Aprendizaje Estadístico 2026

Lista 1

31.enero.2026

1. Dar un ejemplo de dos distribuciones distintas que tengan exactamente la misma entropía (de Shannon).

2. a) Investigar en Python cómo se puede calcular la entropía de una distribución o muestra aleatoria.
   b) Investigar en Python cómo se puede calcular la información mutua entre dos muestras aleatorias.

   En ambos casos, ilustrar con un ejemplo (graficar densidades, aplicar la o las funciones para calcular el estadístico y mostrar los resultados).

3. Escribir un código en Python para simular lanzamientos de una moneda en la computadora. Se debe permitir que el usuario elija parámetro $0 < p < 1$ que indica la probabilidad de obtener un éxito en el lanzamiento de la moneda. Hacer lo siguiente:

   a) Obtener, mediante repeticiones, una estimación de la densidad del número de lanzamientos necesarios para obtener el primer éxito. Por ejemplo, simular un experimento de estos N veces. Usar $N = 1000$.
   b) Elaborar visualizaciones de la función de densidad o masa, y cómo varía en función de p.

4. Elaborar una función en Python que permita comparar dos muestras (puede ser dos muestras provenientes de distribuciones teóricas, una teórica y una a partir de datos, o dos muestras provenientes a partir de datos). La función debe mostrar

   a) Las funciones de densidad f_1 y f_2.
   b) Las funciones de distribución F_1 y F_2.
   c) Una gráfica PP (prob-prob).
   d) Una gráfica QQ (quantil-quantil).

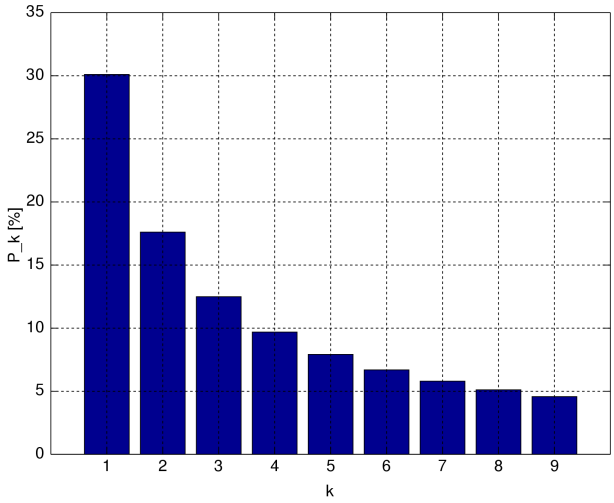
   Además, debe calcular la distancia de Kolmogorov-Smirnov (KS), e ilustrar en las gráficas de densidad y de distribución, el punto donde se alcanza esta distancia KS.

   Usar alguno de los experimentos del ejercicio anterior (con un valor p y N fijo), y comparar la distribución obtenida del experimento, contra una muestra generada aleatoriamente de la distribución geométrica

   i) $Geom(p)$,
   ii) $Geom(q)$, para $q = 1.2p$ (cuidar que $0 < q < 1$).

5. Em adición a la comparación visual anterior. Calcular una prueba de hipótesis de Kolmogorov-Smirnov, para comparar sus distribuciones del Ejercicio 4.
   En Python, una forma de aplicar la prueba de Kolmogorov-Smirnov es llamar a la función
   `from scipy.stats import ks_2samp`.

   Aplicar la prueba de comparación de Kolmogorov-Smirnov a los items (i) y (ii) del Ejercicio 4, y escribir sus conclusiones.



6. La **ley de Benford**, (o ley de Newcomb-Benford, también conocida como la ley del primer dígito), asegura que, en gran variedad de conjuntos de datos numéricos que existen en la vida real, la primera cifra es 1 con mucha más frecuencia que el resto de los números. Además, según crece este primer dígito, menos probable es que se encuentre en la primera posición. Esta ley empírica establece que la probabilidad que el dígito d $(1 \leq d \leq 9)$ aparezca como el primer dígito no-nulo en un conjunto de datos está dada por

$$\mathbb{P}(X = d) = \log_{10}\left(1 + \frac{1}{d}\right) = \log_{10}(d+1) - \log_{10}(d), \qquad 1 \leq d \leq 9.$$

   El archivo `areas.csv` contiene información de las áreas de todos los países.

   Aplicar las comparaciones del Ejercicio 4, así como la prueba estadística de Kolmogorov-Smirnov para determinar si los datos del primer dígito no-nulo en el conjunto de áreas se comporta de acuerdo a la ley de Benford o no. Explique sus conclusiones.

7. Generar una muestra aleatoria de una distribución gaussiana multivariada de dimensión n (con $n \geq 4$), con una media $\mu \in \mathbb{R}^n$ y covarianza $\Sigma \in \mathbb{R}^{n \times n}$ especificadas por el usuario.
   A partir de la muestra, graficar un `pairplot` que permita visualizar todas las densidades de cada variable y todas las nubes de puntos o densidades bivariadas entre pares de variables.

8. Considere el conjunto de datos **winequality-red**, dentro del conjunto de datos wines:
   `https://archive.ics.uci.edu/dataset/186/wine+quality`.

   Para las variables **Ph** y **fixed_acidity**, encontrar la distribución continua que mejor se ajuste al conjunto de datos. Mostrar entre varias distribuciones candidatas por qué la distribución elegida esta es la más adecuada.

   En cada caso, mostrar visualizaciones que contrasten la distribución empírica de los datos contra el modelo ajustado, y las métricas obtenidas.