-Prediccion

En este tema se repaso la técnica de arboles y bosques aleatorios, los arboles se partieron en 2:

1.- Arboles de regresión: Consiste en mediante particiones de los datos (por medio de preguntas del tipo “si”) partir el conjunto entero de datos en diferentes categorías numéricas, si bien es bueno prediciendo falla en el ámbito de que aunque predice una variable continua sus predicciones serán “categóricas” osease, muchas observaciones pueden caer en exactamente la misma predicción simplemente por que es la que más se aproxima a su valor real.

2.- Arboles de clasificación: Mediante el mismo método anteriormente mencionado se parten las variables en categorías especificas , la ventaja de esto es que ya no presentaremos el problema de pasar de variable continua a categórica debido a que la variable a predecir será categórica.

Bosques de regresión:

Los bosques de regresión son una extensión de los arboles, consiste en el hecho de hacer un muestreo aleatorio con reemplazo múltiples veces y en cada una de estas observaciones generar un árbol (Ya sea de regresión o de clasificación dependiendo lo que corresponda), una vez que ya se tienen todos los arboles deseados cuando se quiere predecir un nuevo valor se le pregunta a todos los arboles y se regresa la predicción que haya elegido la mayoría.

Para medir la efectividad de estos modelos se utilizan dos técnicas dependiendo si es categórica o continua la variable a predecir.

De ser categórica utilizaremos la curva ROC y de ser continua utilizaremos el MSE (Error cuadratico medio)

-**Regresion Lineal**

Consiste en aproximar una serie de observaciones mediante una línea recta que capture la mayor parte de la varianza de las observaciones. Esto es en su forma más simple (con un solo predictor) tu quieres aproximar todas tus observaciones con una ecuación de la forma y=m*x+b donde es m y b los valores que tu estas intentando averiguar, este modelo se basa en la minimización del MSE (error cuadratico medio) que es simplemente el promedio de los errores al cuadrado ¿Por qué al cuadrado? Supongamos que yo tengo una observación predicha de la forma x_2 y estoy intentando predecir una observación x el error seria x-x_2 sin embargo si x_2 fuera mucho mayor que x el error seria negativo y al sumarlo junto con los demás errores puede que ese error negativo se “coma” a un error positivo, por lo tanto elevamos al cuadrado para evitar que esto suceda y hacer que todas nuestras observaciones tengan una medida de error positiva.

Otra métrica para medir la efectividad de este modelo es el coeficiente R^2 ajustada, el cual nos dice que porcentaje de la varianza explica nuestro modelo, buscamos que este porcentaje sea lo mayor posible aunque con un ajuste del 80% ya es debatiblemente un modelo muy bueno.

Cabe mencionar que si bien se llama regresión lineal se puede generalizar para hacer

regresiones ya sean exponenciales , logarítmicas o incluso parabólicas mediante lo que se conoce como modelos linealizables, lo cual es simplemente una transformación que le haces a una función para que esta pueda ser expresada como una línea recta, por ultimo aunque se discutió el caso simple, no hay que limitarnos debido a que la regresión lineal también acepta múltiples predictores, en este caso ya no estarías prediciendo una función de la forma y=m*x+b si no una función con múltiples variables estilo y=m1*x1+m2*x2… +m(n)*x(n)+e donde e representa al error esperado de la regresión (el cual no se puede reducir).

-Reglas de asociación

El análisis de reglas de asociación es la predicción en su nivel más primitivo/puro, consiste en hacer preguntas de tal forma que llegues a una relación que consiste en "Si A=>B" esto con el fin de

1: medir relaciones entre objetos que ya conozcamos

2: Encontrar relaciones fuertes entre objetos que desconozcamos

(Punto importante aquí "objeto" se refiere a lo mismo que en muestreo (Todo lo que podamos estudiar es un objeto)) existen diferentes tipos de asociación

1: Asociación cuantitativa :

- 1.1: Asociación booleana : Asociaciones entre presencia o ausencia de algo
- 1.2: Asociación cuantitativa : Asociaciones entre cantidades de algo

2: Asociación multidimensional:

- 2.1: Asociación unidimensional: Si A => B
- 2.2 : Asociación multidimensional: SI A Y B Y C …. => Z

Nos interesan principalmente 3 metricas:

1.- Soporte: Supongamos que tenemos una regla de la forma A => B el soporte de esa regla serán todas las observaciones donde se cumple esa regla entre todas las observaciones.

2.- Confianza: Siguiendo con la misma regla , su confianza es el soporte de la regla / el soporte de A

3.- Lift : Soporte de A => B / (Soporte (A) * Soporte (B)) , básicamente es "las veces donde ocurrió nuestra regla/ todas las veces que pudo haber ocurrido"

Clustering

El clustering engloba un conjunto de técnicas cuya finalidad es encontrar grupos de observaciones que se comporten de manera similar, si bien este comportamiento puede ser estudiado por el analista cabe mencionar que los algoritmos de clustering la respuesta primaria que te dan es el hecho de cuales objetos comparten una relación no la relación en si, de esta manera puedes encontrar diferentes grupos que se comportan de manera similar pero entender como se comportan y por que se comportan así ya es trabajo para el analista.

El algoritmo más básico y por lo tanto el preliminar en este tema es el de k-medias, consiste en primero seleccionar un numero k este numero será cuantos grupos nosotros creemos que existen de manera preliminar, después de esto se generan k puntos

aleatorios sobre el plano/hiperplano que estemos estudiando , una vez generados estos puntos se hace un mapa de voronoi dentro del hiper plano, donde cada observación es “enlazada” / asociada con el punto al cual pertenezca su zona en el mapa de voronoi (osease se asocia con el punto más cercano) una vez que todos los objetos están asociados a un punto se obtiene el centro de masa de todos los objetos(lo cual es solo un “punto promedio” de todos los objetos) ese será el nuevo punto, se repite esto para los k puntos y se regresa al primer paso, este proceso iterativo se repite hasta que los puntos ya no se muevan de sus posiciones con respecto al primer paso.
Una vez terminado el algoritmo uno puede estudiar la varianza del sistema para determinar que tan buena decisión fue la k elegida, por ejemplo nosotros queremos que dentro de cada grupo la varianza sea mínima debido a que esto significaría que todos se parecen lo más posible sin embargo no podemos simplemente minimizarla por que la forma más directa para hacerlo es: si tenemos n puntos elegimos k=n de tal manera que cada cluster sea un punto por si mismo y así todos los clusters tengan varianza 0, para elegir la k adecuada lo que se hace es ir aumentadola hasta que la varianza del sistema entero no haya aumentado gran porcentaje con respecto a la iteración anterior (este método se conoce como el método del codo)

Visualización

La visualización de datos es una parte clave dentro del análisis de datos, debido a que muchos análisis tienen ya sea como respuesta o como análisis preliminar una visualización, esto nos permite saber a simple vista un resumen rápido ya sea de lo que vamos a estudiar o de lo que acabamos de estudiar.
Existen muchos tipos de visualización (incluso algunas que son interactivas) por lo general se suele empezar con tres, el diagrama de dispersión (el cual es una graficacion directa en caso de tener solo dos variables (si tenemos más variables se suele hacer un diagrama de dispersión para cada par)).
El histograma: Este nos permite conocer tanto la distribución de los datos como si hay picos de observaciones, en general esto nos da un resumen muy bueno de la probabilidad de que exista una nueva observación tal cual como la vimos.
El boxplot: Este es tal cual una grafica que nos permite conocer la existencia de datos atípicos para así poder clasificarlos (pero se hablara más de esto en el tema de ouliers).
Existen diversos estándares para la visualización de datos y por lo general actualmente todo se hace en línea por lo tanto los siguientes lenguajes pertenecen al estándar para presentar tus datos: HTML5 (Dibujar/graficar en 2d) , CSS3 (Permite diferenciar contenido) SCV(Utilizado para crear graficas en 2d), WEBGL (utilizado para graficas 3d).
La visualización es un área muy importante en el análisis de datos debido a que es parte clave en diversos algoritmos / predicciones, por ejemplo la regresión lineal parte de observar una relación lineal en las variables, osease desde la grafica tu como analista puedes decidir si vale la pena o no hacer una regresión lineal, o incluso desde la visualización uno es capaz de conocer vagamente temas como la correlación de dos variables o incluso la varianza total de las observaciones o el sesgo o la curtosis.

Clasificación

El tema de clasificación consiste básicamente en la predicción de variables cualitativas, esto es, dada una observación tu quieres poder predecir si va a pertener a una clase especifica o no (o incluso la probabilidad de que pertenezca).

Con este fin se utilizan diversos enfoques tales como los arboles de clasificación mencionados anteriormente, la regla de bayes la cual consiste en ir actualizando la probabilidad de la existencia de una relación del tipo A=> B conforme se encuentran más observaciones de A ya sea con B o sin B todo este enfoque basado en la regla de bayes de la probabilidad que existe en la estadística bayesiana, también se pueden utilizar redes neuronales estas consisten en mediante calculo encontrar el punto donde tu predicción sea lo más correcta posible, por ejemplo tu le puedes proporcionar a la maquina una serie de observaciones con la respuesta correcta por ejemplo una cantidad enorme de datos sobre clientes junto con la variable binaria "compro" la cual los clasifica entre compradores o no compradores, después de esto la maquina se "entrena" así misma para intentar reconocer las relaciones entre los que si compraron y los que no compraron (esto puede llevarse mediante diversos métodos ya sea reducción del gradiente o aprendizaje supervisado o por incentivos) posteriormente la maquina ya conoce un modelo el cual no es conocido por el que la programo debido a que este es un enfoque de caja negra (osease tu no puedes ver lo que hay adentro (no ves el modelo como tal)) solo ves los resultados del mismo.

Existen muchos más métodos para clasificación sin embargo los mencionados anteriormente son usualmente los más usados.

Patrones secuenciales

Este tema es una extensión/ una generalización del tema de reglas de asociación , en este caso nos interesa encontrar reglas de asociaciones más largas, osease ya no buscamos un A => B , ahora se podría decir que buscamos A=> B => C => D … Nos interesa buscar un patron que se repita para así poder hacer predicciones desde el propio comienzo del patrón , al ser una generalización de reglas de asociación las mismas métricas usadas allá son validas aquí por lo tanto no se hará énfasis en eso, sin embargo este tema tiene aplicaciones distintas al de reglas de asociación, por ejemplo gracias a patrones secuenciales es que Outlook reconoce cuando un correo es spam o no , debido a que busco diversas características que se repiten a lo largo de todos los correos que si son spam, también se puede estudiar las relaciones en este tipo de patrones debido a que se pueden expresar como función de tiempo , supongamos que nosotros tenemos una relación estándar A => B , en este le ingresamos el factor tiempo como función , dígase que el evento A(T) significa que el evento A sucedió en el instante T bueno , si tenemos una relación de la forma A=> B los patrones secuenciales nos permiten generalizarlo a una función de la forma A(t) => B(t+n) donde t es una variable pero n es un numero conocido, por eso es que nos permite prepararnos para eventos futuros debido a que nos permite dar con una métrica para cuando sucederán los eventos futuros con base en los eventos

del presente, esta área es usada ampliamente en zonas como la de los deportes o la farmacobióloga, o incluso en GPS.

Outliers

Los outliers (También llamados datos atípicos) son datos que se salen de lo estándar, por ejemplo si tenemos una secuencia de 10000 datos positivos y de repente nos aparece un dato negativo pues eso seria un outlier, o por ejemplo si tenemos 100 1's y de repente aparece un 2 tambien seria un dato atipico, el problema chace en dos ámbitos de esto:

1.- ¿Qué tan atipico se puede ser? ¿Cómo podemos medirlo? Los ejemplos que puse son claros y es muy sencillo de ver pero cuando la secuencia a estudiar no es tan homogénea esta línea entre "atipico" y "todavía no descubro la secuencia entera" no es tan fácil de ver.

2.- ¿Qué tanto nos afectan? Esta claro que si todas las observaciones entran en un rango del 1 al 10 y de repente una es 10,000 y hago un modelo con todos esos datos la de 10,000 me va a cargar el modelo hacia ella sin embargo si la observación no es tan atípica puede que no cargue al modelo tanto peso ¿Dónde dibujamos la línea?

Para responder estas preguntas utilizamos lo que se conoce como pruebas estadísticas no paramétricas, para determinar si un punto es un outlier o no y en caso de serlo si vale la pena o no eliminarlo del modelo.

Si bien las pruebas estadísticas están bastante bien un enfoque directo podría ser hacer un boxplot, este mediante el rango intercuartilico (la distancia entre el primer y tercer cuartil multiplicado por 1.5) nos dice si algo es atipico o no, por ejemplo si mi primer cuartil es 10 y mi tecero es 15 entonces mi rango intercuartílico es de 5 (15-10) así, los limites de aceptación para los valores para ser considerados normales serian Q1-5*1.5 , Q3+5*1.5 (donde Q1 es el primer cuartil (ósea 10) y Q3 es el tercero (ósea 15)) entonces mis rangos aceptables serian 10-7.5 y 15+7.5 osease (2.5,22.5) todo lo que este fuera de este rango seria considerado atipico, si bien como dije anteriormente el boxplot es un enfoque un poco muy simple, sirve para darse una idea general de los datos antes de aplicar las pruebas no paramétricas.