-se no gráfico 1.1 como os pontos ficaram distribuídos no espaço após a transformação dimensional feita no conjunto de dados, e o que vem à tona que a clusterização obedece a distribuição de pontos no espaço os mudando os grupos, logo no método LLE os pontos ficam concentrados numa linha só o que indica que é o pior método para se usar neste caso.

Gráfico 1.1 Distribuição de pontos por método de redução de dimensionalidade.

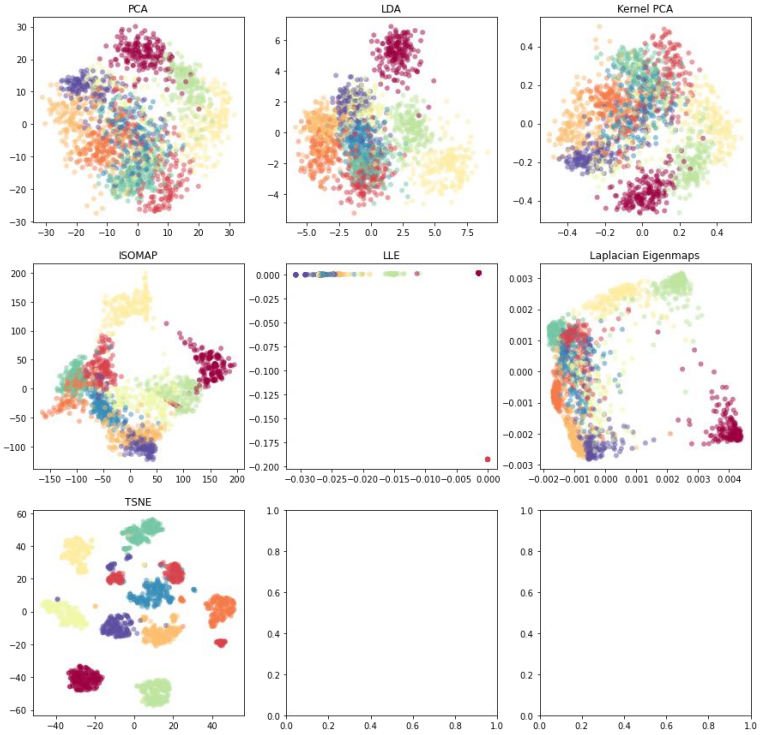
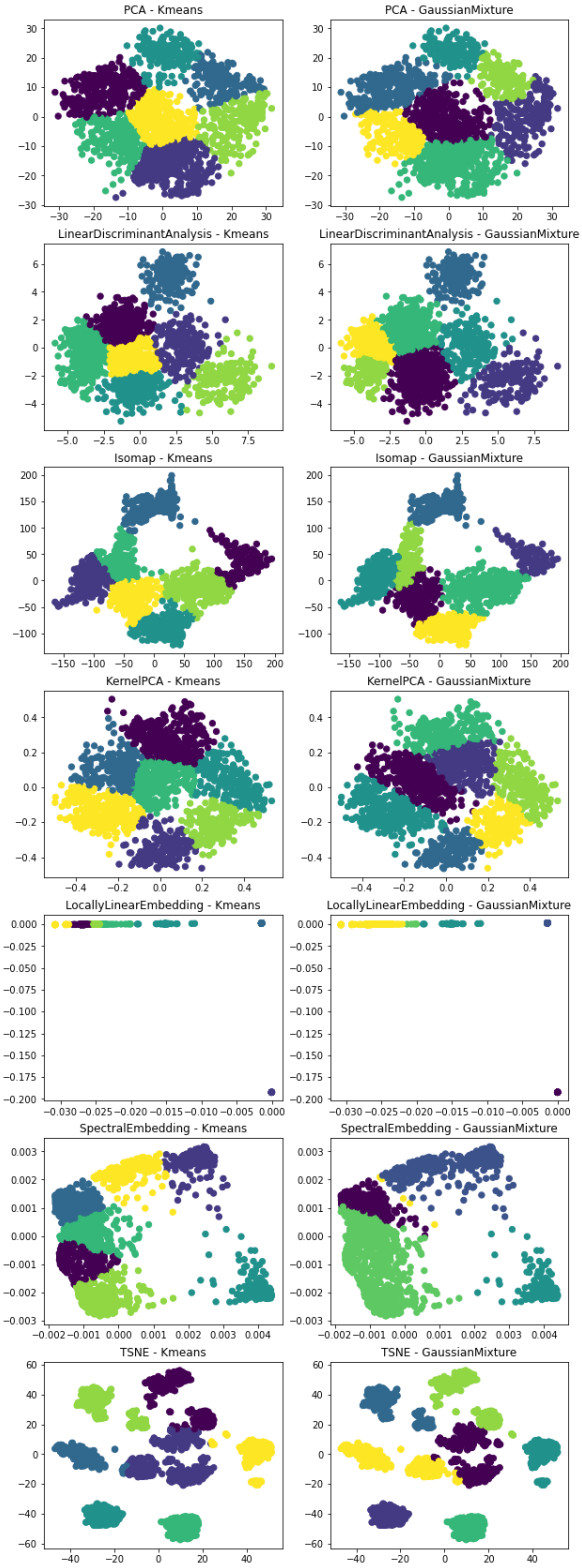


Gráfico 1.2. Métodos Kmeans e GaussianMixture por método de redução



Conclusões

Pode-se dize que para a biblioteca digits do SCI KIT LEARN o melhor método de redução é o TSNE pois como é uma imagem de um número tem multidimensionalidade o que precisa que se preservem as dimensões, já que a representação de dados não tem nenhum tipo de relação porque eles só representam um pixel dentro de uma imagem. O que os outros métodos não respeitam pela naturalidade de cada método. Para outros tipos de dados, o resultado vai mudar.

Referências

[1] Material de Aula

[2] https://scikit-learn.org/0.21/documentation.html