ISTEA

Minería de Datos I

Examen final de curso

Modelo de expresión genética

Contexto:

Trabajas en un laboratorio biomédico que recopila datos de expresión génica de pacientes con diferentes condiciones de salud. El objetivo es identificar patrones subyacentes en la expresión génica y agrupar a los pacientes en clusters que compartan características genéticas similares. Esto podría ayudar en la identificación de subtipos de enfermedades o en la personalización de tratamientos.

Datos:

El conjunto de datos incluye las siguientes columnas:

- Variables de Expresión Génica: Cada columna representa la expresión de un gen específico. Existen 10 genes para cada muestra.
- Edad: Edad del paciente.
- Género: Género del paciente.
- Condición de Salud: La condición de salud del paciente:
    - 0 significa salud normal
    - 1 al 9 especifica un tipo de enfermedad.

Objetivo:

Utilizar técnicas de aprendizaje no supervisado, específicamente PCA y K-Means, para agrupar a los pacientes en clusters basados en sus perfiles de expresión génica, edad y género.

ISTEA

Minería de Datos I

Examen final de curso

Pasos:

Exploración de Datos:

En un notebook de Google Colab, carga el conjunto de datos y realiza una exploración inicial para entender la estructura y la distribución de las variables.

Preprocesamiento de Datos:

Maneja los valores nulos y realiza una normalización de las variables de expresión génica.

Codifica la variable categórica del género y condición.

Análisis de Componentes Principales (PCA):

Aplica PCA para reducir la dimensionalidad de las variables de expresión génica. Decide el número apropiado de componentes principales a retener.

Selección de Características:

Selecciona las características relevantes, incluyendo las componentes principales de PCA, edad y género.

Clustering con K-Means:

Aplica el algoritmo K-Means para agrupar a los pacientes en clusters. Experimenta con diferentes valores de k.

Evaluación del Modelo:

Evalúa la calidad de los clusters obtenidos utilizando el índice de silueta.

Interpretación y Visualización:

Analiza e interpreta los clusters identificados. Utiliza gráficos y visualizaciones para representar la distribución de los pacientes en el espacio reducido por PCA.

ISTEA

Minería de Datos I

Examen final de curso

Recomendaciones Biomédicas:

Proporciona recomendaciones basadas en los clusters identificados. ¿Se pueden inferir subtipos de enfermedades? ¿Cómo podría esto informar sobre enfoques de tratamiento personalizado?

Notas Adicionales:

- Considera la posibilidad de utilizar la técnica de codo para seleccionar el número óptimo de clusters en K-Means.
- Documenta cada paso y realiza comentarios explicativos en tu código para facilitar la comprensión.

Link al conjunto de datos:

https://drive.google.com/file/d/15tjH0emTVAlqVL_qaBWLrrj8fC9YEQyM/view?usp=sharing

Suerte