Ciencias de datos: Intermedio

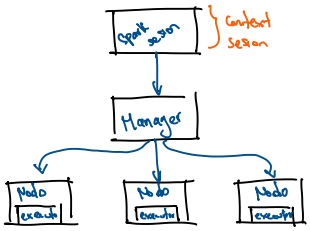
sábado, 5 de noviembre de 2022 8:01

5/11/2022

DataBricks:
Se pueden instalar las librerias desde los notebooks o desde el cluster.

- Si se instala la librería desde el notebook otro notebook no la tendra instalada.
- Si se instala la librería desde el cluster todos los notebooks pueden acceder a ellas, pero, cada vez que se inicie nuevamente el cluster todas las librerías serán instaladas.
- Si a nivel de notebook no deja instalar la librería, intentar a nivel de cluster

Buena práctica: Cada 6 meses hacer upgrade al runtime y revisar el codigo.
Arquitectura Spark:



12/11/2022
Machine Learning
-> Regresion (Regresion lineal, regresion no lineal)
-> Clasificacion (regresion logistica, arboles de decision, random forest, boosting)
-> Clustering (agrupamiento, KMeans)

Regresion **->** Numerica Continua
Clasificacion **->** binarias, binomiales

1. Primero se calcula una regresion.
2. Luego se lleva a una *Funcion sigmoidal*: Se lleva una respuesta analoga (numerica) a una respuesta binaria.

Clustering **->** Agrupamientos de caracteristicas.

Para el clustering se hace uso de PCA (reduccion de la dimensionalidad)

Problema lineal: Lo puedo dividir por medio de una recta, una recta me soluciona el problema.

Problema no lineal: No se puede dividir por medio de una linea recta, no se puede solucionar por medio de una recta.

Nivel explicativo de los algoritmos: Que tan interpretable es la respuesta que me está dando el algoritmo.

Entre más baja el nivel explicativo se entra más en algoritmos caja negra.

Dimensiones: Variables que se tienen
Tener muchas dimensiones a veces no es bueno para el modelo (la maldición de la dimensionalidad)

Test	Train
80	20
70	30
60	40
75	25

Un algoritmo por debajo siempre trata de llegar a un mínimo global.
Por adentro el algoritmo resuelve una ecu para hallar el mínimo global. Una vez llega para el entrenamiento.

Función de costo: Representa como el algoritmo va a resolver el problema.

Algoritmo	Función de costo
Regresión lineal	MSE (El valor real menos el valor predicho al cuadrado)

Descenso del gradiente: Se calcula la derivada para encontrar la inclinación de la función, el mínimo global es cuando la inclinación es igual a 0, para hallar el mínimo global sigue calculando derivadas hasta que por más derivaciones que se haga siempre va a ser 0.

Random_state: Se usa para que los datos sean reproducibles a múltiples modelos, lo que hace es que la partición de los datos siempre va a empezar en un punto determinado.
Se pone un numero definido cualquiera (en python el más usado es el 42).
Mismas condiciones experimentales entre modelos

RMSE: Desfase o umbral de la prediccion (+- 400); 400 hacia arriba o 400 hacia abajo.

19/11/2022
Técnicas de preprocesamiento de los datos:
La idea de estas técnicas es llevar todos los datos a una escala que el modelo pueda interpretar, quitando las equivalencias numéricas (Millones, edad).
Normalizar: Escala entre 0 y 1
Estandarizar: Media 0, desviación estándar 1.
Min Max Scaler: Crea un rango dependiendo de la variable, no un rango en comun.

Pandas.Dataframe.select_dtypes -> selecciona las columnas con un tipo de dato especifico.

Map() aplica una funcion a una lista aplica

Df.query() es más optimizado (veloz) que el slicing del dataframe normal.

Standard Scaler o Z score: La media de cada columna sea igual a 0 y la desviacion estandar 1 para c/u de las columnas.

26/11/2022
Regresión logística: Modelo de clasificación binaria; existen variaciones de la misma que clasifican más de 2 clases (regresión logística multinomial).
Al agregar variaciones el modelo va a tener que identificar patrones que no sean tan relevantes.
El problema de los modelos de clasificación es el desbalanceo de las clases (balanceo de datos).

Las técnicas de balanceo crean muestras sintéticas, crean datos adicionales a partir de la base de datos que ya hay; el problema es que aumentan el sesgo de la predicción del modelo.

Función de costo: Función interna del modelo (J) la cual la gran mayoría de modelos tratan de minimizar al máximo, cuando se minimiza es porque el modelo se ajusta a los datos.

Cada modelo en particular tiene su propia función de costo.

Regresión lineal: MSE
Regresión logística:

$$J = -\frac{1}{m}\sum_{i}^{m} y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)$$

Descenso del gradiente:
Se calculan las derivadas parciales de la función de coste de cada modelo para poder minimizarla. Un modelo converge cuando logra encontrar el mínimo global.

Una vez el modelo converge le asigna unos pesos respectivos a cada una de las variables (var más importante a menos

importante).

Los valores de Y se reemplazan en la función dando un resultado numérico.

Ya que el resultado no es una clasificación, pasan por una función sigmoide la cual transforma dicho resultado numérico a uno binario.

SkLearn tiene la forma de balancera las clases: class_weight = True

Azure - Databricks: Es una unidad de cómputo la cual es muy buena para hacer predicciones en Batch.

Azure - ML: Por su parte Azure ML funciona muy bien para predicciones en tiempo real (real time).

Azure ML tiene 2 opciones:
Container instances: Una imagen de docker fija.
Kubernetes: Tambien conteneriza con docker y ademas es elastico.

4/12/2022
Azure ML SDK: https://learn.microsoft.com/en-us/training/paths/build-ai-solutions-with-azure-ml-service/

1. Registrar el set de datos
2. Enviroment
3. ScriptConfig
4. Run
5. Registrar modelo

Azure ML tiene dos opciones para el despliegue, ASI y AKS
ASI para modelo con pocas solicitudes diarias.
AKS para modelo con altas solicitudes diarias.

Blob: Almacenamiento de más bajo consumo que un DataLake

La semilla garantiza que los modelos estén trabajando bajo las mismas condiciones; la semilla es el punto donde el modelo va a empezar a evaluar la función de error.

Se tiene que poner una misma semilla para diferentes modelos para que las condiciones de experimentación sean las mismas.