SolidQ

Think Big. Move Fast.

Modulo 4 Ingeniería de Características

Raw data: pixel grid		
Better features: clock hands' coordinates	{x1: 0.7, y1: 0.7} {x2: 0.5, y2: 0.0}	{x1: 0.0, y2: 1.0} {x2: -0.38, 2: 0.32}
Even better features: angles of clock hands	theta1: 45 theta2: 0	theta1: 90 theta2: 140

¿Por qué son importantes?

- Los mejores modelos son modelos simples que encajan bien con los datos
- Necesitamos balancear entre precisión y simplicidad
- Los modelos simples
  - tienden a predecir mejor
  - Se interpretan mejor por parte de los humanos
  - Más sencillo hacer predicciones a partir de ellos

Seleccionar Características

Conceptos de Negocio

- Aplicar conceptos de negocio a los datos en bruto
- Ayuda a reducir características
- Ejemplo: Score RFM
  - Recency
  - Frequency
  - Monetary

Técnicas

- Selección hacia atrás
  - Empieza con todas las características
  - Encuentra la característica que menos disminuye el poder de predicción y elimínala
  - Continua el proceso hasta que empiece a dañar tu precisión
- Selección hacia adelante
  - Empieza sin características
  - Busca la características que por si sola es el mejor modelo
  - Mantenla y añade otra.
  - Así hasta que no mejores tus predicciones

Escalado de Características

Escalado de Características

- Mal escalado puede llevar al algoritmo a dar más peso a unas características que a otras sin que tengan relevancia
- Complica la interpretación de coeficientes
- Ensucia la regularización
- En algoritmos que trabajan con distancias, las distorsiona

Escalado con scikit-learn

- StandardScaler
  - Asume que los datos siguen una distribución normal dentro de cada característica por lo que centra la distribución en 0 con una desviación estandar de 1
- MinMaxScaler
  - Reduce el rango de valores en 0 y 1 (o -1 y 1 si tiene negativos)
- RobustScaler
  - Como MinMax pero utilizando un rango de quartiles
- Normalizer
  - Escala cada valor dividiéndolo por su magnitud en un espacio de n dimensiones siendo n el número de características

Demo 04 Importancia Escalado

Importancia del Escalado de Características

Extracción de características en scikit-learn

Interpretando las Características

Correlación

- Existe la creencia muy extendida de que cuantas más características mejor
- Se crean características “artificiales” que son inútiles para el modelo
  - Y además hacen complicado interpretar los coeficientes
- La mejora en la precisión suele ser muy alta cuando se eliminan o se combinan

www.solidq.com

info@solidq.com