Categorización de productos con Deep Learning

Keras, Tensorflow, Pandas, Numpy, etc

Agustín Sarasúa / agustin.sarasua@mercadolibre.com

Pablo Zamudio / pablo.zamudio@mercadolibre.com



Agenda

- Intro / Problema
- Conceptos generales
 - Al vs ML vs DL
 - ¿Por que Deep Learning?
 - Tipos de algoritmos
- Redes Neuronales
 - Cost & Loss Function
 - Gradient Descent
- Intro Natural Language Processing
 - Tokenization
 - Vectorization
- Workflow
- Intro a Python, Numpy y Pandas
- Entrenar una NN

Intro / Problema

"Desmalezadora Bordeadora Gardenplus Gp Naftera"



Accesorios para Vehículos



Accesorios para Vehículos Alimentos y Bebidas Animales y Mascotas Arte y Antigüedades Bebés Cámaras y Accesorios Celulares y Telefonía Coleccionables Computación Consolas y Videojuegos Deportes y Fitness Electrodomésticos y Aires Ac. Electrónica, Audio y Video Herramientas y Construcción Hogar, Muebles y Jardín

Industrias y Oficinas



Hogar, Muebles y Jardín > Jardín y Exterior > Máquinas para el Jardín > Desmalezadoras y Repuestos



Accesorios para Vehículos Alimentos y Bebidas Animales y Mascotas Arte y Antigüedades Bebés Cámaras y Accesorios Celulares y Telefonía Coleccionables Computación Consolas y Videojuegos Deportes y Fitness Electrodomésticos y Aires Ac. Electrónica, Audio y Video Herramientas y Construcción Hogar, Muebles y Jardín Industrias y Oficinas

Baño
Cocina y Bazar
Comedor
Decoración
Dormitorio
Escritorio
Iluminación para el Hogar
Jardín y Exterior
Lavadero y Limpieza
Living
Otros

Buzones
Calefactores de Exterior
Cercas
Contenedores de Residuos
Decks
Escaleras
Gazebos
Herramientas de Jardín
Hornos y Parrillas
Jardinería
Máquinas para el Jardín
Miscola de Jardín
Piscinas
Riego
Sombrillas

Accesorios para Piscinas

Bordeadoras y Repuestos Cortacercos Cortadoras de Césped Desmalezadoras y Repuestos Fumigadores Hidrolavadoras Motosierras y Repuestos Sopladoras y Aspiradoras Otras Máquinas para el Jardín



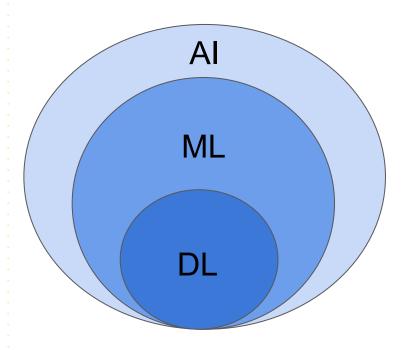
Intro / Problema

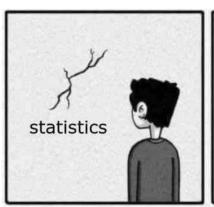
- 20 árboles de categoría (1 por país)
- +3K categorías
 - o E.g. CELLPHONES, TABLETS, BICYCLES, etc.

Agenda

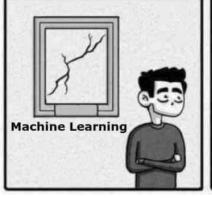
- Intro / Problema
- Conceptos generales
 - Al vs ML vs DL
 - ¿Por que Deep Learning?
 - Tipos de algoritmos
- Redes Neuronales
 - Cost & Loss Function
 - Gradient Descent
- Intro Natural Language Processing
 - Tokenization
 - Vectorization
- Workflow
- Intro a Python, Numpy y Pandas
- Entrenar una NN

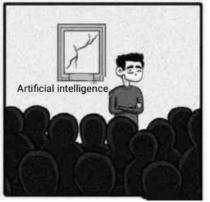
Al vs ML vs DL



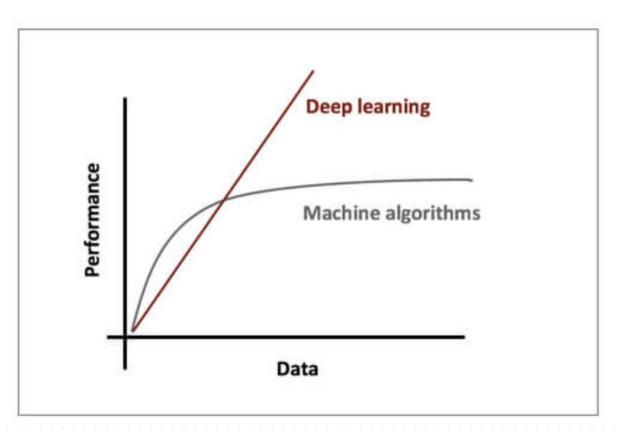








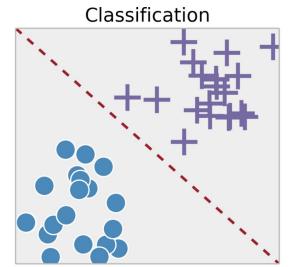
¿Por que Deep Learning?

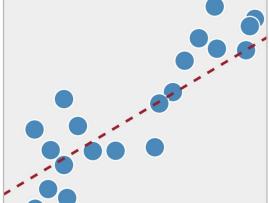




- Supervisados
 - Clasificación
 - Regresión

- No supervisados
 - Clustering
- Reinforcement Learning
- etc...



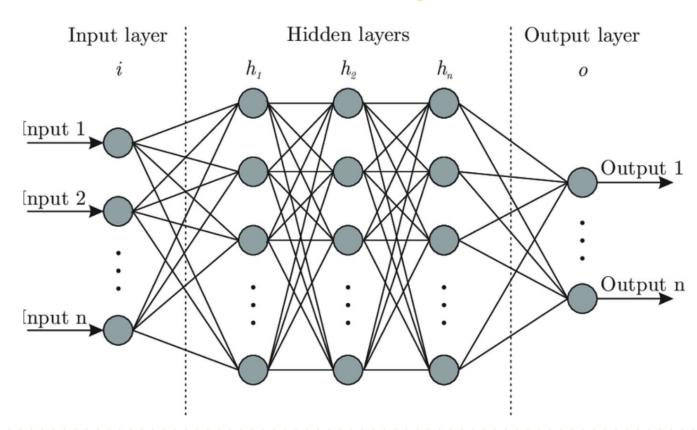


Regression

Agenda

- Intro / Problema
- Conceptos generales
 - Al vs ML vs DL
 - ¿Por que Deep Learning?
 - Tipos de algoritmos
- Redes Neuronales
 - Cost & Loss Function
 - Gradient Descent
- Intro Natural Language Processing
 - Tokenization
 - Vectorization
- Workflow
- Intro a Python, Numpy y Pandas
- Entrenar una NN

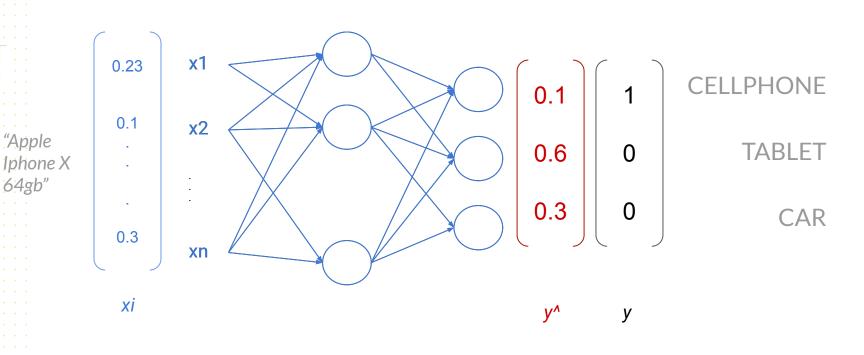
Redes Neuronales - Arquitectura



Redes Neuronales - Conceptos

- Parameters: Los aprende el modelo
- Hyperparameters:
 - Lo que podemos ajustar. Ej: #unidades #capas, learning rate
- Loss Function:
 - Aplicada a un solo ejemplo del training set
- Cost Function:
 - Aplicada al training set completo
 - Efectividad de los parámetros aprendidos sobre training set
- Features: El input que usamos para alimentar el modelo

Visualizando una predicción



Loss Function: Categorical Cross Entropy

$$L(y, \hat{y}) = -\sum_{i}^{C} y \log(\hat{y})$$

$$L(y,y^{\wedge}) = -1^{*}\log(0.1) = 2.302$$

$$L(y,y^{\wedge}) = -1^{*}\log(0.9) = 0.105$$

Cost Function

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y \log(\hat{y})$$

Para 1 dato de entrenamiento

Para todos los datos de

$$J(\theta) = \frac{1}{m} \sum_{i}^{m} L(y_{i}, \hat{y}_{i})$$

Softmax Activation

Softmax: Squashes a vector in the range of (0,1) and sum(vector) = 1

$$f(s)_i = \frac{e^{s_i}}{\sum\limits_{j}^{C} e^{s_j}}$$

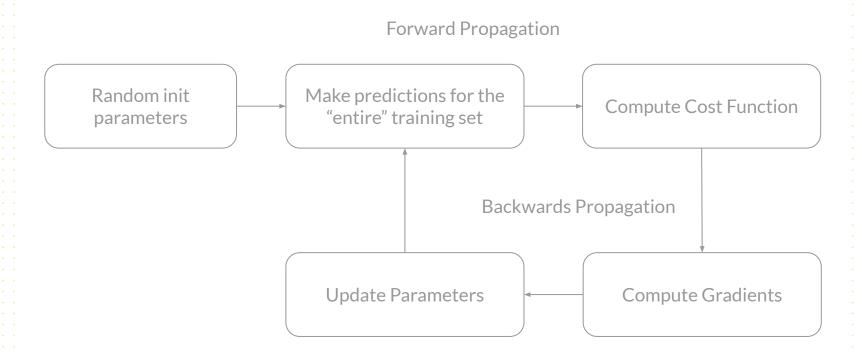
$$0.01 \text{ Class 1}$$

$$0.14 \text{ Class 2}$$

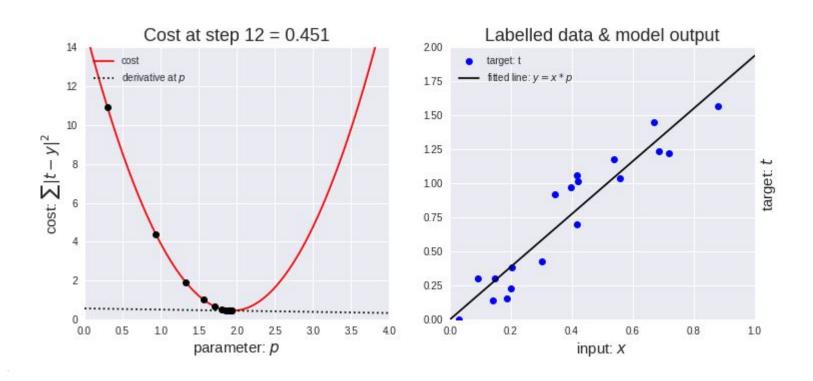
$$0.85 \text{ Class 3}$$

Output layer Multi-class classification

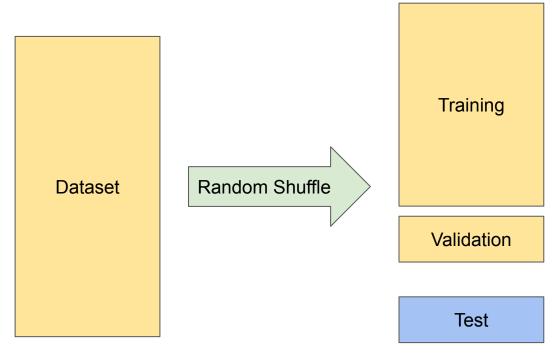




Gradient Descent



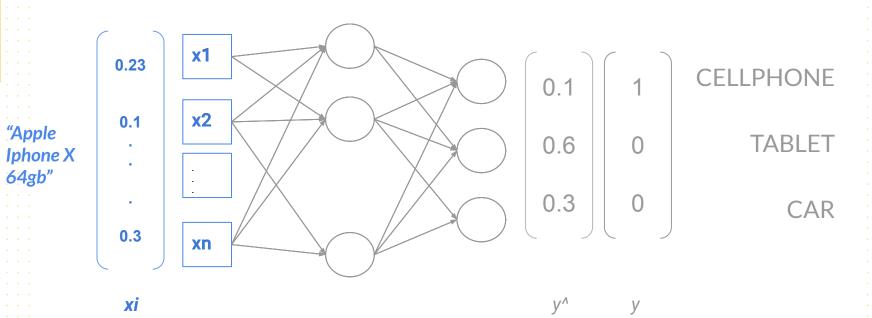
Split de Datasets



Agenda

- Intro / Problema
- Conceptos generales
 - Al vs ML vs DL
 - ¿Por que Deep Learning?
 - Tipos de algoritmos
- Redes Neuronales
 - Cost & Loss Function
 - Gradient Descent
- Intro Natural Language Processing
 - Tokenization
 - Vectorization
- Workflow
- Intro a Python, Numpy y Pandas
- Entrenar una NN

Preparar los datos





Tokenization

"Oferta!!! Cómoda 4 Cajones Blanco Miel Dormitorios."

- ¿Que es el texto?
 - Secuencia de caracteres
 - Secuencia de palabras
 - Secuencia de frases
 - Secuencia de oraciones
 - Secuencia de párrafos

Tokenization

"Oferta!!! Cómoda 4 Cajones Blanco Miel Dormitorios."



oferta comoda 4 cajones blanco miel dormitorios





Token normalization

- Stemming
 - eliminar sufijos de las palabras para retornar una nueva forma de la palabra (stem).
 - o ej: perros, perras, perritos -> perr
- Lemmatization
 - usando un vocabulario retornar el significado de diccionario de la palabra (lemma).
 - o ej: peces -> pez



Vectorization

Bag of Words

Vocabulario →

Oferta! Celular apple iphone 8 64gb +
funda apple de regalo

Funda para samsung galaxy oferta

Tablet apple ipad con funda

	apple	funda	samsung		celular	
	2	1	0		1	•••
•••	0	1	1	••••	0	•••
	1	1	0		0	



N-grams

Problema: Vocabulario muy extenso (muchas features)

celular apple iphone 8	
funda samsung galaxy oferta	
tablet apple ipad con funda	

Celular apple	apple	Apple iphone	 funda
1	1	1	 0
0	0	0	 1
0	1	0	 1

Vectorization

- **TF-IDF**: Term frequency Inverse Document Frequency
- TF: cantidad de ocurrencias del término t en el documento d
- IDF: importancia del término **t** en el corpus (todos los documentos)

Ejemplo: "Vendo iphone de primera de 64gb de color negro"

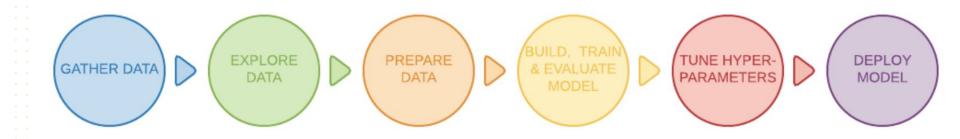
TF-IDF("iphone") = TF("iphone") * IDF("iphone") = (1/9) * (número grande) TF-IDF("de") = TF("de") * IDF("de") = (3/9) * (número pequeño)

$$\operatorname{idf}(t,D) = \log rac{N}{|\{d \in D: t \in d\}|}$$

Agenda

- Intro / Problema
- Conceptos generales
 - Al vs ML vs DL
 - ¿Por que Deep Learning?
 - Tipos de algoritmos
- Redes Neuronales
 - Cost & Loss Function
 - Gradient Descent
- Intro Natural Language Processing
 - Tokenization
 - Vectorization
- Workflow
- Intro a Python, Numpy y Pandas
- Entrenar una NN

Workflow



https://developers.google.com/machine-learning/guides/text-classification/

Agenda

- Intro / Problema
- Conceptos generales
 - Al vs ML vs DL
 - ¿Por que Deep Learning?
 - Tipos de algoritmos
- Redes Neuronales
 - Cost & Loss Function
 - Gradient Descent
- Intro Natural Language Processing
 - Tokenization
 - Vectorization
- Workflow
- Intro a Python, Numpy y Pandas
- Entrenar una NN



Repositorio en Github:

http://bit.ly/workshop-categorizacion-productos

Notebooks:

- 1. Intro Python, Numpy, Pandas
- 2. Train a NN for Products Categorization
 - a. Gather & Explore Data (solución)
 - b. Prepare the data for training (solución)
 - Build, train & evaluate model (<u>solución</u>)

Próximos pasos

#MeLiDataChallenge!!!

- Workshop de fast.ai con baseline (incluye grabación y notebooks)
- <u>Inscribirse</u> y participar para ganar entradas para <u>Khipu</u>!!

Toda la info en:

https://ml-challenge.mercadolibre.com/

Muchas gracias

