Práctica 9 de abril
PCA

Análisis de Componentes Principales

Con los datos contenidos en “iris.csv”

- Haga gráficos de dispersión 2D entre las variables (p.e., Petal.Length vs. Petal.Width). El objetivo es explorar posibles correlaciones entre variables. Comente sus resultados. En caso de hacer varios gráficos de dispersión en una misma figura, use diferentes colores.
- Construya un gráfico de dispersión 3D con las variables menos correlacionadas.

Análisis de Componentes Principales

- Construya un histograma para cada característica (Petal.Length, Petal.Width, Sepal.Length, Sepal.Width) en donde ponga un color diferente para cada especie.
- Haga el procesamiento necesario a cada columna (variable/característica) para que la distribución de datos tenga media cero y varianza unitaria

$$z_i = (x_i - \bar{x})/s,$$

donde $\bar{x}$ es el valor promedio y s es la desviación típica (esto aplica para cada columna/variable del dataset).

Análisis de Componentes Principales

- Construya la matriz de covarianzas

$$\mathbf{\Sigma} = \frac{1}{n-1}\left((\mathbf{x}_i - \overline{\mathbf{x}})^T(\mathbf{x}_i - \overline{\mathbf{x}})\right),$$

donde el vector promedio (vector de promedios)

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x} \in \mathbb{R}^m.$$

Recordando que m es el número de variables (en este punto, el vector de promedios debería ser "cercano" al vector m-dimensional nulo)

La dimensión de la matriz $\mathbf{\Sigma}$ es $m \times m$ (en este caso $m = 4$).
Verifique el resultado de sus operaciones usando el método **cov** de **numpy** (el método **cov()** SÓLO como verificación).

Análisis de Componentes Principales

- Calcule los eigenvectores y eigenvalores de la matriz de covarianza calculada. Una opción es el método **eig()** de **numpy.linalg**. Imprima en pantalla e inlcuya en su reporte los eigenvalores.
- ¿Cómo podría verificar que los eigenvectores obtenidos son una base ortonormal de $\mathbb{R}^4$?
- Organice de forma descendente los eigenvalores (y sus correspondientes eigenvectores).
- Consulte cómo se relacionan los eigenvalores con el "porcentaje de variabilidad explicada" por cada componente principal.

Análisis de Componentes Principales

- Construya la matriz $\mathbf{W} \in \mathbb{R}^{4\times 2}$ con los dos eigenvectores que corresponden a los mayores eigenvalores. ¿Por qué se usan en este caso sólo 2 componentes principales?
- Proyecte los datos al nuevo espacio

$$\mathbf{Y}_{150\times 2} = \mathbf{X}_{150\times 4}\mathbf{W}_{4\times 2}$$

- Haga un gráfico de dispersión (gráfico 2D) para el conjunto de datos proyectado (use un color diferente para cada especie: *setosa*, *virginica* y *versicolor*). ¿Que puede observar de este nuevo conjunto de datos?