Etapas de Proyecto Machine Learning

Dr. José Eduardo Ochoa Luna

Universidad Católica San Pablo | Departamento de Ciencia de la Computación

29 Octubre 2023

Conjuntos de Datos

- UC Irvine Machine Learning Repository
- Kaggle datasets
- Amazon’s AWS datasets
- http://dataportals.org/
- http://opendatamonitor.eu/
- http://quandl.com/
- Quora.com question
- Datasets subreddit

Etapas Proyecto Machine Learning

1. Ver el panorama general del problema.
2. Obtener los datos.
3. Explorar los datos para obtener información.
4. Preparar los datos para exponer mejor los patrones a los algoritmos de ML.
5. Explorar modelos diferentes y seleccionar los mejores.
6. Ajustar los modelos y combinarlos en una gran solución.
7. Presentar la solución.
8. Iniciar, supervisar y mantener el sistema.

1. Ver el panorama general del Problema

- Definir el objetivo en términos de negocio.
- ¿Cómo se utilizará su solución?
- ¿Cuáles son las soluciones/soluciones alternativas actuales (si las hay)?
- ¿Cómo debería plantear este problema (supervisado/no supervisado, online/offline, etc.)?
- ¿Cómo se debe medir el desempeño?
- ¿La medida de desempeño está alineada con el objetivo comercial?

1. Ver el panorama general del Problema

- ¿Cuál sería el desempeño mínimo necesario para alcanzar el objetivo de negocio?
- ¿Cuáles son los problemas comparables? ¿Se pueden reutilizar experiencias o herramientas?
- ¿Hay experiencia humana disponible?
- ¿Cómo solucionarías el problema manualmente?
- Enumerar las suposiciones que usted (u otros) han hecho hasta ahora.
- Verificar suposiciones

2. Obtener datos

- Enumerar los datos que necesita y cuántos necesita.
- Encontrar y documentar dónde puede obtener esos datos.
- Comprobar cuánto espacio ocupará.
- Consultar las obligaciones legales y obtener autorización si es necesario.
- Cree un ambiente de trabajo (con suficiente espacio de almacenamiento).
- Obtener los datos.

2. Obtener datos

- Convertir los datos a un formato que pueda manipular fácilmente (sin cambiar los datos en sí).
- Asegúrarse de que la información confidencial se elimine o proteja (por ejemplo, anonimizada).
- Comprobar el tamaño y tipo de datos (series temporales, datos geográficos, etc.).
- Hacer el muestreo de un conjunto de prueba, separlo y dejarlo de lado (¡sin espiar datos!)

3. Explorar los datos

- Crear una copia de los datos para su exploración
- Crear un Jupyter para mantener un registro de la exploración de datos.
- Estudiar cada atributo y sus características:
  - Nombre
  - Tipo (categórico, int/float, acotado/ilimitado, texto, estructurado, etc.)
  - % de valores faltantes
  - Ruido y tipo de ruido (estocástico, valores atípicos, errores de redondeo, etc.)
  - Tipo de distribución (gaussiana, uniforme, logarítmica, etc.)

3. Explorar los datos

- Para tareas de aprendizaje supervisadas, identifique el atributo objetivo.
- Visualizar los datos.
- Estudiar las correlaciones entre atributos.
- Estudiar cómo se resolvería el problema manualmente.
- Identificar transformaciones que quizás desee aplicar.
- Identificar datos adicionales que serían útiles.
- Documentar lo aprendido.

4. Preparar datos para ML

- Trabajar con copias de los datos
- Escribir funciones para todas las transformaciones:
  - Para preparar fácilmente los datos la próxima vez que obtenga un conjunto de datos nuevo.
  - Para aplicar estas transformaciones en futuros proyectos
  - Para limpiar y preparar el conjunto de prueba
  - Para limpiar y preparar nuevas instancias de datos una vez que la solución esté activa

4. Preparar datos para ML

Limpieza de datos:

- Corregir o eliminar valores atípicos .
- Completar los valores faltantes (por ejemplo, con cero, media, mediana...) o eliminar sus filas (o columnas).

Selección de características (opcional):

- Eliminar los atributos que no proporcionen información útil para la tarea.

4. Preparar datos para ML

Ingeniería de características:

- Discretizar características continuas.
- Descomponer características (por ejemplo, categóricas, fecha/hora, etc.).
- Agregar transformaciones prometedoras de características (por ejemplo, log(x), sqrt(x), x^2, etc.).
- Agregar características nuevas y prometedoras.

Escalar variables: estandarizar o normalizar variables.

5. Listar modelos prometedores

- Si los datos son muy grandes, es mejor tomar muestras de conjuntos de entrenamiento más pequeños para poder entrenar muchos modelos diferentes en un tiempo razonable
- Intentar automatizar estos pasos tanto como sea posible.

5. Listar modelos prometedores

- Entrenar modelos de diferentes categorías (por ejemplo, lineales, Naive Bayes, SVM, random forests, redes neuronales, etc.) utilizando parámetros estándares.
- Medir y comparar su desempeño.
- Para cada modelo, utilizar N-fold cross validation, calcular la desviación estándar de la medida de rendimiento.

5. Listar modelos prometedores

- Analizar las variables más significativas para cada algoritmo.
- Analizar los tipos de errores que cometen los modelos.
- ¿Qué datos habría utilizado un humano para evitar estos errores?
- Realizar una ronda rápida de selección e ingeniería de características.
- Realizar una o dos iteraciones rápidas más de los pasos anteriores.

5. Listar modelos prometedores

Hacer una lista corta de los tres a cinco modelos más prometedores, prefiriendo modelos que cometan diferentes tipos de errores.

6. Afinar el sistema

- Utilizar la mayor cantidad de datos posible para este paso, especialmente a medida que avanza hacia el final del ajuste.
- Automatizar el proceso.
- Ajustar los hiperparámetros mediante validación cruzada.
- Tratar las opciones de transformación de datos como hiperparámetros (por ejemplo, ¿debería reemplazar los valores faltantes con cero o con el valor mediano? ¿O simplemente eliminar las filas?).

6. Afinar el sistema

- A menos que haya muy pocos valores de hiperparámetros para explorar, preferir la búsqueda aleatoria a la búsqueda en grid.
- Probar los métodos Ensemble. Combinar los mejores modelos a menudo funcionará mejor que ejecutarlos individualmente.

6. Afinar el sistema

- Una vez que esté seguro de su modelo final, medir el rendimiento en el conjunto de prueba configurado para estimar el error de generalización.
- Advertencia: no modificar el modelo después de medir el error de generalización: simplemente se comenzaría a sobreajustar el conjunto de prueba.

7. Presentar la solución

- Documentar lo que ha hecho.
- Crear una presentación.
- Resaltar primero el panorama general.
- Explicar por qué su solución logra el objetivo comercial.
- Presentar los puntos interesantes que haya notado en el camino.
- Describir qué funcionó y qué no.
- Enumerar sus suposiciones y las limitaciones de su sistema.

8. Iniciar

- Preparar la solución para producción (conectar a las entradas de datos de producción, escribir pruebas unitarias, etc.).
- Escribir un código de monitoreo para verificar el rendimiento online del sistema a intervalos regulares y activar alertas cuando caiga.
- Monitorear la lenta degradación: los modelos tienden a “empeorar” a medida que evolucionan los datos.

8. Iniciar

- Medir el desempeño puede requerir intervención humana supervisar la calidad de las entradas (por ejemplo, un sensor defectuoso que envía valores aleatorios o la producción de otro equipo se vuelve obsoleta).
- Volver a entrenar los modelos periódicamente con datos nuevos (automatizar tanto como sea posible).