Minería de Datos

Lab. #5

Ivan Saavedra, Ph.D.

saavedrai@uninorte.edu.co

Universidad del Norte
División de Ingenierías
Dpto. Ingeniería de Sistemas

202030

1. Naïve Bayes Classifier

Aplique el clasificador de Bayes para clasificar oraciones entre si son preguntas o no. Considere las siguientes listas de oraciones para entrenar al modelo.

Metodología

- Preparar y adecuar el conjunto de datos de entrada para poder alimentar el modelo
- Entrenar el modelo
- Encontrar predicciones
- Evaluar precisión del modelo
- Comparar precisiones para las dos listas de oraciones

Lista 1:

this is my book
they are novels
have you read this book
who is the author
what are the characters
This is how I bought the book
I like fictions
what is your favorite book
this is your book
who is your favorite author
what are my favorite novels
have they read this book
I bought you the book
who like the novels
how they like the book characters

Lista 2:

this is my book
they are novels
have you read this book
who is the author
what are the characters
This is how I bought the book
I like fictions
what is your favorite book
this is your book
who is your favorite author
what are my favorite novels
have they read this book
I bought you the book
who like the novels
how they like the book characters
you bought the book
they bought the book
who bought the book
I like novels
who like novels
who like the author
they like the author
who read the author
have you read this author
this is your favorite author
this is your favorite book
they are the characters

2. Classifiers

NB, DT, KNN

Aplique los diferentes clasificadores vistos en clase para clasificar el tipo de fruta basado en atributos como el tamaño, el color, entre otros.

Considere el conjunto de datos suministrado para entrenar al modelo.

https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/fruit_data_with_colors.txt

Metodología

- Realizar una análisis descriptivo exploratorio
  - Analizar frecuencia de frutas en general
  - Analizar por medio de un diagrama de cajas cada uno de los atributos
  - Analizar por medio de histogramas cada uno de los atributos
  - Analizar matriz de correlación
- Entrenar los modelos
  - Visualizar el árbol de decisión encontrado para el modelo de Árbol de Decisiones
- Encontrar predicciones
- Evaluar precisión de los modelos
- Comparar precisiones de los modelos. Cual es el modelo que presenta mejor precisión ?

3. Classifiers

NB, DT, KNN

Aplique los diferentes clasificadores vistos en clase para clasificar si una persona puede tener cáncer o no dependiendo del tipo de la masa encontrada afectada (esta puede ser benigna o maligna). Esta decisión se encuentra basada en atributos como Uniformidad del tamaño y forma de las células, la cromatina, los núcleos, entre otros.

Considere el conjunto de datos suministrado para entrenar al modelo.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

Metodología

- Realizar una análisis descriptivo exploratorio
  - Analizar frecuencia de frutas en general
  - Analizar por medio de un diagrama de cajas cada uno de los atributos
  - Analizar por medio de histogramas cada uno de los atributos
  - Analizar matriz de correlación
- Entrenar los modelos
  - Visualizar el árbol de decisión encontrado para el modelo de Árbol de Decisiones
- Encontrar predicciones
- Evaluar precisión de los modelos
- Comparar precisiones de los modelos. Cual es el modelo que presenta mejor precisión ?

Modelos de Clasificación

Entregable:

- Archivos de Excel con datos de entrada
- Un archivo de Jupyter Notebook con el desarrollo del análisis
- Las conclusiones y respuestas al objetivo del análisis deben ser contestadas en el mismo notebook.
- Se sugiere que comente las secciones de manera adecuada para una mejor interpretación de su análisis.
- La fecha de entrega es el Viernes 25 de Octubre del 2020 vía catalogo web enlace de laboratorios.