Lista de Tareas

27 de septiembre de 2022

PCA

glass.csv

Usando la base de datos de vidrios ya estudiada y curada correctamente:

1. Varíen la cantidad de componentes principales con las que se quedan, para graficar el error cuadrático medio en función de la cantidad de componentes principales. Repita el mismo procedimiento para el porcentaje de energía.

2. A partir de la componente principal, genere 5 datos sintéticos de vidrio (simular). Para ello estime la distribución de la componente principal mediante un histograma. Puede usar "plt.hist(?,bins=?,density=True)".

prostata_data.csv y prostata_label.csv

Usando la base de datos de cancer ya estudiada y curada correctamente:

1. Varíen la cantidad de componentes principales con las que se quedan, para graficar el error cuadrático medio en función de la cantidad de componentes principales. Repita el mismo procedimiento para el porcentaje de energía.

2. Grafique las dos componentes principales, marcando con colores distintos cada clase. ¿Que clase aparenta ser más fácil de clasificar?

MNIST

Utilizando las bases de datos estandard MNIST:

1. Utilice los datos de entrenamiento para definir los autovectores de PCA y utilice los datos de testeo para variar la cantidad de componentes principales con las que se quedan para graficar el error cuadrático medio en función de la cantidad de componentes principales. Repita el mismo procedimiento para el porcentaje de energía.

2. Utilice los datos de entrenamiento para definir los autovectores de PCA y haga la reconstrucción sobre los datos de testeo. A ojo estime cuantas componentes principales se necesitan como mínimo para distinguir los dígitos.

3. *Manifold Learning:* La idea es explorar el manifold encontrado por un PCA con 2 componentes principales. Utilice los datos de entrenamiento para entrenar el módulo de PCA, transforme los mismos datos al espacio latente de dimensión dos y anote los valores mínimos y máximos de cada componente. Ahora defina una grilla regular de 10×10 entre los valores mínimos y máximos obtenidos. Cada punto de esa grilla (100 en total) debe ser reconstruido en una imagen y mostrar los 100 dígitos reconstruidos en una grilla de 10×10 representativa del espacio latente. No sea muy optimista al respecto.

4. *Detección de anomalías:* Utilice los datos de entrenamiento para entrenar el módulo de PCA con la cantidad de componentes principales encontrada en el item 2. Arme una nueva base de datos combinando el conjunto de testeo de MNIST con el conjunto de testeo de FASHION-MNIST (estas van a hacer las veces de imágenes anómalas). Construya un detector de anomalías comparando el error cuadrático de la reconstrucción contra un umbral a definir. En lugar de elegir un umbral, grafique la curva ROC y calcule el *equal error rate* (el error para el umbral que hace iguales a los dos tipos de errores). Va a tener que efectuar una búsqueda bibliográfica para entender las curvas ROC (haga su propia implementación!). Repita para 1 componente principal y para 10 y grafique las 3 ROC superpuestas. Extraiga conclusiones.

5. *Pre-processing:* Elija una cantidad de componentes para que la clasificación sea aún nítida (pero no sea aburrido, full no vale) y entrene un módulo PCA con los datos de entrenamiento. Utilice las representaciones latentes para desarrollar un clasificador logístico. Repita el procedimiento pero usando las reconstrucciones. Compare ambos clasificadores entre ellos y con el que desarrollo anteriormente (sin PCA). Extraiga conclusiones.

BuzzFeed

Usando la base de datos BuzzFeed-Webis Fake News Corpus 2016 (la de artículos de izquierda, derecha, etc) y la vectorización de fasttext entrene dos módulos de PCA de dos componentes principales cada uno (uno con los datos de derecha y otro con los datos de izquierda). Elija 20 palabras entre las más frecuentes y proyectelas en cada una de las representaciones. Extraiga conclusiones.