Machine Learning Fundamentals

Rodrigo Alfaro Pinto

Sobre mí

- Más de 15 años en el mundo TI.
- Ingeniero Civil en Computación, Universidad de Chile.
- MBA, Universitat Politècnica de València (UPV).
- Me esfuerzo por promover la innovación para crear ventajas competitivas. Viajero, soñador y speaker.
- Co Fundador de Social Tech Chile
- www.linkedin.com/in/ralfcl

Software a utilizar

- Anaconda.
- Jupyter.
- Python

Agenda del día

- Machine Learning.
- Matriz de confusión.
- Creación de algoritmos para Machine Learning.
- Tensor Flow
- Ejercicios.

Repositorio del curso

http://bit.do/fcq8U

Machine learning

Machine learning

Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente.

Aprender en este contexto quiere decir identificar patrones complejos en millones de datos.

La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros.

Machine Learning

UML (unsupervised learning)

- **Reducción de la dimensionalidad (Dimensionality reduction),** donde el objetivo es identificar patrones en las características de los datos. La reducción de la dimensionalidad se usa a menudo para facilitar la visualización de los datos, así como un método de preprocesamiento antes del aprendizaje supervisado.

Machine Learning

UML (unsupervised learning)

UML presenta desafíos y beneficios específicos:

- No hay un solo objetivo en UML
- Generalmente hay muchos más datos sin etiquetar disponibles que los datos etiquetados.

Machine Learning

SML (supervised learning)

En el aprendizaje supervisado (SML), el algoritmo de aprendizaje se presenta con entradas de ejemplo etiquetadas, donde las etiquetas indican la salida deseada.

SML en sí está compuesto de clasificación, donde el resultado es cualitativo, y regresión, donde el resultado es cuantitativo.

Matriz de confusión

Matriz de confusión

Toolbox del Data Science

La matriz de confusión de un problema de clase n es una matriz nxn en la que las filas se nombran según las clases reales y las columnas, según las clases previstas por el modelo.

Sirve para mostrar de forma explícita cuándo una clase es confundida con otra. Por eso, permite trabajar de forma separada con distintos tipos de error.

Matriz de confusión

Toolbox del Data Science

El problema radica en que al medir la precisión del algoritmo de forma binaria no distinguimos entre los errores de tipo falso positivo y falso negativo, tratandolos como si ambos tuvieran la misma importancia.

Para esto existe la matriz de confusión.

Matriz de confusión

Toolbox del Data Science

Matriz de confusión

- a es el número de predicciones correctas de clase negativa (negativos reales).
- b es el número de predicciones incorrectas de clase positiva (falsos positivos).
- c es el número de predicciones incorrectas de clase negativa (falsos negativos).
- d es el número de predicciones correctas de clase positiva (positivos reales).

Matriz de Confusion		Predicho			
		Negativo	Positivo		
Real	Negativo	a	b	***Verdadero Negativo (True negative rate)***	a/(a+b)
	Positivo	c	d	***Exactitud***	d/(c+d)
		Sensibilidad	***Especifidad***		
		d/(d+c)	a/(a+b)	**Precisión**=(a+d)/(a+b+c+d)	

Matriz de confusión

Precisión o “Accuracy” (AC)

Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud.

Cuanto menor es la dispersión mayor la precisión.

Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones.

$$AC = \frac{a + d}{a + b + c + d}$$

Matriz de confusión

Exactitud o “Precision”(P)

Se refiere a lo cerca que está el resultado de una medición del valor verdadero.

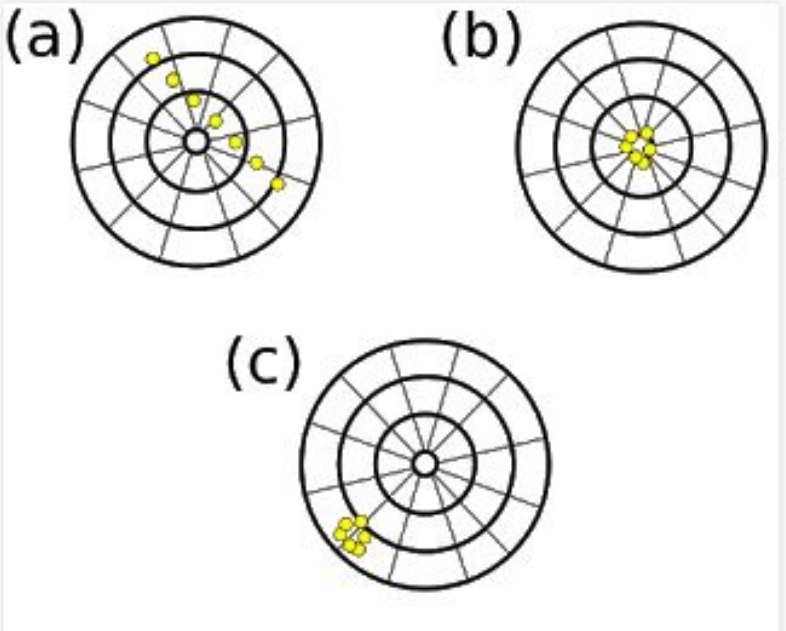
En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación.

También se conoce como Verdadero Positivo (“True positive rate”).

Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

$$P = \frac{d}{b + d}$$

Matriz de confusión



La figura (b) representa un resultado exacto y preciso.

La (C) es preciso, pero no exacto.

(a) no es ni una cosa ni la otra.

Matriz de confusión

Sensibilidad y Especificidad

Son dos valores que nos indican la capacidad de nuestro estimador para discriminar los casos positivos, de los negativos.

La sensibilidad es la fracción de verdaderos positivos, mientras que la especificidad, es la fracción de verdaderos negativos.

Matriz de confusión

Sensibilidad o “Recall”(TP, True Positive Rate)

También se conoce como Tasa de Verdaderos Positivos (True Positive Rate, TP).

Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

$$TP = \frac{d}{c + d}$$

Matriz de confusión

Especificidad o Specificity

Es la Tasa de Verdaderos Negativos (true negative rate TN).

Se trata de los casos negativos que el algoritmo ha clasificado correctamente.

$$TN = \frac{a}{a + b}$$

Matriz de confusión

La exactitud y la sensibilidad nos están indicando la relevancia de los resultados.

Por ejemplo, un algoritmo muy exacto, (P alto) nos dará muchos más resultados relevantes que irrelevantes, mientras que un algoritmo muy específico (TP alto), será el que detecte la mayoría de resultados de interés.

Algoritmos para Machine Learning

Algoritmos para Machine Learning

Demostración en pantalla

- K-means clustering.
- Hierarchical clustering.
- t-Distributed Stochastic Neighbour Embedding.
- Classification performance.
- Random forest.
- Decision trees.

TensorFlow

TensorFlow

¿Qué es?

TensorFlow es el sistema de aprendizaje automático de segunda generación de Google Brain, liberado como software de código abierto en 9 de noviembre de 2015. Mientras la implementación de referencia se ejecuta en dispositivos aislados, TensorFlow puede correr en múltiple CPUs y GPUs (con extensiones opcionales de CUDA para informática de propósito general en unidades de procesamiento gráfico).

TensorFlow está disponible en Linux de 64 bits, macOS, y plataformas móviles que incluyen Android e iOS.

TensorFlow

¿Qué es?

Los cómputos de TensorFlow están expresados como stateful dataflow graphs.

El nombre TensorFlow deriva de las operaciones que tales redes neuronales realizan sobre arrays multidimensionales de datos.

Estos arrays multidimensionales son referidos como "tensores".

Ejemplos en pantalla

www.linkedin.com/in/ralfcl
Gracias por su atención

Bibliografía

- Data Science & Big Data Analytics, Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015 by John Wiley & Sons, páginas 270-290.
- Big Data for dummies, 2013, Judith S. Hurwitz, Alan F. Nugent, páginas 203-227.
- An introduction to Data Science, 2013, Jeffrey Stanton, Syracuse University, páginas 75-95.