Aprendizaje Supervisado

Lautaro Estienne

Junio de 2020

En estas notas vamos a explicar las bases para aprender a diseñar algoritmos de aprendizaje automático que se entrenan en forma supervisada. Además, se explican algunos de los algoritmos que se utilizan comunmente en la práctica y que suelen dar buenos resultados, así como también los fundamentos teóricos por los cuales éstos funcionan. Se busca que, luego de haber entendido el tema, sea posible aplicar estos algoritmos y entender cómo adaptarlos al problema particular.

1. Introducción

Los algoritmos de aprendizaje automático (*Machine Learning*) tienen por objetivo **aprender de los datos**. Dada la facilidad con la que se pueden disponer de gran cantidad de datos actualmente, este área ha crecido mucho en los últimos años y se utiliza actualmente en industria e investigación, presentando muy buenos (a veces, los mejores) resultados. Ahora, ¿qué significa "aprender de los datos"? Más adelante analizaremos una aproximación formal a esta definición, pero la idea general es que las personas y la naturaleza generan eventos cuantificables de los cuales pueden llevarse un registro. Aquí vemos algunos ejemplos:

- Ejemplo 1...

En principio, es posible pensar que puede extraerse información útil de cada uno de estos registros de datos, pudiendo incluír en ella algunas características de los datos que aún no se han registrado. Este proceso por el cual se extrae información de un conjunto de datos, que también sirve para caracterizar otros datos similares que aún no se han visto se conoce como **aprendizaje**, o "aprender de los datos". Y esto es precisamente lo que buscamos hacer.

Vamos a separar en dos tipos de conjuntos de datos: los que contienen una etiqueta por cada ejemplo o muestra y los que no. En los ejemplos anteriores...

Comenzamos la formalización de todas estas ideas exponiendo un ejemplo concreto de los dos tipos de problemas que componen el aprendizaje supervisado: regresión y clasificación.

1.1. Regresión y Clasificación

Consideremos que se tiene un registro de cuán enojada está una persona en función del volumen en el que habla. Es decir, para una cierta medición i, se registró que el volumen de voz de una persona era $x^{(i)}$ (en decibeles) y dicha persona se encontraba enojada un valor $y^{(i)}$ en una escala continua de 1 a 10 (donde 1 es

"nada enojada" y 10 es "extremadamente enojada"). Esta tarea es un ejemplo de **detección de emociones**, muy común en el área de procesamiento de señales de voz.

En la figura **??** se muestra un gráfico de un conjunto de N mediciones de este tipo. En el eje horizontal se representa la variable x, conocida como **entrada** y en el eje vertical, la variable y, comunmente llamada **etiqueta**. Cada punto del gráfico representa una medición o **ejemplo**, y el conjunto de todos los ejemplos constituye el **conjunto de entrenemiento** $\mathcal{S} = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$. Cuando se tiene un conjunto de estas características (es decir, compuesto por una entrada y su etiqueta) se dice que el aprendizaje es **supervisado**.

Figura 1: Conjunto de datos detección de emociones

Nuestro objetivo será, entonces, predecir los valores de enojo de nuevas entradas, es decir, de valores de x que no aparecieron en el conjunto de entrenamiento. Mas formalmente, queremos obtener una función $\hat{h}(x)$ a partir de $\mathcal{S}$ que devuelva un valor continuo entre 1 y 10 que represente el valor de enojo de una persona cuyo volumen de voz medido fue x decibeles. Además, vamos a querer definir una medida de "cuán buena" es la función encontrada. Más adelante veremos que hay muchas maneras de definir el grado de desempeño de esta función, pero siempre se rá en función de nuevos ejemplos, no de los que pertenecen al conjunto de entrenamiento.

Los problemas en que se buscan predecir valores pertenecientes a un espacio continuo, como es el ejemplo anterior, son problemas de **regresión**. Notemos que en esta definición no consideramos las propiedades de la variable de entrada x, la cual, en principio, puede pertenecer a un espacio continuo o discreto, unidimensional o multidimensional, o incluso categórico. Por ejemplo, siguiendo con el ejemplo de detección de emociones, podríamos haber dispuesto desde un principio, no sólo del volumen de la persona sino también de la frecuencia pico máxima detectada. En ese caso, la entrada correspondería con un vector $\mathbf{x} \in \mathbb{R}^2$ en donde la primera componente representa el volumen en decibeles de la medición y la segunda, la frecuencia pico máxima en hertz. Cuando el vector $\mathbf{x}$ está compuesto por una serie de variables que representan una magnitud concreta, suele denominarse vector de **features** o características, y el espacio al que pertenece, **espacio de features**. Además, vale la pena mencionar que es muy comun disponer de un conjunto de datos en el que estas características fueron extraídas manualmente de una medición. Es decir, es más común obtener disponible un conjunto de datos de señales de voz y su correspondiente valor de enojo, que directamente el volumen o frecuencia pico máxima de dicha señal. Por eso, el proceso de predicción (o sea, de obtención de la función $\hat{h}(x)$) viene precedido por una etapa de **extracción de características** de los datos, que muchas veces tiene un peso muy importante en el desempeño final al algoritmo.

Ahora supongamos el mismo problema que antes, con la única diferencia que en lugar de querer obtener un valor de enojo en una escala continua, se desea tomar la decisión de si el el hablante se encuentra enojado o no. Es decir, el conjunto de datos de entrenamiento disponible ahora consiste en un conjunto $\mathcal{S} = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$ en donde $\mathbf{x}$ es un vector de características como antes y el escalar y ahora puede adoptar dos únicos valores: "enojado" o "no enojado".

En la figura 2 se muestra un conjunto de ejemplos utilizados en este problema. Los ejes del gráfico representan las componentes del vector $\mathbf{x}$ y cada punto pertenece a exactamente una de las categorías antes mencionadas.

Figura 2: Conjunto de datos para detección de emociones

El caso en que el espacio al cual se mapean los valores de la función $\hat{h}(x)$ es categórico se conoce como **problema de clasificación**. En este caso, el problema sigue siendo de aprendizaje supervisado, puesto que las etiquetas siguen estando.

1.2. Ejemplos conocidos: cuadrados mínimos y Discriminante de Fisher

Hasta ahora hemos planteado el problema y la nomenclatura que vamos a usar, pero no dijimos cómo obtener el algoritmo de aprendizaje.

1.3. Formalización del problema

Definición de learning en general. Modelización de caja negra para el caso supervisado. Diferentes aproximaciones para encontrar el algoritmo: modelos probabilísticos (ideales, como el filtro de Wienner y aproximados, como los de ERM), modelos por discriminantes como Fisher o SVM (esto es así???). Decir al final de todo que nosotros queremos resolver siempre un problema de regresión, ya que el problema de clasificación es equivalente a uno de regresión y después agregarle una función de decisión.

2. Modelos

- Modelos probabilísticos sin aproximaciones (filtro de wienner, etc.)
- Modelos lineales generalizados
- GDA / LDA
- Naive Bayes
- SVM
- Redes neuronales

3. Learning Theory

Acá empezaría la parte de learning theory para aprendizaje supervisado...