Minería de Datos y Modelización Predictiva

Introducción y conceptos básicos

Autor: Guillermo Villarino

Actualizado: **Otoño 2023**

INTRODUCCION.

En este módulo de Minería de Datos abordaremos los aspectos primordiales del tratamiento de bases de datos estructuradas con el objetivo último de generar un modelo predictivo con la "suficiente" fiabilidad para ser implementado y pasar a la cadena de producción en un esquema de negoció de cualquier empresa que pueda considerarse "Data Driven Company".

"You can have data without information, but you cannot have information without data." - Daniel Keys Moran

---

Delimitaremos entonces nuestras responsabilidades como Data Scientists Modelizadores, de manera que quede perfectamente estipulado cual es nuestro punto de inicio y qué se espera como producto final de nuestra aportación al flujo de negocio.

En el esquema habitual, los modelizadores trabajaremos con conjuntos de datos relativamente estructurados cuyos requerimientos serán trasladados al equipo encargado de la gestión de las grandes bases de datos. Por lo tanto, nuestro esquema mental debería ser algo así como:

¿Cuál es el objetivo de mi trabajo? ¿Qué necesito predecir? (Vamos pensando en posibles modelos a aplicar..)

Aquí normalmente definiremos una variable objetivo sobre la que finalmente trataremos de realizar predicciones de su comportamiento futuro o para nuevas instancias desconocidas.

¿Qué necesita mi modelo para alimentarse? ¿Se dispone o se puede disponer de esta información? ¿Cómo debe estructurarse?

En este punto, una vez tenemos claro el objetivo y cual es la variable que mide dicho objetivo, cabe plantearse cuales son las características (variables independientes o predictores) que pueden tener influencia sobre ese famoso objetivo. De esta forma tendremos instancias que recogerán el valor que toma cada una de estas características junto con el valor que toma la variable objetivo asociada a esa instancia (cada instancia puede representar una persona, un producto etc…). Bien, ya tenemos los datos estructurados!

Ahora empieza la parte del trabajo que consiste en "mancharse las manos" y dar vueltas a los datos, jugando con ellos de distintas formas para evaluar su "calidad" y es esta una palabra muy genérica que en ocasiones puede crear confusión. Por este motivo trataremos de caracterizar esa "calidad" a través de distintos conceptos que representan el pan de cada día y a los que terminaremos cogiendo incluso cariño.

¿Como podemos evaluar la calidad del dato?

Podemos abordar esta cuestión desde el punto de vista de las variables (normalmente en columnas) y desde el punto de vista de las instancias o registros (normalmente en filas) que componen nuestro conjunto de datos.

Atendiendo a las variables (a nivel individual) nos gustaría saber:

- Tipo de variable (continua, categórica/nominal, fecha, ID, objetivo…)
- Valores que puede tomar (rango esperado, número de categorías o niveles..)
- Distribución que presenta la variable. Aquí hay mucho que decir…pero centraremos nuestros esfuerzos en:
    - Evaluación de medidas de tendencia central (media, mediana, moda…)
    - Evaluación de medidas de dispersión (rango, varianza, desviación típica, asimetría, curtosis ..)
- Valores atípicos (outliers). Resulta fundamental evaluar la incidencia de outliers en las variables puesto que son una fuente de posibles problemas a la hora de modelizar.
- Valores perdidos (missing). De la misma forma los valores perdidos pueden introducir un sesgo importante en las predicciones de nuestros modelos y es crucial controlar este aspecto.

Atendiendo a las variables (a nivel conjunto) nos gustaría saber:

- Relación existente entre los pares de variables de nuestro archivo.
    - **Relación entre predictores**. Es importante conocer el nivel de relación que presentan las variables que utilizamos para predecir nuestra variable objetivo pues pueden causar confusión en la interpretación de los modelos. Centraremos nuestros esfuerzos en evitar la famosa *colinealidad* que se produce cuando existe una relación casi perfecta entre predictores y provoca que el modelo de Regresión se encuentre mal especificado y la estimación de los parámetros resulte poco fiable debido a las "interferencias" que se producen por tan alta dependencia. En un caso grave, cuando hay *colinealidad perfecta*, directamente no se pueden estimar los parámetros por imposibilidad de inversión de la matriz de diseño del modelo.
    - **Relación entre predictores y la variable objetivo**. En este punto buscamos justamente lo contrario que en el anterior ya que, como hemos dicho, nuestros predictores deben tener "influencia" en la variable objetivo. Pues bien, esa influencia, en primera aproximación, la podemos valorar como la relación que presenta con la variable objetivo (en términos del coeficiente de correlación lineal u otros indicadores que nos permiten evaluar relaciones de tipo no lineal). Este es el primer paso para nuestro proceso de selección manual de variables ya que frecuentemente (exceptuando los patrones más profundos que pueden existir en los datos que resultan inobservables a nivel descriptivo) las variables que presentan relación a nivel descriptivo tendrán capacidad de discriminación sobre la respuesta, lo que se traducirá en un estimador

estadísticamente significativo y una aportación elevada al R2 de nuestro modelo.

Atendiendo a las instancias o registros nos gustaría evaluar:

- Desde la perspectiva de la **estadística clásica** y el diseño de experimentos, es fundamental conocer y asegurar la representatividad de la muestra de datos recogida con respecto a la población objetivo del estudio. Por ejemplo, si diseñamos una encuesta debemos asegurar que la proporción de gente mayor de 45 años consultada sea aproximadamente igual a la que presenta la población, de lo contrario se pueden producir sesgos de selección que posteriormente son complicados de tratar. Es decir, el objetivo es dejar que el azar juegue lo menos posible en nuestro análisis y nuestro esfuerzo se centrará en "controlarlo" a través del control de distintos factores que puedan introducir indeterminación en el diseño. No hay que olvidar que en este contexto manejamos el término *inferencia* que representa la capacidad de extrapolar las conclusiones sobre la muestra analizada a la población en general. Muestra -> análisis -> inferencia a la población.
- Desde el punto de vista de la **Minería de Datos y el Big Data**. En este punto, el paradigma cambia en lo fundamental puesto que, con frecuencia, no somos nosotros quienes diseñamos el experimento ni controlamos la muestra y más bien nos basamos en la ingente cantidad de datos de los que se dispone (tráfico de uso de la web, base de datos de clientes de bancos, operadores de telefonía…). En estos casos no se tiene claro cual debe ser la muestra o en ocasiones nos encontramos con que tenemos a la población completa…el concepto de inferencia se puede llegar a difuminar… Es por ello que el modelo de estadística clásica se enfrenta a serios problemas en su concepción primigenia. Aquí es donde se hace fundamental manejar los conceptos estrella del machine learning como son el ajuste de modelos basado en entrenamiento/validación/prueba o re-muestreo por medio de validación cruzada repetida. Digamos que estos son nuestros nuevos conceptos de inferencia..

Básicamente centraremos los esfuerzos en asegurar que las instancias tengan suficientes datos válidos recogidos, imaginemos un individuo que solamente tiene el 10% de la información (medida en proporción de variables con valor válido) regida adecuadamente… Es difícil pensar que nos resultará útil para que nuestro modelo "aprenda" de sus características y más bien puede llegar a provocar una gran confusión para el pobre modelo que tratará de buscar patrones en vano.

Tenemos que saber que en este punto, **la imputación** de datos perdidos juega un papel fundamental ya que se trata de una herramienta que permite asignar valores de comportamiento "promedio" a características no observadas y que se basa en el análisis de la base de datos en su conjunto (desde la más pura simplicidad de asignar la media de la variable hasta los más sofisticados procesos que tienen en cuenta los individuos con un comportamiento similar en otras características asignando "vecinos" en los que basar las imputaciones o métodos por cadenas de Markov o imputación múltiple etc.).

Estas herramientas son muy útiles pero se deben utilizar con conocimiento y al menos valorar las consecuencias que puede tener su utilización. Rescatando el ejemplo de ese individuo con solamente el 10% de la información recogida, pues resultará difícil pensar que las imputaciones (por cualquiera de los métodos) resultarán fiables ya que no se tiene suficiente palanca de información para saberlo...finalmente tendremos una instancia (individuo) con el 90% de la información "inventada"…esto es un riesgo!!

Todos estos aspectos componen nuestro trabajo inicial de estudio descriptivo y depuración de los datos estructurados de partida, lo que se conoce como **Data Understanding and Preparation.**

"Input quality data, output quality performance." - Christian Baloga

---

La evaluación de la calidad del dato es un aspecto fundamental en el modelado para la toma de decisiones puesto que los modelos no saben discriminar entre realidades y pequeñas mentirijillas que, de manera más que habitual, son añadidas (o no suficientemente penalizadas) por el humano en la base de "conocimiento" en la que el modelo basa su construcción. Se habla de que este trabajo puede suponer más del 80% del tiempo de dedicación en toda la fase de modelado.

Este trabajo de depuración de los datos es válido para cualquier proceso de modelización predictiva en el que podamos pensar, se trate de modelos de regresión clásicos o de los más sofisticados algoritmos basados en Redes Neuronales o *Gradient Boosting*. Prestemos mucha atención a esto porque puede marcar la diferencia entre un modelo mediocre y un buen modelo.

Esta será nuestra primera "lucha" en el módulo de Minería de Datos y le dedicaremos la atención que merece. Decir en este punto que existen multitud de estrategias en la depuración de datos, así como multitud de herramientas para llevar a cabo esta labor. Es imposible abarcar todo, por lo que nos centraremos en los aspectos comunes y solamente algunas de las herramientas disponibles. Por supuesto, cualquier duda adicional puede ser consultada por el foro y estaré encantado de extender información o remitir a la documentación apropiada.

Vale. Ya tenemos los datos muy bonitos y depurados. ¿Ahora que hacemos con ellos?

Pues vamos a abordar ya lo que supone la fase de modelización predictiva para la toma de decisiones.

¿Qué pretendemos?

El objetivo principal tiene que ser el siguiente: **Dado un nuevo registro/instancia con sus características observables correctamente medidas** (nos referimos al valor que toma esa instancia en las diferentes variable independientes o predictores), **¿cuál es el**

valor que se estima para la variable objetivo? (normalmente no puede medirse al menos al mismo tiempo que las características)

Abordaremos esta pregunta desde el punto de vista de la **modelización o aprendizaje supervisado**, y es importante saber que quiere decir esto de "supervisado". No significa que tienes a la jefa detrás supervisando tu trabajo (menos mal!) sino más bien a que tu variable objetivo ha tomado valores observables y medibles al menos en un periodo de tiempo/subconjunto de instancias, por lo que puedes construir modelos cuya función objetivo sea cometer el menor error posible con respecto a ese valor observado (que supervisa el aprendizaje de tu modelo).

La idea es aprender de lo observado para predecir lo que ocurrirá en el futuro o en otra muestra de instancias no etiquetadas en su variable objetivo.

Es el caso más habitual de modelos predictivos ya que frecuentemente se quiere predecir un comportamiento medible y se realizan los esfuerzos necesarios para medir el evento en al menos una ventana temporal/subconjunto de la población objetivo.

Gran parte de la Inteligencia Artificial se basa en este tipo de modelos, por supuesto sin desmerecer las capacidades de modelos de carácter no supervisado (no existe variable respuesta, o no es observable o no se ha podido medir) como el clustering o el análisis factorial que tienen un amplio abanico de aplicaciones.

Bien, dicho esto, nuestro objetivo es hacer predicciones que sean fiables y cometan el menor error posible y que se basen en cierto conjunto de características para la estimación de nuestro objetivo. En un enfoque general lo que necesitamos es una función que dado un vector de características sea capaz de asignar un valor a la variable objetivo de tal forma que si Y representa nuestro objetivo y disponemos de un vector de n características, X=(x1,x2,….xn), necesitamos algo así como:

$$Y\_estimada = f(x1, x2,\ldots,xn)$$

Siendo f algún tipo de función ya sea lineal o no lineal que opere sobre los valores de los predictores para "adivinar" el valor de la repuesta. De hecho f no necesariamente tiene que ser una función como tal, pudiendo ser una base de reglas sobre los predictores (como hacen los árboles de decisión o su "digievolución" el omnipotente Random Forest), puede ser que en la construcción de esa f se haga necesario un auténtico lío de relaciones entre capas ocultas con neuronas (como en Redes neuronales) o se haga depender de un parámetro de regularización que hace que el algoritmo vaya aprendiendo en sucesivas iteraciones (como pasa con el Gradient Boosting).

Sea como fuere, este es el concepto básico de la modelización predictiva y la cuestión ahora es buscar una buena f.

Nuestro objetivo se centra en los modelos de **Regresión Lineal** (para variables objetivo continuas) y **Regresión Logística** (para variables objetivo binarias aunque puede extenderse a categóricas de más de dos niveles).

Extenderemos entonces los modelos matemáticos que se encuentran detrás de estos dos tipos de modelos predictivos y haremos una pequeña discusión sobre la diferencia entre su aplicación en un contexto de estadística clásica (recordemos aquello de la inferencia y el diseño de experimentos) y en el nuevo paradigma de la minería de datos y el big data.

Algo relevante a tener en cuenta y que resulta general sea cual sea el modelo predictivo a aplicar es el esquema de ajuste a llevar a cabo. En la estadística clásica decíamos, si la muestra es representativa y los valores de ajuste del modelo son buenos (generalmente contrastes de hipótesis sobre significación del modelo, de sus parámetros, evaluación de premisas en cuanto a la normalidad, heterocedasticidad o independencia de los residuos etc) entonces podemos inferir a la población objetivo con cierto nivel de confianza. Es decir, no había esquema más allá del diseño inicial y de conseguir un set de variables que cumplieran estas premisas que comentábamos.

Ahora la cosa cambia y en minería de datos, con frecuencia se hace difícil o imposible el cumplimiento de ciertas hipótesis del modelo por lo que nuestro esquema de ajuste del modelo ha de tener unas cuantas fases para cumplir con los estándares de calidad. Así, para conseguir algo parecido a la seguridad que nos daba la inferencia basada en premisas, lo que se estila es llevar a cabo un buen proceso de validación del modelo para evaluar su comportamiento ante distintas perturbaciones.

¡¡OJO con el sobreajuste!! Esto es algo crucial en este mundillo…el hecho de que un modelo ajuste casi a la perfección a nuestro conjunto de datos no asegura en absoluto (frecuentemente todo lo contrario) que tenga buena capacidad de generalizar a datos nuevos. Esto es lo que se conoce como overfitting y cuanto más complejo es el modelo, más adolece de este mal endémico al ML. Cuando un modelo extrae patrones demasiado ad-hoc para los datos que conoce y sabido que esa muestra de datos no necesariamente es representativa de la población objetivo, es fácil que aparezcan nuevas instancias que sean un poco distintas y se le adjudique un valor muy lejano a su realidad…

Todo esto motiva el esquema de trabajo que introduciremos en el módulo y que será de aplicación para futuras modelizaciones por complejas que sean. Nuestro truco será no enseñarle al modelo todos los datos disponibles y dejar algunos de reserva para luego preguntarle los valores que pronostica, teniendo así una idea de su capacidad de generalización ante datos desconocidos. Esto se conoce como esquema training/test. El modelo se ajusta en training y se prueba en test, se comprueba el error cometido y la diferencia de errores en ambos conjuntos (lo que nos da una idea del posible sobreajuste) y se vuelve a reajustar el modelo (tal vez sin tal variable, con otra transformación…) Y este proceso se repite hasta asegurar la capacidad de generalización de las estimaciones.

Un pequeño paso más… Vale, podemos pensar en que este proceso de training/test está bien, parece buena idea, pero ciertamente tiene una gran dependencia de la elección de esa partición de los datos….has tenido buena suerte y tu test es muy parecido al training..o todo lo contrario, no se parecen en nada y será compleja la generalización…finalmente es algo que queda en manos del traicionero azar. ¿Cómo

podemos mejorar un poquito esto? Pues esta es precisamente la motivación subyacente a la aparición del esquema de aprendizaje basado en remuestreo y validación cruzada.

La idea filosófica de la validación cruzada (Cross Validation) es tratar de evitar esa dependencia en la partición training/test. Para ello podemos pensar que, dado que tenemos buenas máquinas que operan con gran velocidad, ¿por qué no repetir ese proceso de particiones n veces? Así pues, partimos el conjunto de datos por filas, de tal forma que tenemos por ejemplo 5 partes (20%) de las instancias en cada una de ellas. Por lo que, tentativamente tenemos 5 esquemas training/test implícitos ya que puedo utilizar 4 partes contra la restante de 5 formas distintas. Bien, esto es un buen paso puesto que el modelo puede aprender de todos los datos pero nunca simultáneamente con lo que el sobreajuste necesariamente se reduce así mismo tenemos 5 conjuntos de test diferentes para poder evaluar el sesgo (error) pero también la varianza de ese error….es decir funciona muy bien en alguno de ellos pero fatal en otro…varianza elevada. (Este trade off entre sesgo y varianza es un cásico del ML y lo abordamos en el material del módulo).

Vale, casi estamos. Aún tenemos cierta dependencia de esa partición de datos puesto que se realiza una sola vez y aún hay mangan ancha para que el azar juegue una mala pasada. Oye pues entonces repetimos este proceso con distintos puntos de partida para seccionar el conjunto de datos. Esto es lo que se conoce como validación cruzada repetida y representa el estándar de evaluación en el contexto de modelos de ML hasta tal punto que no existe un algoritmo que vea la luz sin un reproducible esquema de CV repetida en datasets conocidos por la comunidad científica. Y este es el esquema que vamos a utilizar como resultado final de nuestra modelización.

Ahora ya no tendremos solamente n esquemas training/test sino que tendremos nxm esquemas, siendo m el número de repeticiones. En el caso anterior teníamos n=5 y podríamos tener m=5, con un total de 25 ajustes por modelos. Con este resultado evaluaremos la relación sesgo varianza y lo pondremos en balanza con algunos otros aspectos como la sencillez de la fórmula empleada (principio de parsimonia) o la buena especificación de las regresiones en nuestro caso mediante los p-valores asociados a los contrastes de hipótesis sobre parámetros y bondad de ajuste del modelo.

Estas son las ideas que abordaremos con la mayor profundidad posible dentro del limitado tiempo disponible en el módulo.

El objetivo no es tanto ser una experta en modelización en 2 semanas sino más bien interiorizar estos conceptos y estar atentas para evitar lo que yo llamo el "mal del botón" y es que, dada la gran oferta de herramientas automáticas para el análisis de datos, encuentro que mucha gente le da al botón sin mucho conocimiento de lo que esa funcionalidad hace y extrae las más extravagantes conclusiones…

¡No es solamente darle al botón sino saber cuándo y porque le damos al botón!

Mucho ánimo y recordad

Practice Makes Perfect