

PREDICCIÓN

OMAR ALEJANDRO DELGADO LOZANO
JOSE PEDRO TREVIÑO HERNANDEZ





El análisis predictivo consiste en la tecnología que aprende de la experiencia para predecir el futuro comportamiento de individuos para tomar mejores decisiones
– Eric Siegel



ANALISIS PREDICTIVO

El análisis predictivo es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado, presente o futuro



27% anual

La demanda de especialistas
en Análisis Predictivo crece
en EEUU

11%.

la media en dicho país para
el resto de demandas





Las cifras anteriores ilustra perfectamente la enorme importancia que administraciones, empresas y organizaciones están otorgando al Análisis Predictivo. Esta tendencia al uso del Análisis Predictivo es consecuencia de la nueva cultura que se ha generalizado con respecto a los datos. La capacidad real de almacenar y procesar grandes conjuntos de datos, ligada a los avances experimentados por las TI, ha permitido generar archivos masivos de datos de todo tipo, susceptibles de ser analizados en busca de tendencias.

ANÁLISIS PREDICTIVO

Para llevar a cabo el análisis predictivo es indispensable disponer de una considerable cantidad de datos, tanto actuales como pasados, para poder establecer patrones de comportamiento y así inducir conocimiento.

Los datos son la fuente de la que se obtienen las variables, las relaciones entre ellas, el conocimiento inducido o los patrones de comportamiento identificados, convirtiéndose en un elemento vital de todo análisis predictivo.





En la actualidad se crean más
datos en un día de los que se
crearon en toda la humanidad
hasta el año 2.000

– Andreas Weingend



PROVEDORES DE DATOS

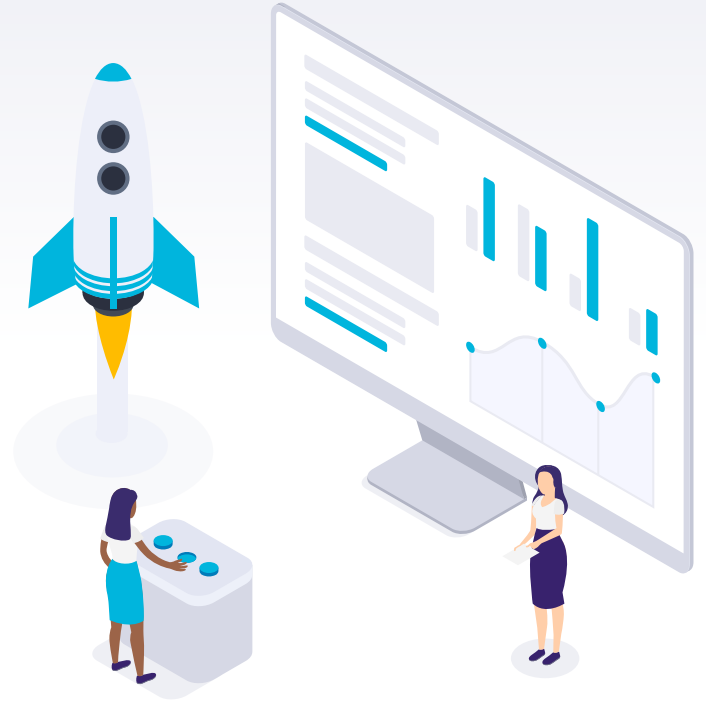


Con la generalización de las Tecnologías de la Información ha aparecido una nueva dimensión en la que contemplar a las personas. Si antes podían ser vistas como ciudadanos, contribuyentes o consumidores (entre otras visiones), las TI permiten contemplar a las personas como proveedores de datos.

Actos como conducir o caminar con un dispositivo capaz de geoposicionar a su usuario, pagar con una tarjeta de crédito o ver una serie online, generan información susceptible de ser explotada. Enviar correos electrónicos, interactuar en las redes sociales o, simplemente, utilizar motores de búsqueda, también genera datos.

1

Predicciones



MODELO PREDICTIVO

Este modelo predictivo se podrá utilizar para predecir qué probabilidades hay de que una persona –en función de los datos que se disponga de la misma– reaccione de una manera determinada (si comprará un producto, si cambiará de voto, si contratará un servicio...). 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.



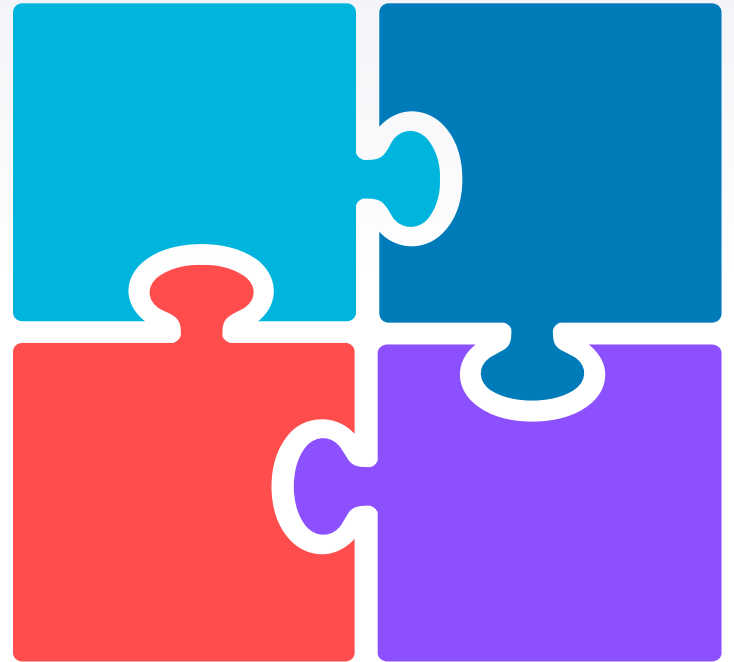
Técnicas aplicables al análisis predictivo

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



1

Técnicas de regresión



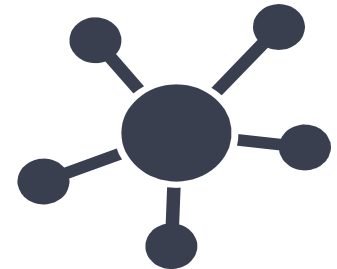
Regresión lineal




El modelo de regresión lineal analiza la relación existente entre la variable dependiente o de respuesta y un conjunto de variables independientes o predictoras. Esta relación se expresa como una ecuación que predice la variable de respuesta como una función lineal de los parámetros. Estos parámetros se ajustan para que la medida de ajuste sea óptima

Árboles de clasificación y regresión

- ▶ Los árboles de clasificación y regresión (Classification And Regression Trees, CART) 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.
- ▶ Los árboles de decisión están formados por una colección de reglas basadas en variables en el conjunto de datos de modelado:



- 
- ▶ Las reglas basadas en valores de variables se seleccionan para obtener la mejor división para diferenciar observaciones basadas en la variable dependiente.
 - ▶ Una vez que se selecciona una regla y divide un nodo en dos, se aplica el mismo proceso a cada nodo "secundario", es decir, es un procedimiento recursivo.
 - ▶ La división se detiene cuando CART detecta que no se pueden realizar más ganancias o se cumplen algunas reglas de parada preestablecidas.

Cada rama del árbol finaliza en un nodo terminal. Cada observación cae en un nodo terminal, y cada nodo terminal es definido de manera única por un conjunto de reglas.

► Curvas de regresión adaptativa multivariable



- Las curvas de regresión adaptativa multivariable (Multivariate Adaptive Regression Splines, MARS) son una técnica no paramétrica que construye modelos flexibles al ajustar regresiones lineales por piezas. Un concepto importante asociado con curvas de regresión es el de un nudo. Un nudo es donde un 20 Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso Carlos Espino Timón modelo de regresión local da paso a otro y por lo tanto es el punto de intersección entre dos curvas.

- ▶ En las curvas de regresión adaptativa multivariante, las funciones de base son la herramienta utilizada para generalizar la búsqueda de nudos. Las funciones básicas son un conjunto de funciones utilizadas para representar la información contenida en una o más variables.
- ▶ La curva de regresión adaptativa multivariable es un modelo que primero realiza un sobreajuste y luego hace una poda para obtener un modelo óptimo. El algoritmo es computacionalmente muy intensivo y en la práctica se requiere especificar un límite superior en el número de funciones de base.

1

Técnicas de aprendizaje computacional



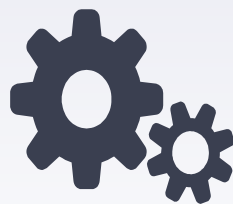


Redes neuronales



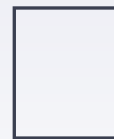
- ▶ Las redes neuronales son técnicas de modelado no lineal sofisticadas que son capaces de modelar funciones complejas. Pueden aplicarse a problemas de predicción, clasificación o control en un amplio espectro de campos como las finanzas, la psicología cognitiva/neurociencia, la medicina, la ingeniería 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

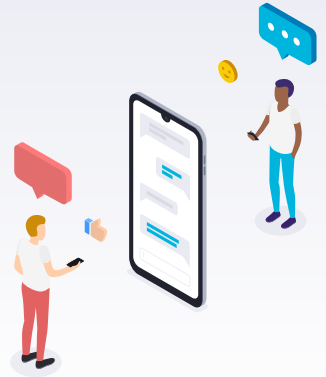
▶ Naïve Bayes




- ▶ 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. El método no impone a priori ninguna suposición sobre la distribución de la que se extrae la muestra de modelado. 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 k-vecino 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 • el número de vecinos utilizados para clasificar la nueva muestra.

► Principales aplicaciones del Análisis Predictivo

- ▶ 1. Segmentación de clientes
- ▶ 2. Personalización de la oferta
- ▶ 3. Detectar el riesgo de que el cliente abandone la relación comercial.
- ▶ 4. Conocer cuáles son los clientes más propensos a responder a las iniciativas de comunicación publicitaria
- ▶ 5. Conocer la tasa de deserción;

► Principales aplicaciones del Análisis Predictivo

- Publicidad predictive
- Predictor de embarazos
- Persuasión del voto en campañas electorale
- Detector de fraude
- Compraventa de acciones
- Estimación del valor hipotecario
- Reducción de reincidencia
- Retención de clients



Gracias..!

