TEC | Tecnológico de Costa Rica

Maestría en Ingeniería Electrónica

Reconocimiento de patrones

Apuntes de clase 1

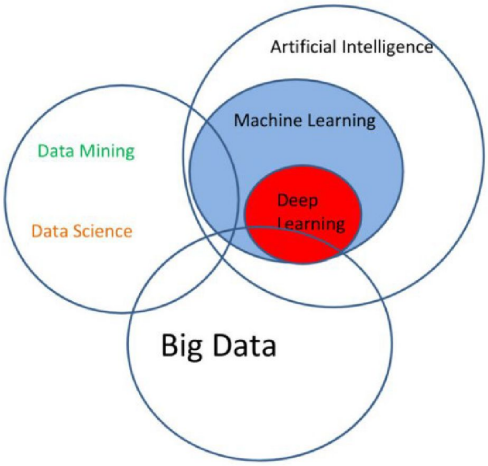
Estudiante: Esteban Martínez Valverde

Profesor: Felipe Meza Obando

Junio 2018

INTRODUCCIÓN

Aprendizaje Automático (Machine Learning - ML)



El aprendizaje automático es un campo de la informática que a menudo utiliza técnicas estadísticas para dar a las computadoras la capacidad de "aprender" con datos, sin estar explícitamente programadas. El nombre de aprendizaje automático fue acuñado en 1959 por Arthur Samuel. (Wikipedia)

- Conjunto de mecanismos que permiten convertir los datos en información/conocimiento útil
- Utiliza técnicas de los campos de computación, estadística, entre otros
- Construcción de programas que mejoren automáticamente con la experiencia.

Algunos ejemplos de aplicación son:

- Carros que se manejan solos
- Reconocimiento de texto y voz
- Generación de rostros

Ejemplos de aplicaciones reales

- SIRI y Cortana
- Facebook (reconocimiento de rostros)
- Google Maps (Al utilizar la locación de los smartphones sugiere rutas alternas)
- Google Search (Recomienda y sugiere basado en búsquedas previas)
- Gmail (Sugiere respuestas a los correos)
- Paypal (Utiliza Deep learning para evitar fraudes)
- Netflix (Sugiere películas y series de acuerdo con los gustos del usuario)
- UBER (Predice los tiempos de llegada, lugares de recogida)

- Spotify (Sugiere canciones de acuerdo con las reproducciones del usuario)

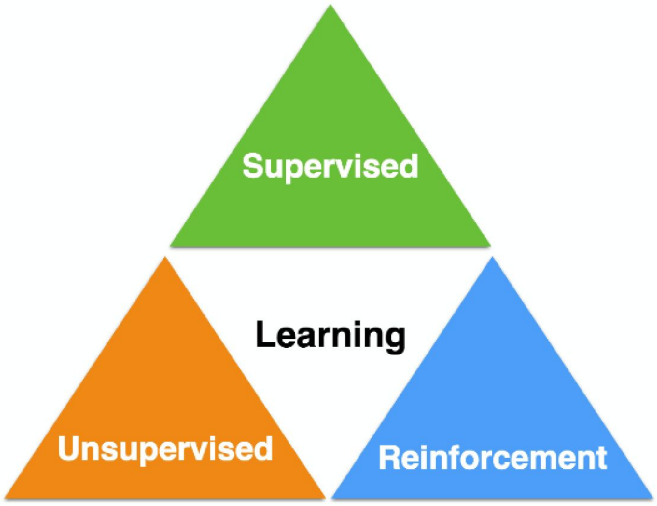
Estadística?

Busca la relación entre variables, mediante ecuaciones matemáticas constituyendo una herramienta útil (matemáticas) . Se encarga de estimar, realizar una hipótesis, obtener muestreos de la información y obtener resultados.

ML?

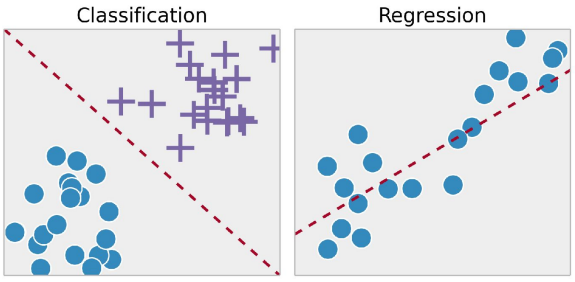
Busca aprender de los datos mediante algoritmos, constituyendo una herramienta vital (Inteligencia artificial). Se encarga de aprender, clasificar, crear instancias y etiquetas.

Tipos de Aprendizaje.



Supervisado

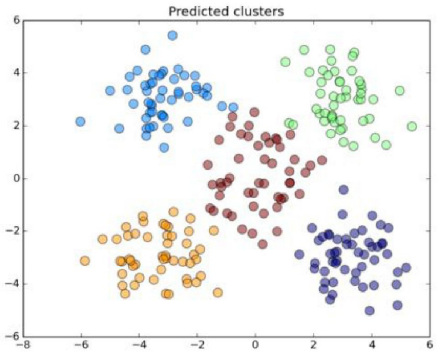
- Los datos están etiquetados y se predice una salida futura
- Clasificación -> Predicción (regresión)



No Supervisado

- Los datos no están etiquetados y se encuentra una estructura oculta

- Agrupamiento (clustering): Se encarga de encontrar los centros



Refuerzo
Se etiquetan los datos y se predice una salida futura

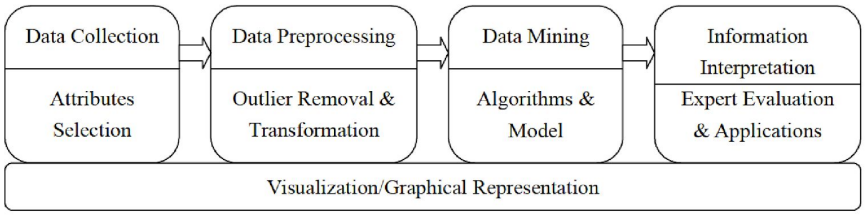
Instancias, Atributos y Clases.

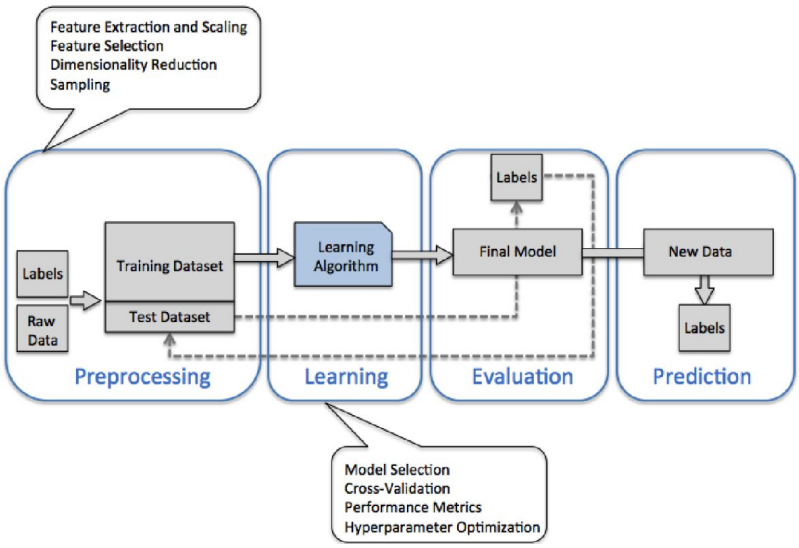
Instancias
Muestras, observaciones, número de datos

Atributos
Características, valores de los datos (inputs)

Clases
Etiquetas, objetivos

Metodología de Diseño.





Métricas para evaluar lo aprendido.

- Exactitud de la clasificación.

• Pérdidas logarítmicas.

• Curva ROC.

• Matriz de confusión.

• MSE.

• R^2.

AMBIENTE PYTHON

pandas

CONDA

matplotlib

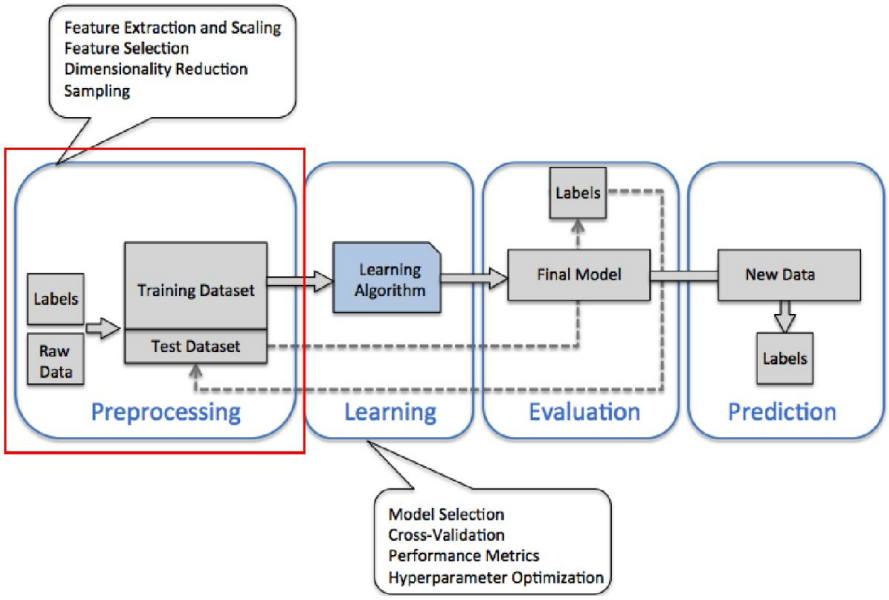
jupyter

python

scikit learn

PRE-PROCESADO

Metodología de diseño



Pre-procesado

(Data preparation, data cleaning, pre-processing, wrangling)

Se realiza un análisis del conjunto de datos para identificar los componentes que sean incompletas, imprecisas, incorrectas, o irrelevantes y puedan ser reemplazadas, removidas o modificadas.

Parte de las modificaciones se incluye:

- Transformación de los datos “puros” a formatos que faciliten el manejo en los algoritmos de minería de datos.
- Reducción de los datos a menores dimensiones para facilitar el procesamiento

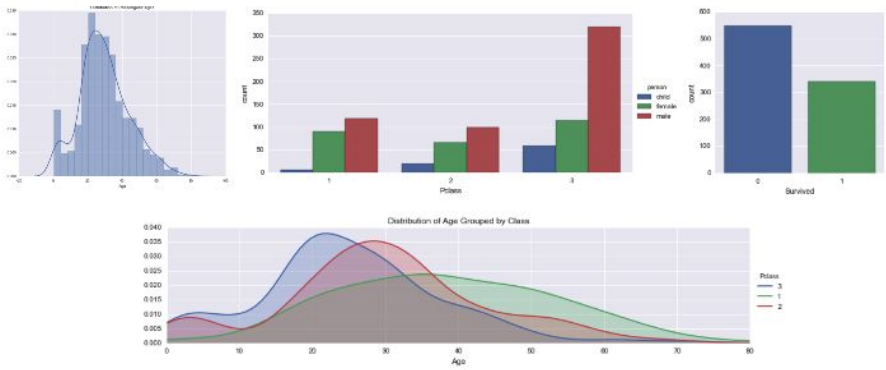
Preparación de datos

En las primeras etapas de los modelos de diseño se lleva a cabo la selección, pre-procesado o transformación de los datos. Esa preparación de los datos no es un componente integral en los algoritmos de aprendizaje, sin embargo, se requiere de bastante tiempo (80% a 90%), por lo que es importante su consideración,

En ambientes de desarrollo como **PYTHON**, se usan librerías como **PANDAS** para la preparación de los datos

Tareas del Pre-Procesado

Análisis exploratorio de los datos (EDA)



Métodos cuantitativos y visuales para comprender mejor un conjunto de datos sin tener que asumir hechos. Arrojar el conjunto de datos a un algoritmo y esperar los mejores resultados, no es la mejor estrategia.

- Visualización de un resumen estadı´stico del conjunto de datos.
- Exploración visual de cualquier relación que pueda tener cada atributo con la clase que nos interesa predecir.
- Mediante diagramas de dispersión observar cualquier tipo de agrupamiento que se pueda presentar en los datos.

Valores Faltantes

Algunas técnicas comunes son:

- Eliminar instancias.
- Eliminar atributos.
- Calcular “media” del atributo faltante.
- Calcular “mediana” del atributo faltante.
- Calcular “moda” del atributo faltante.
- Usar regresión para estimar el valor del atributo faltante.

Se utilizan las funciones ***dropna*** y ***fillna*** de la herramienta ***pandas*** como ejemplo de completar los valores faltantes de un Dataset

Outliers

Un *Outlier* es un dato que se presenta como “atípico” dentro del conjunto de datos. Dependiendo de la naturaleza de los datos, hay ocaciones es necesario mantenerlos y en otras más bien se busca eliminarlos. Algunas técnicas comunes son:

- Removerlos usando desviación estándar (**PYTHON**), importando la librería ***numpy***.
- Removerlos usando percentiles (**pandas**).

Datos no-balanceados

Ocurre cuando una clase de datos en el conjunto, posee una mayoría importante de la cantidad de datos e.g un conjunto de datos de 2 clases, donde: CLASE1 = 98% y CLASE2 = 2%.

Algunas técnicas comunes son:

- Usar otras métricas diferentes al porcentaje de exactitud, por ejemplo:
  - Precision/Specificity: cuantas instancias seleccionadas son relevantes.
  - Recall/Sensitivity: cuantas instancias relevantes son seleccionadas.
  - F1 score: media harmónica de “precision” y “recall”.
- Muestreo de datos:
  - sub-muestreo: Eliminar instancias abundantes (sólo si hay suficientes datos).
  - sobre-muestreo: Generar instancias faltantes (mediante métodos de repetición o generación, solo en caso de que sea posible)
- Descomponer el conjunto de datos en subconjuntos.
- Hacer clustering de grupo abundante.

Transformación de datos

- Ocurre cuando transformamos un valor zi en yi mediante una función f () de forma tal que yi = f (zi).
- Se hace con el fin de alinear los datos con alguna suposición estadistica, mejorar la interpretación de los datos o bien obtener gráficos de mejor apariencia.
- Técnica muy común: One Hot Encode
  - Permite convertir datos categóricos en numéricos (vectores binarios).

Reducción de dimensiones (PCA)

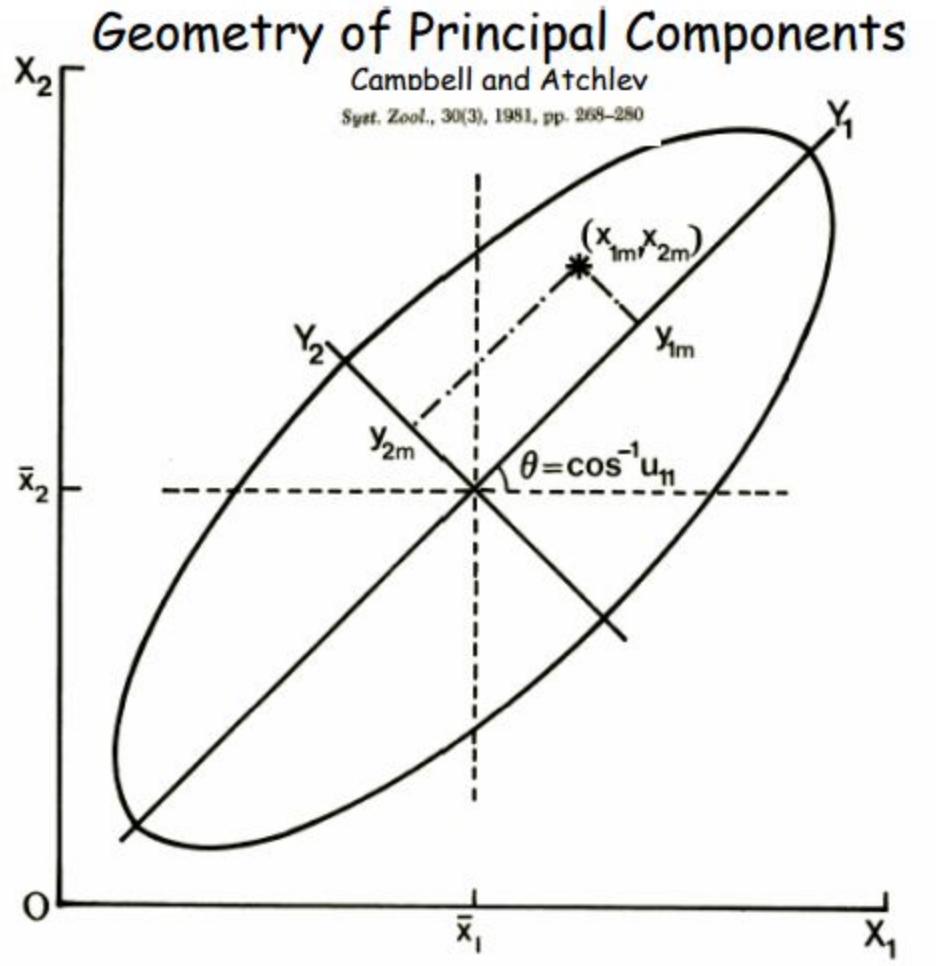
Mediante la detección de correlación entre variables y análisis de máximas varianzas, es posible proyectar datos de grandes dimensiones en sub-espacios de menor dimensión, de manera tal que se conserve la información más relevante.Se lleva a cabo en los siguientes pasos:

- Estandarizar de los datos (en especial si poseen escalas distintas).
- Calcular la matriz de covarianza o correlación:

$$\Sigma = \frac{1}{n-1}\left((\mathbf{X} - \bar{\mathbf{x}})^T (\mathbf{X} - \bar{\mathbf{x}})\right)$$

- Extraer los valores propios (eigenvalues) y ordenarlos.

- Seleccionar los top-k.



El PCA se puede considerar como una rotación de los ejes del sistema de coordenadas de la variable original a unos nuevos ejes ortogonales, llamados "ejes principales". Como se observa en la Figura , el punto y_{1m} es la proyección del punto (x_{1m},x_{2m})$ en el eje definido por Y_1.

Términos y Definiciones

- **Training set:** Colección grande de N datos usado para entrenar los parámetros de un modelo adaptativo
- **Test Set:** Colección de datos para probar el entrenamiento.
- **Target vector:** Una categoría escogida para identificar el tipo de cada dato.
- **Training/learning phase:** se utiliza para determinar la forma de una función **Y(x)** donde **x** es cada dato nuevo de entrada (dígitos)
- **Generalization:** La capacidad de categorizar correctamente nuevos ejemplos que difieren de los utilizados para la capacitación se conoce como generalización.
- **Preprocesado / Feature extraction:** las variables de entrada son transformadas en un nuevo espacio de variables, donde se espera el problema de reconocimiento de patrones sea más fácil.
    - Por lo general se realiza una reducción dimensional de las variables.
    - Hay que tener cuidado para que no se pierda información necesaria para el criterio de selección en el ML
- **Supervised Training:** Son las aplicaciones donde los datos de entrenamiento comprenden los vectores de entrada y sus correspondientes vectores de destino **(Target)**
    - **Classification:** Cada vector de entrada se le asigna un número finito de categorías discretas
    - **Regression:** La salida consiste en una o mas variables continuas
- **Unsupervised Training:** El *training set*, consiste en vectores de entrada **x** sin ninguna categoría definida como **target**
    - **Clustering:** Descubrir grupos de ejemplos similares dentro de los datos de entrada
    - **Density Estimation:** Determinar la distribución de datos dentro del espacio de una entrada
    - **Visualization:** Proyectar datos de una dimensión mayor a un espacio de 2 o 3 dimensiones
- **Reinforcement Learning:** se relaciona con el problema de encontrar acciones adecuadas para tomar en una situación dada con el fin de maximizar una recompensa.