Introduction to Multivariate Statistics

EXPECTED VALUE AND MEAN

Es el valor promedio de una variable aleatoria X, el cual, usa la notación siguiente:

$$E[X]$$

La esperanza es calculada como la suma ponderada de las probabilidades.

$$E[X] = \sum x_1 \times p_1, x_2 \times p_2, \dots, x_n \times p_n$$

La media aritmética puede calcularse en Python tal como indica el Ejemplo 1

VARIANCE AND STANDARD DEVIATION

La varianza de una variable aleatoria X es una medida sobre qué tanto los valores de la distribución varían en promedio con respecto a la media. Su notación y cómo es calculada es:

$$Var[X] = E[(X - E[X])^2]$$

Podemos calcular la varianza en Python como indica el Ejemplo 2

La desviación estándar es calculada como la raíz de la varianza, esto es:

$$\sigma = \sqrt{Var[X]} = \sqrt{E[(X - E[X])^2]}$$

Podemos calcular la desviación estándar en Python como indica el Ejemplo 3

COVARIANCE AND CORRELATION

La covarianza es la medida de la probabilidad conjunta de dos variables aleatorias y describe cómo cambian las dos variables de manera conjunta. Su notación y cómo es calculada es:

$$Cov(X,Y) = E[(X - E[X]) \cdot (Y - E[Y])]$$

Cuando la covarianza sea igual a 0, significa que las variables son independientes.

Podemos calcular la covarianza en Python como indica el Ejemplo 4

La covarianza puede ser normalizada para que esté entre -1 y 1 dividiéndola entre la desviación estándar de X y Y. Este resultado es conocido como el coeficiente de correlación de Pearson.

Podemos calcular el coeficiente de correlación de Pearson en Python como indica el Ejemplo 5.

COVARIANCE MATRIX

La matriz de covarianzas es una matriz cuadrada simétrica que describe la covarianza entre dos o más variables aleatorias. La diagonal de la matriz es la varianza de cada variable aleatoria.

Podemos calcular la matriz de covarianzas en Python como indica el Ejemplo 6.

Principal Component Analysis

WHAT IS PRINCIPAL COMPONENT ANALYSIS

Es un método de reducción de dimensión. Puede ser pensado como un método de proyección donde datos con m columnas son proyectados a un subespacio con m o menos columnas, manteniendo la esencia de los datos originales.

Veamos los pasos de esta operación.

El primer paso es calcular la media de cada columna.

$$M = media(A)$$

Después, es necesario centrar los valores en cada columna y esto lo hacemos restando la media de las columnas.

$$C = A - M$$

El siguiente paso es calcular la matriz de covarianzas de la matriz centrada C.

$$V = Cov(C)$$

Finalmente, calculamos la eigendescomposición de la matriz de covarianza V.

$$valores, vectores = eig(V)$$

Los eigenvalores representan la dirección hacia el espacio reducido B, mientras que los eigenvalores representan las magnitudes de las direcciones. Los eigenvectores pueden ser ordenados a través de los eigenvalores en orden descendiente para darles un ranking. Si todos los eigenvalores tienen un valor similar, entonces sabemos que la representación existente puede ya estar considerablemente simplificada y una proyección nos aportaría poco. Si hay eigenvalores cercanos a cero, pueden ser descartados. De esta manera, se deben escoger m o menos elementos para conformar B.

$$B = seleccionar(valores, vectores)$$

Una vez seleccionados, los datos pueden ser proyectados vía una multiplicación de matrices.

$$P = B^T \cdot A$$

CALCULATE PRINCIPAL COMPONENT ANALYSIS

Podemos realizar el Análisis de Componentes Principales en Python como indica el Ejemplo 1

PRINCIPAL COMPONENT ANALYSIS IN SCIKIT-LEARN

Podemos realizar el Análisis de Componentes Principales en Python con Scikit-Learn como indica el Ejemplo 2.

Linear Regression

WHAT IS LINEAR REGRESSION

Es un método para modelar la relación entre dos valores escalares: la variable de entrada x y la variable de salida y. El modelo asume que y es una función lineal o una suma ponderada de la variable de entrada.

$$y = b_0 + b_1 x_1 + \cdots$$

El objetivo es encontrar los valores para los coeficientes b que minimicen el error en la predicción de la variable de salida y.

MATRIX FORMULATION OF LINEAR REGRESSION

Podemos expresar la regresión linear en notación matricial:

$$y = X \cdot b$$

Donde X son los datos de entrada y cada columna es una característica de los datos, b es un vector de coeficientes desconocidos y y es el vector de las variables de salida para cada fila en X.

Para encontrar la solución minimizando el error se utiliza el método de mínimos cuadrados y ésta tiene una única solución siempre y cuando las columnas sean independientes.

LINEAR REGRESSION DATASET

Definiremos una base de datos en Python para visualizar un scatter plot de ellos. Ver BASE DE DATOS

SOLVE VIA INVERSE

Podemos encontrar los valores de b mediante la ecuación:

$$b = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

Aplicando en Python, ver RESOLVIENDO MEDIANTE LA INVERSA.

SOLVE VIA QR DECOMPOSITION

Usando la descomposición QR, podemos obtener el valor de b.

$$b = R^{-1} \cdot Q^T \cdot y$$

Aplicando en Python, ver SOLUCIÓN CON DESCOMPOSICIÓN QR.

SOLVE VIA SVD AND PSEUDOINVERSE

Usando la descomposición SV y la pseudoinversa, podemos obtener el valor de b.

$$b = X^{+} \cdot y \qquad\qquad X^{+} = U \cdot D^{+} \cdot V^{T}$$

Aplicando en Python, ver SOLUCIÓN CON SVD Y PSEUDOINVERSA.

SOLVE VIA CONVENIENCE FUNCTION

Ver SOLUCIÓN CON FUNCIÓN DE CONVENIENCIA.