



Aprendizaje de Máquina



Objetivo

- Obtener un panorama del aprendizaje de máquina
- Entender el funcionamiento de algunas de las técnicas de esta disciplina
- Obtener experiencia en su uso



Motivación

- Cada vez más datos digitalizados
 - Medicina, economía, sensores, política, publicidad...
- Cada vez máquinas más poderosas
- Necesitamos algoritmos para aprovechar las nuevas oportunidades
- Emular y automatizar algunos de los procesos cognitivos del ser humano
 - Las computadoras como extensión de la inteligencia
 - Video <https://www.youtube.com/watch?v=7Pq-S557XQU>



Bibliografía

- *The Elements of Statistical Learning*
 - Hastie, Tibshirani y Friedman
- *Pattern Recognition*
 - Duda, Hart y Stork
- Bishop, C. M., *Pattern Recognition and Machine Learning*
- Marsland, S., *Machine Learning: An Algorithmic Perspective*
- Artículos varios y material en internet



Herramientas

- Python
- R
- Excel
- Datos
 - <http://archive.ics.uci.edu/ml/datasets.html>
 - <http://www.quandl.com/>
 - <http://catalog.data.gov/dataset>
 - <http://datos.gob.mx/>
 - <https://www.kaggle.com/>



Enfoque

- Nuestra aproximación a estas técnicas será desde el punto de vista del minado de datos
 - También sirven, por ejemplo, para programar robots autónomos



Etapas para el Minado de Datos (1)

- Limpieza
 - Remover datos ruidosos, inconsistentes, etc. Adjudicar valores.
- Integración
 - Combinar las diferentes fuentes de datos
- Selección
 - Seleccionar el subconjunto de los datos relevante para el estudio. Si hay suficientes datos, guardar un subconjunto de estos para probar el modelo resultante
- Transformación
 - Seleccionar atributos, generar atributos agregados, convertir tipos de variables, etc



Etapas para el Minado de Datos (2)

- Minado
 - Utilizar técnicas de clasificación y regresión (una tarea de minado involucra, por lo general, varias técnicas)
- Evaluación
 - Identificar los resultados (patrones) interesantes (¿Qué es interesante?)
- Presentación
 - Usar técnicas de visualización para presentar los resultados obtenidos
 - Generar un sistema o protocolo para repetir el proceso con nuevos datos (de ser necesario)
- Nota el proceso no es necesariamente lineal pues en ocasiones es necesario regresar a etapas anteriores



Etapa de Minado

Aprendizaje

- ¿Qué es aprender?
- ¿Para qué hacer que una computadora aprenda?
 - Buscamos entender, encontrar relaciones, similitudes, diferencias, invariantes. Buscamos predecir
 - Buscamos modelar un fenómeno sin tener el conocimiento explícito del proceso subyacente
- Al aprendizaje de máquina consiste de varias técnicas (familias de funciones) para aproximar el proceso que genera lo observado



Etapa de Minado

Aprendizaje

- Una definición (Mitchell, Machine Learning)
 - Se dice que un programa de computadora aprende de su experiencia E con respecto a una tarea T y función de evaluación F , si su desempeño en la tarea T (con respecto a la evaluación F) mejora con la experiencia E
- En general un problema de aprendizaje debe precisar:
 - La clase de tareas a las que se refiere
 - La función de evaluación
 - La fuente de experiencia
- Ya definido esto podemos seleccionar un modelo (a.k.a función objetivo) y ajustarlo para maximizar su desempeño



Ejemplos

- Encontrar la tendencia de una acción
 - Tarea: Predecir el precio futuro de una acción
 - F.e: Ganancias
 - Experiencia: Movimientos históricos de un año
- Predecir si un día es bueno para jugar tenis
 - Tarea: Clasificar días como buenos o malos para jugar tenis
 - F.e: Porcentaje de días bien clasificados y porcentaje de días mal clasificados
 - Experiencia: Historia de un mes



Ejemplos

- Segmentar un grupo de clientes de supermercado
 - Tarea: Crear n categorías de clientes. Clasificar a cada cliente como miembro de una categoría
 - F.e: Que los grupos tengan sentido estratégico. Disminuir los costos de publicidad e incrementar ventas. (Evaluación indirecta)
 - Experiencia: Transacciones en el supermercado



Tipos de Problemas

- Lo que queremos aprender es una función que dado un ejemplo (un dato) nos entregue un valor
 - Si el valor es numérico se conoce como **regresión**
 - El valor de una acción
 - Si el valor es categórico se conoce como **clasificación**
 - Si un día es bueno o no para jugar tenis



Tipos de Técnicas

- Dependiendo de si tenemos disponible el valor de la función objetivo para los ejemplos de entrenamiento, las tareas de aprendizaje se dividen en:
 - Aprendizaje Supervisado
 - Se utilizan los datos de entrenamiento y el valor correcto para cada uno de ellos de la función objetivo (la función que intentamos aprender)
 - Árboles de decisión, redes neuronales
 - Aprendizaje No-supervisado
 - Sólo se le presentan datos, se desconoce el valor objetivo de los ejemplos de entrenamiento
 - Técnicas de agrupamiento (“clustering”)
 - K-medias, EM, redes neuronales



Lo que hay que hacer

- Ya definido el problema debemos:
 - Determinar los ejemplos con los que vamos a entrenar (fuente de experiencia) y ponerlos a modo
 - Escoger los atributos que utilizaremos de cada ejemplo
 - Normalizar, escalar
 - Escoger el algoritmo de aprendizaje. Qué tan expresiva queremos que sea nuestra representación?
 - Si es muy expresiva necesitaremos muchos ejemplos para poder aprender (y distinguir entre las distintas hipótesis)
 - Si es poco expresiva habrá conceptos que no se pueden representar
 - Evaluación de desempeño
 - Presentar resultados



Detalle Importante Acerca de la Fuente de Experiencia

- En que grado representan los ejemplos utilizados la realidad
 - Puede sesgar mucho el resultado
 - Lo ideal es que la los ejemplos que se usan para entrenar (aprender) sigan la misma distribución que los ejemplos que se encontrará en el mundo. Pero en ocasiones es necesario sesgar
 - e.g. Difícilmente podremos crear una buena segmentación de clientes de supermercado, si entrenamos sólo con datos de un día de la semana



Procedimiento (ideal) para generar un modelo

1. Dividir los datos en un conjunto de entrenamiento y uno de prueba
 - El conjunto de prueba no hay que usarlo para nada!! Ni siquiera para normalizar. Nada, solo para probar al final
 - La regla de dedo es 75% para entrenamiento y 25% para pruebas
2. El conjunto de entrenamiento puede a su vez contener un conjunto de validación
 - El conjunto de entrenamiento (sin el de validación) se usa para ajustar el modelo.
 - El conjunto de validación sirve para comparar diferentes parámetros del modelo (para los que los tienen). Por ejemplo queremos decidir entre si ponerle 4 u 8 neuronas a la capa intermedia de una RN
 - La extracción del conjunto de validación se hace por cada experimento (validación cruzada)
3. Ya elegido el modelo y sus parámetros. Entrenar con todo el conjunto de entrenamiento y reportar desempeño con el conjunto de prueba

Nota: Si no hay suficientes datos para tener un conjunto de prueba intacto entonces se utiliza sólo el paso 2 para seleccionar parámetros. El modelo se reentrena finalmente con todos los datos.



Comentario

- La mayoría de las fallas en la aplicación del aprendizaje de máquina son aquí, en el procedimiento. Es muuuy fácil equivocarse y entrenar, de alguna forma, con información de los datos de prueba



Temario

- Técnicas de aprendizaje supervisado: cubriremos la mayoría de las siguientes:
 - Métodos clásicos:
 - Regresión lineal
 - Regla del perceptrón
 - Redes Neuronales (algo de deep learning)
 - Máquinas de Soporte Vectorial (SVM)
 - Árboles de decisión
 - K-vecinos cercanos
 - Métodos de ensamble
- Selección de modelos
- Aprendizaje no-supervisado
 - K-medias
 - Jerárquico