Universidad Nacional de General Sarmiento

ABRIL 2025

Laboratorio de Construccion de Software

TP Inicial - Parte tres

Implementación del modelo K-Means

PRESENTADO POR:

GRUPO 3

Alfaro, Ezequiel
Ibacache, Maria Jose
Torales, Cecilia
Silva, Agustina

DOCENTES

Dikenstein, Leandro
Orozco, Francisco

Implementación del modelo K-Means

Introducción

El presente informe tiene como objetivo presentar el progreso alcanzado hasta la fecha en el desarrollo de un modelo de clustering basado en K-Means para agrupar datos de empleados según su nivel y tipo de habilidad. Se detallan las etapas completadas, los resultados preliminares obtenidos y los próximos pasos a seguir para consolidar el análisis.

Progreso Actual:

- **Carga y Preprocesamiento de Datos:**

Se cargó el conjunto de datos de empleados desde un archivo CSV (data.csv).
Para poder trabajar con los datos de manera efectiva, primero organizamos la información sobre las habilidades de los empleados. Como las habilidades están expresadas como palabras (por ejemplo, "Assembler", "Java", "Python"), las convertimos en números para que el modelo pueda interpretarlas. Así, asignamos un valor numérico a cada habilidad: por ejemplo, "Assembler" se convirtió en 1, "Java" en 2, y así sucesivamente.
Por otro lado, la columna que indica el "Nivel" de los empleados ya estaba en formato numérico, por lo que no fue necesario hacer cambios en ella. Sin embargo, para asegurarnos de que tanto los valores de "Habilidad" como los de "Nivel" tuvieran la misma importancia en el análisis, ajustamos todos los números a una escala común entre 0 y 1. Esto nos ayuda a evitar que una característica domine sobre la otra solo porque sus valores sean más grandes.

- **Modelado con K-Means:**

Para organizar a los empleados en grupos con características similares, creamos un modelo utilizando la técnica K-Means. Este algoritmo agrupa los datos en función de las columnas "Habilidad_Encoded" y "Nivel_Encoded", que representan el tipo de habilidad y el nivel de cada empleado, respectivamente. Decidimos formar 12 grupos (o clusters), y para asegurarnos de que los resultados sean consistentes cada vez que ejecutemos el modelo, establecimos una semilla aleatoria (random_state=12). Una vez entrenado el modelo, cada empleado fue asignado automáticamente a uno de estos 12 grupos según sus habilidades y nivel.

- **Interpretación de Clusters:**

Para facilitar la comprensión de los resultados, dimos un significado claro a cada grupo. Por ejemplo:
El Cluster 0 representa empleados con conocimientos intermedios en Python.
Estas etiquetas permiten interpretar los resultados del clustering de manera intuitiva.

- **Visualización de Resultados:**

Se creó un gráfico de dispersión para visualizar los clusters generados. Cada cluster se representó con un color distinto, y se añadió un pequeño "jitter" (ruido) para mejorar la claridad visual.Los centroides de cada cluster se destacaron con un marcador especial y un borde negro.La leyenda del gráfico incluye las etiquetas descriptivas de los clusters, facilitando la interpretación de los resultados.

- **Resultados Preliminares:**

El modelo ha logrado agrupar a los empleados en 12 clusters distintos, cada uno asociado con un perfil específico de habilidad y nivel.
La visualización muestra una clara separación entre los clusters, lo que sugiere que el algoritmo está capturando patrones significativos en los datos.

Conclusión:

Es importante señalar que esta es una etapa preliminar y se requiere un análisis más profundo para validar completamente la precisión y utilidad del modelo. Sin embargo, creemos que, en términos generales, el proyecto va por buen camino.