Aprendizaje Automático sobre Grandes Volúmenes de Datos

Clase 1 - 11 de Agosto 2014

Pablo Ariel Duboue, PhD

Universidad Nacional de Córdoba,
Facultad de Matemática, Astronomía y Física

Qué es el aprendizaje automático sobre grandes volúmenes de datos

- Aprendizaje Automático: un nuevo paradigma de programación
- Esta materia: cuando los datos y modelos no entran en RAM / disco de una sola máquina
- Importante para América latina porque no hay muchas máquinas / recursos

A quiénes está dirigida esta materia

- Estudiantes avanzados de carreras de grado
- Estudiantes de posgrado
- Profesionales del campo
- Prerequisitos:
  - Conocimientos de programación
  - Álgebra (particularmente álgebra matricial).
  - Probabilidad y Estadística
  - Redes y Sistemas Distribuídos (o similar, al menos Sistemas Operativos).

Estructura del curso

Tres partes:

1. Aprendizaje Automático (teórico)
2. Computo Distribuido (teórico)
3. Práctica (mahout/hadoop)

Parte I

- Modelos, Ingeniería de Features.
- Clasificación
  - Árboles de decisión
  - Regresión logística
  - SVMs
- Clustering
  - kMeans
  - Clustering estadístico
- Recomendación

Parte II

- Conceptos de Cómputo Distribuido
  - Map/Reduce
  - Teorema CAP
  - Operaciones Matriciales Distribuidas
  - Gradiente
  - Búsqueda distribuida
  - Algoritmos actualizables
  - Colas, shared memory
- Paralelizando algoritmos de Aprendizaje Automático

Parte III

- Implantación
  - Hadoop
    - Map
    - Reduce
  - Mahout
    - Recomendación
    - Clustering
    - Clasificación
  - ActiveMQ e Híbridos
- Casos de estudio

Casos de estudio

- Delicado equilibrio entre lo factible y lo útil
  - Datos disponibles
  - Problemas interesantes
- Clasificación: nombres para métodos compilados (http://keywords4bytecodes.org)
- Recomendación: paquetes para Debian
- Clustering: identificación de páginas co-editadas en Wikipedia

Evaluación

- Estudiantes presenciales
  - Prácticos
  - Parcial
  - Proyecto / monografía
- Oyentes / estudiantes remotos
  - Multiple opción en línea
  - Proyecto
- Ambos: competencia kaggle in class en trámite

Acerca del docente

- Licenciado en Computación UNC-FaMAF
- Doctorado "Indirect Supervised Learning of Strategic Generation Logic"
  - Defendido Enero 2005, Columbia University, NYC
- IBM Research (2005-2010)
  - Sistema DeepQA Watson (Jeopardy! Grand Challenge)
  - Systems team
  - Subsistema de aprendizaje automático ("A framework for merging and ranking of answers in DeepQA", IBM Journal of R&D)
- Consultoria
  - LinkedIn / FB / Legal / Inmobiliario / Soporte técnico
- Software Libre
  - Thoughtland (http://thoughtland.duboue.net)

Aprendizaje Automático

- ¿Nuevo paradigma de programación?
  - La vuelta al concepto de Soft Computing de los años 1980-1990

Algoritmos con error intrínseco

- ¿Qué hacer con un programa que falla aún habiendo sido programado correctamente?
- No todos los problemas pueden ser abordados vía Aprendizaje Automático
- Incluir el error dentro del modelo de uso

Datos

- Limpieza de datos es fundamental
- La tarea que más trabajo lleva en una implantación de Aprendizaje Automático
- Hay una diferencia infinita entre "tenemos datos" y "estos datos son útiles y listos para hacer Aprendizaje Automático"

Aprendizaje Automático como compilación

- El Aprendizaje Automático puede ser parte de un sistema de compilación por lotes (*build system*)
- Sin embargo, las necesidades de cómputo de un build system son muy inferiores a las de un sistema de Aprendizaje Automático
- Los detalles de ingeniería de software relacionados con la implantación de sistemas de Aprendizaje Automático son claves y muchas veces dejados de lado

Algoritmos vs. teoría

- A medida que el campo va pasando de investigadores a profesionales, el enfoque cambia de ventajas teóricas a practicas
- Popularización de sistemas híbridos
- Ingeniería de features
- No-free lunch theorem

Clasificación

- El Aprendizaje Automático sin calificar
  - Aprender lo que uno ya sabe
- Tratar de aprender una función $f(x_1, \ldots, x_n) \rightarrow y$ donde
  - x_i son las caracterísita de aprendizaje (*features*) de entrada
  - y es la clase objetivo
- La clave es *extrapolación*, queremos que la función **generalize** a entradas nunca vistas.
  - Interpolación lineal es en sí una forma de hacer Aprendizaje Automático supervisado.

Una visión como desarrolladores

- Entrenamiento/Estimación/"compilación":
  - Entrada: vectores de *features*, incluyendo la clase objetivo
  - Salida: un **modelo** entrenado
- Ejecución/Predicción/"interpretado":
  - Entrada: vectores de features, sin la clase objetivo, más el modelo entrenado
  - Salida: la clase objetivo predicha

Ejemplo: Biología

- "Disambiguating proteins, genes, and RNA in text: a machine learning approach," Hatzivassiloglou, Duboue, Rzhetsky (2001)
- El mismo término se usa para referirse a genes, proteinas y RNA mensajero (mRNA):
  - "By UV cross-linking and immunoprecipitation, we show that SBP2 specifically *binds* selenoprotein *mRNAs* both in vitro and in vivo."
  - "The SBP2 *clone* used in this study generates a 3173 nt transcript (2541 nt of coding sequence plus a 632 nt 3' UTR truncated at the polyadenylation site)."
- Esta ambigüedad es tan fuerte que en muchos casos los autores del texto insertan la palabra "gen", "proteina" o "mRNA" ellos mismos para desambiguar
  - Esto ocurre sólo en el 2.65 % de los casos

Biología: *features*

- Nos fijamos en un contexto de palabras alrededor del término y usamos el número de veces otras palabras aparecen en dicho contexto como *features*
- Llevamos la cuenta de cuantas veces aparece cada palabra con la clase objetivo:

term	gene	protein	mRNA
PRIORS	0.44	0.42	0.14
D-PHE-PRO-VAL-ORN-LEU		1.0	
NOVAGEN	0.46	0.46	0.08
GLCNAC-MAN	1.0		
REV-RESPONSIVE	0.5	0.5	
EPICENTRE		1.0	
GENEROUSLY	0.33	0.67	

Biología: resultados

- Se realizó una ingeniería de features bastante extendida
- Se quitaron mayúsculas/minúscula, se lematizó, se filtró por tipo de palabras y se adjunto información posicional
  - Se cambió también el problema de clasificar entre tres clases a clasificar entre dos clases
- Árboles de decisión y Naive Bayes produjeron resultados comparables (76% en 2 y 67% en 3).
- De los árboles de decisión salieron reglas interesantes:
  - after ENCODES is present
    before ENCODES is NOT present
    $\Longrightarrow$class gene [96.5%]

Ejemplo: Etiquetado morfosintáctico del francés

- Trabajo realizado para KeaText, una empresa especializada en extracción de información bilingüe
- Se transformó un etiquetador morfosintáctico para el francés que esta escrito como una mezcla entre Python y Perl
  - En vez de hacerle ingeniería inversa, se lo ejecutó sobre un gran conjunto de documentos en francés
  - Se entrenó un modelo de MaxEnt sobre él
- En menos de dos días de trabajo se obtuvo un etiquetador en Java con una caida del 5% respecto del original

Francés: *features*

- Etiquetado morfosintáctico es parecido a la tarea de desambiguación anteriormente descrita
- El problema es más complicado, porque hay más clases y todas las palabras deben ser etiquetadas
- Algunas *features* utilizadas:
  - La palabra misma
  - Sufijos, incluyendo hasta las últimas cuatro letras de la palabra
  - Prefijos, incluyendo hasta las primeras cuatro letras de la palabra
  - Palabras anteriores, con sus etiquetas ya identificadas
  - Si la palabra tiene caracteres especiales o es sólo números o mayúsculas

Aprendizaje sin supervisación

- Descubrir lo que uno no sabe
  - También llamado minería de datos
- Concepto clave: función de distancia entre las instancias
- Concepto clave: el valor (sorpresa) de los datos descubiertos
  - Cerveza al lado de pañales descartables
- Se trata de aplicar una estructura a los datos
  - El tipo de estructura está dado por el algoritmo, y de la estructura se pueden leer características importantes de los datos

Ejemplo: recursos humanos

- Trabajo realizado para MatchFWD, un micro-emprendimiento en Montreal en el 2011
- Queremos agrupar empresas en función de que tan similares son sus culturas corporativas
- Distancia: qué tan similares son dos empresas basado en la gente que trabajo en ambas:

$$distancia(compañía_1, compañía_2) = \frac{|gente\ trabajó\ para\ ambas|}{|gente\ trabajó\ para\ cualquiera|}$$

Recursos humanos: resultados

- 17 mil empresas en 3 mil grupos
- Muchos grupos tenían sentido
- Difícil de evaluar
- Faltaban datos para obtener un valor práctico
  - Pensábamos usar los clusters en la función de matching de personas con trabajos pero sólo eran útiles en un 10% de los casos

Thoughtland

- Visualizando superficies de error n-dimensionales
- Aprendizaje Automático con Weka (nube de errores cros-validada)
- Clustering con Apache Mahout (usando clustering estadísitico)
- Generación de texto (usando OpenSchema y SimpleNLG)
- http://thoughtland.duboue.net
  - Scala
  - Software libre: https://github.com/DrDub/Thoughtland

Recomendación

- Un problema en el medio entre supervisado y no supervisado
- Dado un conjunto de personas y objetos, recomendar nuevos objetos similares a los que la persona elige, pero que no conoce
- Amplia utilidad práctica

Aprendizaje por refuerzo

- No lo vamos a ver en la clase, pero es otro tipo de Aprendizaje Automático
- La función de evaluación está disponible sólo al final de una serie de acciones
- Muy útil para entrenar inteligencias artificiales para juegos de video

Recopilación de datos

- El recopilado de datos es crucial
- Puede requerir mucho esfuerzo y cambio de procesos complejos
  - Por ejemplo, evitar la destrucción de un envase utilizado en un procedimiento médico y su posterior análisis
- El concepto de Garbage-In-Garbage-Out se aplica aquí más que nunca
- Gran diferencia entre investigación (conseguir datos, hacer experimentos, comunicar resultados) y uso profesional (donde la adquisición de datos es continua)

Anotación

- Muchas veces la clase objetivo tiene que ser calculada a mano por grupos de personas designadas para la tarea
- Guías de anotación
- Cros y auto concordancia
- Tareas aburridas, inclusive a veces más fáciles para computadoras que para personas

Entrenamiento

- Manteniendo datos y modelos en sincronía
- Proceso por lotes vs. tiempo real
- Entendiendo como funciona el modelo en la práctica
  - Una base de datos es un sistema de Aprendizaje Automático muy pobre
  - Esto depende del poder expresivo del modelo, el tema de nuestra próxima clase

Entendiendo el error

- Como los sistemas de Aprendizaje Automático funcionan por extrapolación es fácil confundirse
  - Un sistema puede funcionar correctamente en los casos disponibles y aún así exhibir un comportamiento patológico en la práctica que lo deja inútil.
- Para ello se utilizan conjuntos separados de entrenamiento y testeo
  - Pero eso no es suficiente, con el tiempo el profesional del Aprendizaje Automático desarrolla intuiciones sobre el conjunto de testeo. A partir de cierto punto esas intuiciones son erróneas y hay que cambiar de conjunto.

La democratización del cómputo

- Algunas ideas inspiradas en la presentación de Alistair Croll durante la semana de Bigdata en Montreal
  - http://www.slideshare.net/Tiltmill/cycle-time-trumps-scale-big-data-as-the-organizational-nervous-system-montreal-big-data-week-2014
- Computo, lleva a automatizar cosas, las redes llevan a interconectar pero el gran volúmen de datos lleva a predecir y cambiar cosas
- Antes había que elegir sólo dos de entre volúmen, velocidad y variedad
  - Bibliotecas: gran cantidad de datos variados pero lentas
  - Máquina de ordenar monedas: gran cantidad de monedas y rápido pero no son variadas

Los resultados inesperados de la abundancia

- Los estudios y algoritmos que estamos usando no son nuevos
  - Pero su uso indiscriminado lo es
- Antes existían soluciones específicas para grandes volúmenes de datos, a un costo muy elevado
  - Censo
  - Bancos
- Eficiencia $\Longrightarrow$ menores costos $\Longrightarrow$nuevos usos $\Longrightarrow$ $\Longrightarrow$mayor demanda $\Longrightarrow$mayor consumo.
  - Con más poder de cómputo, las necesidades de procesamiento de grandes volúmenes de datos están disparándose
  - La gente tiene necesidad de acceder a tecnología antes reservada para gobiernos y empresas multinacionales

Hadoop

- La parte práctica de la materia va a ser sobre Apache Hadoop
- Un sistema distribuido de proceso por lotes (y mejoras recientes fuera del proceso por lotes) con un sistema de archivos distribuido que prioriza el cómputo en datos locales a cada nodo
  - Implementa el paradigma Map/Reduce
- Muy estable, muy utilizado, muy conocido
- Muchos detractores

Un nodo hadoop

- Según Dirk deRoos et al (2014), un nodo hadoop cuenta en la actualidad con (página 216):
  - de 6 a 12 terabytes de disco y
  - de 48 a 96 gigabytes de RAM
- Una máquina con estas características (capaz con menos disco) está disponible para alquilar en Amazon Web Services por menos de 2 dólares la hora

Representando una instancia

- Representaciones planas
  - Nuevos sistemas permiten representar árboles o grafos
- Tipos de features: booleanas (true, false), numéricas (4, 1, -5), de punto flotante (0.5, 1, 9.2), enumeraciones (rojo,azul,verde,amarillo)

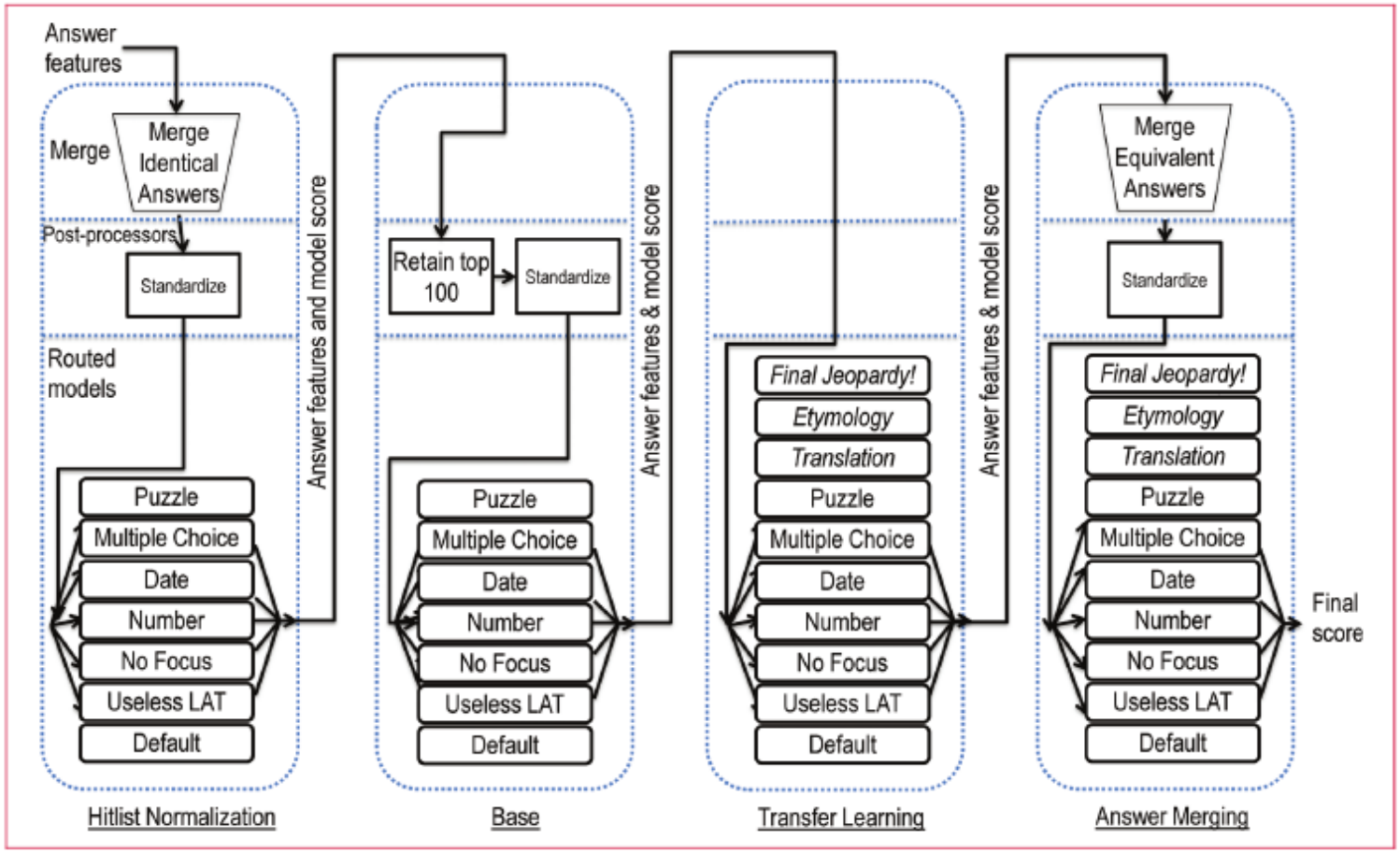
Representando valores complejos

- Texto
  - Cada palabra es un elemento en una enumeración
  - Cada palabra es un feature binario
  - Cada palabra es un feature numérico dependiendo de cuantas veces aparece
- Conjuntos
  - Features binarios por la presencia de cada elemento
  - Clase objetivo de conjuntos: entrenar un sistema independiente por cada elemento

Inventando features

1. Probar con todo lo que se le ocurra a uno
2. Reflexionar
    1. ¿Qué información utilizaría **usted** para resolver ese problema?
    2. Mire trabajos publicados
        - Artículos de investigación: http://aclweb.org/anthology-new/
        - Blogs
        - Proyectos de software libre
3. Agregue *features* computables
    - ¡Aprender a sumar requiere montañas de datos!

Feature engineering



Las primeras cuatro fases de aprendizaje y ranqueo, de Gondek, Lally, Kalyanpur, Murdock, Duboue, Zhang, Pan, Qiu, Welty (2012)

Algunas librerías de Aprendizaje Automático

- Scikit-learn (Python)
- R packages (R)
- Weka (Java)
- Mallet (CRF, Java)
- OpenNLP MaxEnt (Java)
- Apache Mahout (Java)
- ...
- ...