UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Universidad Autónoma de Nuevo León

Facultad de Ciencias Fisico Matematicas

Mineria de Datos

Resúmenes: Tecnicas de mineria de datos

Alumno: Edgar Bladimir Lopez Alonzo #1753141

Grupo: 012

Maestro: Mayra Cristina Berrones Reyes

Regresion Lineal.

Usado para aproximar la relación de dependencia entre una variable dependiente, las variables independientes y un término aleatorio.

En el caso de una regresión lineal asumimos que Y es una función lineal de x, y entonces el modelo lineal se escribe como: Ye = $\alpha + \beta$ * x.

Y la diferencia entre el valor real y el estimado se puede escribir como:

ei = (Yi - Ye(Xi)).

El objetivo es minimizar la suma de los errores al cuadrado sobre todos los puntos del data set: X = {(Xi,Yi)}.

$$\min \sum_{i=1}^{n} ei^2 = \sum_{i=1}^{n} (Yi - Ye(Xi))^2 = \sum_{i=1}^{n} (Yi - (\propto + \beta * Xi))^2$$

$$\beta = \frac{\sum_{i=1}^{n} (Xi - x_m)(Yi - y_m)}{\sum_{i=1}^{n} (Xi - x\_m)^2} \qquad \alpha = y_m - \beta * x_m$$

Siempre se tendrá un componente de error o residuo E Ye = $\alpha + \beta$ * x + E. El residuo E será una variable aleatoria con distribución normal.

El p valor: el modelo presenta una relación lineal entre x e y, para comprobar la existencia de la relación, se plantea el contraste de hipótesis.

$$f(x) = \begin{cases} H_0 : \beta = 0 \\ H_0 : \beta \,!= 0 \end{cases}$$

Si el p valor resultante es menor que el nivel de significancia, se rechaza la hipótesis nula y se acepta que existe una relación lineal entre x e y.

Error estándar residual: RSE es la desviación estándar del término del error (desviación de la parte de datos que el modelo no es capaz de explicar por falta de información o datos adicionales).

En el caso de un regresión lineal simple: $RSE = \sqrt{\frac{\sum (Yi - Y(Xi))^2}{n-2}} = \sqrt{\frac{SSD}{n-2}}$

En el caso de una regresión lineal múltiple: $RSE = \sqrt{\frac{\sum (Yi - Y(Xi))^2}{n-k-2}} = \sqrt{\frac{SSD}{n-k-2}}$

Outliers.
Observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente.

Tipos de outliers.
Casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificar como datos ausentes.
Observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.
Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.
Datos extraordinarios para los que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados.
Los métodos de detección: Se pueden dividir en univariados y multivariados. Los multivariantes son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de aquellas. Son más difíciles de identificar que los outliers unidimensionales, dado que no pueden considerarse “valores extremos”, como sucede cuando se tiene una única variable bajo estudio.

Método de detección - desviación estándar: Si se tiene algún punto de datos que sea más de 3 veces la desviación estándar, es muy probable que esos puntos sean anómalos o atípicos.

Método de detección - boxplots: Los diagramas de caja son una representación gráfica de datos numéricos a través de cuantiles. Cualquier punto de datos que se muestre por encima o por debajo de los bigotes, puede considerarse atípico o anómalo.

Método de detección - DBScan Clustering: *Core Points*, min_samples es simplemente el número mínimo de puntos centrales necesarios para formar un grupo, eps es la distancia máxima entre dos muestras para que se consideren como en el mismo grupo. *Border Points*, se encuentran en el mismo grupo que los puntos centrales, pero mucho más lejos del centro del grupo.

Todo lo demás se denomina Puntos de ruido (Noise Points), son puntos de datos que no pertenecen a ningún grupo. Pueden ser anómalos o no anómalos y necesitan más investigación.

Clustering.
Es una técnica dentro de la disciplina de Inteligencia Artificial, identifica de manera automática agrupaciones de acuerdo a una medida de similitud entre ellos.

Métricas de distancia: Una métrica de distancia es una función $\boldsymbol{d\,x, y}$ que especifica la distancia entre elementos de un conjunto de números reales no negativos, dos elementos son iguales bajo una métrica particular si la distancia entre ellos es cero.
Distancia euclídea: Este tipo de distancia es usada principalmente para calcular distancias. La distancia entre dos puntos en el plano con coordenadas (x, y) y (a, b) según la fórmula de la distancia euclidiana viene dada por:

$$\text{Euclidean dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

Distancia de Manhattan: Este tipo de distancia es definida como la suma de las longitudes de las proyecciones del segmento de línea entre los dos puntos en los ejes de coordenadas. La fórmula está dada por:

$$\text{Manhattan dist}((x, y), (a, b)) = |\, x - a \,| + |\, y - b \,|$$

Clustering jerárquico: El clustering jerárquico puede realizarse tanto en forma divisiva o aglomerativa, y permite analizar alternativas para distintos números de grupos.
Clustering jerárquico aglomerativo: Se comienza con tantos clusters como individuos y consiste en ir formando grupos según su similitud.
Clustering jerárquico de división: Se comienza con un único clúster y consiste en ir dividiendo clústeres según la disimilitud entre sus componentes.

Clustering de partición: Se encarga de dividir un conjunto de datos en una pequeña cantidad de agrupaciones o particiones, basado en sus atributos.
Consiste en agrupar los elementos en torno a elementos centrales llamados centroides a cada clúster. Definimos el centroide de un clúster como aquel elemento que minimiza la suma de las similitudes al resto de los elementos del clúster.

Algoritmo k-means: En el algoritmo k-means, n objetos se agrupan en k agrupaciones en función de características, donde $k < n$ y k es un número entero positivo. La agrupación de objetos se realiza minimizando la suma de cuadrados de distancias, es decir, una distancia euclidiana entre los datos y el centroide del grupo correspondiente.

Reglas de asociación.

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro un conjunto de transacciones, en concreto, ítems o atributos que tienden a ocurrir de forma conjunta.

A cada uno de los eventos o elementos que forman parte de una transacción se le conoce como ítem y a un conjunto de ellos itemset. Una transacción puede estar formada por uno o varios ítems, en el caso de ser varios, cada posible subconjunto de ellos es un itemset distinto. Podemos representar una BDD Transaccional con las siguientes métricas de interés: lista, representación vertical y representación horizontal.

Lista: Básicamente representa cada transacción como una fila, cada fila lista los artículos comprados por el consumidor y cada fila es una transacción por lo que cada fila puede tener un número diferente de columnas.
Representación vertical: Es la forma más eficiente de guardar los datos de tamaño más industrial o comercial, este ocupa solo 2 columnas, indica el número o ID de la transacción y el artículo.
Representación horizontal: Se representa como una matriz binaria, cada fila de una matriz representa una transacción, y cada columna representa un articulo, si un artículo está presente se representa como 1 si un artículo está ausente se representa como 0.

Soporte (Frecuencia relativa): Dada regla si A=>B, el soporte de esta se define como el número de veces o la frecuencia relativa con que A y B aparecen juntos en una BDD Transaccional.

Confianza (Probabilidad empírica): Dada una regla si A=>B, la confianza de esta regla es el cociente de soporte de la regla y soporte del antecedente solamente. Confianza (A=>B) = Soporte (A=>B)/Soporte (A) Si el soporte mide la frecuencia, Confianza mide la fortaleza de la regla. Confianza (A=>B) = P(B/A).

:
Lift (A→B) = Soporte(A→B)/(Soporte (A) * Soporte(B)) , si Lift = 1 o muy cerca a 1, indica que la relación es producto del azar de lo contrario, indica que la relación es realmente fuerte.

Apriori: uno de los primeros algoritmos desarrollados para la búsqueda de reglas de asociación tiene dos etapas: Identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado límite (itemsets frecuentes). Convertir esos itemsets frecuentes en reglas de asociación.

Prediccion.
Consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento.

Modelo predictivo: Este modelo predictivo se podrá utilizar para predecir qué probabilidades hay de que una persona reaccione de una manera determinada. Una vez introducidos los datos de la persona y se aplique el modelo predictivo se obtendrá una calificación que indicará la probabilidad de que se produzca la situación estudiada por el modelo.

Técnicas aplicables al análisis predictivo:
Técnicas de regresión: regresión lineal, Árboles de clasificación y regresión, Curvas de regresión adaptativa multivariable.
Técnicas de aprendizaje computacional: Redes neuronales, Máquinas de vectores de soporte, Naïve Bayes, K-vecinos más cercanos.

Árboles de clasificación y regresión: Los árboles de clasificación y regresión son una técnica de aprendizaje de árboles de decisión no paramétrica que produce árboles de clasificación o regresión, dependiendo de si la variable dependiente es categórica o numérica, respectivamente.
Cada observación cae en un nodo terminal, y cada nodo terminal es definido de manera única por un conjunto de reglas.

Técnicas de aprendizaje computacional:
Redes neuronales: Las redes neuronales son técnicas de modelado no lineal sofisticadas que son capaces de modelar funciones complejas. Pueden aplicarse a problemas de predicción, clasificación o control en un amplio espectro de campos como las finanzas, la psicología cognitiva/neurociencia, la medicina, la ingeniería y la física. Las redes neuronales se utilizan cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y de salida. Una característica clave de las redes neuronales es que aprenden la relación entre los valores de entrada y salida a través del entrenamiento.

Máquinas de vectores de soporte: Las máquinas de vectores de soporte (SVM) se usan para detectar y explotar patrones complejos de datos agrupando, ordenando y clasificando los datos. Son máquinas de aprendizaje que se utilizan para realizar clasificaciones binarias y estimaciones de regresión. Usualmente usan métodos basados en kernel para aplicar técnicas de clasificación lineal a problemas de clasificación no lineal. Hay una serie de tipos de SVM tales como lineal, polinomial, sigmoide, etc.

Naïve Bayes: El clasificador bayesiano ingenuo se basa en la regla de probabilidad condicional de Bayes, que se utiliza para la tarea de clasificación. El clasificador bayesiano asume que los predictores son estadísticamente independientes, lo que hace que sea una herramienta de clasificación eficaz que sea fácil de interpretar. Se emplea mejor cuando se enfrenta al problema de la "maldición de la dimensionalidad", es decir, cuando el número de predicciones es muy alto.

K-vecinos más cercanos: El algoritmo vecino más próximo k-NN (Nearest Neighbor) pertenece a la clase de métodos estadísticos de reconocimiento de patrones. Se trata de un conjunto de entrenamiento con valores positivos y negativos. Una nueva muestra se clasifica calculando la distancia al vecino más cercano del conjunto de entrenamiento. El signo de ese punto determinará la clasificación de la muestra. En el clasificador kvecino más cercano, se consideran los k puntos más cercanos y se utiliza el signo de la mayoría para clasificar la muestra. El rendimiento del algoritmo k-NN está influenciado por tres factores principales: la medida de distancia utilizada para localizar a los vecinos más cercanos, la regla de decisión usada para derivar una clasificación de los k-vecinos más cercanos y el número de vecinos utilizados para clasificar la nueva muestra.