Organización de Datos - Cátedra Rodriguez

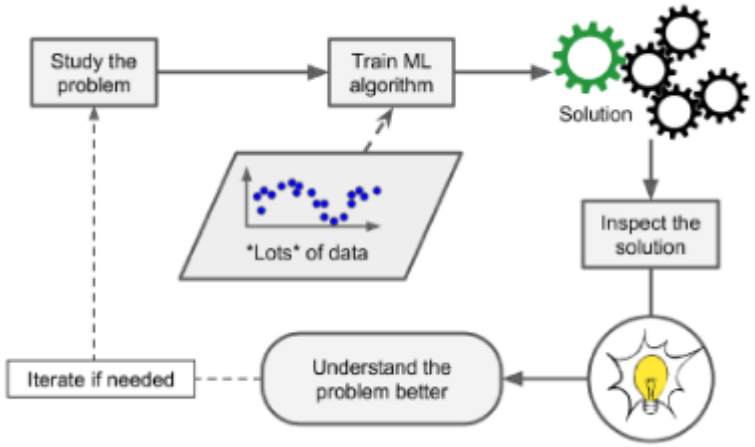
Introducción a la materia

Data Scientist

Capaz de:
- Obtener, interpretar, procesar y filtrar los datos.
- Llegar a conclusiones a partir del análisis anterior.
- Construir soluciones para los problemas que se presentan.

Data Scientist != Data Engineer

Machine Learning (Aprendizaje Automático)



Metodología

1. Entender el problema
2. Recolectar los datos
3. Procesar los datos
4. Explorar los datos
5. Analizar los datos
6. Comunicar los resultados

Visualización de datos - para Machine Learning

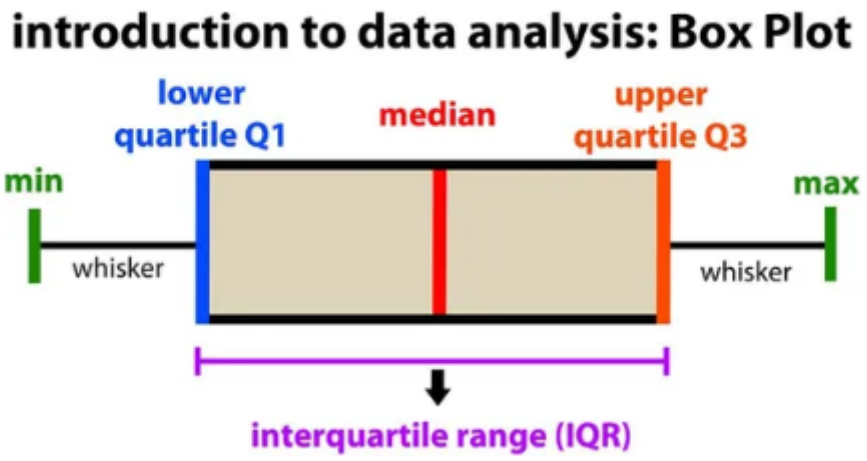
Análisis inicial de los datos:
- Examinar si los datos satisfacen los supuestos requeridos.
- Complicaciones inesperadas (valores atípicos o no linealidad)
- Evaluar ajuste del modelo:
- Predicho vs Observado
- Análisis de residuos

Plots:
- Distribución continua o discreta.
- De relación.
- Series de tiempo.
- Etc.

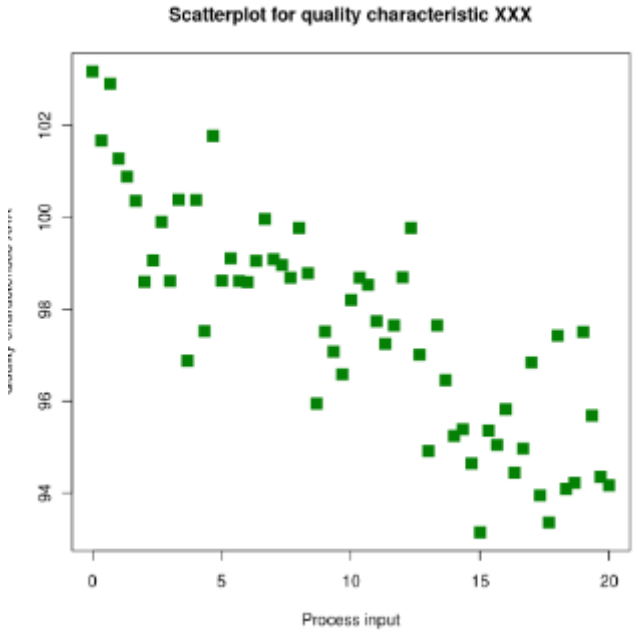
Números útiles:
- Media: promedio
- Mediana: valor en la mitad de la población
- Cuartil: son los valores límite que dejan el 25% de la población entre ellos (deja fuera el 25%).
- Rango intercuartílico: el rango entre el cuartil 1 y el cuartil 3.

Tipos de gráficos

Boxplot
Muestra la distribución de los datos numéricos y la asimetría mediante la visualización de los cuartiles (o percentiles) y los promedios de los datos.



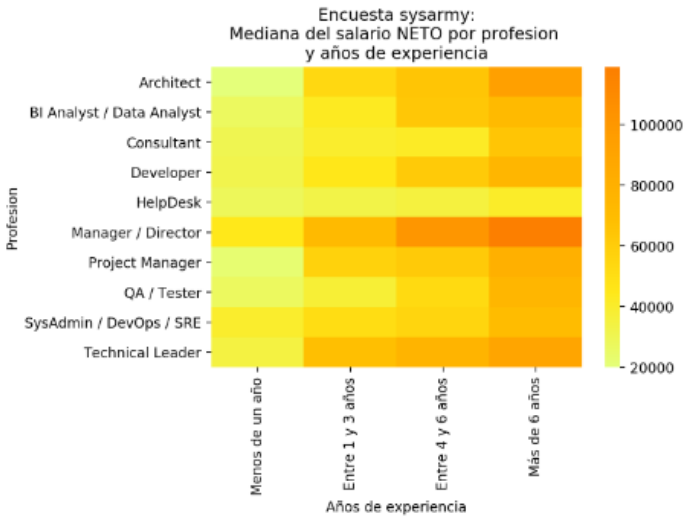
Scatter Plot
Utiliza las coordenadas cartesianas para mostrar los valores de 2 variables



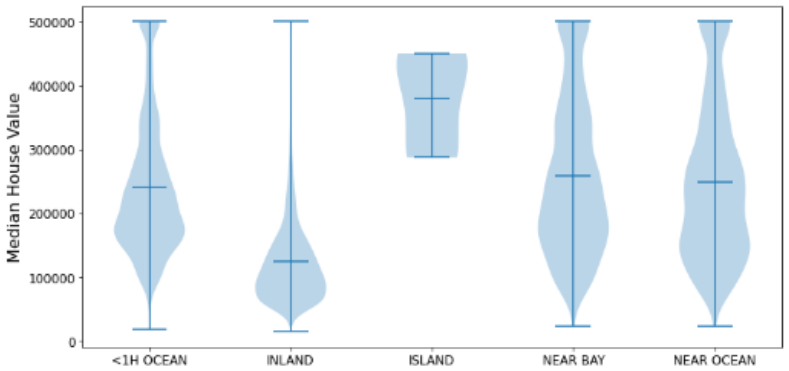
Correlación de Pearson
Función que mide cuán relacionadas están 2 variables de forma líneal.

- Si da 0 NO existe correlación
- Si da 1 Están relacionadas linealmente de forma perfecta (todos los puntos están en una línea)
-
- Si da -1 Existe una correlación negativa perfecta.

Heatmap
Comparar distribuciones en donde ambos ejes son discretos y un tercero de profundidad numérico.



Violin Plots
Similar a los boxplots excepto que también muestran la densidad de probabilidad de los datos en distintos valores.



Introducción a la ciencia de datos

Variables:
- Independientes (entradas)
- Dependientes (salidas, categorías) -> *Problema de Agrupamiento*

Variables Independientes:
Cualitativas (de texto o numéricas) -> *Problema de Clasificación*:
- Nominales: estas variables solo se distinguen por ser diferentes, no se puede establecer un ordenamiento entre ellas (ej: color de pelo, tipo de auto, género)
- Ordinales: se pueden ordenar, pero no se puede establecer una distancia entre ellos (ej: calificación de un examen, estadío de una enfermedad)
Cuantitativas -> *Problema de Regresión*:
- Discretas: toman valores numéricos siendo que entre 2 valores consecutivos no existen valores intermedios (ej: contar billetes, cantidad de materias aprobadas)
- Continuas: toman valores numéricos, asociadas al proceso de medir (ej: peso, edad, tiempo).

Correlación de Variables

Dos variables están correlacionadas cuando varían de igual forma sistemáticamente. La correlación puede ser positiva, negativa o nula.

Correlación NO implica causalidad.

Varianza: promedio de la diferencia entre todas las observaciones, respecto de su media.
Desvío estándar: cuantifica variación o dispersión de un conjunto de datos numéricos.
Covarianza: indica grado de variación conjunta de 2 variables aleatorias a sus medias.

Métodos de Regresión

Buscamos predecir un valor en un rango continuo para ciertos valores de entrada.

Regresión lineal o ajuste

- Modelo lineal simple: para el vínculo de 2 variables, se dice simple ya que vincula una sola variable predictora con la variable de respuesta.
- Modelo lineal múltiple: con más de una variable predictora.

El error producido por una predicción de regresión líneal se calcula como y'-y siendo y' el valor predicho e y el real.

Métricas para medir el error

MAE (Error Medio Absoluto)

Mean absolute error (MAE)

$$\mathrm{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^{m} \left| h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right|$$

MSE (Error Cuadrático Medio)

Mean Square Error

$$\mathrm{MSE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^{m} \left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right)^2$$

RMSE (Raíz del Error Cuadrático Medio)

Root Mean Square Error (RMSE)

$$\mathrm{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right)^2}$$

Métodos de Clasificación

Buscamos para ciertos datos de entrada, una categoría c de un conjunto C de categorías posibles. Son categorías finitas que conocemos de antemano.

Regresión Logística

Quiero clasificar, categorizar. Dado una serie de puntos quiero encontrar una función que separe los puntos en 2 conjuntos. Y una vez que lo encuentre puede determinar para cualquier valor X futuro, el conjunto al cual pertenece.

Métodos de Clusterización (Agrupamiento)

Clustering: agrupar datos de tal forma que queden definidos N conjuntos distinguibles, aunque no necesariamente se sepa qué significan los conjuntos. Agrupamiento por características similares.

K Means

1. El usuario decide la cantidad de grupos
2. K-Means elige al azar K centroides
3. Decide qué grupos están más cerca de cada centroide. Esos puntos forman un grupo
4. K-Means recalcula los centroides al centro de cada grupo
5. K-Means vuelve a reasignar los puntos usando los nuevos centroides. Calcula nuevos grupos
6. K-means repite punto 4 y 5 hasta que los puntos no cambian de grupo.

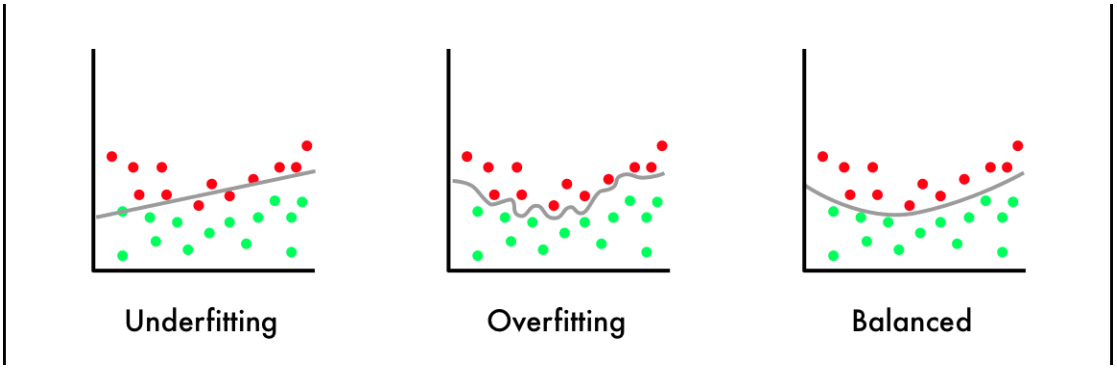
Entrenamiento

A la hora de analizar datos, partimos nuestro dataset en un conjunto de train o entrenamiento con el cual entrenamos los modelos y otro conjunto de test o prueba que serán utilizados para probar los modelos sobre datos que no vieron.

Overfitting (Sobreentrenamiento): cuando el modelo aprendió tan bien (como de memoria) los datos del train que después es muy malo con las nuevas observaciones. Se da por oversampling (muchas muestras de clase minoritaria).

Underfitting: se le dieron muy pocas muestras por lo que el modelo no pudo aprender. Se da por undersampling (muchas muestras de una clase mayoritaria pero pocas de clases minoritarias).

Balanced (balanceado): cuando no tenemos ni over ni underfitting.

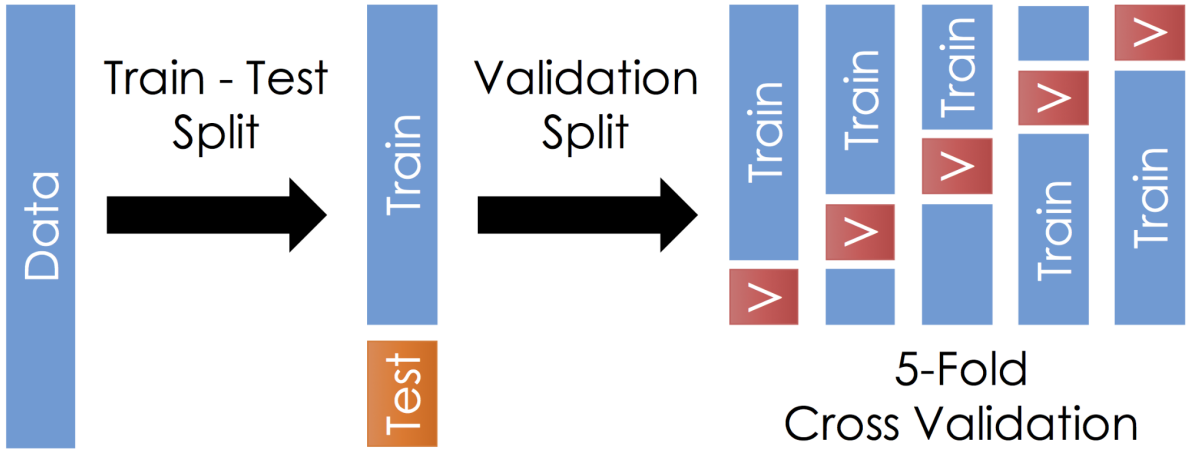


Formas de partir en train-test

Porcentajes
Ej: 70-30 (70 train y 30 test)

Cross Validation
El conjunto de entrenamiento se divide en k conjuntos más pequeños. Para cada uno de los k 'folds' se hace lo siguiente:

- Se entrena un modelo utilizando los k folds como datos de entrenamiento.
- El modelo resultante se valida con la parte restante de los datos.
- La medida de rendimiento informada por la validación cruzada de k veces es el promedio de los valores calculados en el bucle o el mejor de todos.



Métricas

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$