Facultad de Matemática, Astronomía y Física y Computación
Universidad Nacional de Córdoba

Ciencia de Datos

Práctico N°5: LDA y PCA

Problema 1: Estudiar las implementaciones de Linear Discriminant Analysis (LDA) y Principal Component Analysis (PCA) provistas por scikit-learn.

a) Implementar ambos análisis sobre el `iris` data set:

```
from sklearn.datasets import load_iris
iris = load_iris()
```

b) Calcular la fracción de varianza explicada por las primeras componentes de cada método. Tener en cuenta que en LDA: ***n_components cannot be larger than min(n_features, n_classes-1)***

c) Discutir las siguientes afirmaciones:

i) PCA identifies the combination of attributes (principal components, or directions in the feature space) that account for the most variance in the data.

ii) LDA tries to identify attributes that account for the most variance between classes.

iii) LDA, in contrast to PCA, is a supervised method, using known class labels.

Problema 2: Estudiar el dataset sobre cáncer de mama provisto por scikit-learn:

```
from sklearn.datasets import load_breast_cancer
cancer = load_breast_cancer()
print(cancer.DESCR)
```

a) Identificar los nombres y números de clases y atributos. ¿Cuántos ejemplos tiene el dataset?

b) Calcular la fracción de varianza explicada por las primeras 10 componentes de PCA. En base a lo calculado, establecer un criterio de corte para seleccion de atributos.

c) Graficar los datos proyectados sobre el plano definido por las dos primeras componentes. ¿Son suficientes estas dos componentes para separar las clases?

d) ¿Por qué no puede implementarse LDA sobre este dataset?

Problema 3: Estudiar el dataset `Labeled Faces in the Wild` provisto por scikit-learn:

```
from sklearn.datasets import fetch_lfw_people
faces = fetch_lfw_people (min_faces_per_person=60)
```

a) Determinar el número de `features` `samples` y `classes`

b) Implementar PCA sobre este dataset usando 150 componentes y la opción `svd_solver='randomized'` para acelerar el algoritmo.

c) Graficar las primeras componentes (eigenfaces).

d) Graficar la fracción de varianza acumulada en función del número de componentes y escoger el número de componentes que explica aproximadamente el 80 % de la varianza acumulada.

e) Reconstruir las caras con las 150 componentes de PCA y graficar en dos filas las primeras 10 caras originales y reconstruidas para comparar las imágenes.

f) Reservar el 25 % del dataset para testing y clasificar las caras usando Naïve Bayes. Previo al entrenamiento, estandarizar las imágenes restando la media y dividiendo por el desvío estándar. Reportar accuracy y analizar la matriz de confusión. ¿Resulta este modelo mejor que el mero azar?
g) Con el número de componentes elegido en el ítem **(d)**, para explicar el 80 % de la varianza acumulada, repetir el análisis de los items **(e)** y **(f)**.

FaMAF 2023