Machine Learning for research

August 10, 2019

0.0.1 CICATA

8 de Agosto del 2019

Dr. Eric Dolores,
matemático en NewSci labs.
ericd@newsci.ai

Aquí les comparto lo que en mi experiencia es el camino mas sencillo para que un estudiante de postgrado aprenda ML.

0.0.2 Motivación.

En FSU creamos una clase de ML para estudiantes de doctorado con nociones básicas de programación.

https://mendozacortesgroup.github.io/MachineLearningForHumans/

Abajo vemos dos imagenes, en una removimos objetos usando ML en 2 segundos.

[2]:
```
from IPython.display import Image
```

[2]:
```
Image("img/carlos.png")
```

[2]:

[3]:
```
Image("img/Carlo.jpg")
```

[3]:

Ejemplo de edición de imagen.
https://www.nvidia.com/research/inpainting/selection
Lo importante es que el proceso no requiere conocimientos de herramientas de edición y fue relativamente rapido.

Abajo vemos un ejemplo de un algoritmo que traduce, solo que en ves de traducir de un idioma a otro, traduce de selfie a anime, o de caballo a cebra, gato a perro, etc.

[21]:
```
Image("img/gan.jpg")
```

[21]:

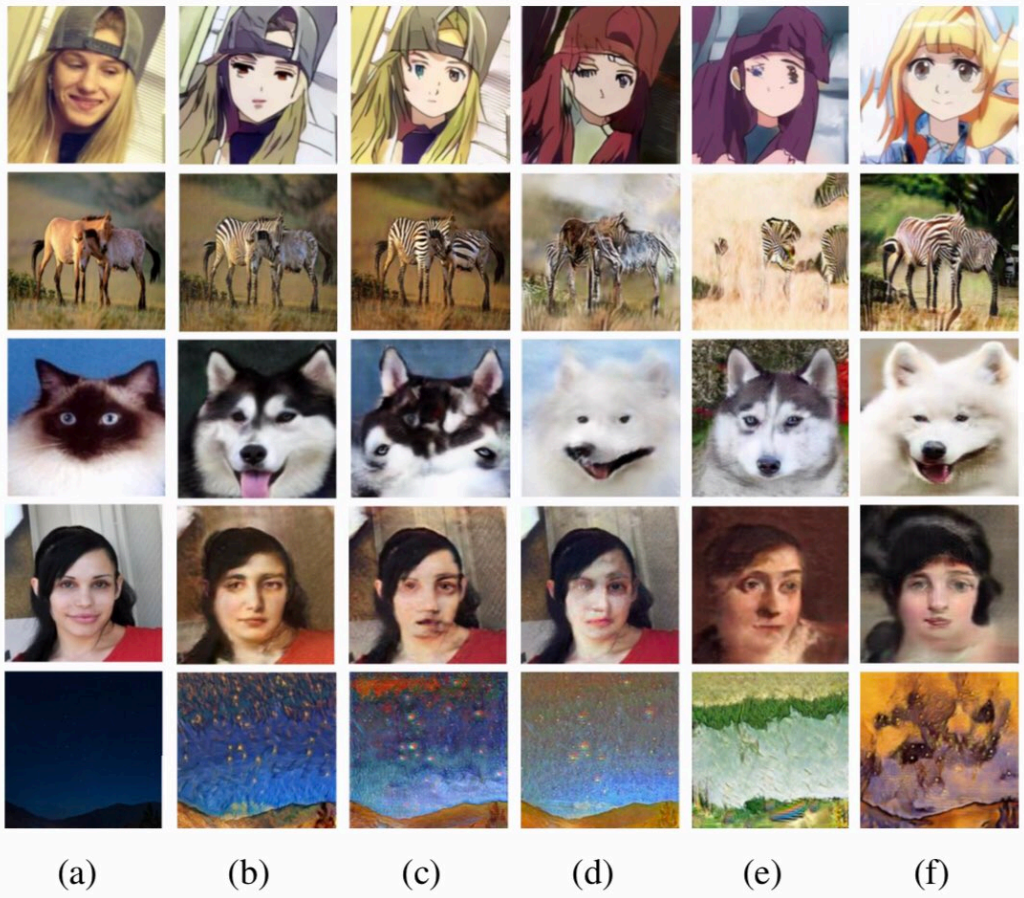


Figure 6. Visual comparisons on the five datasets. From top to bottom: selfie2anime, horse2zebra, cat2dog, photo2portrait, and photo2vangogh. (a) Source images, (b) U-GAT-IT, (c) CycleGAN, (d) UNIT, (e) MUNIT, (f) DRIT.

Traducción de imagen.

[Photo credit: @roadrunning01]

Quizas una de las aplicaciones mas conocidas, ha sido en el ajedrez, donde a un algoritmo (voy a mentir un poco en estas notas) se le dieron las reglas de ajedrez, y el algoritmo jugo consigo mismo. Se le dió una manera de evaluar su juego y el algoritmo buscaba maneras de mejorar la evaluación.

Como el algoritmo nunca vió juegos de humanos, no aprendió de nuestros errores.

Sus juegos se han descrito como raros, porque al usar sus propias estrategias hace cosas que no habiamos considerado. Abajo vemos dos videos donde Maestros del ajedrez se maravillan por las jugadas del algoritmo. Los grandes Maestros pasan horas aprendiendo de esta maquina.

[1]:
```
from IPython.display import YouTubeVideo
YouTubeVideo('YgZEaP6Qte0?t=249')
```

[1]:
```
<IPython.lib.display.YouTubeVideo at 0x27c9de90588>
```

[2]:
```
YouTubeVideo('lFXJWPhDsSY?t=654')
```

[2]:
```
<IPython.lib.display.YouTubeVideo at 0x27c9deb7a58>
```

Se ha calculado que el Nivel ELO de el algoritmo es 3200, un valor superior a lo que los mejores ajedrecistas jamás han alcanzado.

Elo:

```
2882    Magnus Carlsen
3200    Alpha Zero
```

0.0.3 Ejemplos en Ciencias

El mismo equipo que trabajó con el algoritmo de ajedrez, usó sus técnicas para atacar un problema imporante para la ciencia.

Si sabemos que aminoacidos se necesitan en una proteina, £podemos predecir la estructura de la proteina? esto es muy importante porque muchas enfermedades son resultado de protenias con errores en su estructura.

https://deepmind.com/blog/alphafold/

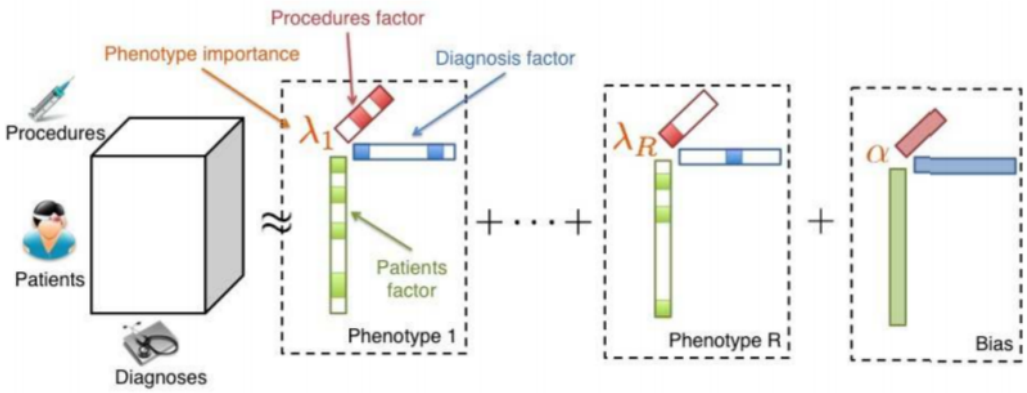
Abajo tenemos el resultado de trabajo realizado en la Universidad de Northwestern. El algoritmo usa nociones avanzadas de algebra, pero la explicacion es sencilla:

Durante dos años anotamos la interacción de pacientes con problemas en el corazoón. En una tabla ponemos como columnas los medicamentos y como renglones el diagnóstico.

Si tenemos un paciente con enfermedad 'y' y se le dió la medicina 'x', entonces ponemos 1 en la intersección del renglón 'y' y columna 'x', y ponemos zero en los demas valores. Asi a cada paciente le asigmanos una matriz con la mayoría de las entradas zero. Y ahora consideramos un nuevo eje, el de el paciente, y ponemos las matricies por cada paciente. Terminaremos con una matriz de tres dimensiones.

[3]:
```
Image("img/tensor.png")
```

[3]:



A esta matriz la llamamos tensor e intentamos factorizarla en suma de tensores irreducibles.

Factorizacion en tensores de rango uno muestra dianósticos concurrentes con medicación, este proceso es llamado fenotipo. El estudio de fenotipos en machine learning permitió el descubrimiento de grupos distintos de Hearth Failure with preserved Ejection Fractions (HFpEF), esos grupos diferian marcadamente en características clínicas, de estructura/función, para mas detalles ver:

```
Phenomapping for Novel Classification of Heart Failure With Preserved Ejection Fraction Sanji
```

[Photo credit: Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization]

A continuación vemos los últimos avances de ML en Ciencia de Materiales:

https://community.apan.org/wg/afosr/w/researchareas/22949/scientific-autonomous-reasoning-agent-sara-integrating-materials-theory-experiment-and-computation/

También les recomiendo visitar el siguiente Webinar:

Machine Learning, AI, and Data Driven Materials Development and Design.

[3]:
```
YouTubeVideo('DBknkNvYlcE')
```

[3]:

El siguiente artículo busca usar ML en el estudio de estructuras de cristales: Crystal Structure Prediction via Deep Learning

https://pubs.acs.org/doi/10.1021/jacs.8b03913

Aquí el problema es que la información consiste de coordenadas y etiquetas (atomo de H en la posicion (1,.003,4), etc). Así que primero se construyen imagenes que son invariantes de la elección de la celda unitaria y de ahi se usan metodos tradicionales de procesamiento de imagenes.

0.0.4 ML en acción.

El siguiente tutorial les muestra la creación de un clasificador.

http://www.r2d3.us/una-introduccion-visual-al-machine-learning-1/

Si ya saben que es una red neuronal, aqui pueden mejorar su intuición. http://playground.tensorflow.org/

0.0.5 £De verdad necesito ML?

[8]:
```
Image("img/fomo.jpg")
```

[8]:

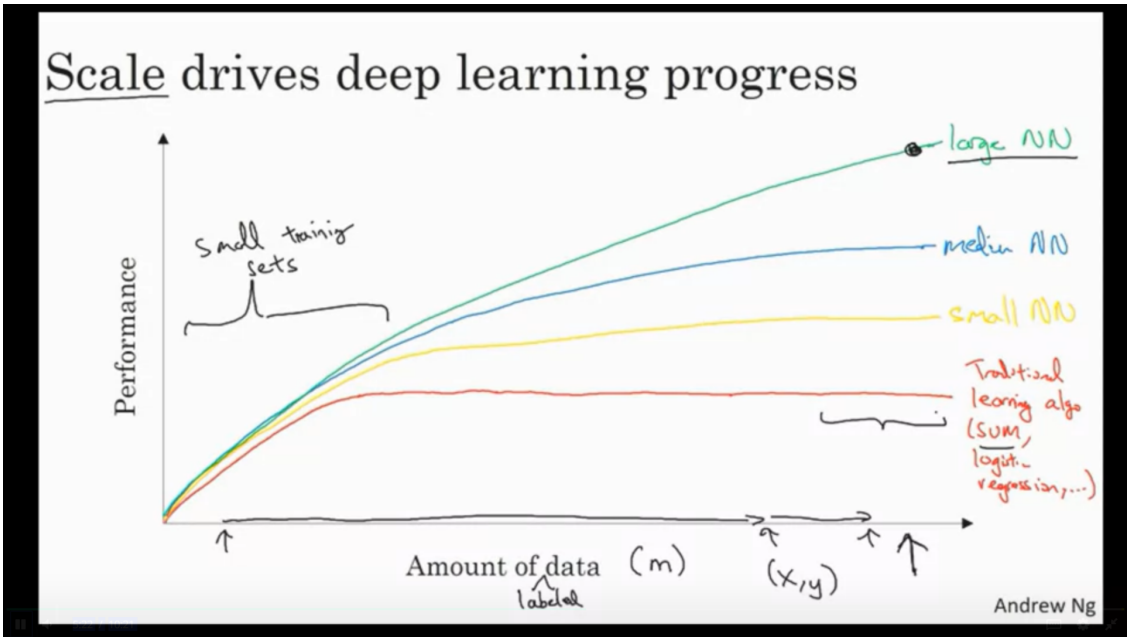
[Photo credit: PlusLexia.]

£Le conviene a los estudiantes aprender sobre ML? si el estudiante decide no seguir por el rumbo académico, podrá trabajar como científico de datos.

£Necesito machine learning? Yo respondo esa pregunta con otra pregunta: £Tienes una cantidad considerable de datos o planeas obtener miles-millones de datos? Los metodos usuales como la regresion o SVM son suficientemente buenos cuando uno trabaja con pocos datos. Las redes neuronales tienen un mejor desempeño cuando se maneja big data, miles o millones de datos:

[9]:
```
Image("img/nn.png")
```

[9]:



[Photo credit: Andrew Ng]

£Qué matemáticas necesito estudiar? Cálculo de varias variables, álgebra lineal, probabilidad...básicamente los primeros dos años de la ESFM, quizás un poco menos.

Esta es un area muy informal, pero hay trabajo de físicos intentando crear las bases por ejemplo:

https://arxiv.org/abs/1608.08225 usando física se ha intentado explicar el funcionamiento de las redes neuronales.

£Qué deberian saber mis estudiantes?

Esto es muy subjetivo, yo sugiero:

[10]:
```
Image("img/pylogo.png")
```

[10]:

(Python)

El curso mas sencillo de Python.

https://runestone.academy/runestone/books/published/thinkcspy/index.html

[11]:
```
Image("img/sklearn.png")
```

[11]:

80% de los problemas se resuelven con sklearn.

Si estas en el caso de big data, no recomiendo usar redes neuronales en sklearn, es mejor usar KERAS:

[13]:
```
Image("img/keras.png")
```

[13]:

Keras es una libreria que te permite crear redes neuronales.

Una vez que estes familiarizado puedes experimentar con pytorch, que tiene mensajes de error mas sencillos de entender y otras propiedades que dan flexibilidad.

[14]:
```
Image("img/pytorch.png")
```

[14]:

Pytorch.

Y ya que estoy dando recomendaciones, les sugiero el siguiete sofware para crear graficas interactivas, vean el link:

[15]:
```
Image("img/altair.png")
```

[15]:

Altair. https://altair-viz.github.io/gallery/scatter_linked_brush.html

Para estar al día en ML estas son las cuentas en Twitter a seguir:

```
@drfeifei : Fei-Fei Li - Director of Stanford AI Lab, creator of ImageNet
@ylecun: (Yann LeCun) Leading Facebook AI/ML Research
@karpathy : Andrej Karpathy - Teaches Deep Learning at Stanford, Research Scientist at OpenAI
@AndrewYNg ( Andrew Ng) : Led ML research in google, now leads AI research in Baidu
@Kdnuggets ( Gregory Piatetsky) : Leading KDnuggets - tweets/retweets lots of relevant stuff.
@OpenAI
@googleresearch
@BaiduResearch
```

Aqui hay una sugerencia de projectos posibles en Mexico:

Calentamiento global.

https://arxiv.org/pdf/1906.05433.pdf Este archivo nos dice las posibles areas donde se puede empezar a trabajar y los correspondientes algoritmos que se podrian usar.

(Denuncia anónima) un Doctor del IMSS opera a pacientes para implantar prótesis de rodilla en casos innecesarios. Si alguien tiene accesso a informacion del IMSS se puede usar SVM para ver si su comportamiento es abnormal, y con evidencia estadística pedir que otros medicos revisen si se esta comentiendo alguna arbitrariedad.

(Busqueda de fraude) se puede analizar si los miembros del SNI cambian sus patrones de publicación, y si lo hacen de manera abrupta, verificar sus artículos para evitar fraude. Esto es muy comun en areas como la química donde las revistas no pueden reproducir los experimentos, pues muchos de esos experimentos toman años.

0.0.6 Posibles problemas.

En las noticias decian que de cada 20 projectos de IA, solo tres son exitosos. Algunas razones son:

- Plausibilidad. £Es posible resolver tu problema con ML? Esto se soluciona hablando con expertos, como NewSci.
- Ya tienes los datos o necesitas a alguien que los capture. Esta es la parte mas delicada y que consume mas tiempo.
- Análizis y preprocesamiento de datos. Hay que establecer contacto con quien recogió los datos, ver que fue de manera objetiva, que notación usarón, etc.
- Equipo. £Tienes la capacidad de trabajar con big data?
- Infra estructura. Si es una interacción internacional, quizas crear una aplicación, etc.
- Interpretación de resultados. Es importante tener una fundación de matemáticas que te permita entender los resultados. Esta es la razón por la que los p-values estan cayendo de moda en la estadística. Nunca dierón un si o un no, sino que nos daban evidencia.

En la parte de recolección de datos, se necesita verificar que se hace de manera objetiva, por ejemplo, nosotros descubrimos prejuicios en los algoritmos de un escaner en 3d cuando creabamos la exhibición 'adaptation'. https://www.sciartmagazine.com/straight-talk-justus-harris.html

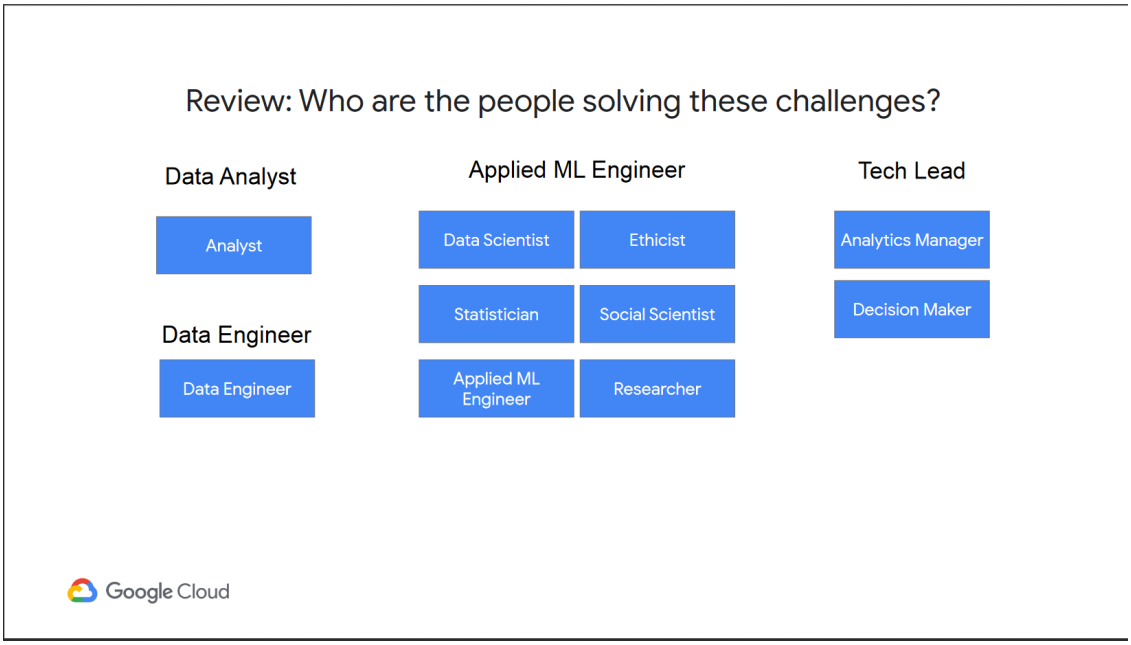
Básicamente, no pudimos escanear la imagen de una mujer por su tez obscura. El algoritmo también se usa en jugetes, y es muy triste que un niño de tes cafe no pueda usar el jugete mientras un niño blanco si. Es peligroso que un coche no reconozca a los mexicanos solo porque fue creado en EU y solo les intereso que reconozca a la gente de piel blanca.

Recomendamos tomar el curso en equidad de Google https://developers.google.com/machine-learning/crash-course/fairness/video-lecture?utm_campaign=mle-outreach&utm_medium=blog&utm_source=keyword-blog&utm_content=mlcc-fairness

En tu projecto es importante delimitar responsabilidades, pues enterder algoritmos de ML y crear la base de datos requiren conocimientos distintos:

[16]:
```
Image("img/roles.png")
```

[16]:



[Photo credit: Google]
Yo recomiendo (sin que ellos me lo pidieran) GCP.

[17]:
```
Image("img/gcp.png")
```

[17]:

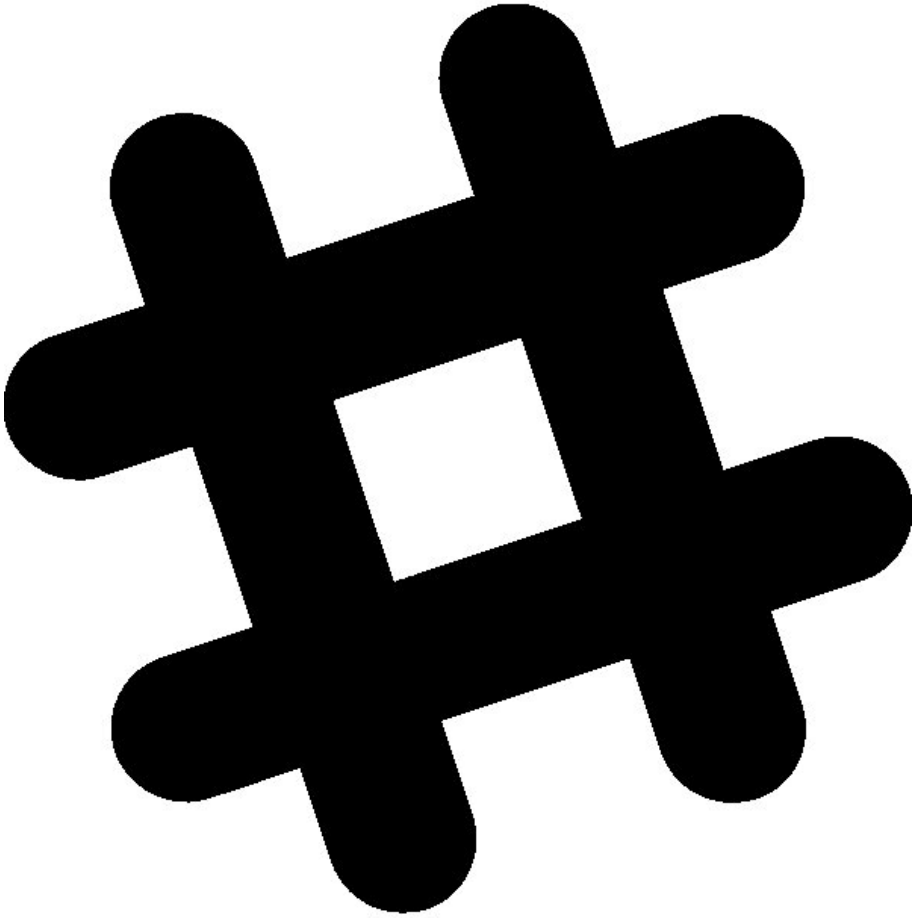
Google Cloud Platform tiene herramientas de machine learning con big data en tres niveles.

```
-Automatico para gente sin conocimientos de programacion.
-Sql te permite usar algoritmos predeterminados.
-TensorFlow para gente con experiencia en ML.
```

Y ellos tienen planes de descuento para investigadores:
https://edu.google.com/intl/es-419/why-google/higher-ed-solutions/?modal_active=none

[18]:
```
Image("img/slack.jpg")
```

[18]:



Nosotros tenemos un grupo en slack sobre ML en la ciencia. Aquí la gente compila aplicaciones y artículos científicos que usan ML:
https://tinyurl.com/FSUMachineLearning

[19]:
```
Image("img/NewSci.png")
```

[19]:

£Dudas sobre la plausibilidad de su projecto? contactenos ericd@newsci.ai