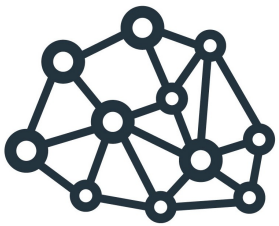
Visión General del Aprendizaje Automático (Machine Learning)

José Moreno

April 11, 2025



Visión General del Aprendizaje Automático

- **¿Qué es el Aprendizaje Automático?** Aprender a partir de datos sin programación explícita.
- **Tipos de Aprendizaje Automático:**
  - Aprendizaje Supervisado (Datos Etiquetados)
  - Aprendizaje No Supervisado (Datos No Etiquetados)
  - Aprendizaje Semi-Supervisado (Mezcla de Datos Etiquetados/No Etiquetados)
  - Aprendizaje por Refuerzo (Aprendizaje a través de la Interacción)
- **Consideraciones Clave:** Son cruciales para construir modelos de aprendizaje automático efectivos.
  - Métricas de Evaluación: Medir el rendimiento del modelo.
  - Ajuste de Hiperparámetros: Optimizar la configuración del modelo.
  - Ingeniería de Características: Seleccionar y transformar características relevantes.
  - Sesgo y Equidad: Garantizar resultados equitativos.

Clasificación - Prediciendo Categorías

Clasificadores Probabilísticos:

- **Naïve Bayes**: Simple, rápido. Asume independencia de características. Variantes: Gaussiana, Multinomial, Bernoulli.
- **Regresión Logística**: Predice probabilidades para clasificación binaria. Utiliza la función sigmoide y regularización (L1/L2).

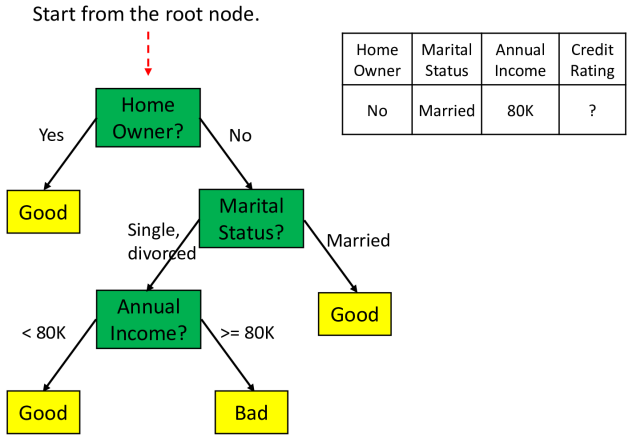
Aprendizaje Basado en Instancias:

- **K-Vecinos Más Cercanos (KNN)**: Clasifica basándose en la clase mayoritaria de sus k vecinos más cercanos. Requiere una cuidadosa escala de características.

Clasificación - Métodos Basados en Árboles y Kernel

Métodos Basados en Árboles:

- **Árbol de Decisión**: Interpretable, pero propenso al sobreajuste.
- **Bosque Aleatorio (Random Forest)**: Conjunto de árboles de decisión, reduciendo la varianza y mejorando la precisión.
- **GBM (XGBoost, LightGBM, CatBoost)**



Métodos Kernel:

- **Máquina de Vectores de Soporte (SVM)**: Efectiva en espacios de alta dimensión. El Truco del Kernel permite la clasificación no lineal.

Clasificación - Redes Neuronales

Redes Neuronales (Aprendizaje Profundo):

- **MLP**: Red feedforward con múltiples capas. Funciones de Activación (ReLU, Sigmoide, Tanh) introducen no linealidad.
- **CNNs**: Capas Convolucionales y de Pooling sobresalen en el reconocimiento de imágenes.
- **RNNs**: LSTM, GRU manejan datos secuenciales como texto o series temporales. Abordan el Problema del Desvanecimiento/Explosión del Gradiente.

Regresión - Prediciendo Valores Continuos

Modelos Lineales:

- **Regresión Lineal Simple**: Modela la relación entre dos variables con una línea recta.
- **Regresión Lineal Múltiple**: Extiende la regresión lineal simple a múltiples características.
- **Regresión Polinómica**: Captura relaciones no lineales utilizando términos polinómicos.

Regresión - Modelos Regularizados y Métodos Basados en Árboles

Modelos Lineales Regularizados:

- **Lasso (L1)**: Realiza la selección de características reduciendo algunos coeficientes a cero.
- **Ridge (L2)**: Reduce los coeficientes hacia cero, disminuyendo el sobreajuste. Elastic Net combina la regularización L1/L2.

Métodos Basados en Árboles:

- Regresor de Árbol de Decisión
- Regresor de Bosque Aleatorio (Random Forest)
- GBM (XGBoost, LightGBM, CatBoost) - Potente y versátil para tareas de regresión.

Regresión - Redes Neuronales

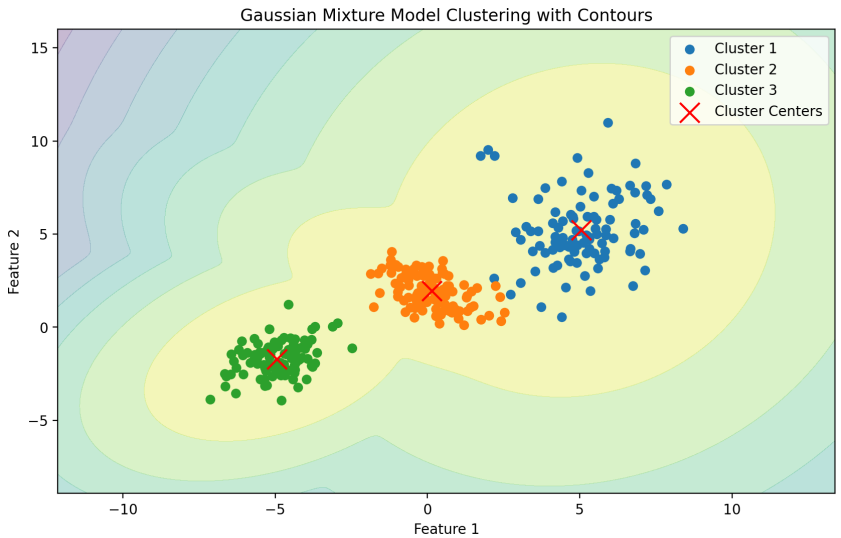
Redes Neuronales (Aprendizaje Profundo):

- MLP para Regresión: Puede modelar relaciones no lineales complejas.

Agrupamiento - Agrupando Puntos de Datos Similares

Agrupamiento Basado en Particiones:

- **K-Means**:
  Simple y eficiente, pero sensible a los centroides iniciales. El Método del Codo ayuda a determinar el número óptimo de clústeres.
- **Mini-Batch K-Means**: Versión más rápida para conjuntos de datos grandes.



Agrupamiento - Métodos Basados en Densidad y Probabilísticos

Agrupamiento Basado en Densidad:

- **DBSCAN**: Puede descubrir clústeres de formas arbitrarias. Requiere ajustar los parámetros Epsilon & MinPts.

Agrupamiento Probabilístico:

- **Modelos de Mezcla Gaussiana (GMM)**: Asume que los datos se generan a partir de una mezcla de distribuciones Gaussianas. Utiliza el Algoritmo EM para la estimación de parámetros.

Agrupamiento - Jerárquico e ICA

Agrupamiento Jerárquico: Enfoques Aglomerativos (de abajo hacia arriba) y Divisivos (de arriba hacia abajo). Los Dendrogramas visualizan la jerarquía de agrupamiento.
Análisis de Componentes Independientes (ICA): Separa una señal multivariante en subcomponentes aditivos. Útil para el procesamiento de señales.

Minería de Reglas de Asociación

- **Apriori**: Encuentra conjuntos de elementos frecuentes basados en soporte, confianza y lift.
- **Eclat**: Utiliza un formato de datos vertical para la minería eficiente de conjuntos de elementos frecuentes.
- **FP-Growth**: Algoritmo eficiente para grandes conjuntos de datos mediante la construcción de un Árbol de Patrones Frecuentes.

Reducción de Dimensionalidad

- **PCA**: Reduce la dimensionalidad preservando la mayor varianza en los datos.
- **t-SNE**: Excelente para visualizar datos de alta dimensión en 2D o 3D.
- **Autoencoders**: Redes neuronales que aprenden una representación comprimida de los datos de entrada.

Detección de Anomalías - Identificando Valores Atípicos

Métodos Estadísticos:

- Algoritmo Z-Score: Identifica valores atípicos basándose en su distancia a la media.

Métodos Basados en Árboles:

- Algoritmo Isolation Forest: Aísla anomalías particionando aleatoriamente el espacio de datos.

One-Class SVM: Entrenado solo con datos normales para identificar desviaciones.

Redes Neuronales (Deep Learning): Autoencoders (Error de Reconstrucción). VAEs.

Aprendizaje Semi-Supervisado - Mezcla de Datos Etiquetados y No Etiquetados

- **Autoentrenamiento**: Entrena un modelo con datos etiquetados, predice etiquetas para los datos no etiquetados, agrega predicciones confiables al conjunto de entrenamiento.
- **Co-Entrenamiento**: Utiliza múltiples clasificadores con diferentes subconjuntos de características para aprovechar los datos no etiquetados.
- **Propagación de Etiquetas**: Propaga las etiquetas a través de un grafo construido a partir de los datos.

Aprendizaje por Refuerzo - Aprendiendo a Través de la Interacción

Métodos Sin Modelo:

- **Q-Learning**: Aprende la función valor óptima acción-valor.
- **SARSA**: Algoritmo de aprendizaje On-Policy.
- **Optimización de Políticas**: REINFORCE, métodos Actor-Crítico optimizan directamente la política.

Aprendizaje por Refuerzo - Métodos Basados en Modelos y Deep RL

Métodos Basados en Modelos: Aprende un modelo del entorno (probabilidades de transición y recompensas). Planifica acciones óptimas utilizando programación dinámica.
Deep Reinforcement Learning: DQN, DDPG, PPO, SAC combinan redes neuronales profundas con algoritmos de aprendizaje por refuerzo.

Consideraciones Clave: Métricas de Evaluación

¡Elegir la Métrica Correcta es Crítico!

- **Clasificación:**
  - **Precisión**: En general, predicciones correctas (puede ser engañosa con conjuntos de datos desequilibrados).
  - **Precisión/Exhaustividad**: Compromiso entre falsos positivos y falsos negativos. Importante cuando los costos de los errores difieren.
  - **Puntuación F1**: Media armónica de Precisión y Exhaustividad. Buena para una evaluación equilibrada.
  - **AUC-ROC**: Área bajo la curva ROC (Receiver Operating Characteristic). Mide la capacidad de distinguir clases.
- **Regresión:**
  - **ECM/RMSE**: Error Cuadrático Medio / Raíz del Error Cuadrático Medio (sensible a valores atípicos).
  - **R-cuadrado**: Proporción de la varianza explicada por el modelo.
  - **MAE**: Error Absoluto Medio (menos sensible a valores atípicos)
- **Clustering:** Puntuación de Silueta, Índice de Davies-Bouldin, Índice de Calinski-Harabasz.

Consideraciones Clave: Ajuste de Hiperparámetros y Preprocesamiento de Datos

- **Ajuste de Hiperparámetros:**
  - **Búsqueda en Grilla**: Búsqueda exhaustiva en un espacio de parámetros predefinido.
  - **Búsqueda Aleatoria**: Muestrea aleatoriamente los parámetros, a menudo más eficiente que la búsqueda en grilla.
  - **Optimización Bayesiana**: Utiliza modelos probabilísticos para guiar la búsqueda de hiperparámetros óptimos. Herramientas como Optuna o Hyperopt.
- **Preprocesamiento de Datos:**
  - **Escalado/Normalización**: Escalado Min-Max, Estandarización (puntuación Z). Importante para algoritmos sensibles a las escalas de características (por ejemplo, KNN, SVM, Redes Neuronales).
  - **Codificación de Características Categóricas**: Codificación One-Hot, Codificación Etiqueta.
  - **Manejo de Valores Faltantes**: Imputación (media, mediana, moda), eliminación.

Consideraciones Clave: Ingeniería de Características y Sesgo/Equidad

Creación de Características y Garantía de Resultados Éticos

- **Ingeniería de Características:**
  - ¡El conocimiento del dominio es clave! Cree nuevas características a partir de las existentes.
  - Características polinómicas, términos de interacción.
  - Técnicas de selección de características (por ejemplo, utilizando la importancia de las características de los modelos basados en árboles).
- **Sesgo y Equidad:**
  - ¡Tenga en cuenta los posibles sesgos en sus datos! (Histórico, muestreo, etc.)
  - Métricas de equidad: Impacto Dispar, Igualdad de Oportunidades.
  - Técnicas de mitigación: Re ponderación, remuestreo, des sesgo adversarial.

Consideraciones Clave: Selección del Modelo, Sobreajuste/Subajuste y Explicabilidad

Elegir el Modelo Correcto y Comprender sus Predicciones

- **Selección del Modelo:** No hay almuerzo gratis. Diferentes modelos funcionan mejor en diferentes conjuntos de datos. ¡Experimente!
- **Sobreajuste/Subajuste:**
  - Utilice la validación cruzada para evaluar el rendimiento de generalización.
  - Las técnicas de regularización (L1, L2) pueden ayudar a prevenir el sobreajuste.
  - Más datos a menudo ayudan a reducir el sobreajuste.
- **IA Explicable (XAI):**
  - **Valores SHAP**: Explique las predicciones individuales atribuyendo importancia a cada característica.
  - **LIME**: Explicaciones locales interpretables y agnósticas al modelo: aproxima el modelo localmente con un modelo más simple e interpretable.
  - Comprender **por qué** su modelo hace ciertas predicciones es crucial para la confianza y la responsabilidad.

Conclusión

¡Gracias! ¿Preguntas?