Introducción al aprendizaje supervisado

(ver 2019-2020)

Emilio Letón y Elisa M. Molanes-López

27-Oct-2019

Página legal

INTRODUCCIÓN AL APRENDIZAJE SUPERVISADO

Madrid, versión 2019-2020

ISBN electrónico: xxx.
Edición digital (epub): xxx.

Prefacio

Este material está diseñado con el paradigma del grupo de innovación docente miniXmodular de generar material en formato mini y modular, según se puede ver en la página web www.minixmodular.ia.uned.es, donde se introduce, entre otros, el concepto de mini-libro electrónico modular.

En este mini-libro se verán algunos conceptos básicos relativos al aprendizaje supervisado que está indicado en la modelización de una variable respuesta Y (también llamada dependiente, explicada, "output" o resultado) a través de múltiples variables explicativas $X_1,\ X_2, \cdots, X_p$ (también llamadas independientes, predictores, características, "input" o entradas). Existe también un aprendizaje no supervisado que está indicado en el caso de que no haya variable respuesta y sí haya múltiples variables explicativas $X_1,\ X_2, \cdots, X_p$. Este tipo de nomenclatura (supervisado / no supervisado) es habitual en el ámbito de la minería de datos (ver, por ejemplo, el capítulo 2 de [2]). A lo largo de este mini-libro se verá que, dentro del análisis supervisado, lo que determina el modelo a utilizar no es la naturaleza de las variables explicativas sino la naturaleza de la variable respuesta.

Se recomienda refrescar los conceptos básicos de esperanza y varianza de una variable aleatoria (v.a.) en, por ejemplo, [3].

En el mini-capítulo 1 se mencionan los principales modelos dentro del análisis supervisado y en el mini-capítulo 2 aspectos relevantes a la hora de la construcción de un modelo cualquiera.

Mini-capítulo 1

Tipos de modelos

En este mini-capítulo se enumeran los principales modelos del análisis supervisado según la naturaleza de la variable respuesta. Estos modelos son:

- Y dicotómica: regresión logística / análisis discriminante
- Y continua: regresión lineal.
- Y nominal: regresión logística politómica / análisis discriminante politómico.
- Y ordinal: regresión de riesgos proporcionales ("odds proportional regression").

Conviene mencionar que en muchos textos del ámbito de la minería de datos (ver, por ejemplo, el capítulo 2 de [2]) se suelen enumerar los principales modelos del análisis supervisado según la dicotomía:

- Regresión: en el caso de que la variable repuesta sea continua.
- Clasificación: en el caso de que la variable respuesta sea nominal.

Esta clasificación tiene un inconveniente pedagógico frente a la clasificación anterior y es que no se sabría muy bien, por ejemplo, dónde tendría cabida la regresión logística pues aunque lleva en el nombre la palabra regresión, sin embargo sirve para clasificar, ya que la variable respuesta es nominal. Este hecho se menciona en [2], donde se llega a decir "However, the distinction is not always that crisp".

Mini-capítulo 2

Construcción del modelo

La mayoría de los modelos establecen una relación de Y en términos de $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ a través de una ecuación del tipo

$$Y = f(X_1, X_2, \cdots, X_p) + \mathcal{E} = f(\boldsymbol{X}) + \mathcal{E}$$

donde f es una función fija pero desconocida, que representa la información sistemática de las variables explicativas, y $\mathcal{E}$ es una v.a. independiente de $\boldsymbol{X}$ con $E[\mathcal{E}] = 0$, que representa una perturbación de la información sistemática de X. Evidentemente, el objetivo es determinar $\widehat{f}$ para estimar f y poder predecir Y a través de $\widehat{Y} = \widehat{f}(\boldsymbol{X})$.

Una medida del error que se comete cuando se utiliza el modelo estimado es

$$E[Y - \widehat{Y}]^2.$$

Esta medida del error se puede descomponer de forma que aparezcan dos sumandos según se detalla a continuación:

$$\begin{aligned} E[Y - \widehat{Y}]^2 &= E[f(\boldsymbol{X}) + \mathcal{E} - \widehat{f}(\boldsymbol{X})]^2 = E[(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + \mathcal{E}^2 + 2(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))\mathcal{E}] \\ &= (f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + E[\mathcal{E}^2 + 2(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))\mathcal{E}] = (f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + E[\mathcal{E}^2] \\ &= (f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + V[\mathcal{E}] \end{aligned}$$

Algunos autores (ver, por ejemplo, el capítulo 2 de [2]) denominan error reducible al primer sumando de dicha descomposición y error irreducible al segundo, ya que aunque se llegue a estimar f de forma perfecta, siempre existará el error debido a la perturbación $\mathcal{E}$.

2.1. Codificación de las variables explicativas

La codificación de las variables explicativas para que formen parte de los modelos de este mini-libro se hace de la forma siguiente:

- Si son continuas no es necesario hacer ninguna codificación. Se introducen en el modelo directamente con sus valores numéricos.
- Si son nominales con k categorías, es necesario crear $k-1$ variables ficticias (variables indicador, auxiliares o "dummy"). Existen varias formas de crear dichas variables. Uno de los métodos es el dado en la tabla 2.1.

Variable Nominal	Dummy_2	Dummy_3	...	Dummy_k
Nivel 1	0	0	...	0
Nivel 2	1	0	...	0
Nivel 3	0	1	...	0
...	...	...	...	...
Nivel k	0	0	...	1

Tabla 2.1: Generación de variables "dummy"

- Si son ordinales, una posibilidad es tratarlas como nominales (se estaría perdiendo información), sin embargo lo más adecuado sería usar transformaciones ortogonales (ver, por ejemplo, [1]). No sería correcto introducirlas como si fueran continuas con valores numéricos asociados a los niveles ordenados.

A menudo también se suelen considerar transformaciones de variables explicativas (por ejemplo transformación cuadrática o logaritmo neperiano) e interacciones entre variables explicativas que se codifican como el producto de éstas.

2.2. Ajuste del modelo

En los modelos es habitual, cuando f es paramétrica, tener que estimar los parámetros o coeficientes del modelo, de forma puntual y a través de sus intervalos de confianza. Los coeficientes del modelo dan una medida del efecto de cada variable, aunque en algunas ocasiones hay que transformarlos, como en la regresión logística que, a la hora de interpretar los coeficientes, hay que considerar e elevado a los coeficientes ya que en ese caso representan los "odds ratio".

La estimación de los parámetros se suele hacer a través de distintos métodos, como el de mínimos cuadrados o máxima verosimilitud. El proceso de cosntrucción de un modelo tiene que seguirse a través de una estrategia de modelización ("model building") en el que bien de forma manual o de forma automática se seleccione un modelo candidato que habrá que evaluar si se comporta de manera razonable. En este proceso hay que dividir la muestra de forma aleatoria en un conjunto de datos de entrenamiento (con el que se construye el modelo candidato) y un conjunto de datos de validación (con el que se evalúa su rendimiento).

En los modelos con variable respuesta continua es habitual utilizar el error cuadrático medio ("mean square error") en el conjunto de entrenamiento como medida del rendimiento del modelo. Cuando la variable respuesta es dicotómica se utilizan la sensibilidad, la especificidad y la curva *roc*. En el caso de que la respuesta sea nominal, se utiliza la matriz de confusión.

En los modelos con variable respuesta continua, la fórmula siguiente

$$E[y_0 - \widehat{f}(\boldsymbol{x_0})]^2 = V[\widehat{f}(\boldsymbol{x_0})] + (sesgo(\widehat{f}(\boldsymbol{x_0})))^2 + V[\mathcal{E}],$$

expresa el error cuadrático medio esperado en el conjunto de validación, para una observación $(y_0, \boldsymbol{x_0})$ de dicho conjunto.

La fórmula anterior está relacionada con el compromiso sesgo-varianza (ver, por ejemplo, el capítulo 2 de [2]). Dicho compromiso refleja el hecho de que cuanto más flexible sea el modelo, la varianza será menor pero el sesgo será mayor y recíprocamente para modelos menos flexibles.

Ejemplo

1. Demostrar la fórmula del error cuadrático medio esperado en el conjunto de validación.

   SOLUCIÓN:

   En la expresión $E[y_0 - \widehat{f}(\boldsymbol{x_0})]^2$ se suma y se resta dentro de ella el valor $E[\widehat{f}(\boldsymbol{x_0})]$ y se opera, es decir

$$\begin{aligned}
E[y_0 - \widehat{f}(\boldsymbol{x_0})]^2 &= E[y_0 - E[\widehat{f}(\boldsymbol{x_0})] + E[\widehat{f}(\boldsymbol{x_0})] - \widehat{f}(\boldsymbol{x_0})]^2 \\
&= E[(y_0 - E[\widehat{f}(\boldsymbol{x_0})])^2] + E[(E[\widehat{f}(\boldsymbol{x_0})] - \widehat{f}(\boldsymbol{x_0}))^2] \\
&+ 2E[(y_0 - E[\widehat{f}(\boldsymbol{x_0})])(E[\widehat{f}(\boldsymbol{x_0})] - \widehat{f}(\boldsymbol{x_0}))] \\
&= E[y_0^2] + E^2[\widehat{f}(\boldsymbol{x_0})] - 2E[y_0]E[\widehat{f}(\boldsymbol{x_0})] \\
&+ E^2[\widehat{f}(\boldsymbol{x_0})] + E[\widehat{f}^2(\boldsymbol{x_0})] - 2E[\widehat{f}(\boldsymbol{x_0})]E[\widehat{f}(\boldsymbol{x_0})] \\
&+ 2E[y_0 E[\widehat{f}(\boldsymbol{x_0})] - y_0\widehat{f}(\boldsymbol{x_0}) - E^2[\widehat{f}(\boldsymbol{x_0})] + \widehat{f}(\boldsymbol{x_0})E[\widehat{f}(\boldsymbol{x_0})]] \\
&= E[f^2(\boldsymbol{x_0})] + V[\mathcal{E}] + E^2[\widehat{f}(\boldsymbol{x_0})] - 2E[f(\boldsymbol{x_0})]E[\widehat{f}(\boldsymbol{x_0})] \\
&+ E^2[\widehat{f}(\boldsymbol{x_0})] + E[\widehat{f}^2(\boldsymbol{x_0})] - 2E^2[\widehat{f}(\boldsymbol{x_0})] + 0 \\
&= f^2(\boldsymbol{x_0}) + V[\mathcal{E}] - 2f(\boldsymbol{x_0})E[\widehat{f}(\boldsymbol{x_0})] + E[\hat{f}^2(\boldsymbol{x_0})] \\
&= f^2(\boldsymbol{x_0}) - 2f(\boldsymbol{x_0})E[\widehat{f}(\boldsymbol{x_0})] + V[\widehat{f}(\boldsymbol{x_0})] + E^2[\widehat{f}(\boldsymbol{x_0})] + V[\mathcal{E}] \\
&= V[\widehat{f}(\boldsymbol{x_0})] + (E[\widehat{f}(\boldsymbol{x_0})] - f(\boldsymbol{x_0}))^2 + V[\mathcal{E}] \\
&= V[\widehat{f}(\boldsymbol{x_0})] + (sesgo(\widehat{f}(\boldsymbol{x_0})))^2 + V[\mathcal{E}]
\end{aligned}$$

Referencias

[1] Hosmer Jr, D.W., Lemeshow, S., & Sturdivant, R.X., *Applied logistic regression*, John Wiley & Sons (New York), 2013.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R., *An introduction to statistical learning*, Springer (New York), 2013.

[3] Ramos, E., Vélez, R., Hernández, V., *Modelos probabilísticos y optimización*, Sanz y Torres (Madrid), 2019.

Acerca de

Emilio Letón

Figura 2.1: Emilio Letón

Emilio Letón nace en Madrid en 1966. Es Licenciado en Matemáticas por la UCM en 1989 y doctor en Matemáticas en la misma universidad en 2002. En la actualidad es profesor contratado doctor de la UNED en el departamento de Inteligencia Artificial, al que se incorporó en 2009. Anteriormente fue profesor del departamento de Estadística de la UC3M durante 5 años. Asimismo, ha trabajado durante 15 años en departamentos de Planificación y Estadística dentro del sector bancario y de la industria farmacéutica. Sus líneas de investigación incluyen el Análisis de Supervivencia, tests no paramétricos, PLS, Meta-Análisis, Bioestadística y B-Learning. Ha participado en más de 30 proyectos de innovación docente (siendo coordinador en más de 10 de ellos) colaborando con distintas universidades: UNED, UC3M, UCM y UPM. Ha recibido 1 premio en excelencia en publicaciones científicas (UC3M) y 5 premios en excelencia docente (1 en UC3M, 3 en UNED y 1 en OCW Consortium). En @emilioleton se pueden encontrar sus tweets y su página web personal con información ampliada de su curriculum. https://twitter.com/emilioleton

Elisa M. Molanes-López

Figura 2.2: Elisa M. Molanes-López

Elisa M. Molanes López nace en Vigo en 1976. Se licencia en Matemáticas por la Universidad de Santiago de Compostela en 2000 y consigue el grado de doctora, con acreditación europea, por la Universidad de A Coruña en 2007. Entre 2003 y 2007, al amparo de una beca FPI, realiza varias estancias de investigación

en las siguientes universidades extranjeras: Universiteit Hasselt (Bélgica), Université Catholique de Louvain (Bélgica) y The University of Texas (EE.UU.). Durante 8 años, entre 2007 y 2015, es profesora visitante en el departamento de Estadística de la Universidad Carlos III de Madrid. Actualmente, desde octubre de 2015, es profesora ayudante doctora en la Unidad Departamental de Bioestadística del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid. Durante su etapa docente ha participado en 10 proyectos de innovación docente. En diciembre de 2013, recibe un accésit a la mejor práctica docente en los Premios del Consejo Social de la UNED por su participación en el MOOC "Mini-vídeos docentes modulares: un elemento crítico en el diseño de un MOOC", y en abril de 2014, el OpenCourseWare Consortium le otorga un premio de excelencia por su participación en el curso OCW "Mini-vídeos docentes modulares para diseñar un MOOC". Sus líneas de investigación incluyen la estadística no paramétrica, el análisis de supervivencia, las curvas ROC y las funciones cópula. En la página web personal de Elisa M. Molanes-López, se puede encontrar información más detallada. https://elisamariamolanes.wixsite.com/emolanes

Índice general

Índice de figuras

Índice de tablas