Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2022

23 de Noviembre, 2023

1. **Un cliente que vende artículos online le proporciona un dataset con reseñas de los productos que vende en su plataforma. El cliente le comenta que quiere diferenciar los productos buenos de los malos en base a las reseñas. Para eso él etiqueto manualmente una gran porción de las reseñas por lo que el dataset indica los textos y sus respectivas categorías binarias. Con sus conocimientos de NLP que pasos seguiría para entrenar un clasificador binario?**

Pasos para entrenar un clasificador binario utilizando nuestros datos de texto:

* Elimine cualquier información irrelevante, como caracteres especiales, números o etiquetas HTML.
* Manejar los valores faltantes si están presentes.
* Tokenice el texto en palabras o unidades de subpalabras.
* Convierte texto a minúsculas.
* Elimine las stop words (palabras comunes como "y", "el", etc.).
* Aplicar stemming o lemmatization para reducir palabras a su forma básica.
* Convierta los datos de texto preprocesados ​​a formato numérico utilizando técnicas como Bag-of-Words (BoW) o Term Frequency-Inverse Document Frequency (TF-IDF).
* Dividir el conjunto de datos en train y test.
* Elija un modelo de clasificación binaria.
* Entrene el modelo .
* Evalue el modelo utilizando una metrica de clasificación.

1. **Qué problemas pueden surgir si uno imputa valores faltantes utilizando la moda? Qué problemas pueden surgir si uno imputa valores faltantes utilizando la media?**

Los problemas surgirán dependiendo la distribución de los datos de la columna a imputar. En el caso de la moda, una distribución multi-modal (como la bimodal por ejemplo) puede tener mas de un valor de moda por lo que podría llegar a no ser representativo. Asimismo, la moda no tiene en consideración la varianza de la distribución y puede llegar a introducir sesgos en nuestro conjunto de datos.

En el caso de la media o promedio, las distribuciones que presenten outliers pueden influenciar negativamente al correr la media en la dirección del outlier. Al igual que en el caso de la moda se pueden introducir sesgos a favor de la tendencia central reduciendo la varianza original.

1. **Ilustre con un ejemplo que problemas pueden surgir si uno no realiza un muestreo estratificado a la hora de crear sus conjuntos de train y test. Asuma que su dataset esta desbalanceado.**

Si su variable objetivo está desbalanceada, lo que significa que una clase es mucho más frecuente que las demás, una división aleatoria puede dar como resultado que uno de los subconjuntos tenga una distribución de clases significativamente diferente a la del conjunto de datos original. Esto puede conducir a una evaluación sesgada del modelo, donde el modelo podría funcionar bien en la clase mayoritaria pero mal en la clase minoritaria. Puede parecer que el modelo tiene una gran precisión cuando, en realidad, no se generaliza bien entre la clase minoritaria.

Sin estratificación, las métricas de rendimiento (como exactitud, precisión, recuperación) calculadas en el conjunto de validación pueden no reflejar con precisión qué tan bien se generaliza el modelo a todo el conjunto de datos. Las métricas de evaluación engañosas pueden dar lugar a una mala selección de modelos y potencialmente conducir a la implementación de un modelo que no funciona bien en todas las categorías o subgrupos relevantes.

En conjuntos de datos más pequeños, las divisiones aleatorias pueden dar lugar a subconjuntos con propiedades estadísticas significativamente diferentes.

En algunos casos, es posible que tenga subgrupos dentro de sus datos que es importante representar con precisión tanto en los conjuntos de entrenamiento como en los de validación. Sin estratificación, existe el riesgo de perder representación de ciertos subgrupos en uno de los subconjuntos.

1. **Usted entrena un modelo lineal para predecir la cantidad de likes de una publicación en Instragram utilizando una regresión lineal. El modelo le devuelve los parámetros del mismo, por lo que la ecuación resulta:**

**Usted sabe que los coeficientes son y y . Interprete el modelo a través de sus parámetros.**

Por cada seguidor adicional, el modelo predice un aumento de 0,15 me gusta en la publicación de Instagram. El coeficiente positivo sugiere una relación positiva entre el número de seguidores y el número de me gusta.

Por cada comentario negativo adicional, el modelo predice una disminución de 0,6 me gusta en la publicación de Instagram. El coeficiente negativo indica que los comentarios negativos están asociados con una disminución en el número de me gusta.

El término de intercepción, que representa la cantidad prevista de Me gusta cuando tanto los seguidores como los comentarios negativos son cero. En términos prácticos, sugiere que incluso si no hay seguidores ni comentarios negativos, el modelo predice 40 me gusta como punto de partida.

1. **Suponga que usted tiene dos datasets 2-dimensionales y decide graficar scatterplots para cada uno obteniendo las siguientes figuras:**

**A screenshot of a computer game

Description automatically generated**

**Usted decide utilizar dos algoritmos de clustering para categorizar los datos: K-Means y DBSCAN. Para que dataset utilizaría cada uno? Justifique su respuesta.**

DBSCAN debe ser utilizado para el dataset izquierdo y K-Means para el derecho. La razón es que el dataset izquierdo no es linealmente divisible por lo que K-means nunca podría capturar los clusters del circulo externo e interno. Conviene utilizar DBSCAN para este ejemplo ya que puede capturar clusters por densidad. En el caso derecho, conviene utilizar K-Means ya que visualmente se pueden apreciar 3 claros clusters que son linealmente divisibles. DBSCAN en este caso podría llegar a clasificar el grupo azul y naranja dentro de un mismo grupo por proximidad.

1. **Suponga que usted tiene un dataset con variables independientes *X* y una variable dependiente *y.* Usted realiza 3 experimentos y mide la performance utilizando un modelo lineal:**
   1. **En el primer experimento no realiza ningún pre-procesamiento.**
   2. **En el segundo decide usar un MinMaxScaler (normaliza/escala las columnas).**
   3. **En el tercero decide usar un StandardScaler (estandariza las columnas).**

**Usted no encuentra diferencias de performance entre el experimento 1 y 2 pero si encuentra diferencia respecto al 3. Por qué razón sucede esto?**

Las diferencias de performance existentes entre los experimentos 1 y 2 y el experimento 3 se deben a que en los experimentos 1 y 2 no se distorsiona la distribución original y en el experimento 3 si se distorsiona. Un escalamiento de datos (o no escalamiento) es una transformación lineal que mantiene la distribución invariante, es decir, se modifican los valores pero no las relaciones/distribucion entre ellos. Aplicar una estandarización en cambio si cambia la distribución ya que la vuelva “mas Gaussiana”. Al utilizar un modelo lineal estos cambios se reflejan en la metrica de performance.

1. **Usted trabaja en el área comercial de un supermercado. El supermercado tiene muchas áreas que se dedican a la venta de productos distintos. Usted es el gerente comercial y le pide a los encargados de área que le reporten sus niveles de venta a fin de mes. El gerente general a su vez le solicita a usted que cada área llegue a un nivel determinado de ventas a fin de mes. Qué gráfica cree usted apropiada para mostrarle al gerente general los resultados mensuales? Dibuje la gráfica y justifique su respuesta.**

En esta pregunta existen diversas respuestas, pero lo principal es recalcar los niveles esperados y contrastarlos contra los niveles alcanzados distinguiendo claramente los distintos sectores y permitiendo una comparación visual simple.

Algunas graficas podrían ser: grafico de rendimiento radial, bulletgraph, grafico de barras con línea de benchmark, etc.

1. **Series de tiempo:**
   1. **Qué significa que una serie de tiempo sea estacionaria?**
   2. **Analizando gráficamente, son las dos series de tiempo que se muestran a continuación estacionarias?**

A graph of different types of lines

Description automatically generated with medium confidence

Una serie estacionaria es aquella cuyas propiedades estadísticas son constantes en el tiempo. No significa que la serie de tiempo no cambie con el tiempo sino que la forma en que cambia permanece invariante con el tiempo. Esto implica que no sólo el primer y segundo momento (media y varianza) son constantes, sino que toda la función de distribución de probabilidad es invariante en el tiempo.

La imagen de la izquierda presenta una serie de tiempo no estacionaria ya que la varianza de la misma aumenta con el tiempo mientras que la de la derecha es estacionaria ya que su media y varianza permanecen constantes.

1. **Qué diferencias existen entre supervised y unsupervised learning? De ejemplos de cómo serían los datasets, mencione un modelo y métricas que utilizaría para cada tipo de tarea.**

La diferencia principal es el tipo de problema. En supervised learning uno utiliza un modelo para descubrir una función de mapeo entre variables independientes y variables dependientes. Es decir, en supervised learning uno intenta descubrir la relación subyacente entre las variables de entrada y la variable de salida a predecir. En unsupervised learning uno intenta descubrir patrones, relaciones y grupos dentro de nuestro conjunto de datos.

Respecto a los datasets, en supervised learning es necesario una variable objetivo/dependiente *y* a predecir mientras que en unsupervised no existe tal variable. En otras palabras, en supervised tenemos datos etiquetados mientras que en unsupervised no.

En supervised learning, dependiendo si el problema es de regresión o clasificación, podemos encontrarnos con modelos como Linear Regression, SVM, Redes Neuronales, XGBoost Regressor, etc (caso regresión), y modelos como Logistic Regression, Decision Trees, Random Forests, XGBoost Classifier, Redes Neuronales (caso clasificación). En unsupervised nos podemos encontrar con modelos de clustering como K-Means, DBSCAN, OPTICS, Birch, Hierarchical Clustering, etc.

En supervised learning, podemos encontrarnos con métricas como MAE, RMSE (caso regresión), y métricas como Accuracy, F1-Score, ROC-AUC, etc (caso clasificación). En unsupervised nos podemos encontrar con métricas como Silhouette, Davies-Bouldin Index, Clainski Harabasz Index, etc.

1. **Usted trabaja en una empresa de software de streaming televisivo. Quiere recomendar a los usuarios de la plataforma contenido similar a lo consumido por las personas con gustos mas similares a uno.**

**Su empresa recaba información de las películas que cada usuario ha visto en el pasado y lo almacena en una tabla como se muestra en la figura.**

1. **Como transformaría los datos para poder ser utilizados por un modelo?**
2. **Que modelo entrenaría?**
3. **Como haría las recomendaciones?**

A table with numbers and letters

Description automatically generated

Nota: asumir que la frecuencia es proporcional/es el rating. Usted debería realizar un pivot de la matriz de ratings, donde los usuarios son filas y las películas son columnas y los ratings son las entradas de la matriz. Puede normalizar por usuario o por fila para balancear los efectos de películas muy populares o usuarios muy cinéfilos. Como la letra sugiere recomendaciones a partir de usuarios similares podría usar KNN y encontrar las películas vistas por los K usuarios mas similares. Para recomendar las películas podría encontrar los K usuarios mas similares y recomendar una película no vista por el usuario en cuestión pero que si haya obtenido una gran frecuencia o puntaje por los usuarios mas similares. Puede hacer voto mayoritario o ponderar por similitud.