INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

SITUACIÓN PROFESIONAL 3

CLASE 3 (continuación)

Prof. Ricardo Piña

Ver en Video:

Título: [IA] Clase 3
link: https://www.youtube.com/watch?v=YvbFfM3OIU8&t=1237s (https://www.youtube.com/watch?v=YvbFfM3OIU8&t=1237s)
desde: 34:43
hasta: final

Out[5]:

A esta altura de la carrera Ud. todavía no sabe programar en Python, así que en este archivo hemos ocultado las celdas que contienen código para facilitar su lectura. Si Ud. quiere ver el código u ocultarlo, haga click aquí.

Out[6]:

Situación Profesional: Marketing Bancario

Un banco Portugués está realizando una campaña de Marketing Directo en la cual ofrece a sus clientes realizar una operación de **Plazo Fijo (term deposit)**.
El banco le solicita que aplique técnicas de IA para determinar en función de los datos recogidos qué clientes tomarán el Plazo Fijo y cuáles no lo harán.

Nota: los datos son reales.

El problema de los datos

Recordemos que en Machine Learning el objetivo es aprender a partir de los datos y que como dijimos con anterioridad:

> si entra basura, saldrá basura.

Hay muchos problemas que se nos pueden presentar con los datos, por ejemplo:

Errores en los dispositivos de medición

Supongamos que una variable que deseamos medir es la altura de las personas, puede ocurrir que comencemos midiendo con una cinta métrica y tiempo después utilicemos otra cinta métrica, es muy posible que los resultados fueran dsitintos con uno u otro instrumento de medición.
Por otro lado aún usando la misma cinta métrica si las mediciones son llevadas a cabo por distintas personas es posible que los resultados fueran distintos!

Ruido

Ruido es un término genérico y no siempre hace referencia a ruido sonoro, sino que debe interpretarse más como interferencia producida generalmente por algún elemento distinto al instrumento de medición.

Observaciones con valores faltantes

Puede ser que por diversos motivos nos encontremos con que algunas observaciones estén incompletas, puede deberse a muchos motivos, quizá los datos eran cargados por operarios humanos y se olvidaron de cargar algunos valores o quizá los iban a cargar mañana, o si provenían de un dispositivo informático se produjo una falla, el problema es que tenemos una tabla como la siguiente:

Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Credito
Si	F	-	Si	Si
-	M	No	No	Si
No	-	No	No	No

Hemos marcado la falta de datos con - .

Datos mal cargados

Observe el desastre que puede representar datos cargados de esta manera:

Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Credito
Si	F	Si	Si	Si
Si	M	Sí	No	Si
No	fem	SI	departamento	No
sin dato	Masculino	trabajador	Si	Si

Por suerte problemas como estos se pueden solucionar creand interfaces de carga para humanos que validen los datos que se cargan.

Otros problemas similares pueden darse al utilizar unidades de medidas distintas para valores de la misma columna, suponga que en una columna correspondiente a la altura de una persona, puede ser que alguien cargue 1,78 y otra persona cargue 178. Ambos valores son correctos, pero el primero corresponde a una altura medida en metros y el otro en centímetros!

Los problemas anteriores pueden tener soluciones relativamente simples.

Datos inconsistentes

Este ya es un problema más grave:

Tiene_Deuda	Genero	Trabaja	Propietario	Dar_Credito
Si	F	Si	Si	Si
Si	F	Si	Si	No
No	M	Si	No	No

Observe los primeros dos registros, tienen exactamente los mismos valores para las variables explicativas, pero en un caso la variable a pronosticar tiene valor Si y en el otro valor No. Esos dos registros u observaciones podrían ser inconsistentes porque por ejemplo alguien se equivocó al cargar los datos o bien podría ocurrir que en realidad nos falta utilizar alguna variable más que "separe" estas dos observaciones, como se puede ver a contuinuación:

Tiene_Deuda	Genero	Trabaja	Propietario	Nueva_Variable	Dar_Credito
Si	F	Si	Si	Si	Si
Si	F	Si	Si	No	No
No	M	Si	No	No	No

Esto plantea preguntas sobre cómo descubrir qué otra variable necesitamos?

Los problemas anteriores serán analizados en diversas materias de la carrera, pero ahora quiero referirme a un problema intrínseco a nuestra disciplina y que **siempre** nos acompañará:

Datos escasos

Recordemos el caso del médico que nos pedía que con los datos del análisis de sangre pronosticáramos si el paciente tendría o no determinada enfermedad, supongamos que el médico consiguió datos de mil pacientes, algunos que tenían la enfermedad y otros que no.
Mil datos de pacientes pueden parecernos una buen cantidad, cada paciente tiene una combinación distinta de valores de x_1 y x_2 y nuestro gráfico en dos dimensiones se verá bastante lleno de puntos.
Pero supongamos que el médico era de la Ciudad de Córdoba ... cuántas personas viven en Córdoba? Más de un millón de personas, así que sólo contaríamos con información de una persona cada mil, es decir del 0,1% de los casos posibles en ese momento ya que cada persona tendrá una combinación distinta de valores de estas sustancias.

Ahora nuestros mil casos, parecen ser bastante escasos!
A no ser casos excepcionales,

> **nuestros datos siempre corresponderán a una muestra del Universo posible de datos.**

Y lamentablemente **siempre habrá casos fuera de nuestros datos que presenten algún patrón de comportamiento distinto** a los que forman parte de los que poseemos.

Es por eso que nuestros pronósticos **nunca** darán resultados perfectos al aplicarse al la población en general,

> deberemos admitir que **nuestras observaciones no abarcan todos los casos** posibles y que por lo tanto **las predicciones que pueda hacer nuestro modelo**, que ahora consideraremos como una **hipótesis**, estarán sujetas a **errores en el pronóstico** al aplicarlas a casos **no observados** del resto de la población .
>
> Por lo tanto, **entrenar** a nuestro modelo y no someterlo a algún tipo de test con respecto al resto de la población es por lo menos arriesgado y **nunca** lo deberemos hacer.

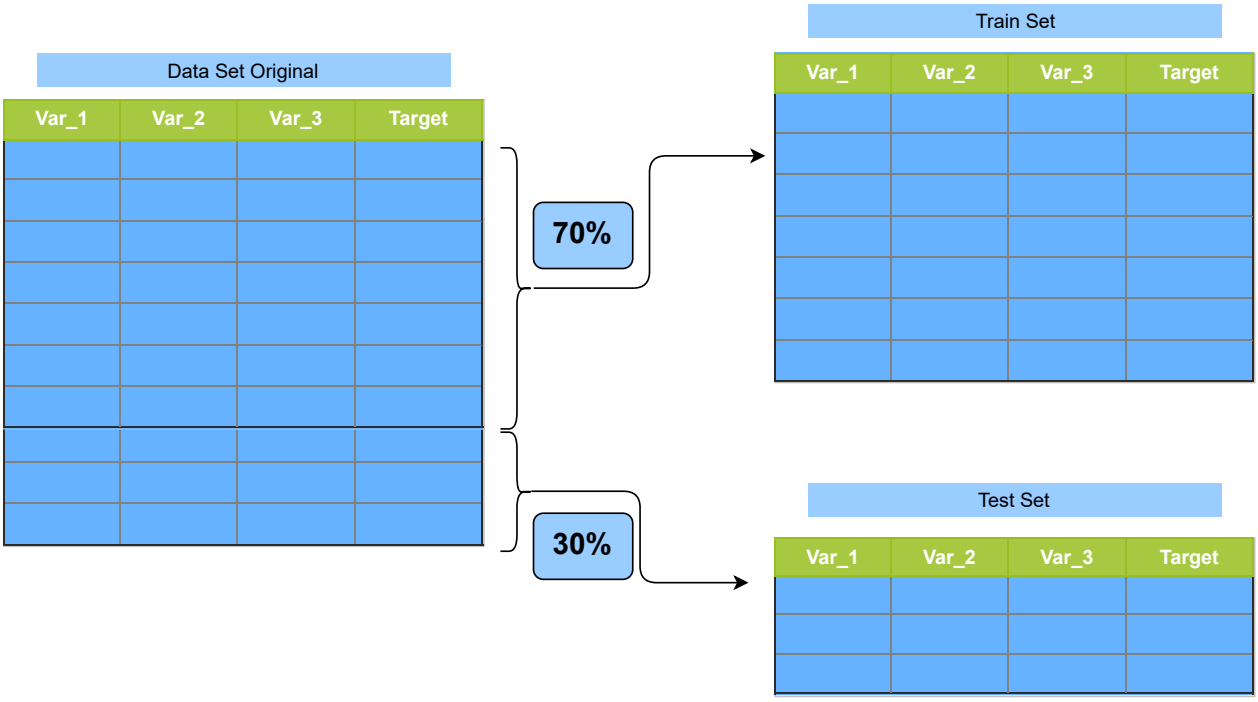
Ahora la pregunta es, cómo podemos saber cómo se comporta nuestro modelo con los casos que no tenemos a disposición? Es decir cómo podemos estimar qué tan bien o tan mal pronosticamos los casos u observaciones que **no conocemos** ?

TESTEO DE HIPÓTESIS (TESTING) en Aprendizaje Supervisado

Existen diversos métodos aplicables al Aprendizaje Supervisado; el que vamos a investigar ahora se usa prácticamente siempre, existen muchas variantes adecuadas para diversos casos, pero que mantienen una misma idea:

- **entrenaremos** a nuestro modelo con un conjunto de observaciones o datos, se los suele denominar **"Train Set"**, con el cual elaboraremos nuestro modelo o hipótesis.
- **testearemos** al modelo o hipótesis con **otros** datos que **no** formaron parte del entrenamiento; se los suele denominar **"Test Set"**.
- **Nunca debemos testear nuestro modelo con los datos de los que aprendió**
- utilizaremos algún criterio **previamente establecido** para **medir** qué tan bien pronosticó nuestra hipótesis en el test set y si no alcanzó los valores deseados tomaremos alguna medida para cambiar el modelo o hipótesis.

Ahora bien, de dónde saldrán los datos del Test Set?



En la figura anterior observamos que dividimos el Conjunto de Datos Original en dos partes:

- un **70%** de las observaciones las utilizaremos como Train Set, es decir para **entrenar** a nuestro modelo y obtener una hipótesis
- el **30%** restante lo **apartamos** y lo reservaremos para testear el modelo hipótesis obtenido anteriormente. Estas observaciones representarán los casos no observados. Lo denominaremos **Test Set**.

- Esta división suele hacerse en forma **aleatoria** (random sample) es decir **al azar**, aunque en algunos casos se suelen tomar otras consideraciones que veremos más adelante. La mayoría de los paquetes de software científico traen esta opción, entre ellos Orange3.
- Las proporciones de 70% y 30% o 66% / 33% o 75% / 25%, 80% / 20% son valores bastante **razonables** para la mayoría de los casos, aunque en algunos casos especiales se pueden tomar otras proporciones en consideración.

Por ahora tomaremos estos criterios como válidos.