CAPÍTULO 02: EVALUACIÓN DE MODELOS

Universidad Industrial de Santander

Escuela de Química

Código: 27661

Recursos: tutoriales, mini-cursos, etc.

Medium
Daily Digest

https://medium.com/

Recursos: tutoriales, mini-cursos, etc.

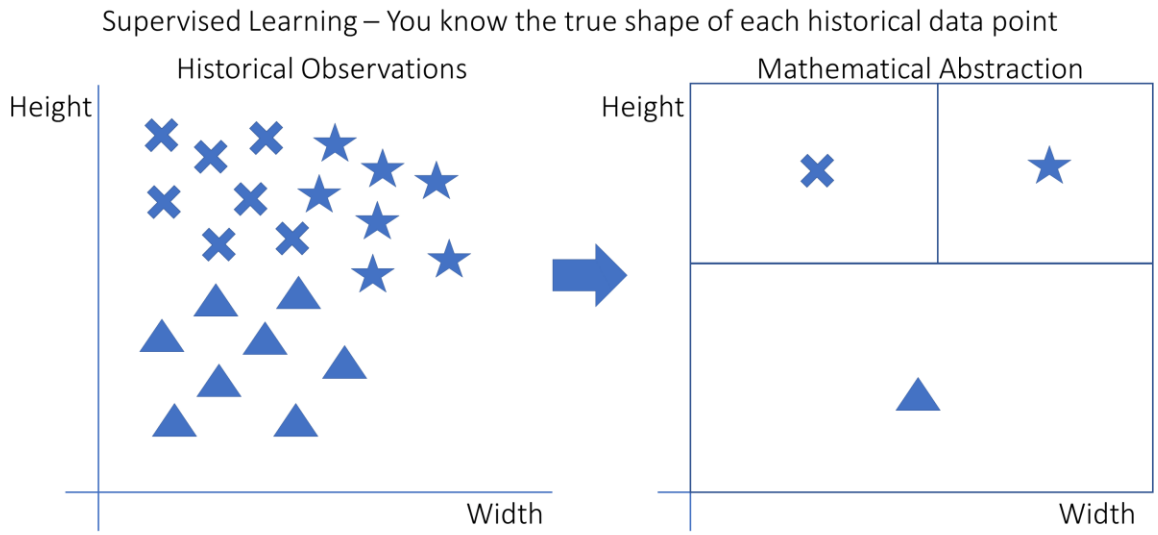
https://realpython.com/

MODELOS PREDICTIVOS

Se denomina Modelo Predictivo a un método de análisis de datos y estadísticas para definir hipótesis o deducir resultados o sucesos futuros y tomar decisiones. El modelado genera predicciones con un grado de probabilidad según las variables analizadas

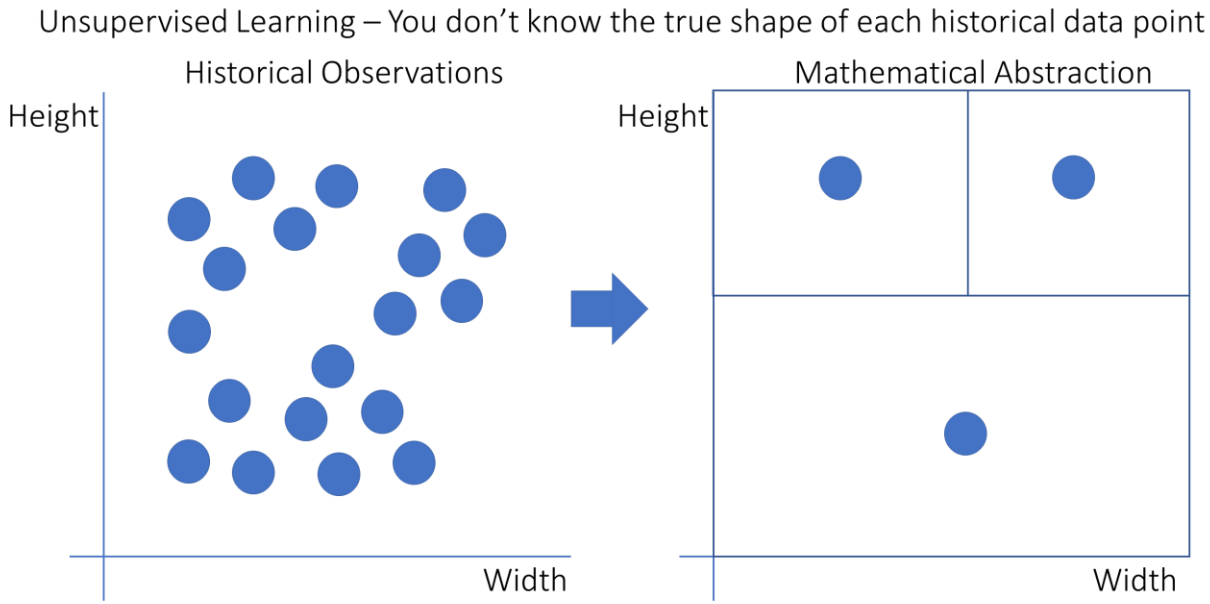
Tipos de Modelo

Los modelos predictivos se pueden dividir en dos clases; **los modelos supervisados y los no supervisados**. La diferencia entre los modelos supervisados y no supervisados es el planteamiento del problema. En los **modelos supervisados**, se tienen dos tipos de variables al mismo tiempo: Una variable objetivo, que también se llama la variable dependiente o la variable ***y*** y la(s) variables independientes, que también se conocen como variables *x*, variables explicativas o predictores.



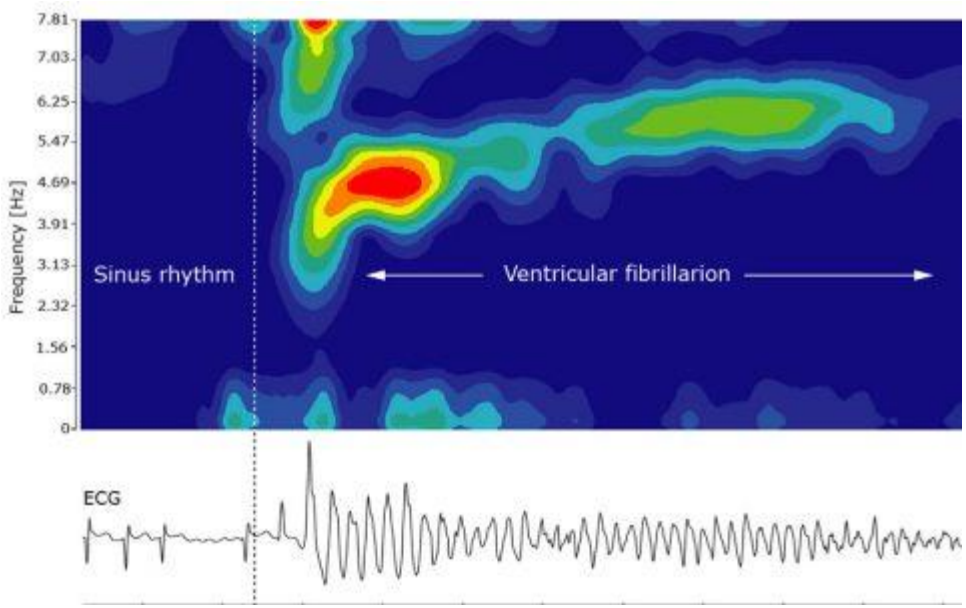
Tipos de Modelo

En los **modelos no supervisados**, no hay una división entre variables objetivo y variables independientes. En **modelos no** supervisado se trata de agrupar los datos evaluando su similitud. Como se puede ver en el ejemplo, nunca se puede estar seguro de que los puntos de datos agrupados pertenezcan fundamentalmente juntos, pero siempre que la agrupación tenga sentido, puede ser muy valiosa en la práctica.



https://realpython.com/knn-python/

Pregunta: Existe un modelo perfecto?
Respuesta: No existe



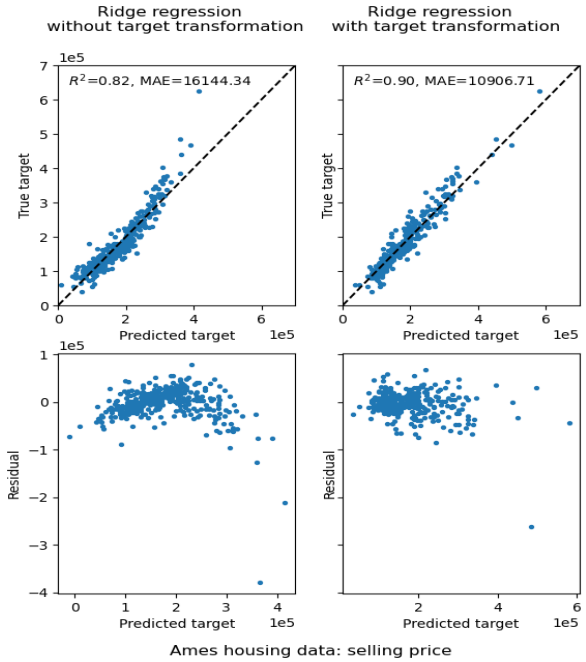
- No hay un modelo único
- Un modelo puede resultar útil para responder a algunas preguntas sobre un conjunto de datos determinado

Calidad del Ajuste

- *A menudo nos preguntamos: ¿qué tan bueno es un modelo?*
- *Lo que queremos decir es: ¿en qué medida coinciden las predicciones del modelo con las observaciones?*
    - ***¿Cómo elegimos el modelo más adecuado?***

Métricas Estadísticas

La selección de los modelos más adecuados se basan en métricas estadísticas que permiten responder las preguntas de arriba. La métricas más usadas son: **Error cuadrático medio (mean_squared_error), Error medio absoluto (mean_absolute_error), Variancia explicada (explained_variance_score) and Coeficiente de regresión (r2_score)**.



Error cuadrático medio (MSE)

La función MSE calcula el error cuadrático medio y corresponde a una métrica que evalúa el riesgo de error del modelo. Si y_{ip} es el valor predicho de la muestra $iesima$ y y_i es la medida de esa misma muestra, entonces el error cuadrático medio (MSE) sobre un número de muestras $n_{samples}$ está definido por:

$$MSE(y_i, y_{ip}) = \frac{1}{n_{samples}} \sum_{i=0}^{n-1} (y_i - y_{ip})^2$$

Training MSE

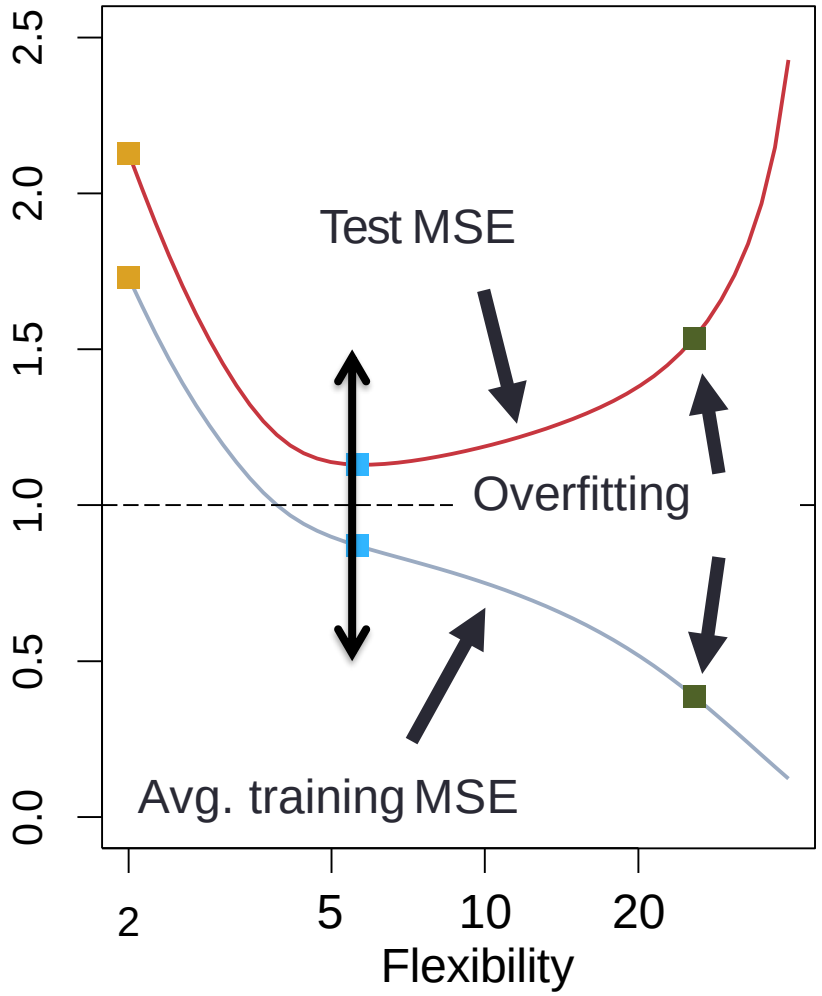
- Esta versión de MSE es calculado usando los **datos de entrenamiento** para el ajuste del modelo

Test MSE

- Evalúa el modelo con datos que no se tuvieron en cuenta en la construcción del modelo

Training vs. test MSE

- flexibilidad ↑:
  - monotono ↓ en training-MSE
  - Forma de U en test-MSE
- **Dato curioso:** se produce independientemente de los datos o del método estadístico
- Esto es llamado **overfitting**



Compromiso entre el sesgo (bias) y la varianza

- La forma de U del Test MSE es el resultado de dos propiedades que compiten: el **sesgo** (bias) y la **varianza**
- **Varianza**: La varianza es lo que varía la estimación en torno a su media.
- **Sesgo (Bias):** el sesgo es lo lejos que está el modelo de la media del valor cierto.
- El **reto**: encontrar un método para el cual, ambos, variancia y sesgo sean bajos
- Este **compromiso** es uno de los temas más importantes en modelamiento.

Error Logarítmico Cuadrático Medio (MSLE)

El MSLE calcula el error logarítmico cuadrático medio y corresponde a una métrica que evalúa el riesgo de error del modelo, especialmente cuando la respuesta tiene un crecimiento exponencial con la variable independiente o predictores. Si y_{ip} es el valor predicho de la muestra $iesima$ y y_i es la medida de esa misma muestra, entonces el error logarítmico cuadrático medio (MSLE) sobre un número de muestras $n_{samples}$ está definido por:

$$MSLE(y_i, y_{ip}) = \frac{1}{n_{samples}} \sum_{i=0}^{n-1} \left(log_e(1 + y_i) - log_e(1 + y_{ip}) \right)^2$$

Error Absoluto Medio (MAE)

El MAE es el error absoluto medio y corresponde a una métrica que evalúa el riesgo de error del modelo. Si y_{ip} es el valor predicho de la muestra $iesima$ y y_i es la medida de esa misma muestra, entonces el error absoluto medio (MAE) sobre un numero de muestras $n_{samples}$ está definido por::

$$MAE(y_i, y_{ip}) = \frac{1}{n_{samples}} \sum_{i=0}^{n-1} |y_i - y_{ip}|$$

Error Máximo (Max Error)

El Max Error calcula el máximo error residual, esta métrica captura el peor error entre la medida y la propiedad. Si y_{ip} es el valor predicho de la muestra $iesima$ y y_i es la medida de esa misma muestra, máximo error está definido por:

$$Max\ Error(y_i, y_{ip}) = \max(|y_i - y_{ip}|)$$

Coeficiente de determinación (R^2)

El coeficiente de determinación, que puntualiza la regresión, conocido como R^2, representa la proporción de varianza (de *y*) que ha sido explicada por la(s) variables independientes del modelo. Por ello esta métrica provee una buena indicación del ajuste y por lo tanto, es una medida de que tan probable es que el modelo prediga las muestras desconocidas, a través de la proporción de la varianza explicada. R^2 con valores cercanos a 1 (incluso negativo) indica un mejor ajuste del modelo y por el contrario su cercanía a cero indicaría lo contrario.. Si y_{ip} es el valor predicho de la muestra $iesima$ y y_i es la medida de esa misma muestra, entonces el R^2 sobre un numero de muestras $n_{samples}$ está definido por::

$$R^2(y_i, y_{ip}) = 1 - \frac{\sum_{i=0}^{n}(y_i - y_{ip})^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}$$

Muestreo

La construcción de un modelo predictivo consiste en crear un programa que sea capaz de generalizar a muestras de entrada que nunca ha visto antes los resultados que puede producir. Esta tarea requiere exponer el modelo - durante el entrenamiento (ajuste) a un cierto número de variaciones de ejemplos de entrada, lo que probablemente conducirá a una mejor precisión del modelo. Esto incluye múltiples pasos por los que el modelo tiene que pasar antes de estar disponible para su uso:

La estrategia tradicional consiste en dividir en 3 partes en el conjunto de datos una para entrenar *(training dataset),* una segunda para probar el modelo (*test dataset*) y una final para validar (*validation dataset*).

Training dataset - se utiliza para entrenar el modelo y ajustar los parámetros del modelo
Test dataset - se utiliza para evaluar si el modelo es lo suficientemente generalizado como para funcionar correctamente con datos con los que no se ha entrenado.
Validation dataset - Se utiliza como la comprobación final de que el modelo es capaz de generalizar con datos previamente desconocidos.

Pre-procesamiento

En cualquier proceso de aprendizaje automático *(machine learninig) o desarrollo de modelos predictivos,* el **pre-procesamiento** de datos es el paso en el que los datos se transforman, o se codifican, para llevarlos a un estado tal que puedan ser analizarlos fácilmente. En otras palabras, las características de los datos pueden ser interpretadas fácilmente por el algoritmo.

Veamos cuales son los métodos de pre-procesamiento de datos mas generalmente usados. No todos los métodos son aplicables para cada problema, depende en gran medida de los datos con los que estamos trabajando, por lo que tal vez sólo unos pocos pasos pueden ser necesarios con su conjunto de datos. Generalmente son :

- **Evaluación de la calidad de los datos**
- **Agregación o agrupamiento**
- **Muestreo**
- **Reducción de la dimensionalidad**
- **Codificación**
- **Separación de datos**

Pre-procesamiento

- **Evaluación de la calidad de los datos**

Debido a que los datos se toman a menudo de múltiples fuentes no es realista esperar que los datos sean perfectos. Puede haber problemas debidos a errores humanos, limitaciones de los dispositivos de medición o fallos en el proceso de recogida de datos. Los principales son, ***datos perdidos***, ***valores inconsistentes***, ***valores duplicados***

- **Agregación o agrupamiento**

Se realizan para tomar los valores agrupados con el fin de poner los datos en una mejor perspectiva. Esto permite reducir el consumo de memoria y el tiempo de procesamiento. Las agregaciones nos proporcionan una visión de alto nivel de los datos, ya que el comportamiento de los grupos o agregados es más estable que el de los objetos de datos individuales.

- **Muestreo**

El muestreo es un método muy común para seleccionar un subconjunto del conjunto de datos que estamos analizando. El principio clave en este caso es que el muestreo debe realizarse de forma que la muestra generada tenga aproximadamente las mismas propiedades que el conjunto de datos original, lo que significa que la muestra es representativa.

Pre-procesamiento

- **Reducción de la dimensionalidad**

La mayoría de los conjuntos de datos del mundo real tienen un gran número de características variables. Por ejemplo, si consideramos un problema de procesamiento de imágenes, es posible que tengamos que lidiar con miles de variables, también llamadas dimensiones. Como su nombre indica, el objetivo de la reducción de la dimensionalidad es reducir el número de características (variables), pero no simplemente seleccionando una muestra de variables del conjunto total de éstas, que es otra cosa: la selección de variables. La dimensión se refiere al número de planos geométricos en los que se encuentra el conjunto de datos, que puede ser tan elevado que no pueda visualizarse con lápiz y papel. Cuanto mayor sea el número de estos planos, mayor será la complejidad del conjunto de datos. Lo que hace esencialmente la reducción de la dimensión es que mapea el conjunto de datos a un espacio de menor dimensión. Reducir la dimensionalidad de un conjunto de datos crea nuevas variables que son una combinación de las antiguas. El espacio de variables de mayor dimensión se convierte en un espacio de menor dimensión. El **análisis de componentes principales** y la **descomposición de valores singulares** son dos técnicas ampliamente aceptadas.

Pre-procesamiento

- **Codificación**

La codificación consiste básicamente en realizar transformaciones en los datos de manera que puedan ser aceptados fácilmente como entrada para los algoritmos de modelado sin perder su significado original.

- **Separación de datos**

Una vez realizada la codificación de características, nuestro conjunto de datos está listo para los emocionantes algoritmos (scripts) de modelado. Pero antes de empezar a decidir el algoritmo que debe utilizarse, siempre se aconseja dividir el conjunto de datos en 2 o, a veces, 3 partes. Los algoritmos de aprendizaje automático, o cualquier otro algoritmo, tienen que ser entrenados primero en la distribución de datos disponible y, a continuación, validados y probados, antes de que puedan ser desplegados para tratar con datos del mundo real. La realción en que se dividen los datos depende en gran medida del tipo de modelo que estemos construyendo y del propio conjunto de datos. Si nuestro conjunto de datos y nuestro modelo son tales que se requiere una gran cantidad de entrenamiento, entonces utilizamos una parte más grande de los datos sólo con fines de entrenamiento (suele ser el caso).

https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

Preguntas?