Seminario de Proyectos I 2022-1

UNIDAD 5

Definición y construcción del modelo analítico

Por José Florentino Chavira Sánchez

13 de marzo de 2022

Antecedentes

En la actualidad, los desarrollos tecnológicos en el campo de procesamiento de imágenes digitales son ampliamente utilizados en muchos ámbitos laborables. Así, existen herramientas de edición, tales como Photoshop, Freehand, etc., que permiten alterar o modificar una imagen; con ello, se puede ocultar fácilmente alguna información significativa o útil para hacer imágenes falsificadas digitalmente, las cuales son difíciles de reconocer su alteración a simple vista. Para dar solución a ello, se pueden analizar las imágenes digitales utilizando herramientas de dibujo forense como FotoForensic, Ghiro, Forensically y Jpegsnoop. Sin embargo, existen otras técnicas como es el Análisis de Niveles de Error (ELA), los Metadatos y las técnicas de Compresión JPEG para verificar la autenticidad e integridad de la imagen digital. Todo ello se ha convertido en algo importante, especialmente cuando las imágenes juegan un papel importante como fuente de información o evidencia en diferentes organismos sociales: tribunales, documentos financieros, uso médico, sector del transporte, etc

Aprendizaje Supervisado

Para entender el aprendizaje supervisado de forma intuitiva usaremos un ejemplo cotidiano. Todos hemos ido al doctor y alguna vez le hemos dicho que nos duele la garganta, que hemos tenido dolor de cabeza y fiebre. Éste nos hará unas cuantas preguntas más y luego nos dirá qué enfermedad podríamos tener y qué tratamiento seguir.

Intuitivamente sabemos que el doctor tuvo que **entrenar** inicialmente a partir de clases y libros donde muestran casos pasados, y estudiar qué síntomas son señal de cada enfermedad. Luego, empezó a **testear** lo aprendido en un grupo de pacientes durante su internado. Finalmente, cuando ya estaba entrenado tuvo licencia para poder **aplicar** este aprendizaje a pacientes en su consultorio u hospital.

Este es un ejemplo de **aprendizaje supervisado** porque el **entrenamiento** se realizó a partir de datos conocidos o *inputs* los cuales están etiquetados (duele la garganta, dolor de cabeza, fiebre) con la finalidad de obtener un resultado o *output* que también era conocido y etiquetado (¿tiene gripe o no?). Cuando un doctor **testea** lo aprendido se sabe los inputs de pacientes y también el output que es dado por un doctor con más experiencia que puede decir qué tan efectivo es su entrenamiento. Cuando el doctor sale a atender pacientes solo tendrá *inputs etiquetados* con la finalidad de **predecir** un *output etiquetado*.

Algoritmos:

- a. Árboles de decisión

El árbol de decisiones **es uno de esos ejemplos de herramientas que facilitan la toma de decisiones**. En base a un diagrama de flujo se visualiza el proceso de toma de decisiones mediante el mapeo de diferentes cursos de acción, así como sus posibles resultados.

Por muy diferentes que sean los **propósitos que motivan la creación del árbol de decisiones.**

Cómo se crea un árbol de decisiones: ejemplos de pasos a seguir para hacerlo

- **Comenzar estableciendo el objetivo general** en la parte superior (raíz). Representa la decisión que se está intentando tomar.
- Dibujar las flechas para cada curso de acción posible. Estas flechas salen de la raíz y deben hacer referencia a los costes asociados con cada acción, así como la probabilidad de éxito.
- Incluir nodos de hoja al final de las ramas. **¿Cuáles son los resultados de cada curso de acción?** Si se debe tomar otra decisión, se dibuja un nodo de hoja cuadrada. Si el resultado es incierto, se dibuja un nodo de hoja circular.
- Determinar las probabilidades de éxito de cada punto de decisión. Al crear un árbol de decisión, es importante investigar, para poder predecir con precisión la probabilidad de éxito.
- Evaluar riesgo vs recompensa. Calcular el valor esperado de cada decisión en el árbol ayuda a minimizar el riesgo y aumentar la probabilidad de alcanzar un resultado favorable.

b. Clasificación de Naïve Bayes

Naïve Bayes (NB), Bayesiano ingenuo o el Ingenuo Bayes es uno de los algoritmos más simples, pero potentes, para la clasificación basado en el Teorema de Bayes con una suposición de independencia entre los predictores. Naive Bayes es fácil de construir y particularmente útil para conjuntos de datos muy grandes. El clasificador Naive Bayes asume que el efecto de una característica particular en una clase es independiente de otras características.

- c. Regresión por mínimos cuadrados

Es un procedimiento de análisis numérico en la que, dados un conjunto de datos (pares ordenados y familia de funciones), se intenta determinar la función continua que mejor se aproxime a los datos (línea de regresión o la línea de mejor ajuste), proporcionando una demostración visual de la relación entre los puntos de los mismos. En su forma más simple, busca minimizar la suma de cuadrados de las diferencias ordenadas (llamadas residuos) entre los puntos generados por la función y los correspondientes datos.

Su expresión general se basa en la **ecuación de una recta y = mx + b**. Donde m es la pendiente y b el punto de corte, y vienen expresadas de la siguiente manera:

$$m = \frac{n \cdot \Sigma(x \cdot y) - \Sigma x \cdot \Sigma y}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

$$b = \frac{\Sigma y \cdot \Sigma x^2 - \Sigma x \cdot \Sigma(x \cdot y)}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

1. Encontrar la recta que mejor se ajusta a la gráfica de los datos proporcionados.

2. Encontrar una recta y = mx + b. Aplicar el método de mínimos cuadrados, entonces, primero encontrar el valor (x · y).

3.- Encontrar el valor x^2.

4. Obtener los valores de las sumatorias de cada columna:

$$\sum x \quad ; \quad \sum y \quad ; \quad \sum (x \cdot y) \quad ; \quad \sum x^2 \quad ; \quad n =$$

5. Sustituir en cada una de las expresiones: m y b

$$m = \frac{n \cdot \Sigma(x \cdot y) - \Sigma x \cdot \Sigma y}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

$$b = \frac{\Sigma y \cdot \Sigma x^2 - \Sigma x \cdot \Sigma(x \cdot y)}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

6. La recta obtenida con el método de los mínimos cuadrados es la siguiente: y=mx+b

7. Graficar

d. Regresión logística

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo. La regresión logística se puede aplicar a un rango más amplio de situaciones de investigación que el análisis discriminante.

Consideraciones sobre datos de regresión logística
Datos. La variable dependiente debe ser dicotómica. Las variables independientes pueden estar a nivel de intervalo o ser categóricas; si son categóricas, deben ser variables auxiliares o estar codificadas como indicadores (existe una opción en el procedimiento para recodificar automáticamente las variables categóricas).

Supuestos. La regresión logística no se basa en supuestos distribucionales en el mismo sentido en que lo hace el análisis discriminante. Sin embargo, la solución puede ser más estable si los predictores tienen una distribución normal multivariante. Adicionalmente, al igual que con otras formas de regresión, la multicolinealidad entre los predictores puede llevar a estimaciones sesgadas y a errores estándar inflados. El procedimiento es más eficaz cuando la pertenencia a grupos es una variable categórica auténtica; si la pertenencia al grupo se basa en valores de una variable continua (por ejemplo “CI alto ” en contraposición a “CI bajo”), deberá considerar el utilizar la regresión lineal para aprovechar la información mucho más rica ofrecida por la propia variable continua.

Procedimientos relacionados. Utilice el procedimiento Diagrama de dispersión para mostrar en pantalla sus datos para multicolinealidad. Si se cumplen los supuestos de normalidad multivariante y de matrices de varianzas-covarianzas iguales, puede obtener una solución más rápida utilizando el procedimiento Análisis discriminante. Si todos los predictores son categóricos, puede además utilizar el procedimiento Loglineal. Si la variable dependiente es continua, utilice el procedimiento Regresión lineal. Puede utilizar el procedimiento Curva ROC para realizar gráficos de las probabilidades guardadas con el procedimiento Regresión logística.

Aprendizaje No supervisado

Mientras que en el aprendizaje **supervisado** tenemos un conjunto de variables que usamos para predecir una determinada clase de salida (sube/baja, renuncia/no renuncia), en el aprendizaje **no supervisado** no tenemos clases de salida esperadas. En el aprendizaje supervisado teníamos data de entrenamiento y data de testeo que nos permitía validar la efectividad del modelo por la cercanía a la clase conocida. En el aprendizaje no supervisado no tenemos *output* predeterminado. Esto genera a su vez un gran reto porque es muy difícil saber si ya culminamos con el trabajo o podemos aun generar otro modelo con el que nos sintamos más satisfechos.

El ejemplo más sencillo para entender este tipo de aprendizaje es cuando tenemos nuestra base de clientes y queremos segmentarlos por primera vez. En ese caso buscamos clientes que se comporten de la misma forma, pero al ser la primera vez no sabemos cuántos segmentos podemos tener. El reto está en determinar el corte de ¿cuántos segmentos buscamos crear?.

Las principales aplicaciones del aprendizaje no supervisado están relacionadas en el agrupamiento o ***clustering*** de datos. Aquí, el objetivo es encontrar subgrupos homogéneos dentro de los datos. Estos alogoritmos se basan en la distancia entre observaciones. El ejemplo de la segmentación de clientes sería un ejemplo de *clustering*.

Los algoritmos más utilizados de agrupamiento son: agrupamiento por k-medias y agrupamiento jerárquico.

a. Algoritmos de Clustering

Un algoritmo de **clustering** tiene como objetivo agrupar los objetos de un dataset según su **similaridad**, de forma que los objetos que hay dentro

de un grupo (**cluster**) sean más similares que aquellos que caen en grupos distintos.

Desde un punto de vista intuitivo, este problema tiene un objetivo muy claro: agrupar adecuadamente un conjunto de datos no etiquetados. A pesar de su intuición, la noción de "clúster/agrupamiento" no puede ser definido con precisión, una de las casusas por las que se ha propuesto un rango tan amplio de algoritmos de clustering.

b. Análisis de componentes principales

Principal Component Analysis (PCA) es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables (X_1, X_2, ..., X_p), es decir, el espacio muestral tiene p dimensiones. *PCA* permite encontrar un número de factores subyacentes $(z<p)$ que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal.

Principal Component Analysis pertenece a la familia de técnicas conocida como *unsupervised learning*. Los métodos de *supervised learning* descritos en capítulos anteriores tienen el objetivo de predecir una variable respuesta Y a partir de una serie de predictores. Para ello, se dispone de *p* características (X_1, X_2 ... X_p) y de la variable respuesta Y medidas en *n* observaciones. En el caso de *unsupervised learning*, la variable respuesta Y no se tiene en cuenta ya que el objetivo no es predecir Y sino extraer información empleando los predictores, por ejemplo, para identificar subgrupos. El principal problema al que se enfrentan los métodos de *unsupervised learning* es la dificultad para validar los resultados dado que no se dispone de una variable respuesta que permita contrastarlos.

El método de *PCA* permite por lo tanto “condensar” la información aportada por múltiples variables en solo unas pocas componentes. Esto

lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, *clustering*... Aun así no hay que olvidar que sigue siendo necesario disponer del valor de las variables originales para calcular las componentes.

A partir de la investigación, elabora una metodología que incluya materiales y métodos para tu proyecto (incorporar en formato de registro o en la plantilla de artículo científico).

*Nota: Aquí no supe qué hacer.

[1] Daniel Paredes Inilupu. (2022). Data Science con R. 13 marzo de 2020, de Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional. Sitio web: https://bookdown.org/dparedesi/data-science-con-r/aprendizaje-supervisado.html

[2] Juan Francisco Pérez Herrera. (6 nov 2020). LEAN CONTRUCCIÓN MÉXICO. 13 marzo 2020, de LCM Sitio web: https://www.leanconstructionmexico.com.mx/post/%C3%A1rbol-de-decisiones-ejemplos-de-ventajas-y-pasos-a-seguir

[3] Miprofe. (xxx). Mínimos cuadrados. 13 marzo 2022, de miprofe.com Sitio web: https://miprofe.com/minimos-cuadrados/

[4] IBM. (1989,2021). Regresión Logística. 13 marzo 2022, de Copyright IBM Corporation 1989, 2021 Sitio web: https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=regression-logistic

[5] Fernando Sancho Caparrini. (2009). Algoritmos de Clustering. 13 marzo 2022, Sitio web: http://www.cs.us.es/~fsancho/?e=230

[6] Joaquín Amat Rodrigo. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. 13 marzo de 2022, de Joaquín Amat Rodrigo Sitio web: https://www.cienciadedatos.net/documentos/35_principal_component_analysis