

Aprendizaje de Máquina con Python

Pandas y scikitlearn

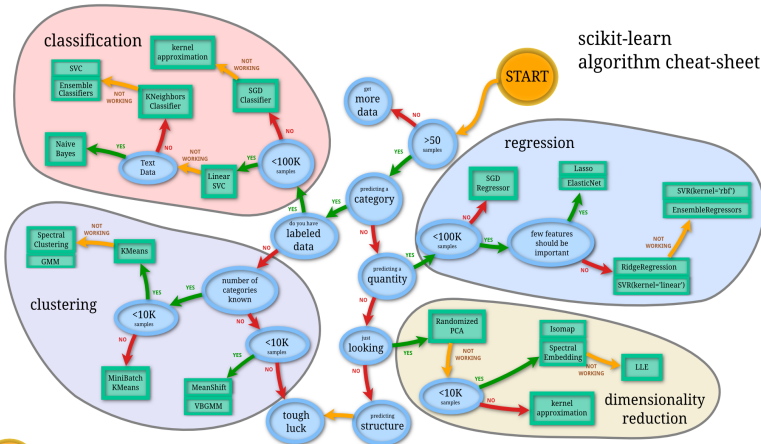
Sergio Morales Esquivel

smorales+charlas@ucenfotec.ac.cr

26 de Febrero de 2019



scikit-learn algorithm cheat-sheet



Agenda

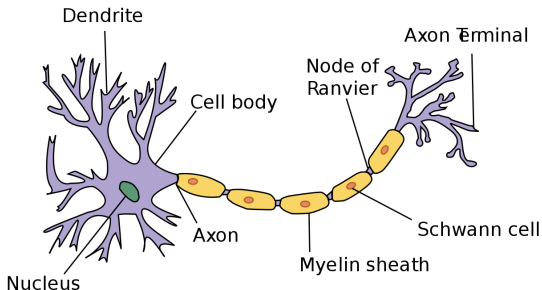
- 1 Toma de Decisiones
 - 2 Aprendizaje de Máquina
 - 3 Datos
 - 4 Entrenamiento
 - 5 Evaluación
- Aprendizaje Supervisado

Aprendizaje No
Supervisado
Regresión vs Clasificación

- 6 Resumen
- 7 Scikit-Learn
- 8 Ambiente
- 9 Preguntas y Discusión

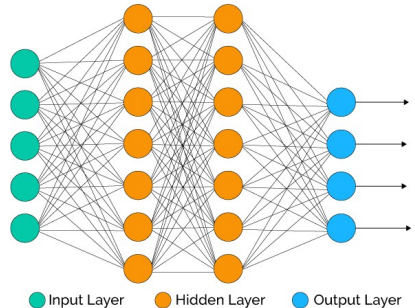
¿Cómo tomamos decisiones?

- Todo nuestro comportamiento está basado en decisiones.
- No sólo decidimos acciones, sino cómo interpretar lo que nos rodea.
- El aprendizaje de máquina se inspira en el caso humano.



```
def recibirSeñales(señales, pesos):
    valores = []
    for s, p in zip(señales, pesos):
        valores.append(s*p)
    suma = sum(valores)
    valActivacion = funAct(suma)
    if suma > umbral:
        propagarSeñal()

def backPropagation(resultado):
    # reajustar lista de pesos
    # de acuerdo a resultados
    # esperados
```





¿Cómo tomamos decisiones?

- La herramienta de **aprendizaje** humano más importante es el **reconocimiento de patrones**.
- Por medio del reconocimiento de indicadores sencillos de detectar y determinar, es posible generar interpretaciones complejas de la realidad.

El aprendizaje de máquina es un área de estudio que busca hacer que un sistema computacional sea capaz de **aprender** sin un sistema de reglas explícito. Esto es, recibir un conjunto de datos y:

- **Clasificar:** Decidir entre "tipos" de muestras.
- **Predecir:** Predecir tendencias.
- **Abstraer:** Reducir datasets complejos a relaciones simples.
- **Generar Insight:** Deducir relaciones e interacciones no explícitas en los datos de entrada.

El aprendizaje de máquina es el área de la Inteligencia Artificial que más éxito ha tenido en el área de negocios:

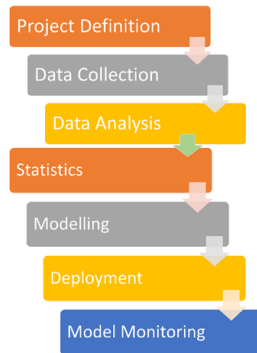
Usos en Negocios

- Reconocimiento de Lenguaje Natural.
- Reconocimiento de Imágenes.
- Auto-Clasificación de Mensajes.
- Recomendaciones de Productos.
- Estimaciones/Proyecciones Financieras.
- Retención de Clientes.



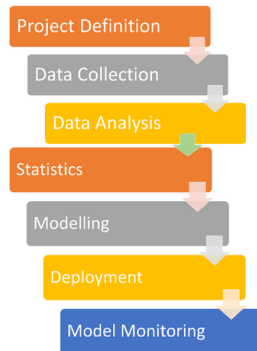
Proceso de Aprendizaje de Máquina:

- **Recuperación de Datos:** ¿En donde están los datos?
- **Transformación y Limpieza de Datos:** Formato de datos, valores nulos, atributos inútiles y de bajo valor.
- **Visualización y Exploración:** Opcional, pero puede arrojar insights sobre el modelo.



Proceso de Aprendizaje de Máquina:

- **Partición en Conjuntos de Entrenamiento y Prueba:** Inicio del proceso de aprendizaje supervisado.
- **Construcción/Entrenamiento del Modelo:** Correr algoritmo de entrenamiento.
- **Prueba del Modelo:** Probar modelo sobre conjunto de prueba.



La efectividad de un proceso de aprendizaje de máquina depende tanto de los datos usados como del sistema de aprendizaje.

Aspectos a considerar:

- **Tamaño:** Se ocupa una cantidad adecuada de muestras.
- **Valores Nulos/Blancos:** Datos faltantes o nulos pueden degradar la calidad de los datos.
- **Valores Nominales vs. Numerales:** Pueden influenciar el modelo de clasificación.
- **Estructura:** Normalmente una tabla, pero pueden ser histogramas u otras estructuras.
- **Atributos:** No necesariamente más es mejor.

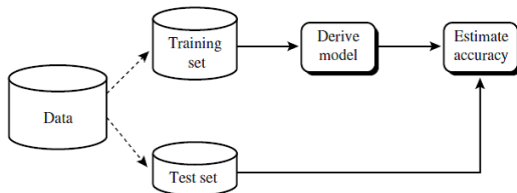
Por lo general, un conjunto de datos es una **serie de muestras**, cada una con **atributos** y una **etiqueta** o **atributo objetivo** (en caso de **aprendizaje supervisado**).

ID Paciente	Edad	# Tumores	Area Promedio	Densidad Prom.	Diagnostico
1	22	2	20	64	Maligno
2	36	5	15	121	Benigno
3	44	3	17	160	Maligno
4	42	7	10	89	Benigno
5	38	2	30	95	Maligno
6	53	3	12	104	Maligno
7	51	9	11	75	Benigno
...
301	22	2	20	64	???
302	36	5	15	121	???
303	44	3	17	160	???

Es muy probable que sea necesario limpiar y transformar o preprocesar datos previo al modelaje del sistema de aprendizaje.

Limpieza y Transformación:

- **Tratamiento de Valores Nulos:** ¿Eliminar muestras? ¿Eliminar atributos? ¿Reemplazar con la media?
- **Seleccionar los mejores atributos:** Eliminar IDs. Agrupar o separar, usar conocimiento previo.
- **Transformar valores:** Ordinales a categorías, normalizar, atributos binarios, etc.
- **Remove etiquetas del conjunto de prueba.**
- **Balancear muestras:** Evita sesgos en el modelo.

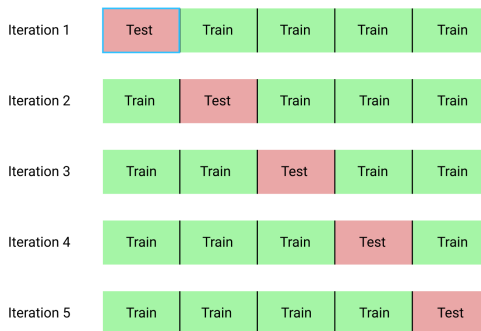


Limpieza y Transformación:

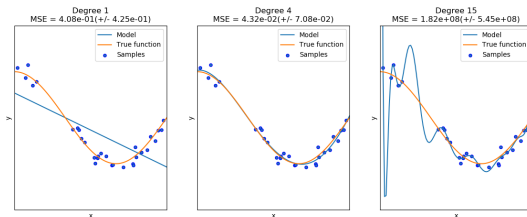
- **Conjunto de Entrenamiento:** Se refiere a un subconjunto de muestras que se utiliza para entrenar al **modelo predictivo**.
- **Conjunto de Prueba:** Subconjunto de muestra que se utiliza para probar un **modelo entrenado**.

Generación de Folds para K-Fold Cross-Validation

Consiste en segmentar el dataset en N partes iguales (**folds**), y entrenar N modelos diferentes, utilizando **$N-1$ folds en cada modelo** como conjunto de entrenamiento, y el fold sobrante (diferente en cada modelo) como conjunto de prueba.

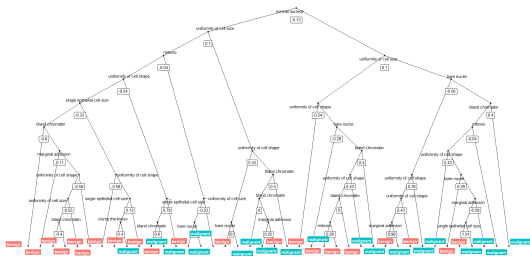


Listos los datos, se entrena al modelo (**model fitting**), utilizando el conjunto de entrenamiento (muestras etiquetadas). Esto genera una función o **modelo de clasificación**/predicción y es el resultado del proceso de **aprendizaje**.



Debe tenerse cuidado de no generar un modelo que sufra de **underfitting** (izquierda), u **overfitting** (derecha).

El resultado del entrenamiento es un **modelo** que puede tomar muchas formas: Una **red neuronal**, una **lista de reglas**, un **árbol de decisiones**, etc. También puede arrojar información relacionada importante, como por ejemplo una lista ordenada de importancia de atributos.



En este punto, puede usarse el **conjunto de prueba** para verificar la precisión del modelo.

Aprendizaje Supervisado:

Se refiere a aquellos modelos que son entrenados utilizando pares de **entrada y salida**. Es decir, la clasificación esperada es provista como ejemplo.

- Es útil para buscar atributos o relaciones importantes.
- Utiliza los conjuntos de entrenamiento y prueba mencionados anteriormente
- **Ejemplos:**
 - Árboles de Decisión
 - Regresión Lineal
 - "Vecinos Cercanos"
 - Máquinas de Soporte Vectorial

Aprendizaje No Supervisado:

Se refiere a aquellos modelos que son entrenados sin proveer "**etiquetas**" o valores de clasificación asociadas a las muestras de entrada.

- Puesto que los datos no poseen etiquetas, no puede "**evaluarse**" el modelo generado.
- Se busca que el modelo arroje información no descubierta acerca de **atributos o relaciones**.
- **Ejemplos:**
 - Clustering
 - Redes Neuronales

Regresión y Clasificación:

Dentro del **aprendizaje supervisado**, se hace una distinción entre **regresión** y **clasificación**.

- **Regresión:** Busca encontrar un modelo que devuelve **valores continuos** para cada entrada provista luego de la generación del modelo.
- **Clasificación:** Busca encontrar un modelo que devuelve **uno de varios valores nominales** para cada entrada provista luego de la generación del modelo.



Scikit-Learn

Librería de Machine Learning/Análisis de Datos/Minería de Datos para Python

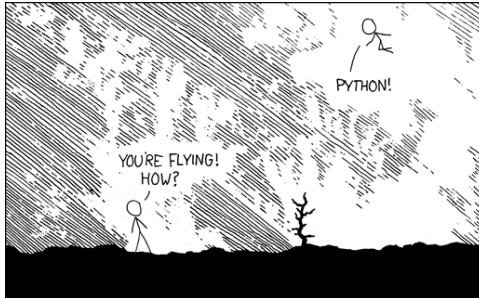
- Basada en NumPy, SciPy, y matplotlib
- Open source, usable comercialmente - licencia BSD





Herramientas

- **Pandas:** Python Data Analysis Library; provee estructuras de datos y herramientas de análisis.
- **matplotlib:** Librería de visualización de datos.
- **statsmodels/seaborn/biopython...**: Otros paquetes del stack científico de Python.



Instalación:

Las librerías pueden instalarse de varias maneras, sin embargo siempre es prudente trabajar en ambientes virtuales usando **virtualenv**, e instalar librerías con **pip**.

```
$ pip install sklearn  
$ pip install pandas  
$ pip install matplotlib
```

Importante: En MacOS, puede que sea necesario correr algunos de estos comandos con la bandera **–ignore-installed numpy**, dependiendo de como y donde esta instalada la distribución de Python.

Anaconda

Anaconda es una plataforma de Análisis de Datos que incluye todos los paquetes esenciales de Python, e incluye una interfaz gráfica para la administración de paquetes, así como un ambiente de desarrollo.



Demostración Práctica

Todo el material de esta charla puede ser descargado desde:
<http://www.fireblend.com/charla-ml.zip>

Muchas gracias

¿Preguntas o comentarios?

“We are trying to prove ourselves wrong as quickly as possible, because only in that way can we find progress”.

Richard Feynman.