**PREWORK**

**SESIÓN 04**

**Objetivo:**

En esta sesión aprenderemos a realizar predicciones y regresiones como algoritmos supervisados.

**Pre-requisito 1: generación de un dataset (opcional)**

En esta sesión es importante que busques dos datasets que puedan ser utilizados para realizar predicciones o regresiones. Es importante que tengas una buena cantidad de datos que tengan cierta relación con el paso del tiempo, para que podamos realizar predicciones.

Estos son los requisitos:

* Tus datos deben ser lo bastante numerosos para que puedas realizar predicciones para ellos. Puedes intentar descargar los datos de https://www.quandl.com como son los tipos de cambio de peso y dólar (<https://www.quandl.com/data/BDM/SF60653-Exchange-rate-pesos-per-US-dollar-Used-to-settle-liabilities-denominated-in-foreign-currency-Date-of-settlement-Available-Period-11-14-1991-09-01-2020>)
* Los datos deben de ser numéricos, si no lo son, debes de mapearlos con Pandas utilizando el comando:

*datos.map({‘dato en texto’:valor numérico})*

Por ejemplo:

*datos.map({‘verdadero’:1})*

**Prerequisito 2: Medición de errores en predicciones (obligatorio)**

En la sesión 02 hablamos acerca de medir errores en clasificación. Sin embargo, como puedes intuir, no es lo mismo clasificar que predecir: clasificar consiste en tomar un dato y asignarle cierta clase o categoría, mientras que predecir consiste en estimar qué valor va a tener cierto dato sin haberlo visto.

En la clasificación podemos saber si una categoría dada fue correcta o no, como si fuese un valor binario (acertaste = 1, no acertaste = 0). Sin embargo, en una predicción el error entre la predicción y el valor real puede ser un valor de punto decimal (digamos, te equivocaste por -0.3832).

Esta discrepancia en la medición de errores requiere que tratemos los errores de forma diferente a la clasificación: no existe una matriz de confusión, true positives, true negatives, false positives ni false negatives en un problema de predicción.

Para esta sesión, que será eminentemente de predecir valores, lo que requerimos es aprender un poco acerca de un concepto llamado ***error cuadrático medio, o MSE*** *(mean squared error)*. El MSE es un estimador de la cantidad de los errores y también es conocida como una función de riesgo que mide la diferencia entre los valores esperados y los valores actuales.

A diferencia de otras métricas de desempeño como la medición de falsos positivos, falsos negativos y matrices de confusión, MSE es utilizado para medir el desempeño de los algoritmos de predicción: Cuando calculas los errores, estos pueden ser valores reales con punto decimal sin mayor problema, ya que todo valor de error, positivo o negativo, contribuye a darnos una buena estimación de que tanto funciona tu algoritmo de predicción.

Otra característica crucial de Mean Squared Error es que es una función derivable: Al ser derivable puede utilizarse para entrenar regresores y deducir los pesos sinápticos con el algoritmo de descenso en gradiente. Esto lo veremos a mayor detalle durante la sesión.

Es también importante que repases la sesión 01 y 02, sobre todo las técnicas de separación de datos en entrenamiento/prueba, ya que lo usaremos mucho.

**Pre-requisito 3: Nociones de series de tiempo**

A diferencia de otros seres, los seres humanos -hasta donde sabemos- somos la única especie que tiene noción del futuro. La noción del futuro nace también de las nociones de que con el conocimiento del pasado, podemos tratar de predecir qué nos depara el futuro. Para ello, tomaremos como base el concepto de una serie de tiempo.

Una serie de tiempo consiste en realizar medidas de un valor que cambia cada cierto tiempo. Por ejemplo, la temperatura, el tiempo de ensamblaje en una fábrica, la productividad, o los indicadores de la economía. Cuando tomamos información cada cierto tiempo, y la distancia (en tiempo) es la misma en todas las mediciones (por ejemplo: cada 5 horas tomas la temperatura ambiental) decimos que has ensamblado una serie de tiempo.

Las series de tiempo deben tener ciertas características para ser útiles. Todas las mediciones deben de tener el mismo lapso de tiempo: de nada sirve que midas la productividad tres días seguidos, te olvides de hacer las mediciones, y dos meses después midas de nuevo. Lo que nos interesa en las series de tiempo es saber si existe una especie de “inercia” que nos diga, dadas las condiciones anteriores, si las condiciones actuales pueden ser predichas.

Como puedes intuir, las predicciones son muy necesarias para tomar decisiones. Desde hace siglos los humanos las necesitamos para la supervivencia: El llevar un registro del clima era de vital importancia para saber cuando era época de cosecha y cuando era época de siembra.

Puedes ensamblar una serie de tiempo de prácticamente lo que desees, y con ella, puedes tratar de predecir cómo se comportará un fenómeno el día de mañana, dados N días anteriores. Sin embargo, tienes que tener en consideración que entre más tiempo en el futuro trates de predecir, más difícil será que tus predicciones sean precisas: variaciones muy sutiles pueden provocar comportamientos enteramente diferentes (como hemos visto en el fenómeno llamado efecto mariposa).

**Quiz:**

**¿Por qué usamos MSE en lugar de una matriz de confusión?**

1. **Porque el MSE permite estimar la eficacia de una predicción.**
2. Porque no hay suficientes datos para usar una matriz de confusión.
3. Porque el MSE nos ayuda a estimar la matriz de confusión.
4. Porque el MSE es preciso.

**¿Para qué sirve la función “map” de pandas?**

1. Transformar un tipo de dato en otro.
2. **Transformar una serie de datos en otros datos numéricos.**
3. Generar un mapa de correlaciones.
4. Todas las anteriores.

**¿Por qué nos interesa que una función sea derivable?**

1. Porque es reducible a cero.
2. **Porque nos permite entrenar un algoritmo de aprendizaje.**
3. Porque nos permite encontrar los errores.
4. Porque nos permite modelar sistemas.