Apunte Organicacion de datos

lcondoriz

May 2023

Índice

1. Introducción al machine learning.

En el contexto del aprendizaje supervisado, se trabaja con variables independientes (también conocidas como características, predictores o variables de entrada) y variables dependientes (también llamadas variables objetivo o target o variables de salida). Las variables independientes son las que se utilizan para hacer predicciones o estimaciones, mientras que las variables dependientes son los resultados que se intentan predecir o modelar. Por ejemplo, en un problema de predicción de precios de viviendas, las características como el número de habitaciones, la ubicación y el tamaño de la propiedad serían las variables independientes, mientras que el precio de venta sería la variable dependiente.

- Variables Independientes (entradas)
    1. Cualitativas
        a) Texto
            - Nominales (categorías, Por ejemplo: países, sexo)
            - Ordinales (poco, mucho, muchísimo, Por ejemplo: nivel de tabaquismos)
        b) Números
            - Nominales
            - Ordinales
    2. Cuantitativas: cuando hablamos de cantidad
        a) Discretas: Por ejemplo año, mes, edad, etc.
        b) Continuas: Por ejemplo altura, peso, etc.
- Variables Dependientes (salidas, categorías)

Las variables **cualitativas** son aquellas que describen características o cualidades y no pueden ser medidas en términos numéricos. Las variables **cuantitativas**, por otro lado, son aquellas que pueden ser medidas en términos numéricos y tienen valores numéricos.

Las **variables nominales** son aquellas que no tienen un orden natural, como el género o el color de ojos. Las **variables ordinales** son aquellas que tienen un orden natural, como el nivel de educación (primaria/secundaria/universidad), nivel de tabaquismos: Clasificamos como leve (1), moderado (2), nivel medio (3), importante (4) y muy importante (5).

Variables y tipos de problemas (video 02a min 9:00)

1. Si la variable dependiente es **cualitativa**, el tipo de problema es de **clasificación**.
2. Si la variable dependiente es **cuantitativa**, el problema es de **regresión**. Por ejemplo si quiero predecir el precio de una propiedad.

3. Si **NO hay variable** dependientes, el problema es agrupamiento.

- **Outliers:** Valores atípicos, pueden ser errores o un dato que se sale de la norma.
- **Correlación:**
  - **Positiva:** Cuando una variable aumenta la otra también.
  - **Negativa:** Cuando una variable aumenta la otra disminuye.
  - **Sin correlación:** Cuando una variable aumenta la otra no cambia.
- **Varianza:** Es la medida de dispersión de una variable respecto a su media. Si la varianza es alta, los datos están muy dispersos, mientras que si la varianza es baja, los datos están muy agrupados.
- **Covarianza:** Es una medida de la relación lineal entre dos variables aleatorias. Indica cómo varían conjuntamente dos variables aleatorias respecto a sus medias. Si la covarianza es positiva, las variables aumentan o disminuyen conjuntamente, mientras que si la covarianza es negativa, una variable aumenta mientras la otra disminuye.

2. Preprocesamiento de datos.

Limpieza de datos.

La limpieza de datos es un paso fundamental en el preprocesamiento de datos en el machine learning. Este proceso implica identificar y corregir o eliminar datos incorrectos, incompletos, duplicados o inconsistentes en un conjunto de datos. Aquí se explica más detalladamente la limpieza de datos:

1. **Identificar datos incorrectos o erróneos:** Durante la recopilación de datos, es posible que se hayan introducido errores o que los datos estén mal formateados. En este paso, se deben identificar y corregir los errores obvios, como valores que están fuera de rango o que no se ajustan al formato esperado. Por ejemplo, si se tiene un conjunto de datos que contiene información sobre el género de los empleados de una empresa, los datos incorrectos serían los que no son ni "masculinos"ni "femeninos".

2. **Tratamiento de datos faltantes:** Los conjuntos de datos a menudo contienen valores faltantes, ya sea porque no se recopilaron o porque se perdieron durante el proceso de almacenamiento o transferencia. Los valores faltantes pueden afectar el rendimiento de los modelos de machine learning, por lo que es importante abordarlos. Esto implica decidir si se deben eliminar las instancias con valores faltantes, imputar los valores faltantes utilizando técnicas de imputación (como el promedio, la mediana o modelos más avanzados), o considerarlos como una categoría separada.

3. **Eliminación de datos duplicados:** En algunos casos, puede haber instancias duplicadas en el conjunto de datos. Estas instancias duplicadas no aportan información adicional y pueden sesgar los resultados. Por lo tanto, es esencial identificar y eliminar las instancias duplicadas para garantizar la integridad y la calidad de los datos.

4. **Resolución de inconsistencias y valores atípicos (outliers):** Los valores atípicos son observaciones que difieren significativamente del resto de los datos. Pueden ser resultado de errores de medición o indicar situaciones inusuales. Es importante evaluar si los valores atípicos deben ser corregidos, eliminados o si contienen información relevante y deben mantenerse.

5. **Normalización y estandarización:** Los datos pueden tener diferentes escalas y rangos, lo que puede afectar el rendimiento de algunos algoritmos de machine learning. En este paso, se pueden aplicar técnicas de normalización o estandarización para asegurar que los datos tengan una distribución más uniforme y comparable

6. **Manejo de datos desbalanceados:** En algunos problemas de clasificación, puede haber una falta de equilibrio entre las clases objetivo, lo que significa que una clase puede tener muchos más ejemplos que las demás. En estos casos, se deben utilizar técnicas de muestreo o ponderación para abordar el desequilibrio y evitar que el modelo se sesgue hacia la clase mayoritaria

Transformación y normalización de datos.

La transformación y normalización de datos son técnicas utilizadas en el preprocesamiento de datos en machine learning para ajustar los datos a una escala o distribución específica. Estos procesos son importantes para garantizar que los datos sean adecuados para su uso en algoritmos de machine learning y que no se vean afectados por diferencias en las unidades o escalas de las características. A continuación, se explica con más detalle la transformación y normalización de datos:

3. Machine Learning.

Classical Machine Learning

1. Supercised Learning

   a) Classification:

      1) Regresión logística.
      2) K-Nearest Neighbors (k-vecinos más cercanos) (k-NN).
      3) Árboles de decisión.

4) Bosques aleatorios.
5) Máquinas de vectores de soporte (SVM)
6) Naive Bayes.

b) Regression: algoritmos y técnicas: regresión lineal, la regresión polinómica, la regresión de bosques aleatorios

2. Unsupervised Learning

a) Clustering
b) Association
c) Dimensionality Reduction

Aprendizaje supervisado: Tenemos datos de entrenamiento con una salida esperada. Validación de resultados. Datos de entrada y salida etiquetados durante la fase de entrenamiento del ciclo de vida del machine learning.

Aprendizaje No Supervisado: No tenemos datos de salida sólo de entrada Cambiar representación de los datos. Facilitar entendimiento. Es el entrenamiento de modelos de datos sin procesar y sin etiquetar.

4. Supercised Learning

El Supercised Learning se puede clasificar en dos grandes grupos:

1. Classification:

a) Regresión logística.
b) K-Nearest Neighbors (KNN).
c) Decision Trees.
d) Random Forest.

2. Regression.

a) Linear Regression.

Existen métodos y algoritmos que se pueden usar en ambos problemas:

Métricas Clasificación:

- Precisión.
- Recall (exhaustividad).
- Valor-F (F_1)

Ver video 03b min 40. Matriz de confucion.

Supervised Learning	Classification	Regression
Logistic Regression	✓	
k-Nearest Neighbors (KNN)	✓	✓
Decision Trees	✓	✓
Random Forest	✓	✓
Linear Regression		✓

Cuadro 1:

4.1. Classification

En el contexto del aprendizaje automático supervisado, la variable predictora (también conocida como variable independiente o característica) es una variable que se utiliza para predecir el valor de la variable objetivo (también conocida como variable dependiente o target). Las variables predictoras son los datos de entrada al modelo y la variable objetivo es el resultado que se desea predecir.

4.1.1. Regresión logística.

La regresión logística es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente para resolver tareas de "clasificación" binarias, aunque contiene la palabra regresión". Se lo utiliza para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las **variables predictoras** o independientes. Un ejemplo de clasificación podría ser la detección de spam: un programa de aprendizaje automático puede aprender a marcar el spam después de recibir ejemplos de correos electrónicos spam marcados (**variable objetivo** "target" sería una variable binaria que indica si un correo electrónico es spam o no) por los usuarios y ejemplos de correos electrónicos regulares no spam (también llamados "ham"). Notebook ejemplo [2].

4.1.2. K-Nearest Neighbors (KNN).

k vecinos más cercanos es un método de clasificación no paramétrico. Video de youtube [3].
Se utiliza principalmente para la clasificación de datos no lineales y para resolver problemas de clasificación en los que los datos son muy complejos y desestructurados.

Resumen:

- Es sensible a conjuntos de datos no balanceados.
- Es sensible aoutliers.

4.1.3. Decision Trees

Existen varios algoritmos ID3, C4.5 y CART. Pagina buena [1]:

• **ID3 - Iterative Dichotomiser 3:** Genera un árbol de decision a partir de un conjunto de ejemplos.
Este algoritmo usa las metricas de *entropía* y *ganancia de la información*. Cuando se usa la librería sklearm *sklearn.tree.DecisionTreeClassifier* en el parámetro *criterion* colocar *entropy*.

- Un nodo principal llamado raíz en la parte superior.
- Nodos terminales. como su nombre lo indica, son nodos donde termina el flujo y que ya no son raiz de ningún otro nodo. Estos nodos terminales deben contener una respuesta, o sea, la clasificación a que pertenece el objeto que ha conducido hasta él.
- Los demás nodos representan preguntas con respecto al valor de uno de los atributos.
- Las líneas nodos representan preguntas con respecto al valor de uno de sus atributos.
- Las líneas representa las posibles respuestas que los atributos pueden tomar.

Algoritmo básico

1. Calcular la entropía para todas las clases.
2. Calcular la entropía para cada valor posible de cada atributo.
3. Seleccionar el mejor atributo basado en la reducción de la entropía. usando el calculo de la ganancia de la información.
4. Iterar, para cada sub-nodo, Excluyendo el nodo raíz, que ya fue usado.

• **C4.5:** Utiliza la metrica *gini*.

1. Mitiga el sobreajuste por que emplea inherentemente el proceso de poda de un solo paso.
2. Funciona para datos discretos y continuos.
3. Es útil para datos incompletos.

Metricas:

- **Entropía:**
  La medida del desorden o la medida de la pureza. Básicamente, es la medida de la impureza o aleatoriedad de los datos.

  Para calcular la entropia de n clases se utiliza la fórmula:

$$H(S) = \sum_{i=1}^{n} -p_i \cdot \log_2(p_i) \tag{1}$$

  Dónde:

  - S: es una lista de valores posibles.
  - p_i: es la probabilidad de los valores.
  - i: cada uno de los valores.

  Importante:

  - Para una muestra homogénea la entropia es igual a cero 0. Si no existe aleatoriedad, es decir, una moneda cargada.
  - La máxima entropía viene dada por $\log_2(n)$, n son los posibles valores de salida. Si $n = 2$ (true o false) entonces, la máxima entropía es 1. O sea es la máxima incertidumbre, ejemplo moneda equilibrada.

- **Ganancia de la Información:** La ganancia de la información se aplica a cuantificar qué característica, de un conjunto de datos dados, proporciona la máxima información sobre la clasificación.

4.1.4. Random Forest - Bosques aleatorios.

Usa una técnica, o meta-algoritmo llamado Bootstrap aggregating. Se crean m tablas reducidas en atributos y para cada una de ellas entrenamos un árbol.

4.2. Regresión

4.2.1. Linear Regression

La regresión lineal es un método estadístico que se utiliza para estudiar la relación entre una variable dependiente (también conocida como variable objetivo) y una o más variables independientes (también conocidas como variables predictoras). El objetivo de la regresión lineal es encontrar la línea que mejor se ajuste a los datos y pueda utilizarse para hacer predicciones.

En el caso de una sola variable independiente, la línea de regresión se puede representar mediante la ecuación $y = a + bx$, donde y es la variable dependiente, x es la variable independiente, a es la intersección en y y b es la pendiente de la línea. Los valores de a y b se determinan utilizando los datos disponibles para minimizar la suma de los errores cuadrados entre los valores observados y los valores predichos por la línea de regresión. Ver ejemplo

4.2.2. Regresión Lineal Múltiple

Consiste en predecir una respuesta numérica y en base a múltiples variables predictoras $x_1, x_2, ..., x_n$, suponiendo una relación lineal.

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_n \cdot x_n + b \tag{2}$$

• **Métricas Regresión:** Las métricas de evaluación son medidas utilizadas para evaluar el rendimiento de un modelo de aprendizaje automático. En el caso de los problemas de regresión en el aprendizaje supervisado, estas métricas nos ayudan a determinar qué tan bien nuestro modelo está haciendo predicciones cuantitativas, como valores continuos.

Metricas para la regresión:

- Raíz del error cuadrático medio **(RMSE)**.
- Error absoluto medio **(MAE)**.
- Error cuadratico medio **(MSE)**.
- Suma Residual de los cuadrados **RSS**.

Para cada una de las métricas:

- m : número de instancias
- h : Funcion hipotesis, es el modelo entrenado. En este caso regresion lineal.
- x : Todos los valores de entrada, todas las columnas.

Error cuadrático medio (MSE):

$$MSE(X, h) = \frac{1}{m} \cdot RSS = \frac{1}{m} \cdot \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2 \tag{3}$$

Error absoluto medio (MAE):

$$MAE(X, h) = \frac{1}{m} \cdot \sum_{i=1}^{m} |h(x^{(i)} - y^{(i)}| \tag{4}$$

Raíz del error cuadrático medio (RMSE):

$$RMSE(X, h) = \sqrt{\frac{1}{m} \cdot \sum_{i=1}^{m} (h(x^{(i)} - y^{(i)}))^2} \tag{5}$$

5. Unsupervised Learning

1. Unsupervised Learning
   a) Clustering
      1) K-means.
      2) Clustering jerárquico.
      3) DBSCAN.
      4) Mezcla de Gaussianas.
   b) Association
      1) Reglas de asociación.
      2) algoritmo Apriori
   c) Dimensionality Reduction

5.1. Clustering

En este tipo de problemas se trata de agrupar los datos. Agruparlos de tal forma que queden definidos N conjuntos distinguibles, aunque no necesariamente se sepa que signifiquen esos conjuntos. El agrupamiento siempre será por características similares.

Cuantos clusters elegir:

1. Regla del codo (Elbow Method). fer
2. Método de Silhouette. fer
3. Estadistica de Hopkins. fer

Regla del codo (Elbow Method) En el grafico buscamos un 'codo', el lugar donde baja abruptamente.

- Elegimos un rango, ejemplo 1 a 10, y para cada valor:
  - Para cada centroide calculamos la distancia promedio.

5.1.1. K-means

1. El usuario decide la cantidad de grupos.
2. K-Means elige al azar K centroides.
3. Decide qué grupos están más cerca de cada centroide. Esos puntos forman un grupo.
4. K-Means recalcula los centroides al centro de cada grupo

5. K-Means vuelve a reasignar los puntos usando los nuevos centroides. Calcula nuevos grupos

6. K-means repite punto 4 y 5 hasta que los puntos no cambian de grupo.

Un ejemplo de este algoritmo seria el conjunto de datos Iris, donde se tiene 5 columnas (Largo de sépalo, Ancho de sépalo, Largo de pétalo, Ancho de pétalo, Especies). La columna Especies no la usamos por que estamos en Unsupervised Learning. Con las columnas restantes tenemos que buscar cuantos clusters hay y luego prodria compara con la columna Especies.

Referencias

[1] Arboles de Decisión. En: (). URL: https://www.ibm.com/es-es/topics/decision-trees.

[2] Colab Regresion Logistica. En: (). URL: https://colab.research.google.com/drive/1JbRUFa5hnijNQdJ_HLywhMJrICkColMJ?authuser=1#scrollTo=sIl68yHmh0yS.

[3] Video de YouTube KNN. En: 31 min (). URL: https://www.youtube.com/watch?v=cH-kUai4Boo.