uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

Presentación e Introducción

En este tema inicial primero se presenta la asignatura y a continuación se comienza con la primera clase introductoria, que introduce los conceptos básicos.

Presentación

Se presenta el aprendizaje automático como un subcampo de la Inteligencia Artificial, aunque relacionando la parte de mayor interés para la titulación, como una disciplina complementaria a otras como el análisis de datos, la estadística o el aprendizaje estadístico.

Se introducen las ideas principales del aprendizaje automático mediante un ejemplo perteneciente al aprendizaje automático supervisado (clasificación). Este ejemplo se basa en un sistema llamado Skycat (clasificación automática de objetos del firmamento) construido para la NASA en los años 90. Aunque se trata de un ejemplo concreto, el esquema que se sigue para resolverlo es similar a otras aplicaciones del aprendizaje automático en otros campos.

El esquema básico consiste en disponer de un conjunto de datos de entrenamiento, que se usa como entrada a un método de aprendizaje automático y cuya salida es un modelo. El modelo propiamente dicho es una « estructura » (por usar un término genérico) que toma una entrada y devuelve una salida. La entrada es un dato y la salida es la predicción para ese dato. En este ejemplo concreto, el conjunto de datos está compuesto de datos. Cada uno de ellos contiene las características de un objeto del firmamento y también la clase a la que pertenece (estrella, galaxia, ...). El modelo toma las características de un objeto nuevo y devuelve la clase a la que (probablemente) pertenece el objeto.

Este esquema básico se irá complicando a lo largo de la asignatura de varias maneras. Por un lado, existen muchos métodos de aprendizaje automático, los cuales dan lugar a una gran variedad de tipos de modelos (reglas, árboles, redes de neuronas, máquinas de vectores de soporte, etc.). Por otro lado, se añadirán otras fases al esquema, tales como el preproceso de los datos, la evaluación del modelo o el ajuste de hiper-parámetros.

Se pueden poner muchos otros problemas que siguen un esquema parecido al expuesto, lo que muestra la aplicabilidad del aprendizaje automático. Por ejemplo, la recomendación de

productos financieros, la recomendación de libros, o la predicción de oferta y demanda eléctrica (entre otros).

A continuación se introducen y motivan los temas de la asignatura :

1. Introducción al aprendizaje automático
1.1. Tareas, modelos y algoritmos
1.2. Conceptos básicos en clasificación.
2. Métodos básicos para clasificación y regresión:
2.1. Modelos basados en similaridad: vecino más cercano (K-Nearest Neighbor KNN)
2.2. Modelos basados en árboles y reglas
3. Metodología en aprendizaje automático (machine learning pipeline)
3.1. Entrenamiento
3.2. Ajuste de hiper-parámetros
3.3. Evaluación (validación cruzada)
4. Pre-proceso de datos
4.1. Imputación, normalización, …
4.2. Selección y transformación de atributos (o predictores, o características)
5. Métodos avanzados para clasificación y regresión. Conjuntos de modelos (ensembles):
5.1. Bagging, Random Forests
5.2. Boosting, Gradient Boosting
5.3. Otros métodos avanzados: redes de neuronas, máquinas de vectores de soporte
6. Clasificación y evaluación con coste y muestras desbalanceadas. Curvas ROC
7. Introducción a técnicas de Big Data:
7.1. MapReduce
7.2. Spark

Por último se describe la herramienta que se utilizará para realizar las prácticas : MLR. MLR es una librería del lenguaje R que permite hacer aprendizaje automático. Por un lado, MLR permite trabajar con muchos métodos de aprendizaje automático disponibles para R, de la misma manera (de otra forma, habría que aprender a usar cada uno de los métodos por separado). En segundo lugar, MLR automatiza algunas tareas de aprendizaje automático (ej : ajuste de hiper-parámetros, evaluación con validación cruzada, selección de atributos, etc.).

Introducción

A continuación se introducen los conceptos básicos de aprendizaje automático (tareas, métodos, modelos). En esta asignatura solo se tratarán los métodos de aprendizaje automático inductivos, es decir aquellos que aprenden a partir de datos o ejemplos (a partir de un conjunto de entrenamiento).

A alto nivel, se distingue entre las tareas de aprendizaje automático supervisado, semi-supervisado, no supervisado y aprendizaje por refuerzo. El aprendizaje automático es supervisado si existe un valor a predecir, y se hablará de clasificación cuando lo que se prediga sea una valor categórico, como si/no o enfermo/sano ; y de regresion cuando lo que se prediga sea un valor numérico. Dentro de la tarea de clasificación se puede tratar un subproblema denominado estimación de probabilidades. La estimación de probabilidades sigue siendo un clasificación, pero aquí el objetivo no es predecir la clase sino la probabilidad de la clase. Ranking y scoring son problemas asociados, aunque lo que se predice no es una probabilidad sino un « score », es decir, un valor numérico, que sin cumplir las propiedades de las

probabilidades, sirve para conocer la seguridad que tiene el modelo en la predicción (a mayor valor, mayor confianza en que la clasificación es correcta).

El aprendizaje semi -supervisado es aquel en qué solo algunos datos están etiquetados, es decir solo se conoce la clase (o el valor contínuo a predecir) de alguno de ellos. Este tipo de tareas es útil cuando es costoso etiquetar los datos. En el aprendizaje no supervisado los datos no disponen de la clase o del valor numérico a predecir. En este caso se pueden realizar dos subtareas: agrupación (o clustering) y asociación (o problema de la cesta de la compra). Por último, existe la tarea del aprendizaje por refuerzo. Está a medio camino entre el aprendizaje supervisado y el aprendizaje no supervisado, y se utiliza sobre todo en robótica.

Como ejemplo de problema de clasificación se puede poner el de concesión de créditos bancarios : un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no van a devolverlo. La entidad bancaria cuenta con una gran base de datos correspondientes a los créditos devueltos (o no) por otros clientes con anterioridad.

Este problema es útil para introducir varios conceptos tales como instancias, atributos, clase, algoritmo de aprendizaje automático, modelo, o dato de test. En este caso el modelo es un conjunto de reglas, pero en la asignatura se verá que puede adoptar otras formas (árboles, redes de neuronas, etc.). Otro concepto importante es el *feature space* o espacio de instancias, el cual es el espacio matemático poblado por las instancias. El espacio de instancias permite entender los problemas de clasificación de otra manera a la vista hasta ahora. Hasta el momento, hemos visto que resolver un problema de clasificación consiste en encontrar un modelo que se adecue a los datos. Ahora podemos entender también que resolver un problema de clasificación consiste en encontrar una separación o frontera en el espacio de instancias.

Como ejemplo de problema de agrupación podemos plantear un problema simple relacionado con libros, en el que se dispone de dos atributos, la longitud media de las palabras del libro y la longitud media de las frases del libro. El algoritmo de clustering encontrará que los datos se agrupan en dos grupos. Estos podrían ser : libros de filosofía (que tienen palabras largas y frases largas), y novelas *best-seller* (que tienen palabras cortas y frases cortas). En cualquier caso, el método de *clustering* sólo informa que existen dos grupos. La interpretación de lo que son esos grupos la tiene que hacer un experto. El segundo tipo de problemas de aprendizaje no supervisado es la asociación. Un ejemplo que lo ilustra es el problema de la cesta de la compra, en el que se intenta encontrar asociaciones entre productos comprados por los usuarios (si compra pañales entonces también compra leche).

Por último, en la tarea de aprendizaje por refuerzo se busca encontrar modelos que representen estrategias (*policies*). Se puede ilustrar con el problema de un robot que debe encontrar su camino en una habitación, a partir de la información proporcionada por tres sensores de distancia. Los modelos a aprender son estrategias, que transforman la información proporcionada por los sensores, en la mejor acción a realizar en ese momento (que en este ejemplo simplificado serían tres : girar hacia la izquierda, hacia la derecha, o avanzar). El aprendizaje por refuerzo está a medio camino entre el aprendizaje supervisado y el no supervisado, porque el conjunto de datos no contiene la clase (izquierda/derecha/avanza) para cada situación, sino que toda la información que se proporciona al método de aprendizaje es un refuerzo, positivo cuando el robot lo hace bien, negativo cuando se equivoca (chocando con

una pared, por ejemplo). En esta asignatura no se estudiarán las técnicas de refuerzo, porque tienen interés sólo en campos muy especializados como la robótica.