Fase 2 - Aprendizaje supervisado

O. Andrés Russi Porras.
oarussir@unadvirtual.edu.co

Universidad Nacional Abierta y a Distancia

Resumen

Este trabajo explora técnicas fundamentales de machine learning a través de cuatro actividades principales. En la primera, se aborda la regresión lineal y logística mediante la creación de un glosario, análisis de gráficos para identificar variables y modelos matemáticos, e interpretación de resultados. La segunda actividad se centra en el descenso del gradiente, explicando su concepto y aplicando el método tanto analíticamente como a través de iteraciones para encontrar mínimos de funciones. La tercera actividad introduce los árboles de decisión, pidiendo diseñar uno a partir de datos dados y discutir su clasificación y construcción. Finalmente, la cuarta actividad examina el método del vecino más cercano (KNN), solicitando un glosario, descripción de métricas para clases numéricas y categóricas, y aplicación práctica en un problema específico de clasificación. Este conjunto de actividades está diseñado para afianzar el entendimiento y la aplicación práctica de conceptos clave en el campo del machine learning.

Palabras clave: Aprendizaje automático, regresión, gradiente, árboles de decisión, KNN.

1. Regresión Lineal y Logística.

1.1. Realice un glosario con los siguientes conceptos:

1.1.1. variable dependiente:

Concepto utilizado en investigación, estadística y matemáticas para referirse a la variable que se estudia para observar cómo es afectada por otras

variables, denominadas variables independientes. Permite medir el efecto de una o más variables independientes en un experimento o modelo. Por ejemplo, en un estudio para determinar el efecto de las horas de estudio (variable independiente) en las calificaciones de un examen (variable dependiente), la variable dependiente sería las calificaciones del examen, ya que estas dependen de la cantidad de horas dedicadas al estudio. La variable dependiente es lo que el investigador espera que cambie como resultado de la manipulación de la variable independiente. [Gru, pag 107-108]

1.1.2. Variable independiente

Es la variable que se manipula o cambia para observar su efecto en la variable dependiente. Es la causa presumida en una relación causa-efecto y se utiliza para determinar si tiene un efecto significativo sobre la variable dependiente. Es comúnmente ajustada por el investigador para ver cómo afecta a una variable de salida (la variable dependiente). Por ejemplo, un experimento diseñado para estudiar el efecto de la cantidad de fertilizante en el crecimiento de plantas de tomate. En este caso, la variable independiente sería la cantidad de fertilizante aplicado a las plantas. Los investigadores variarán esta cantidad para observar cómo afecta al crecimiento de las plantas, medido en términos de altura o biomasa, por ejemplo. Así, la variable dependiente, el crecimiento de las plantas de tomate, se espera que cambie en respuesta a las diferentes cantidades de fertilizante aplicadas. [Gru, pag 107-108]

1.1.3. Descenso del gradiente

Es un algoritmo de optimización utilizado para minimizar alguna función encontrando iterativamente el valor de los parámetros de dicha función que resultan en el mínimo valor posible de la misma. Este método es ampliamente usado en el aprendizaje automático y la inteligencia artificial, especialmente para el entrenamiento de modelos de regresión y redes neuronales.

La idea básica detrás del descenso del gradiente es ajustar los parámetros de la función de manera iterativa, moviéndose en la dirección opuesta al gradiente (o derivada) de la función de costo con respecto a los parámetros en el punto actual, porque esta dirección es la que indica la mayor disminución de la función.

El proceso comienza con valores iniciales para los parámetros y se actualizan repetidamente con la siguiente fórmula:

$$\theta_{nuevo} = \theta_{anterior} - \alpha \nabla J(\theta) \tag{1}$$

donde:
θ son los parámetros de la función
α es la tasa de aprendizaje, un hiperparámetro que controla el tamaño del paso en cada actualización.
$\nabla J(\theta)$ es el gradiente de la función de costo J con respecto a los parámetros θ.

El gradiente indica la dirección en la cual la función aumenta. Por lo tanto, restamos este valor para ajustar los parámetros en la dirección opuesta, es decir, hacia donde la función decrece más rápidamente.

La tasa de aprendizaje determina la magnitud del ajuste a los parámetros en cada iteración. Un valor menor implica que el algoritmo será más preciso, pero a costa de un mayor consumo de tiempo y recursos computacionales. Por otro lado, una tasa de aprendizaje alta puede acelerar el algoritmo, pero con el riesgo de sacrificar precisión y posiblemente sobrepasar el mínimo de la función de costo.

Ejemplo: supongamos que queremos encontrar el punto mínimo de la función $f(x, y) = (x + 3)^2 + (y - 2)^2$ usando el método del descenso de la gradiente. Utilixando una tasa de aprendizaje de 0.1 y valores iniciales de $x_0 = 0$ y $y_0 = 0$, sacamos la gradiente de la función aplicando la derivada parcial sobre cada variable independiente, en este proceso nos daría: $\nabla f(x, y) = (2x + 6, 2y - 4)$ El proceso para la primera iteración sería el siguiente:
$x_1 = 0 - 0{,}1 \cdot (2(0) + 6)$
$x_1 = -0{,}6$
$y_1 = 0 - 0{,}1 \cdot (2(0) - 4)$
$y_1 = 0{,}4$
usamos los valores de x_1 y y_1 para obtener x_2 y y_2 en la segunda iteración:
$x_2 = -0{,}6 - 0{,}1 \cdot (2(-0{,}6) + 6)$
$x_2 = -1{,}08$
$y_2 = 0{,}4 - 0{,}1 \cdot (2(0{,}4) - 4)$
$y_2 = 0{,}72$
Seguimos repitiendo el proceso para acercarnos a la solución. En la iteración 13 tendríamos:

$x_{13} = -2{,}793841569792$ y $y_{13} = 1{,}8900488372224$
Y ya para la iteración 25 tendríamos:
$x_{25} \approx -2{,}99093$ y $y_{25} \approx 1{,}993955$
Muy cercanos a la solución correcta que es $x = -3$ y $y = 2$ [Gru, pag 140-148]

1.1.4. Función de pérdida

También conocida como función de costo, es el componente que mide el grado de discrepancia entre la predicción del modelo y los valores reales de los datos. En otras palabras, cuantifica el error que un modelo de aprendizaje automático comete en sus predicciones. El objetivo principal de un algoritmo de aprendizaje es minimizar esta función de pérdida, lo cual indica que el modelo está mejorando su precisión y reduciendo la diferencia entre las predicciones y los valores reales.

Existen diferentes tipos de funciones de pérdida, cada una adecuada para tipos específicos de tareas de modelado. Dos de las funciones de pérdida más comunes son:

Error Cuadrático Medio (MSE, por sus siglas en inglés): Utilizado principalmente en tareas de regresión. Calcula el promedio de los cuadrados de las diferencias entre los valores predichos y los reales. La fórmula es la siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

Donde:
MSE es el Error Cuadrático Medio
n es el número total de observaciones
y_i es el valor real de la i-ésima observación
$\hat{y}_i$ es la predicción del modelo para la i-ésima observación

Error Absoluto Medio (MAE, por sus siglas en inglés): También usado en regresión. Calcula el promedio de las diferencias absolutas entre las predicciones y los valores reales, proporcionando una medida robusta a los valores atípicos en comparación con MSE. La fórmula es la siguiente:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

donde:
MAE es el Error Absoluto Medio
n es el número total de observaciones
y_i representa el valor real de la i-ésima observación
$\hat{y}_i$ denota la predicción del modelo para la i-ésima observación
[Ribe, pag 3-15]

1.2. PREGUNTAS GRAFICA 1

1.2.1. Identifique la variable dependiente y la variable independiente. ¿Cuál es el propósito de la regresión lineal en este caso específico?

En el gráfico la variable independiente es la latitud y la variable dependiente es la temperatura en enero. La regresión lineal de la gráfica nos muestra el efecto del cambio de latitud en las temperaturas en el mes de enero.

1.2.2. ¿Cuál es el modelo o representación matemática que se obtuvo?

La función lineal obtenida fue la siguiente:

$$y = -2{,}11x + 109 \tag{4}$$

Se trata de una función lineal decresciente.

1.2.3. Interprete la pendiente

La pendiente de la función es -2.11, esto indica que la función es decreciente a medida que la variable independiente crece. En este caso, un mayor valor de la latitud implicará un menor valores de la temperatura. El 2.11 nos indica que por cada cambio de una unidad en la latitud, la temperatura varia en 2.11 unidades.

1.2.4. Mencione dos métodos para calcular las estimaciones de los parámetros del modelo.

Método de Mínimos Cuadrados (Ordinary Least Squares - OLS): El método de mínimos cuadrados es el enfoque más utilizado para estimar los parámetros de un modelo lineal. Este método busca minimizar la suma

de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. En términos matemáticos, si tenemos un conjunto de n observaciones (x_i, y_i) el método de mínimos cuadrados busca encontrar los valores de los parámetros m y b que minimizan la función de coste:

$S = \sum_{i=1}^{n}(y_i - (mx_i + b))^2$

En los mínimos cuadrados, la pendiente se calcula de la siguiente forma:

$m = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$

Y la intersección se calcula de la siguiente forma:

$b = \frac{\sum y_i - m\sum x_i}{n}$

Método de máxima verosimilitud: es otro enfoque popular para estimar los parámetros de un modelo estadístico. Este método se basa en encontrar los valores de los parámetros del modelo que maximizan la probabilidad (o verosimilitud) de observar los datos dados. Para un modelo lineal, el MLE asume que los residuos (las diferencias entre los valores observados y los predichos) siguen una distribución normal y busca maximizar la función de verosimilitud asociada a estos residuos.

Aunque el MLE es más complejo computacionalmente que los mínimos cuadrados y puede requerir el uso de algoritmos numéricos para encontrar la solución, ofrece algunas ventajas, como una mayor flexibilidad en la modelación de los errores y la posibilidad de incorporar explícitamente supuestos sobre la distribución de los residuos.

1.2.5. Cuál es la diferencia entre un modelo lineal simple y un modelo lineal múltiple

La diferencia principal entre un modelo lineal simple y un modelo lineal múltiple radica en el número de variables independientes (predictores) que cada uno utiliza para predecir la variable dependiente.

Modelo lineal simple: utiliza una sola variable independiente para predecir la variable dependiente. Es decir, trata de establecer una relación lineal entre dos variables La ecuación de un modelo lineal simple se puede representar como $y = mx + b$, donde y es el valor de la variable dependiente, m es la pendiente que indica cuánto cambia y por un cambio de una unidad en x, b es la intersección que indica el valor de y cuando x es 0, y x es el valor de la variable independiente. [Weis, pag 19-20]

Modelo Lineal Múltiple: Incorpora dos o más variables independientes para predecir la variable dependiente. Esto permite examinar cómo múltiples

factores influyen simultáneamente en una variable de interés. La ecuación de un modelo lineal múltiple se puede expresar como:

$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$

Donde:

y es la variable dependiente

$x_1, x_2, ..., x_n$ son las variables independientes b_0 es la intercepción de indica el valor de y cuando todas las variable independientes son 0

$b_1, b_2, ..., b_n$ son los coeficientes de las variables independientes, indican cuantos cambia el valor dependiente cuando la variable dependiente cambia una unidad. [Weis, pag 47-48]

1.2.6. ¿Cuál es la función de pérdida que se intenta minimizar? Expréselo en sus palabras (ayuda: ECM)

El Error Cuadrático Medio (ECM) es una función de pérdida, esto es una función que mida la diferencia entre el valor real y el valor predecido por un modelo. El ECM se calcula usando el promedio de los cuadrados de los errores o diferencias entre los valores predichos por el modelo y los valores reales observados. Al elevar al cuadrado cada uno de estos errores antes de promediarlos, el ECM da un mayor peso a los errores más grandes, lo que hace que el modelo sea particularmente sensible a los outliers o valores atípicos en los datos. La fórmula para calcular el ECM es:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

Donde ECM es el Error Cuadrático Medio, n es el número total de observaciones, y_i es el valor real de la i-ésima observación, $\hat{y}_i$ es la predicción del modelo para la i-ésima observación. Por ejemplo, teniendo los siguientes datos reales:

x	-2	-1	0	1	2
y	-2	0	-1	2	4

Y además tenemos el siguiente modelo para los datos:

$$y = \frac{3x}{2} + 1 \tag{6}$$

lo que nos daría los siguiente predicciones del modelo:

x	-2	-1	0	1	2
y	-2	-0.5	1	2.5	4

Entonces el ECM sería:
$ECM = \frac{1}{5}((-2+2)^2 + (-0{,}5-0)^2 + (1+1)^2 + (2{,}5-2)^2 + (4-4)^2)$
$ECM = \frac{1}{5}(0 + 0{,}25 + 4 + 0{,}25 + 0)$
$ECM = 0{,}9$

1.3. PREGUNTAS GRÁFICA 2

1.3.1. ¿Por qué un modelo de regresión lineal simple no es una herramienta apropiada para describir la relación entre edad del paciente y la presencia de infección?

La relación entre la edad del paciente y la presencia de infección es mejor descrita por una función de ajuste logística, lo que indica una relación no lineal entre estas dos variables. En el contexto de la presencia de infección, esto sugiere que la probabilidad de infección no aumenta o disminuye de manera constante con la edad. En cambio, la función logística implica que puede haber un umbral de edad a partir del cual la probabilidad de infección aumenta o disminuye significativamente, lo cual es característico de fenómenos binarios o de ocurrencias que tienen una probabilidad que varía de forma no uniforme con respecto a una variable independiente.

1.3.2. ¿Cuándo se debe utilizar regresión lineal y cuando regresión logística?

La regresión lineal se utiliza cuando la variable dependiente es continua, es decir, puede tomar cualquier valor dentro de un rango. Ejemplos incluyen el precio de una casa, la temperatura, el salario, etc. Se asume que existe una relación lineal entre la(s) variable(s) independiente(s) y la variable dependiente. Esto significa que un cambio en una variable independiente se asocia con un cambio proporcional en la variable dependiente. La regresión logística es apropiada cuando la variable dependiente es categórica, es decir, se divide en dos categorías (binaria) como "sí/no", "éxito/fracaso", o en múltiples categorías (multinomial) como tipos de especies, clasificaciones, etc. Se utiliza especialmente para estimar la probabilidad de que ocurra un evento dado, en función de una o más variables independientes. [Weis, pag 264-265]

1.3.3. Describa un objetivo del análisis de regresión logística en este caso en particular

La regresión logística es ideal para análisis donde la variable de respuesta es categórica, como en casos de resultados binarios (ejemplo: sí/no, éxito/fracaso, presente/ausente). En el escenario presentado por la gráfica, la función logística es particularmente valiosa porque modela cómo la probabilidad de infección varía con la edad. Matemáticamente, la regresión logística no solo identifica el rango de edades con mayor probabilidad de infección, sino que también revela el punto de inflexión: la edad específica donde la probabilidad de infección aumenta o disminuye más rápidamente. Este punto de inflexión es crítico porque indica un cambio significativo en el riesgo de infección, permitiéndonos entender mejor cómo la edad influye en la susceptibilidad a la infección. Además, el modelo logístico, al utilizar la función logística, facilita una interpretación directa de los coeficientes en términos de odds ratio, lo que proporciona insights valiosos sobre la fuerza y la dirección de la relación entre la edad y la probabilidad de infección.

2. DESCENSO MÁXIMO DE LA GRADIENTE

2.1. ¿Qué es el gradiente de una función?, ¿Recuerda alguna aplicación del gradiente?

Formalmente, el gradiente de f se denota como ∇f y se define como el vector de las derivadas parciales de f con respecto a cada una de sus variables independientes. En términos matemáticos: $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}\right)$ El gradiente apunta en la dirección de mayor incremento de la función, y su magnitud indica qué tan rápido cambia la función en esa dirección. Para funciones de una sola variable, este concepto se simplifica a la derivada de la función en ese punto, que indica la pendiente de la tangente a la curva de la función en ese punto. La pendiente de esta tangente es efectivamente la tasa de cambio más grande en ese punto, ya que solo hay una dirección en la que puedes moverte a lo largo de la línea (hacia adelante o hacia atrás a lo largo del eje x).

Para funciones de múltiples variables, la interpretación se expande a más dimensiones. La gradiente es un vector que apunta en la dirección de la

mayor pendiente y su magnitud indica qué tan empinada es la çuesta.[en] esa dirección.

Por ejemplo, para la función:
$f(x, y) = x^2 + y^2$
La gradiente sería:
$\partial f(x, y) = (2x, 2y)$ Entonces, tomado como ejemplo el punto x=3, y=3, tendríamos que la gradiente en dicho punto sería (6,6) esto nos indicaría que la dirección de mayor crecimiento sería aquella en el plano xy que conecta el origen con el punto (6,6). la magnitud de la gradiente en el punto dado, que indica la velocidad a la que crece la función en la dirección de la gradiente, sería:
$\partial\sqrt{(6^2 + 6^2)} = \sqrt{72} \approx 8{,}49$
Esto refleja qué tan rápido está aumentando el valor de la función al moverse desde el punto (3,3) en la dirección del vector gradiente (6,6) en el plano xy.

La gradiente tiene las siguiente aplicaciónes:

Optimización y machine learning: El gradiente es crucial en métodos de optimización, especialmente en problemas donde se busca encontrar máximos o mínimos de funciones. Al conocer la dirección de mayor crecimiento dada por el gradiente, es posible ajustar las variables para acercarse al máximo o mínimo de una función. Esto se aplica por ejemplo en el descenso de la gradiente, un algoritmo muy utilizado en machine learning y deep learning para minimizar una función de costo o pérdida. El algoritmo actualiza iterativamente los parámetros del modelo en la dirección opuesta al gradiente de la función de costo respecto a los parámetros, con el fin de encontrar los parámetros que minimizan la función.

Ciencias e Ingeniería: En el campo de la ciencia, el gradiente es utilizado para describir campos vectoriales, como el campo eléctrico, donde el gradiente de un potencial eléctrico indica la dirección y magnitud de la fuerza que experimentaría una carga puntual en cualquier punto del espacio. En ingeniería, se utiliza para analizar gradientes de temperatura, presión, y otros gradientes de campos escalares.

Gráficos por Computadora: El gradiente se usa para técnicas de iluminación y sombreado, donde los cambios en la intensidad de la luz sobre un superficie se modelan utilizando el gradiente de funciones que representan la intensidad de la luz, permitiendo efectos visuales más realistas.

Geometría y Topografía: En el análisis del terreno y modelado de superficies, el gradiente puede indicar la pendiente y la orientación del terreno, lo cual es crucial para el diseño de infraestructuras, análisis de riesgos en

deslizamientos de tierras, y otros estudios geológicos.

2.2. Suponga que se tiene una función de costo dada por la siguiente expresión $f(x) = 3x^2 - 7x + 1$

2.2.1. Encuentre el mínimo de la función por el método analítico de las derivadas:

Sacamos la derivada de la función:
$f'(x) = 6x - 7$
Igualamos a cero:
$0 = 6x - 7$
$x = \frac{7}{6}$
Entonces en $\frac{7}{6}$ tenemos un punto crítico, para confirmar que se trata de un mínimo usamos la segunda derivada:
$f''(x) = 6$
La segunda derivada de la función es 6, esto indica que la pendiente de la función crece en todo momento (y de forma constante) a medida que el valor x se hace más grande, por lo cual el punto crítico $\frac{7}{6}$
debe ser un mínimo local.

2.2.2. Encuentre una aproximación del mínimo usando el método máximo del gradiente. Para ello suponga una tasa de aprendizaje $\alpha = 0{,}1$ y valor inicial $x_0 = 0$. Calcule las primeras seis iteraciones. Fórmula recurrente del algoritmo:

$x_{k+1} = x_k - \alpha g(x_k)$
donde $g(x)$ es el gradiente de $f(x)$

Al tratarse de una función con una sola variable independiente la gradiente sería igual a la derivada de la función en este caso:
$f'(x) = 6x - 7 = g(x)$
teniendo que $x_0 = 0$ entonces x_1 sería:
$x_1 = 0 - 0{,}1 \cdot (6(0) - 7)$
$x_1 = 0{,}7$
Repetimos el mismo proceso hasta x_6:
$x_2 = 0{,}7 - 0{,}1 \cdot (6(0{,}7) - 7)$
$x_2 = 0{,}98$
$x_3 = 00{,}98 - 0{,}1 \cdot (6(0{,}98) - 7)$

$x_3 = 1{,}092$
$x_4 = 1{,}092 - 0{,}1 \cdot (6(1{,}092) - 7)$
$x_4 = 1{,}1368$
$x_5 = 1{,}1368 - 0{,}1 \cdot (6(1{,}1368) - 7)$
$x_5 = 1{,}15472$
$x_6 = 1{,}15472 - 0{,}1 \cdot (6(1{,}15472) - 7)$
$x_6 = 1{,}16188$
El valor analíticamente correcto del mínimo, obtenido mediante el cálculo de las derivadas y su análisis es $x = \frac{7}{6} \approx 1{,}166667$ entonces al comparar el valor correcto del mínimo y el obtenido usando el descenso de la gradiente obtenemos el siguiente valor del error absoluto:
Error Absoluto $= |1{,}166667 - 1{,}16188| = 0{,}004779$
Y obtenemos además el siguiente error relativo:
Error Relativo $= \frac{|1{,}166667 - 1{,}16188| \cdot 100}{1{,}166667} \approx 0{,}409628$
Esto nos da un error relativo aproximadamente de $0.41\,\%$.
En resumen, el método del gradiente demostró ser eficiente, acercándose significativamente al mínimo exacto con un error relativo de aproximadamente $0{,}41\,\%$ después de seis iteraciones. Esto subraya la capacidad del método para ofrecer resultados precisos con un número limitado de pasos, proporcionando una solución cercana a la obtenida por el método analítico.

3. Árboles de decisión.

3.1. Realice un glosario con los siguientes conceptos:

3.1.1. Nodo

Un nodo representa un punto dentro del árbol de decisión donde se toma una decisión o se evalúa una condición. Existen diferentes tipos de nodos, incluyendo nodos raíz, nodos internos y nodos hoja. El nodo raíz es el punto de partida del árbol, los nodos internos representan las decisiones basadas en ciertos criterios, y los nodos hoja indican el resultado final o la decisión tomada.

3.1.2. Arcos

Las líneas que enlazan los nodos dentro de un árbol de decisión se conocen como arcos. Estos delinean la trayectoria de una decisión a la siguiente,

fundamentándose en las condiciones o elecciones examinadas en el nodo inicial. Los arcos juegan un papel crucial al definir la ruta desde el nodo raíz hasta los nodos hoja, ilustrando las variadas alternativas o rutas que pueden tomarse durante el proceso de decisión.

3.1.3. Hojas

Los nodos finales en un árbol de decisión, conocidos como hojas o nodos hoja, marcan el punto donde no se necesitan decisiones adicionales. Son el reflejo de las acciones finales o resultados determinados por las elecciones y condiciones establecidas anteriormente en el árbol. Dentro de la estructura de un árbol de decisión, estas hojas simbolizan los diversos resultados posibles alcanzados tras recorrer una ruta específica a través de los nodos y arcos.

3.1.4. Ramas

Las ramificaciones dentro de un árbol de decisión son series de nodos y arcos que enlazan el nodo principal con los nodos terminales. Estas secuencias representan las diversas rutas disponibles a lo largo del árbol, cada una llevando a un desenlace específico derivado de consecutivas decisiones. Dichas ramificaciones reflejan las variadas tácticas o escenarios de decisión que pueden ser explorados a través del proceso analizado por el árbol. Este glosario sirve como fundamento para comprender los elementos esenciales de un árbol de decisión y la manera en que interactúan para crear una estructura empleada en la toma de decisiones y el análisis predictivo. [**?**, pag 95-97]

3.2. Con base en los datos de la tabla 1, diseñe un árbol de decisión (manualmente) para clasificar si se realiza o no una actividad al aire libre. La clasificación tiene en cuenta criterios como el nivel de energía requerido, el tamaño de los grupos y si se requiere equipo.

El siguiente árbol de decisión fue realizado con la herramienta orange:

En el árbol de decisión, se utilizó el método del Índice Gini para determinar la mejor manera de organizar los nodos. Este criterio selecciona las divisiones que resultan en la menor impureza posible, buscando nodos

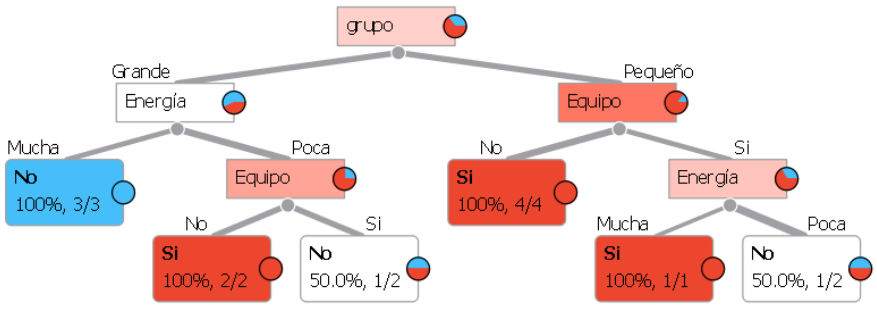

Figura 1: *Arbol de decisión*

con altas tasas de pertenencia de clase homogénea. Para la construcción del árbol se especificó un mínimo de una muestra por nodo, y no se estableció una profundidad máxima del árbol.

3.3. ¿Este árbol es de clasificación o de regresión?, explique

El árbol de decisión creado es de clasificación La razón es que la tarea de predecir la .ªctividad al aire libre.ª partir de tres características implica clasificar la observación en una de varias categorías posibles (en este caso, clasificar el valor de la columa .ªctividad al aire libreçomo sí o no).

Los árboles de clasificación se utilizan cuando el objetivo es predecir a qué categoría (clase) pertenece una observación, basándose en las características de entrada. Por otro lado, los árboles de regresión se usan cuando la variable objetivo es continua y se desea predecir un valor numérico específico, no categorías.

3.4. ¿Con los mismos datos se pueden realizar dos árboles de decisión diferentes? ¿Cómo se decide cuál es el primer nodo?

Es posible generar dos árboles de decisión diferentes utilizando el mismo conjunto de datos. Esto puede ocurrir debido los siguientes factores:

Método de división: Los árboles de decisión pueden usar diferentes criterios para elegir la mejor división en cada nodo. Algunos de los criterios más comunmente usados incluyen: la ganancia de información y el índice Gini. La elección del criterio puede influir en la selección de las características utilizadas para dividir los datos en cada paso, llevando a la creación de árboles diferentes.

• Método gini: este método clasifica las clases por la homogeneid de su distribución, para eso se usa la fórmula:

$1 - \sum_{i=1}^{n} p_i^2$

Donde un valor de 0 indica perfecta homogeneidad, lo que ocurre cuando todos los datos de un nodo pertenecen a la misma clase, mientras que valores más altos indican una mayor heterogeneidad en la distribución de los datos. El objetivo es ordenar los dictintos nodos por su nivel de homogeneidad, empezando por los más homogeneos.

Por ejemplo, supongamos un árbol de decisión donde predecimos la probabilidad de un cierta enfermedad usando dos criterios: adulto y género, ambos de tipo binario. Tenemos que hay dos adultos y dos menores, mientras que en género tenemos 3 mujeres y 1 hombre. El gini de la variable adultos sería:

$1 - ((2/4)^2 + (2/4)^2) = 0{,}5$

mientras que el gini para la opción género sería:

$1 - ((3/4)^2 + (1/4)^2) = 0{,}375$

Esto indica que la homogeneidad de la variable género es mayor (al ser más cercana a 0) y entonces este sería el primer nodo.

• Ganancia de información: para este criterio usamos la fórmula:

$1 - \sum_{i=1}^{n} p_i log(p_i)$

Para el ejemplo anterior, la ganancia de información para la variable adultos sería:

$1 - (\frac{1}{2} \cdot \log_2(\frac{1}{2}) + \frac{1}{2} \cdot \log_2(\frac{1}{2})) = 0$

Y la ganancia de información para género sería:

$1 - (\frac{3}{4} \cdot \log_2(\frac{3}{4}) + \frac{1}{4} \cdot \log_2(\frac{1}{4})) \approx 0{,}189$

Como la ganancia de información es mayor para género entonces este sería el primer nodo. Profundidad máxima del árbol: Limitar la profundidad de un árbol de decisión puede generar árboles diferentes. Un árbol más profundo puede capturar más detalles de los datos, mientras que un árbol menos profundo podría ser más general. Número mínimo de muestras para dividir: Establecer un umbral para el número mínimo de muestras requeridas para considerar una división puede afectar el tamaño del árbol. Un umbral más

alto podría prevenir divisiones que crearían nodos con muy pocas muestras, lo que podría llevar a árboles con menos ramificaciones.

4. Método del vecino más cercano

4.1. Realice un glosario con los siguientes conceptos:

4.1.1. ¿Cómo se determina un vecino más cercano en un dataset?

Para determinar el vecino más cercano primero se elige una métrica de distancia para medir qué tan cerca están las muestras unas de otras. Las métricas de distancia más comunes incluyen:

Distancia Euclidiana: Es la más común y es adecuada para muchas situaciones. Se calcula con la fórmula:
$\sqrt{(\sum_{i=1}^{n}(x_i - y_i)^2)}$
donde x_i y y_i son los valores de la i-ésima característica de las muestras x e y, respectivamente.

Distancia de Manhattan (o distancia L1): Suma las diferencias absolutas de las coordenadas. Se calcula con la fórmula:
$\sum_{i=1}^{n} |x_i - y_i|$

Distancia de Minkowski: Generalización de las distancias Euclidiana y de Manhattan. Se calcula con la fórmula:
$(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$
Nótese cómo con p=1 obtenemos la distancia manhattan, y con p=2 obtenemos la distancia euclidiana.

4.1.2. ¿Qué métricas existen si las clases son numéricas?

Algunas de las métrias más comunes para vriables numéricas son:
Coeficiente de Determinación (R^2, R-squared): Proporciona una medida de cuánto de la variabilidad en la variable dependiente es explicada por el modelo. Un R^2 de 1 indica que el modelo ajusta perfectamente los datos. La fórmula del R^2 es:
$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
donde $\hat{y}_i$ son los valores producidos por el modelo, y $\bar{y}$ es el promedio de los valores reales.

Coeficiente de Correlación de Pearson: Mide el grado de correlación lineal entre los valores reales y los predichos. Un valor de 1 significa una correlación positiva perfecta, -1 una correlación negativa perfecta, y 0 ninguna correlación. La fórmula del coeficiente es:
$\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}$
Error Cuadrático Medio(MSE, Mean Squared Error): Calcula el promedio de los cuadrados de los errores entre los valores predichos y los valores reales. Es muy sensible a los valores atípicos debido al cuadrado de los errores. La fórmula es:
$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

Raíz del Error Cuadrático Medio (RMSE, Root Mean Squared Error): Es la raíz cuadrada del MSE. Proporciona una medida de la magnitud del error en las mismas unidades que la variable de respuesta. La fórmula es:
$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
Error Absoluto Medio (MAE, Mean Absolute Error): Calcula el promedio de los valores absolutos de los errores. Es menos sensible a los valores atípicos en comparación con MSE. La fórmula es:
$\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$

4.1.3. ¿Qué métricas existen si las clases son categóricas?

Algunas de las métricas más comunes para clases paramétricas son:
exactitud: esta métrica se calcula con la siguiente formula:
$\frac{TP+TN}{TP+TN+FP+FN}$
Donde:
TP = verdadero positivo, casos donde se predijo correctamente que un elemento era de una determinada clase.
TN = verdadero negativo, caso donde se predijo correctamente que un modelo no era de una determinada clase.
FP = falso positivo, predicciones fallidas de que un elemento era de una determinada clase.
FN = falso negativo, predicciones fallidas de que un elemento no era de una determinada clase.
Tanto para modelos binarios como multiclases el modelo es el mismo y se puede simplificar como la división de las predicciones correctas sobre el total de predicciones. precisión: Esta métrica se preocupa por cuán precisas son

las predicciones positivas del modelo, es decir, de todas las instancias que el modelo clasifica como positivas, cuántas son realmente positivas. Por ejemplo, si se hacen 15 predicciones que que una clase es de tipo A, y de esas solo 12 eran A entonces la precisión sería del 80 % la formula es:
$\frac{TP}{TP+FP}$
Recall: El recall mide la capacidad del modelo para identificar correctamente las instancias positivas de un conjunto de datos. Por ejemplo, si es hicieron 12 predicciones correctas de que una clase era A, de un total de 20 clases A entonces el recall sería del 60 %. La fomrula es:
$\frac{TP}{TP+FN}$

promedio macro: En este caso para cada clase se calcula la precisión y luego se hace el promedio de estas métricas. Este método da igual peso a cada clase, lo cual es útil cuando las clases están desbalanceadas.

promedio micro: Esta métrica se basa en el mismo principio del promedio macro pero usando el recall en lugar de la precisión. [**?**, pag 105-107]

4.1.4. Describa un problema típico de su disciplina de desempeño profesional donde:

• Deba tomar una decisión a partir de información dada por dos categorías.

• Indique dos características de tipo cuantitativo.

• Describa cómo sería el uso del algoritmo de KNN en este escenario, para clasificar un nuevo registro, si se usa la distancia euclidiana y se elige k=3 como el número de vecinos más cercanos.

En aplicaciones de visión por computadora integradas en RPA, necesitamos identificar dinámicamente elementos en una interfaz de usuario para interactuar con ellos (por ejemplo, hacer clic). Sin embargo, estos elementos pueden variar en texto y ubicación, lo que complica su identificación precisa. Para clasificar los elementos usamos dos criterios, primero la similitud del texto del elemento con un texto objetivo o de referencia. Esta similitud puede medirse mediante técnicas como la distancia de Levenshtein, coincidencias exactas o parciales, o cualquier otro método de comparación de cadenas que cuantifique cuán similar es el texto de un elemento al texto de referencia.
El segundo criterio sería la distancia espacial del elemento respecto a una

ubicación de referencia en la pantalla, que podría estar determinada por coordenadas previas conocidas del elemento objetivo. Para cada elemento interactuable (por ejemplo, botones, enlaces, campos de texto) en la interfaz de usuario, calculamos dos características cuantitativas: la similitud de su texto con el texto de referencia y su distancia espacial a la ubicación de referencia. Cuando necesitamos interactuar con el elemento, calculamos la distancia euclidiana entre el elemento y el objetivo usando las dos características cuantitativas. La fórmula sería:
$\sqrt{x^2_{\text{similitud texto}} + y^2_{\text{distancia espacial}}}$
Aplicando un k=3 podríamos obtener los tres vecinos más cercanos a los criterios del elemento original. Esto podría servir para identifiar si hay múltiples elementos similares al objeto original, en cuyo caso la automatización debería lanzar un error o una advertencia. Esto permitiría identificar situaciones donde los criterios actuales no son suficientes para una selección clara, proporcionando una oportunidad para optimizar y refinar los procesos de automatización.

Referencias

[Weis] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

[Gru] Grus, J. (2015). Data science from scratch: first principles with Python. First edition. Sebastopol, CA, O'Reilly

[Ribe] Ribeiro, M. I. (2004). Gaussian probability density functions: Properties and error characterization. Institute for Systems and Robotics, Lisboa, Portugal.

[Ra] Raschka, S., Mirjalili, V. (2017). Python Machine Learning. Livery Place 35 Livery Street Birmingham B3 2PB, UK: Packt Publishing Ltd..