Práctica 5: Árboles
Análisis de Grandes Volúmenes de Datos CI-
Carlos Espinoza B92786
I-2022

Algoritmos y Modelos

En Machine Learning se suele trabajar con estos dos conceptos. Dentro del área de la computación podemos hasta utilizarlos para referirnos a la misma acción.

Conceptos Básicos

Algoritmos

Es una función que se corre para crear un modelo de ML. Este puede reconocer patrones, aprender y ajustarse a un conjunto de datos

Tipos:

```
- Clasificación: Árboles de decisión, k-vecinos
- Regreción: Reg. Lineal, árboles de decisión
- Segmentación: k-medias, DBSCAN
```

Propiedades:

```
- Se pueden describir usando lenguaje matemático o pseudocódigo
- Se puede medir la eficiencia
- Amplia variedad de lenguajes de programación
```

> **Bibliotecas:** scikit-learn, TensorFlow, PyTorch

Modelos

Son las salidas que brindan los algoritmos de ML que fueron ejecutados sobre un conjunto de datos. También,son las representaciones de lo que los algoritmos lograron aprenden. Además, es aquello que se puede guardar luego de correr los algoritmos sobre datos de entrenamiento. Apartir de un modelo se pueden realizar predicciones.

El modelo presenta reglas, números y demás estructuras para poder realizar las predicciones.

- **Reg. Lineal:** El modelo es un vector de coeficientes con valores.
- **Árbol de Decisión:** Modelo de if-else con valores.

- **Red Neuronal:** El modelo es un grafo con vectores o matrices con pesos y valores.

Árboles de Decisión

Estos nos permiten abarcar los problemas de clasificación. Tenemos varios típos, por ejemplo, Random Forest y Extremely Random Forest

Tipos:

- **Clasificación**
    - Predicen una clase (binaria, multiclase)
    - Las clases deben ser conocidas (valores V-F, Idiomas, Bien-Mal, otras)
- **Regresión:**
    - Predicen un valor contínuo o real (ingresos, acciones, cosechas)
    - Se conocen las variables que son atributos

Algortimo de Bosques Aleatorios

Este algoritmo es de los más utilizados en el aprendizaje de máquina por su excelente rendimiento a la hora de enfrentarse a problemas de modelos de clasificación y regresión.

Otra ventaja que tiene este algoritmo es que es fácil de utilizar dada la poca cantidad de hyperparámentos que este posee.

Claves sobre este algortimo:

- Es un conjunto de árboles de decisión
- Es una extensión de bootstrap
- Utiliza diferentes árboles con diferentes muestras del dataset.
- Estas muestras son tomadas con reemplazo

Comentarios

Cómo se vio en las prácticas, el modelado no es una ciencia exacta, sino, una práctica de prueba y error en la que se van a realizar diferentes combinaciones para obtener el modelo que genere la mayor exáctitud. Además, hay que tener cuidado a la hora de alimentar un modelo para su entrenamiento. Se puede caer en escenarios pesimistas y optimistas. Para tratar esto podemos hacer uso de técnicas como CV (Cross-Validation) que nos permiten iterar sobre splits del conjunto de datos en los que se encuentran tanto los datos para el anális como datos de prueba.

Usos

Algunos usos que le podría dar a este conocimiento sería poder hacer predicciones sobre las duraciones de los trámites según su complejidad o cantidad de componentes. Para esto se usaría la regresión lineal, que permite hacer clasificaciones con una salida de tipo contínua (tiempo).

Problemas que resuelven

Clasificar el rendimiento de los estudiantes tomando en cuenta el contexto en el que se desenvuelven. Por ejemplo, no es lo mismo ver el rendimiento académico de una persona con todas las facilidades (acceso a internet, espacio libre de distracciones, computadora, vehículo propio, entre otras) que ver a alguien que no.