Encuentros 15. Análisis predictivo y la regresión lineal - Avances PIN

Tema 1. ¿Qué es el análisis predictivo? Interpolación y extrapolación
Tema 2. Modelos. Problemas de regresión, clustering y clasificación

Mapa mental del encuentro.
El análisis predictivo, como su nombre lo indica, busca predecir acciones futuras a partir del estudio de determinadas variables. En este módulo nos introduciremos en el análisis predictivo donde veremos los distintos problemas que se pueden resolver.

Distinciones que se adquieren en el encuentro

+ Problemas de regresión, clustering y clasificación

1 hora >> presentación
2 hora >> práctica
3 hora >> hackatón

Análisis predictivo

El **análisis predictivo** es el proceso de **realizar predicciones sobre una población basándonos en una muestra**. El término “análisis predictivo” describe la aplicación de una técnica estadística o de aprendizaje automático para crear una predicción sobre el futuro. En este proceso se crean **modelos predictivos** para predecir eventos futuros.

El **modelado predictivo** utiliza métodos matemáticos y de cálculo para predecir un evento o un resultado. Estos modelos pronostican un resultado en algún estado o tiempo futuros en función de los cambios en las entradas del modelo.

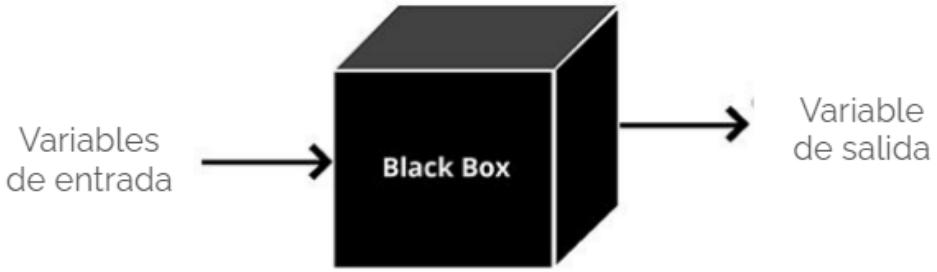
La motivación para modelar es buscar respuestas a preguntas generales:

- Se quiere predecir las ventas de juguetes conociendo la inversión publicitaria.
- Se quiere estimar el riesgo de padecer neumonía a partir de variables clínicas de los pacientes.
- Se desea clasificar mensajes como spam, mediante reconocimiento de patrones.
- Se desea estimar el impacto de una campaña publicitaria de una empresa de bebidas en las ventas.

Para modelar una variable **necesitamos contar con información** de la variable de estudio y de una o varias variables que están relacionadas a nuestra variable de interés. La relación

entre ellas es una **relación causal**: una existe por la otra, entre ellas guardan una relación intrínseca que conocemos y se encuentra

Podemos imaginarnos un modelo como una caja negra, que transforma las variables de entrada para darnos la variable de salida:



Matemáticamente, un modelo es una función matemática *f()* y la podemos escribir de la siguiente forma:

$$y = f(x_1, x_2, x_3.. x_n)$$

donde *y* es la variable que queremos modelar y $x_1, x_2, ... x_n$ son variables que utilizamos para modelar la variable salida. Tanto *y* como *x* reciben distintos nombres en la bibliografía:

y	**x**
Salida	Entrada
Dependiente	Independiente
Respuesta	Predictora
Explicada	Explicativa
Target	Features

Los modelos se construyen para conocer o predecir propiedades de un objeto o sistema, y **son representaciones simplificadas de la realidad** que muestran algunas de sus propiedades. Para ello, utiliza una porción de la realidad empírica, un recorte que se debe hacer con conocimientos específicos del problema.

Es por esto que **la utilidad de los modelos está condicionada principalmente por:**

- una buena selección de **los factores relevantes para el problema**

- una adecuada descripción de **sus relaciones funcionales**

Modelos, supuestos y errores

Como vimos, los modelos son simplificaciones de la realidad. Para hacer esas simplificaciones tuvimos que hacer suposiciones. A veces esas suposiciones son correctas y útiles, pero otras veces pueden ser erradas. Esto puede verse con claridad en el chiste de la vaca esférica..

> La producción de leche de una granja era tan baja que el granjero pidió a la universidad local ayuda académica. La universidad reunió un equipo multidisciplinar de profesores, encabezado por un físico teórico, y estuvieron dos semanas haciendo investigación de campo intensiva. Los científicos volvieron a la universidad, con sus portátiles repletos de datos, y el encargo de escribir el informe se dejó para el líder del equipo. Poco después, el granjero recibió el informe, que empezaba así: «*Tengo la solución, pero funciona solo en el caso de vacas esféricas en el vacío*».
>
> Harte, John (1988), Consider a Spherical Cow: A Course in Environmental Problem Solving, University Science Books, ISBN 978-0935702583.

¿Cuántas vacas esféricas conocen Uds? La vaca esférica representa esos modelos matemáticos de servilleta, demasiado simplificados, quizás algo ingenuos o un poco torpes. También se conocen como modelos de juguete, porque se utilizan para poner en práctica ciertas habilidades pero no para sacar conclusiones trascendentes. Sin embargo, tienen una gran virtud: permiten hacer cálculos y predicciones, evaluar hipótesis rápidamente y, de esta forma, avanzar en nuestro conocimiento.

Construir un modelo es un proceso iterativo: se realiza un modelo sencillo, mediante un conjunto de datos acotado, se prueba y se valida para determinar su precisión, y si no cumple con lo esperado, se realiza otro modelo un poco más complejo hasta obtener el modelo más efectivo. ¿Nos serviría un modelo perfecto? ¿Existe?

Modelizar es representar la realidad con una cantidad menor de información. Por tanto, **existe un error inherente al proceso de modelización que puede ser reducido pero no eliminado**. La eliminación del error implicaría una perfecta identificación del modelo con el objeto real, y por lo tanto, ¡dejaríamos de tener un modelo y tendríamos la realidad con TODA su complejidad!

En este sentido, **debe buscarse un compromiso entre la complejidad del modelo y el error aceptable en los resultados**. No debemos perder de vista para quién y para qué estamos realizando una modelación, y debe ser una de las primeras preguntas que hagamos: ¿qué tan preciso debe ser mi modelo?

La reducción del error puede hacerse por dos caminos complementarios:

- mayor precisión en la medida y mejor selección de los componentes: no implica mayor complejidad del modelo.
- mayor cantidad de componentes -partes e interrelaciones funcionales-: implica una mayor complejidad del modelo.

Las herramientas que se utilizan para evaluar el error de los modelos se conocen como **métricas de error**. Según los tipos de modelos predictivos que estemos utilizando, serán distintas las métricas de error que precisemos. Aquí presentaremos las métricas de error más comunes, aunque existen muchísimas y muy variadas, resultando muy fina la diferencia entre estas. Por lo general, **resulta conveniente utilizar las métricas de error más utilizadas en cada campo, debiendo utilizarse siempre las mismas métricas para comparar distintos modelos**.

Tipos de modelo predictivos

Podemos realizar una categorización sencilla de los modelos matemáticos según el tipo de variable que queramos modelar:

Modelos de **regresión** -> variable **cuantitativa**
Modelos de **clasificación** -> variable **cualitativa**

Sobre el final de esta clase nombraremos algunos modelos que se escapan de esta clasificación. Sin embargo, ahora comencemos con lo más sencillo..

Modelos de regresión

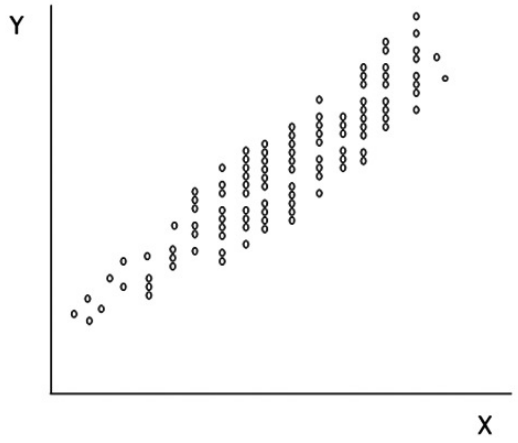
Los modelos de regresión sirven para predecir una variable cuantitativa. Es decir, nos interesa obtener números. Veamos algunos ejemplos de regresión:

- ¿A cuánto se va a vender una propiedad inmobiliaria?
- ¿Cuánto tiempo va a permanecer un empleado en la empresa?
- ¿Cuánto combustible se va a requerir para llegar a destino?
- ¿Cuántos productos se van a vender este mes?
- ¿Cuánto crecerá una acción?

Interpolación y extrapolación

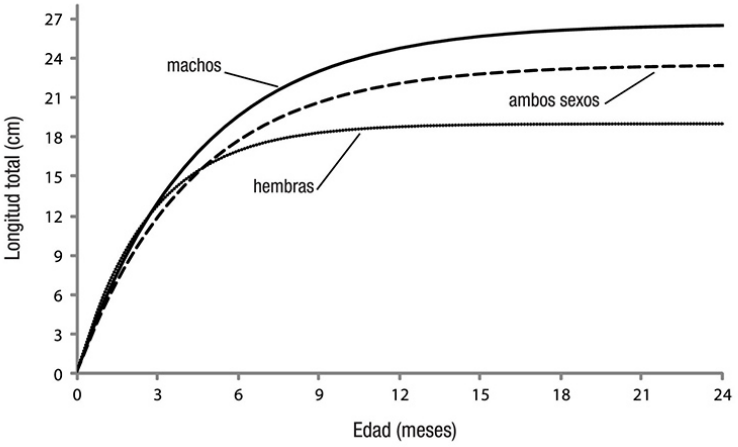
Como mencionamos, para la confección del modelo nos basamos en datos muestrales. Lo ideal es que nuestra muestra se extienda en todo el rango de valores que nuestra variable respuesta puede existir. Es decir, si nos interesa calcular la longitud de un cachorro de perro según su edad, deberemos tener mediciones de cachorros a todas las edades. ¡No una,

muchas! Porque no todas las razas de perros son iguales. Hay perros más grandes, otros más chiquitos. Al juntar muchos datos de cachorros a las distintas edades tendremos una base de datos robusta y de esta forma nos aseguraremos de tener suficientes datos para hacer una buena predicción.



Predecir dentro del rango de valores que consideramos para confeccionar el modelo se conoce como **interpolación**. Mientras que **extrapolación** es la acción de calcular valores fuera del intervalo de los datos conocidos. El riesgo de la extrapolación es que suponemos que fuera del rango estudiado, la variable se comporta igual. Esto no siempre es cierto..

En nuestro ejemplo, los cachorros no crecen siempre de forma constante. La velocidad de crecimiento se va desacelerando a medida que se hacen adultos. Es decir, se va "planchando" la recta, y la gráfica presenta una asíntota que es el tamaño promedio de un perro adulto.



Si nosotros hubiésemos hecho suposiciones para perros adultos, teniendo solo los datos de los cachorros, nos hubiésemos equivocado.

Modelos de clasificación

Mientras que en **los modelos de clasificación obtenemos una clase como resultado**. Los modelos de clasificación sirven para obtener una variable cualitativa. La clase obtenida se encuentra entre un número limitado de clases utilizadas para confeccionar el modelo.

Las clases son categorías arbitrarias definidas según el tipo de problema. Por ejemplo, si queremos detectar si un correo es spam o no, sólo hay 2 clases: “es spam” y “no es spam”. veamos otros ejemplos:

- ¿Comprará el cliente este producto? [sí, no]
- ¿Qué tipo de tumor se observa? [maligno, benigno]
- ¿Subirá el índice bursátil mañana? [sí, no]
- ¿Es este comportamiento una anomalía? [sí, no]
- ¿Nos devolverá este cliente el crédito? [sí, no]
- ¿Obtendrá una historia un número alto de visitas? [sí, no]
- ¿Cuál será el color más vendido? [azul, amarillo, rojo, blanco, verde]

Aunque hay algunas técnicas que son específicas de clasificación y otras de regresión, **la mayoría de los algoritmos pueden utilizarse para ambos tipos de problemas con ligeras modificaciones**.

Matriz de confusión, precisión y sensibilidad

La evaluación de los modelos de clasificación son un poco más complejos que los modelos de regresión (pero tampoco es muy grave!). En estos casos, también utilizaremos como medida del error la exactitud (accuracy), pero también incorporaremos las medidas de precisión (precision) y sensibilidad (recall). Para que su comprensión sea intuitiva, resulta conveniente presentar **la matriz de confusión** para un problema de clasificación binario.

Imaginemos que queremos determinar la capacidad de predecir de un nuevo test de covid. En este caso, nuestro modelo tiene solo dos posibles respuestas: [COVID] y [No-COVID]. A su vez, tenemos un conjunto de sujetos ya examinados por otro método 100% confiable, donde 108 son covid positivo y 100 covid negativos. Luego de examinar a los sujetos, podemos representar los resultados en una matriz de la siguiente forma:

	Predicción		
Observación		Positivos	Negativos
	Positivos	85	23
	Negativos	35	65

donde: 35 son los verdaderos positivos (VP = se predijo de forma correcta que están infectados), 65 los verdaderos negativos (VN = se predijo de forma correcta que no están infectados), 35 son los falsos positivos (FP = = se predijo que estaban infectados cuando no estaban infectados) y 23 los falsos negativos (FN = = se predijo que no estaban infectados cuando sí estaban infectados).

	Predicción		
Observación		Positivos	Negativos
	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

La matriz de confusión nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo. De esta forma, podemos calcular fácilmente las métricas de error:

Exactitud = Número de aciertos (VP + VN) sobre el total de predicciones (VP + FN + FP + VN)

Precisión = Aciertos de positivos (VP) sobre el total de predicciones positivas (VP + FP)

Sensibilidad = Aciertos de positivos (VP) sobre el total observaciones positivas (TP + FN)

Otros tipos de modelos