El aprendizaje de máquina

es el conjunto de técnicas y algoritmos que, basándose en conjuntos de datos pretenden crear modelos que resuelvan principalmente dos tipos de problemas, el problema de regresión y el problema de clasificación. Ya que es conveniente para muchos casos de aplicación, los algoritmos de machine learning abren un panorama a muchas alternativas que con métodos tradicionales serían impensables. Por ejemplo, hoy en día se cuenta con algoritmos en muchas actividades como son: la búsqueda en internet, la clasificación de si un cliente es viable para ofrecerle un crédito, la segmentación de personas para enviarle una publicidad adecuada tal que la posibilidad de vender sea alta, aumentar el tiempo de atención que tiene una red social mediante la exposición de contenidos, filtrar correo no deseado, convertir grabaciones de audio en textos, eliminar el ruido de un mensaje de voz, predecir la siguiente palabra que un usuario escribirá en un teclado de un teléfono inteligente, detectar cuando una persona sonríe para tomar una fotografía, entre muchas aplicaciones más. Revisemos de qué se trata cada uno de los problemas en el machine learning.

Se habla de una regresión lineal,

es decir, un modelo de línea recta. En este caso particular es de interés conocer los parámetros m y b que modelan una única recta en el espacio de dos dimensiones. El ajuste debe permitir que, dado un punto independiente x o medición se pueda encontrar el resultado y con el modelo de la forma más aproximada a la realidad. Cuando se intenta resolver el problema de regresión, lo que se hace es encontrar un modelo matemático y un conjunto de parámetros que debe tener ese modelo tal que el error entre los puntos de datos del modelo y los datos que se intentan modelar sea el mínimo. Es decir, el problema de regresión se puede resolver minimizando el error entre dos valores, los puntos de datos del modelo y los datos que se intentan modelar. Por ejemplo, si se tienen los datos de el precio de las casas y se quiere encontrar el precio que puede tener una casa con base en el número de habitaciones, la ubicación, el tamaño, entre otros. La tarea de encontrar el precio dados los datos es resolver un problema de regresión.

Existen técnicas de modelado de datos como el ajuste de mínimos cuadrados (lineales, no lineales), modelos de máxima verosimilitud, regresión linear bayesiana, y modelos más elaborados como las máquinas de soporte vectorial y las redes neuronales.

TIC

Modelos de clasificación

El problema de clasificación consiste en separar puntos de datos que pertenecen a diferentes categorías. Por ejemplo, en el dataset IRIS se pueden crear modelos que, dados los datos de las medidas del tallo, sépalo y otras mediciones de las flores sea posible clasificar si ciertas mediciones pertenecen a una especie determinada o no. El problema de clasificación consiste en asignar el grupo adecuado en un conjunto de categorías dadas ciertas variables de entrada específicas. Para resolver el problema de clasificación se tienen modelos como: discriminantes lineales, mínimos cuadrados para clasificación, el discriminante de Fisher, modelos de regresión logística, perceptrones, máquinas de soporte vectorial y redes neuronales.

TIC

El problema de regresión

La regresión es una técnica estadística para modelar la relación entre una variable dependiente y una o múltiples variables independientes. Lo que se pretende es encontrar un modelo matemático que mejor se ajuste a los datos observados. Es importante recalcar que los datos son observaciones, o mediciones. Cuando se habla de una observación o una medición siempre se debe pensar que hay ruido asociado al proceso, por lo que los datos están desplazados de su coordenada real. Sin embargo, el valor real de las mediciones no se puede conocer, solamente se puede estimar. Por este motivo, no es posible encontrar un modelo que explique perfectamente los datos, sino que, se debe recurrir al ajuste de variables que tengan en cuenta el ruido inherente a las medidas y minimicen el error al intentar encontrar un modelo que explique los datos. Cuando los datos se explican con un modelo de la forma

$$y = mx + b$$

TIC

Elementos clave en el machine learning

Existen muchos algoritmos de machine learning, de hecho, con el paso del tiempo muchos algoritmos son desarrollados, sin embargo, cada uno de ellos contiene tres componentes:

Representación de conocimiento: Quiere decir cómo se estructura el conocimiento, ya sea como un conjunto de reglas, un modelo gráfico, una red o incluso un grupo de modelos.

Evaluación: Es la forma en la que se evalúa el desempeño de un modelo, por ejemplo, se suelen medir variables como la precisión, la exactitud, la incertidumbre, el error medio cuadrático, la probabilidad, el costo, la entropía y muchas mediciones que ayudan a establecer qué tan bien un modelo hace su trabajo.

Optimización: Es la forma en que un algoritmo mejora sus parámetros internos, es decir, cómo se llega a una superficie de decisión (en el caso de clasificación) o a una curva (en el caso de regresión) que mejor represente los datos con el menor error posible. La optimización ayuda a cambiar gradualmente los parámetros del modelo para que las métricas definidas en la evaluación mejoren. Generalmente el criterio para optimizar es comparando las mediciones establecidas por la evaluación.

TIC

Tipos de aprendizaje de máquina

Existen diferentes enfoques, dependiendo de los datos y cómo están estructurados. Se tiene cuatro técnicas de aprendizaje:

Aprendizaje supervisado: Se trata de emplear datos que tienen etiquetas, es decir, datos que una persona ha marcado de tal forma que los datos con los que se entrenará un modelo incluyan las salidas deseadas. Por ejemplo, este correo es spam y este no. El aprendizaje supervisado se debe emplear principalmente en datos no estructurados, como es el caso de modelos que hacen tareas con imágenes o con archivos de audio.

Aprendizaje no supervisado: En este caso los datos no tienen etiquetas, por lo que no incluyen las salidas deseadas. En este caso es difícil establecer las métricas ya que se debe establecer qué es una buena medida de éxito, sin embargo, se pueden emplear técnicas de agrupamiento de datos para crear etiquetas de forma automática y luego entrenar modelos con esas etiquetas.

Aprendizaje semi supervisado: Se tienen conjuntos de datos con algunas etiquetas y se deben preprocesar los datos para encontrar las etiquetas faltantes. En este caso también se pueden emplear algoritmos de agrupamiento.

Aprendizaje por refuerzo: Se crea una función de costo, la cual crea "recompensas" por ciertas secuencias de acciones. Por ejemplo, se quiere entrenar a un programa para que aprenda a conducir un carro en un juego y la entrada es el puntaje según el recorrido que haga el vehículo. En este caso el programa se entrenará de acuerdo con qué tan lejos llega dadas las salidas del modelo. Se deben repetir muchas veces los experimentos para que el modelo aprenda qué secuencias de acciones debe tomar.