Métodos predictivos: conceptos y características

UCO ONLINE

1. Introducción a los métodos predictivos

En nuestro día a día tomamos decisiones basadas en nuestro conocimiento y experiencias pasadas. A menudo, podemos identificar una serie de factores o características que influyen más que otras en esas decisiones. También somos conscientes de que podemos acertar o equivocarnos al hacer una suposición, pues puede cumplirse o no. Como se suele decir, "de los errores se aprende". Estos mismos conceptos y premisas se dan cuando queremos que sea una máquina la que, por medio de técnicas computacionales, tome decisiones a partir de los datos. En el ámbito de la ciencia de datos, los métodos que se engloban bajo el término *aprendizaje automático* son de gran utilidad para apoyar la toma de decisiones. En este curso en concreto, nos centraremos en aquellos métodos capaces de realizar *predicciones* a partir de un análisis más o menos complejo de unos datos de partida. Podemos pensar en los siguientes ámbitos de aplicación donde se nos plantea la necesidad de realizar predicciones:

- En un contexto meteorológico, nos gustaría ser capaces de predecir qué temperatura hará mañana o qué probabilidad hay de que llueva. Gracias a estas predicciones, podemos adaptar nuestra vestimenta o ajustar sistemas de regadío para que sean más eficientes.
- En el ámbito sanitario es importante predecir qué enfermedad está desarrollando un paciente para proporcionarle un tratamiento. Además, conocer qué indicadores biológicos influyen en el diagnóstico nos puede ayudar a desarrollar mejores técnicas de prevención.
- En el sector económico podemos pensar en la predicción de los valores que pueden alcanzar las acciones bursátiles a un mes o un año vista. A nivel empresarial, también podría interesar hacer predicciones de ventas para una campaña de rebajas o Navidad.

En todos estos escenarios se plantea una misma problemática: necesitamos determinar qué va a suceder en el futuro, esto es a lo que comúnmente llamamos "hacer una predicción". En un mundo completamente analógico, estas predicciones solo las podría realizar un profesional (el meteorólogo, el médico o el economista) en base a su conocimiento. Afortunadamente, los humanos hemos desarrollado la capacidad de analizar nuestro pasado y, en las últimas décadas, disponemos de medios electrónicos para registrar datos y consultarlos. En la era digital, los profesionales han ido apoyándose en los datos históricos para tomar mejores decisiones. Ahora se trata de ir un paso más allá, dejando que también sea la máquina la que nos proporcione una predicción identificando para ello los datos más significativos. Esto es especialmente relevante a medida que crece la cantidad y variedad de datos que almacenamos, pues la máquina goza de una capacidad de análisis más rápida que la mente humana. No obstante, no se debe olvidar que una máquina, por avanzados que sean los métodos que aplique, no tiene la capacidad de comprender lo que infiere y las consecuencias que puede implicar cada decisión.

Debemos ser conscientes de que llevar a la práctica una decisión o implantar una solución propuesta por la máquina seguirá siendo responsabilidad del humano. Por tanto, es importante conocer también las limitaciones de los métodos computacionales aplicados, y saber interpretar adecuadamente sus resultados.

2. Concepto de aprendizaje supervisado

Antes de introducir los principales conceptos que rodean a los métodos predictivos, es importante ubicarlos en un contexto más amplio. La ciencia de datos puede hacer uso de muchos tipos de técnicas, desde aquellas más ligadas a las ramas de las matemáticas y la estadística, hasta sofisticados métodos computacionales. En realidad, estos últimos tienen una fuerte base matemática y estadística, pero han evolucionado tanto que se han hecho merecedores de su propia área de conocimiento y terminología. Hablamos del *aprendizaje automático*, la rama de la inteligencia artificial que se centra en el análisis de grandes cantidades de datos para apoyar la toma de decisiones. Dentro de esta área encontramos diferentes clasificaciones en función del tipo de problema a resolver o la forma en la que se tratan los datos durante el "aprendizaje".

Respecto al tipo de problema a resolver, en este curso abordaremos dos casos: *regresión* y *clasificación*. Aunque serán definidos formalmente más adelante, podemos mencionar aquí su principal diferencia. En un problema de regresión, queremos predecir un valor *cuantitativo*, como la temperatura en los ejemplos puestos al principio. En un problema de clasificación, el objetivo es predecir un valor *cualitativo*, por ejemplo, si el paciente tiene o no una determinada enfermedad.

Desde el punto de vista del tratamiento de los datos, ambas tareas se abordan mediante *aprendizaje supervisado*. Este tipo de aprendizaje consiste en disponer de una serie de ejemplos para los cuales se conoce su predicción. Los ejemplos se le presentan al método sin considerar la predicción real, pero se dispone de ella para ir comprobando si acierta o falla en su predicción. El "aprendizaje" ocurre, por tanto, bajo una estrategia supervisada, pues el método va evaluando su propio rendimiento con el objetivo de ir mejorando progresivamente. Volviendo al ejemplo médico, si disponemos del historial clínico de miles de pacientes diagnosticados por un médico, el método predictivo va a intentar acercarse lo más posible al criterio de ese médico (aunque no sepamos cuál es) para clasificar la enfermedad que sufre un nuevo paciente. Este último matiz es importante, ya que una de las claves principales del aprendizaje supervisado es conseguir una buena capacidad de *generalización*. Esto significa que el mejor método no es aquel que más acierta a la hora de diagnosticar los pacientes de ese historial concreto (los ejemplos "conocidos"), sino aquel que consigue una mejor predicción ante pacientes futuros (los ejemplos "no conocidos"). En otras palabras, no se trata de memorizar lo que el médico diagnosticó en esos casos concretos, sino de simular su proceso de razonamiento para detectar la enfermedad ante unos síntomas dados. El concepto de generalización se desarrollará más formalmente en las próximas semanas.

En contraposición al aprendizaje supervisado, en el aprendizaje no supervisado solo se dispone de la colección de ejemplos, sin asociarles un valor de "salida". Por tanto, este tipo de enfoque es aplicado en tareas de índole descriptiva, no predictiva. El aprendizaje no supervisado se estudiará detalladamente en otro curso de este máster.

Finalmente, el aprendizaje automático nos ofrece una gran variedad de métodos predictivos. Cada uno de ellos se basa en distintos principios matemáticos y formas de representación del conocimiento. En el siguiente apartado se introducen brevemente las principales características de los métodos predictivos, tanto los orientados a regresión como a clasificación. A lo largo de las semanas, se profundizará en muchos de ellos tanto desde una perspectiva teórica como práctica.

3. Características de los métodos predictivos

Un método predictivo se define como una técnica computacional orientada a establecer una relación entre un conjunto de variables de entrada con el fin de *estimar* el valor de una variable de salida (la predicción). Para ello, los métodos predictivos analizan datos históricos sobre situaciones conocidas a fin de extrapolar o inferir una relación de causa-efecto entre las variables de entrada y la de salida. Dicho análisis suele basarse en la estadística, el análisis de tendencias y el descubrimiento de patrones en los datos. Bajo el enfoque del aprendizaje supervisado, los métodos predictivos presentan una serie de características en común:

- Se dispone de una serie de ejemplos descritos a partir de un conjunto de variables de entrada que miden propiedades de dichos ejemplos.
- Para la colección de ejemplos, denominada *conjunto de entrenamiento*, el valor de la variable de salida es conocida, pudiendo ser de naturaleza cuantitativa o cualitativa. Cuando se da esta situación, decimos que los datos están *etiquetados*.
- Un método predictivo trata, por lo general, de *aproximar una función* definida en base a algunas o todas las variables de entrada. A este proceso se le llama *entrenamiento*, y suele ser iterativo. La función resultante puede ser más o menos compleja en cuanto a su formulación matemática.
- La salida de la función es la variable a predecir, de forma que puede utilizarse para obtener su valor estimado dados los valores de las entradas.
- Algunos métodos no expresan su predicción como una función matemática (aunque la construyan internamente), sino como un *modelo*. Un modelo es una representación computacional más compleja (por ejemplo, en forma de reglas o árbol) que nos indica cómo obtener la predicción a partir de los valores de las variables de entrada.
- Es posible evaluar cómo de bien se *ajusta* la función o el modelo a los ejemplos del conjunto de entrenamiento, pues se conoce el valor real de la variable de salida.
- Para comprobar la calidad del modelo inferido se utiliza un conjunto de ejemplos etiquetados que no ha sido utilizado para el entrenamiento, el *conjunto de test*.

4. Tipos de métodos predictivos

Aunque los problemas de regresión y de clasificación tienen formulaciones ligeramente diferentes, existen métodos predictivos aplicables a ambos. Habitualmente, muchos de estos métodos son propuestos para un tipo de problema (normalmente regresión), y luego se generalizan o adaptan al otro. No obstante, también es posible encontrar métodos específicos para cada tipo de problema, como se irá estudiando a lo largo del curso.

Antes de describir brevemente algunos de los principales tipos de métodos predictivos, es conveniente hacer otra aclaración. Los métodos predictivos que se estudian en este curso tienen una naturaleza estática. Esto es, se entrenan sobre un conjunto histórico de datos que se asume que no varía rápidamente en el tiempo. Si se reciben más datos, habría que reentrenar el método de nuevo. En contraposición, existen métodos predictivos de naturaleza dinámica que son capaces de aprender de forma incremental, conforme se van llegando nuevos datos. Algunos de ellos son propuestos como adaptaciones de los métodos estáticos, ya que un escenario dinámico es más complejo que uno estático.

En general, los métodos predictivos pueden agruparse en los siguientes tipos principales:

1. **Métodos lineales**. Son aquellos que establecen una relación lineal entre las variables de entrada y salida. Son aplicables tanto a regresión como a clasificación.
2. **Métodos probabilísticos**. Se basan en asignar una probabilidad a cada valor posible de la variable de salida. Son más habituales para problemas de clasificación.
3. **Métodos basados en reglas.** Generan un conjunto de reglas de la forma *If-Then* para expresar las condiciones que deben darse en las variables de entrada para producir una salida u otra. Tienen gran utilidad para abordar problemas de clasificación.
4. **Métodos basados en árboles de decisión**. Son similares a los anteriores, pero representan el flujo de decisión en forma de árbol, donde cada rama termina en un valor de la variable de salida. Aunque se pueden aplicar a regresión, son más frecuentes en clasificación.
5. **Máquinas de vector soporte.** Son métodos que aplican transformaciones en las variables de entrada hasta encontrar nuevas dimensiones en la que exista una separación máxima entre los ejemplos según su tipo (en el caso de clasificación) o pueda establecerse una relación lineal entre las variables de entrada (en el caso de regresión).
6. **Redes neuronales.** Son métodos inspirados en la composición neuronal del cerebro humano, definida en base a unidades de procesamiento (neuronas) organizadas en capas. Los valores de las variables de entrada se propagan desde la capa de entrada hasta la de salida, sufriendo transformaciones en las capas intermedias. Aunque una red neuronal genera un valor cuantitativo a la salida (aplicable a regresión), también es posible utilizar dicho valor para clasificar.
7. **Métodos de *ensemble*.** Son métodos que se basan en la integración de varios modelos predictivos más básicos y distintos entre ellos de algún modo. La salida final del ensemble se obtendría combinando las salidas de los clasificadores que lo integran. Estos métodos se pueden aplicar tanto a clasificación como a regresión.

5. Terminología y notación

A lo largo de este documento se han ido introduciendo varios términos alrededor del concepto de aprendizaje supervisado. Estos y otros términos se utilizarán a lo largo del curso para formular los problemas tanto de regresión como de clasificación, así como explicar los distintos métodos predictivos. Como finalización a este tema introductorio, conviene recopilar las definiciones de los principales términos que aparecerán en el resto del curso y establecer una notación común.

Término	Símbolo	Definición
Algoritmo	Ω	Secuencia de pasos para ajustar la función o construir el modelo en un método predictivo.
Atributo	x_j	Cada una de las variables de entrada que caracterizan a un ejemplo del conjunto de datos.
Coeficiente o peso	β / ω	Ponderación asociada a un atributo o cálculo, que expresa su influencia en una predicción.
Conjunto de datos	X	Colección de ejemplos destinados al aprendizaje.
Conjunto de entrenamiento	X_{train}	Subconjunto de instancias de X con los que se lleva a cabo el entrenamiento.
Conjunto de test	X_{test}	Subconjunto de instancias de X con los que se lleva a cabo la evaluación.
Entrenamiento		Proceso de ajuste de una función o construcción del modelo de decisión a partir de X_{train}.
Estimación	$\hat{y}$	Valor obtenido por la función o modelo predictivo para una instancia dados los valores de sus atributos.
Etiqueta (de clase)		Valor cualitativo asignado a la variable de salida de un ejemplo en problemas de clasificación
Error	ε	Diferencia entre el valor estimado y el valor real de una variable.
Función predictiva (o de ajuste)	f	Función a aproximar para resolver el problema de regresión o clasificación.
Función base (*kernel)*	ϕ	Función que realiza transformaciones sobre las variables de entrada.
Instancia	X_i	Cada uno de los ejemplos disponibles en X.
Método predictivo		Técnica computacional por la cual se obtiene un modelo o función capaz de estimar una variable de salida.
Modelo predictivo		Estructura de datos que encapsula la forma de predecir la variable de salida.
Variable independiente	y	En regresión, variable de salida a predecir.

6. Librerías software para aprendizaje supervisado

La creciente popularidad del aprendizaje automático y, en particular, el aprendizaje supervisado, ha propiciado la aparición de numerosos recursos computacionales para poner en práctica sus métodos. Actualmente, existen multitud de librerías y herramientas software disponibles en una gran variedad de lenguajes de programación. Estas herramientas nos permiten construir modelos predictivos, tanto de regresión como de clasificación, aplicando algunas de las estrategias de entrenamiento y evaluación que se explicarán a lo largo de este curso.

En general, los ejemplos y tareas propuestas para este curso se han desarrollado en Python, por lo que nos centraremos primero en presentar algunas de las librerías más útiles para el propósito de este curso. De forma complementaria, los estudiantes dispondrán de ejemplos en R a través de recursos externos, de forma que tendrán la oportunidad de aplicar los mismos conceptos en dicho lenguaje. Al final de este apartado se mencionan algunas de las herramientas disponibles en otros lenguajes de programación que gozan de gran popularidad y madurez.

6.1. Librerías para aprendizaje supervisado en Python

Python se ha convertido en el lenguaje de programación por excelencia para el desarrollo de análisis y aplicaciones basadas en aprendizaje automático. Las siguientes librerías serán las utilizadas en este curso para proporcionar ejemplos prácticos y proponer ejercicios a completar en las tareas de asignación:

- ***Statsmodels*** es una librería que inicialmente formaba parte de la librería *SciPy*. Desde 2009, la librería es independiente y ha evolucionado enormemente. Con esta librería se pueden hacer análisis estadísticos (exploración de datos, comprobación de hipótesis), además de una colección de métodos de estimación basados en regresión. Esta librería se utilizará para las lecciones sobre regresión durante las dos primeras semanas del curso. La documentación de la librería puede consultarse en: https://www.statsmodels.org/
- ***Scikit-learn*** es una librería que incluye varios métodos de aprendizaje automático, incluyendo métodos para regresión y clasificación. Comenzó a desarrollarse en 2007, aunque el proyecto se hizo público en 2010. Su implementación se basa en librerías ampliamente extendidas como *NumPy*, *SciPy* y *matplotlib*. Además, cuenta con numerosas extensiones para otros tipos de aprendizaje. Esta librería será empleada a lo largo de todo el curso. La documentación de la librería puede consultarse en: https://scikit-learn.org/

6.2. Librerías para aprendizaje supervisado en R

Como se ha mencionado anteriormente, para este curso se ha elegido Python como lenguaje de programación. La razón principal es que *scikit-learn* aglutina la mayor parte de los métodos que se explicarán. No obstante, R es otro lenguaje que nos proporciona una gran cantidad de recursos para analizar datos aplicando métodos predictivos.

A continuación, se enumeran los paquetes más relevantes que los estudiantes pueden explorar de forma complementaria:

- ***Stats*** es un paquete que contiene una gran cantidad de funciones para análisis estadístico y, más relacionado con este curso, construir modelos de regresión.
- ***Mass*** es otro paquete que ofrece métodos de regresión de diferentes características, incluyendo facilidades para generar gráficos.
- ***Caret*** es un paquete que proporciona una gran variedad de métodos para regresión y clasificación. Incluye funciones útiles para las tareas que rodean al proceso de aprendizaje.

6.3. Otras herramientas para aprendizaje supervisado

En este apartado se describen algunas herramientas basadas en otros lenguajes de programación que también resultan de gran utilidad para realizar tareas de aprendizaje supervisado. Cabe señalar que estas herramientas cuentan en su mayoría con entornos gráficos que pueden facilitar un primer acercamiento a los métodos predictivos desde una perspectiva de usuario final. No obstante, también disponen de métodos avanzados y funciones para realizar experimentaciones.

En primer lugar, ***Weka*** es una herramienta Java que proporciona varios métodos de aprendizaje, incluyendo varios dirigidos a resolver problemas de regresión y clasificación. Cuenta con un largo historial de desarrollo (desde 1999), por lo que se ha convertido en una herramienta muy popular y adecuada para iniciarse en la disciplina. Tanto el software como su extensa documentación pueden encontrarse en su página web: https://www.cs.waikato.ac.nz/ml/weka/

KNIME proporciona una plataforma *open source* para el desarrollo de soluciones basadas en ciencia de datos. También ofrece una solución de pago a nivel servidor, dirigida a la creación de sistemas basados en ciencia de datos con una perspectiva más industrial. Su entorno gráfico es intuitivo, ya que se basa en la construcción de flujos de trabajos. Se puede encontrar más información en su página web: https://www.knime.com/

RapidMiner es otra plataforma para aprendizaje automático basada en flujos de trabajos. Cuenta con diferentes modalidades para la construcción, gestión y despliegue de soluciones basadas en ciencia de datos. Además, permite integrar código desarrollado en Python y R, así como funcionalidades avanzadas para el ajuste de parámetros, la validación de los modelos predictivos y la interpretación de resultados.

KEEL es una librería *open source* desarrollada principalmente por personal de distintas universidades españolas, que permite la creación de flujos de manera visual. No solo permite utilizar algoritmos de aprendizaje supervisado, sino que cubre todo el proceso de descubrimiento del conocimiento. Al contrario que otras herramientas similares, KEEL presta especial atención a aquellos métodos basados en algoritmos evolutivos (los cuales se encuentran fuera del ámbito de este curso). Se puede encontrar más información en su página web http://www.keel.es/

Referencias

S.V. Burger. “Introduction to Machine Learning with R”. O’Reilly, 1ª edición, 212 páginas. 2018. Disponible en: https://github.com/bcaffo/regmodsbook

B. Caffo. “Regression Models for Data Science in R”. Leanpub (CCA-NC 3.0), 129 páginas. 2015.

E. Duchesnay, T. Löfstedt, F. Younes. “Statistics and Machine Learning in Python”. Disponible en: https://duchesnay.github.io/pystatsml/

A. Géron. “Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems”. O’Reilly, 2ª edición, 483 páginas. 2019.

G. Hackeling. “Mastering Machine Learning with Scikit-Learn”. Pactkt Publishing, 221 páginas. 2014.

T. Hastie, R. Tibshirani, J. Friedman. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. Springer Series in Statistics, 2ª edición, 745 páginas. 2017.

G. James, D. Witten, R. Tibshirani, T. Hastie. “An Introduction to Statistical Learning with Applications in R”. Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/

M. Kubat. “An Introduction to Machine Learning”. Springer, 2ª edición, 348 páginas. 2017.

B. Lantz. “Machine Learning with R”. Pactkt Publishing, 1ª edición, 375 páginas. 2013.

A.C. Müller, S. Guido. “Introduction to Machine Learning with Python: A Guide for Data Scientists”. O’Reilly, 1ª edición (3ª impresión), 378 páginas. 2017.