Inteligencia artificial avanzada para la ciencia de datos I

TC3006C

Dr. Esteban Castillo Juarez
TEC de Monterrey, Campus Santa Fe


esteban.castillojz@tec.mx

Tecnológico de Monterrey

Agenda

- Objetivo
- Instrucciones
- Conjunto de datos
- Función de similitud
- Formato de la actividad

Tecnológico de Monterrey

Objetivo

- Reforzar la teoría de KNN aplicando el algoritmo de clasificación sobre un conjunto de datos diferente y con una medida de distancia distinta.

- Practica la teoría detrás del KNN en el contexto del lenguaje de programación Python y Scikit-learn.

Tecnológico de Monterrey

Instrucciones

Para la implementación manual y con Scikit-learn vista en clase hacer lo siguiente:

1. Utilizar el conjunto de datos de enfermedades cardíacas para entrenar una nueva versión del clasificador KNN.
2. Cambiar la medida de similitud euclidiana por la distancia de Manhattan.
3. Proponer otra medida de similitud distinta a la euclidiana y Manhattan.

Tecnológico de Monterrey

Instrucciones

4. Variar el valor de K de 3 a 41 para cada medida de similitud.
5. Obtener la precisión para cada variación.
6. Mostrar el valor de K que proporciona una mejor exactitud para cada variación.
7. Proponer dos gráficos que describan el comportamiento del modelo y sus resultados (distintas a las graficas de la siguiente diapositiva).
8. Describir los hallazgos y proporcionar una conclusión sobre el proceso de experimentación.

Tecnológico de Monterrey

Instrucciones

En el caso de Scikit-learn, hacer adicionalmente lo siguiente:

1. Crear un gráfico que muestre cómo varía la precisión del clasificador KNN al cambiar el valor de K para cada medida de similitud especifica.
2. Crear un gráfico que comparen el rendimiento del clasificador KNN utilizando las tres medidas de similitud propuestas anteriormente.

Tecnológico de Monterrey

Conjunto de datos

Conjunto de datos de enfermedades del corazón

La enfermedad cardíaca describe una variedad de afecciones que afectan el corazón. Las enfermedades que se incluyen en el grupo de enfermedades cardíacas incluyen enfermedades de los vasos sanguíneos, como la enfermedad de las arterias coronarias, problemas del ritmo cardíaco (arritmias) y defectos cardíacos congénitos, entre otros.

Teniendo en cuenta lo anterior, para esta actividad utilizaremos el conjunto de datos de enfermedades cardíacas del repositorio de la UCI.

Tecnológico de Monterrey

Conjunto de datos

El conjunto de datos consta de 303 muestras (242 muestras de entrenamiento y 61 de prueba). Hay 14 columnas (separadas por ",") en el conjunto de datos, que se describen a continuación:

1. **Edad:** muestra la edad del individuo.

2. **Sexo:** muestra el género del individuo utilizando el siguiente formato:
   1 = hombre
   0 = mujer

3. **Dolor en el pecho:** muestra el tipo de dolor que experimenta el individuo utilizando el siguiente formato:
   1.0 = angina tipica
   2.0 = angina atipica
   3.0 = dolor no anginoso
   4.0 = asintomático

Tecnológico de Monterrey

Conjunto de datos

4. Presión arterial: **muestra el valor de la presión arterial en reposo de un individuo en mmHg (unidad).**

5. **Colesterol en suero:** muestra el colesterol en suero en mg/dl (unidad).

6. **Glucosa en ayunas:** compara el valor de glucosa en ayunas de un individuo. Si la glucosa en ayunas > 120 mg/dl, entonces: 1 (verdadero); de lo contrario: 0 (falso).

7. **ECG en reposo:** muestra resultados electrocardiográficos en reposo
   0.0 = normal
   1.0 = Anomalía de la onda ST-T
   2.0 = hipertrofia ventricular izquierda

Tecnológico de Monterrey

Conjunto de datos

8. **Frecuencia cardíaca:** muestra la frecuencia cardíaca máxima alcanzada por un individuo.

9. **Angina inducida por el ejercicio:**
   1.0 = si
   0.0 = no

10. **Depresión del segmento ST inducida por el ejercicio en relación con el reposo:** muestra el valor que es un número entero o flotante.

11. **Segmento ST de ejercicio máximo:**
   1.0 = pendiente ascendente
   2.0 = Plana
   3.0 = pendiente descendiente

Tecnológico de Monterrey

Conjunto de datos

12. **Número de vasos principales (0-3) coloreados por fluorosopía:** muestra el valor como entero o flotante.

13. **presencia de talasemia:**
    3 = normal
    6 = defecto corregido
    7 = defecto reversible

14. **Diagnóstico de enfermedad:** muestra si el individuo padece o no una enfermedad cardíaca:

Ausencia
Presencia

Etiqueta de clasificación

Tecnológico de Monterrey

Conjunto de datos

- Los archivos del conjunto de datos (entrenamiento y prueba) se proporcionan junto con esta presentación.

- Es obligatorio utilizar el conjunto de datos sobre enfermedades cardíacas para esta actividad. Cualquier otro conjunto de datos no será útil.

Tecnológico de Monterrey

Función de similitud

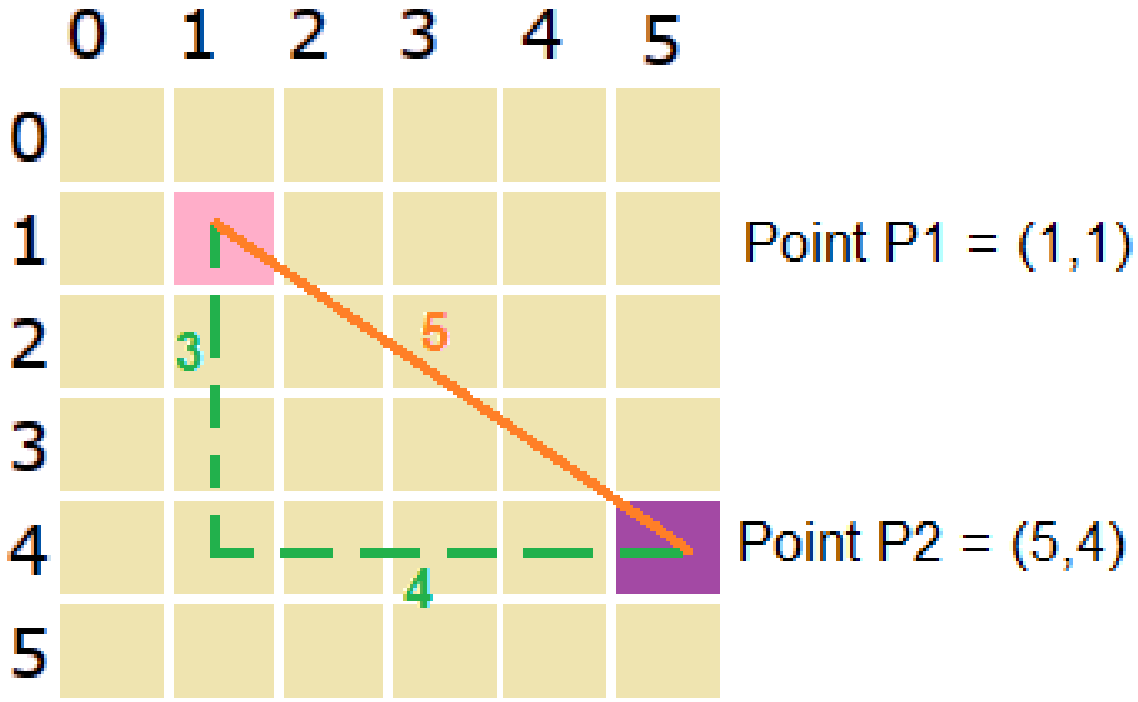
Distancia de Manhattan

La distancia de Manhattan entre 2 vectores es la suma del valor absoluto de la diferencia de sus coordenadas:

$$distance = \sum_{i=0}^{n-1} |(x[i] - y[i])|$$

Tecnológico de Monterrey

Función de similitud



Distancia de Manhattan

$$distance = \sum_{i=0}^{n-1} |(x[i] - y[i])|$$

Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

Tecnológico de Monterrey

Formato de la actividad

- Todas las actividades son individuales.
- La actividad se entregará el día de la siguiente sesión de clases con el profesor.
- Todas las tareas deben estar documentadas en detalle. Las tareas deben entregarse o enviarse como un archivo comprimido (nombre: matricula) con la siguiente información:
  - Archivo de Python, notebook de Jupyter o Google Colab.
  - Documento con explicación del problema e imágenes de su ejecución. Pueden simplemente mostrar un notebook (si lo ocuparon) si toda la documentación está en celdas de markdown, debidamente explicada.

Formato de la actividad

- La actividad deben contener lo siguiente:
  - Portada con información del estudiante (5%).
  - Sección de introducción que explica explícitamente el problema abordado (10%).
  - Sección experimental que explica el enfoque utilizado para resolver el problema (30%).
  - Sección de resultados que recoge los hallazgos hechos(40%).
  - Sección de conclusiones que recoge lo aprendido en la tarea (15%).

Tecnológico de Monterrey

¡Gracias!

Tecnológico de Monterrey