

Contexto del Curso - Administrativas

Marcelo Luis Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina

²Universidad Nacional de la Patagonia Austral, Argentina

e-mails: merreca@unsl.edu.ar, merrecalde@gmail.com



Curso: Minería de Datos
Universidad Nacional de San Luis - Año 2018

Resumen

1 Datos del Curso

- Objetivos del curso
- Organización del Curso
- Evaluación

2 Contexto de nuestro grupo de investigación

- Áreas de Investigación

3 Herramientas y Material de Estudio

- Herramientas
- Material de Estudio

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Colaboradores:

- Leticia Cagnina
- Edgardo Ferretti
- Luis Ávila

Administrativas

Profesor Responsable:

Marcelo Luis Errecalde

Colaboradores:

- Leticia Cagnina
 - Edgardo Ferretti
 - Luis Ávila

Material del curso:

Se enviará via e-mail o en links específicos a lo largo del desarrollo del curso

¿Qué se pretende con este curso?

- Introducir los **conceptos** y **herramientas** básicas fundamentales vinculados al **aprendizaje automático** y la **minería de datos**.

Objetivos del curso

¿Qué se pretende con este curso?

- Introducir los **conceptos y herramientas** básicas fundamentales vinculados al **aprendizaje automático** y la **minería de datos**.
 - Identificar las principales **etapas y procesos** de la minería de datos, los mecanismos **estadísticos** y de **aprendizaje automático** frecuentemente utilizados en esta área.

¿Qué se pretende con este curso?

- Introducir los **conceptos** y **herramientas** básicas fundamentales vinculados al **aprendizaje automático** y la **minería de datos**.
- Identificar las principales **etapas** y **procesos** de la minería de datos, los mecanismos **estadísticos** y de **aprendizaje automático** frecuentemente utilizados en esta área.
- Extender y **aplicar** los contenidos de este curso en **aplicaciones concretas** del **mundo real** de interés para el asistente.

Organización del Curso

4 encuentros:

- **1er encuentro:** 25, 26 y 27 de Octubre. Horarios: 25 y 26 de 18 a 21. 27 de 9 a 13. Sala de Postgrado 2.

Organización del Curso

4 encuentros:

- **1er encuentro:** 25, 26 y 27 de Octubre. Horarios: 25 y 26 de 18 a 21. 27 de 9 a 13. Sala de Postgrado 2.
- **2do encuentro:** 8, 9 y 10 de Noviembre. Horarios: 8 y 9 de 18 a 21. 10 de 9 a 13. Sala de Postgrado 1.

Organización del Curso

4 encuentros:

- **1er encuentro:** 25, 26 y 27 de Octubre. Horarios: 25 y 26 de 18 a 21. 27 de 9 a 13. Sala de Postgrado 2.
- **2do encuentro:** 8, 9 y 10 de Noviembre. Horarios: 8 y 9 de 18 a 21. 10 de 9 a 13. Sala de Postgrado 1.
- **3er encuentro:** 22, 23 y 24 de Noviembre. Horarios: 22 y 23 de 18 a 21. 24 de 9 a 13. Sala de Postgrado 2.

Organización del Curso

4 encuentros:

- **1er encuentro:** 25, 26 y 27 de Octubre. Horarios: 25 y 26 de 18 a 21. 27 de 9 a 13. Sala de Postgrado 2.
- **2do encuentro:** 8, 9 y 10 de Noviembre. Horarios: 8 y 9 de 18 a 21. 10 de 9 a 13. Sala de Postgrado 1.
- **3er encuentro:** 22, 23 y 24 de Noviembre. Horarios: 22 y 23 de 18 a 21. 24 de 9 a 13. Sala de Postgrado 2.
- **4to encuentro:** 13, 14 y 15 de Diciembre. Horarios: 13 y 14 de 18 a 21. 15 de 9 a 13. Sala de Postgrado 2.

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
- **Clase 2:** Aprendizaje Automático Supervisado (AAS)

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
- **Clase 2:** Aprendizaje Automático Supervisado (AAS)
- **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
 - **Clase 2:** Aprendizaje Automático Supervisado (AAS)
 - **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
 - **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
- **Clase 2:** Aprendizaje Automático Supervisado (AAS)
- **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
- **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)
- **Clase 5:** AAD (continuación) - Clustering

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
- **Clase 2:** Aprendizaje Automático Supervisado (AAS)
- **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
- **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)
- **Clase 5:** AAD (continuación) - Clustering
- **Clase 6:** Clustering (continuación)

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
- **Clase 2:** Aprendizaje Automático Supervisado (AAS)
- **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
- **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)
- **Clase 5:** AAD (continuación) - Clustering
- **Clase 6:** Clustering (continuación)
- **Clases 7 y 8:** SVM y Regresión

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
- **Clase 2:** Aprendizaje Automático Supervisado (AAS)
- **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
- **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)
- **Clase 5:** AAD (continuación) - Clustering
- **Clase 6:** Clustering (continuación)
- **Clases 7 y 8:** SVM y Regresión
- **Clase 9:** Aspectos avanzados de MD

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
 - **Clase 2:** Aprendizaje Automático Supervisado (AAS)
 - **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
 - **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)
 - **Clase 5:** AAD (continuación) - Clustering
 - **Clase 6:** Clustering (continuación)
 - **Clases 7 y 8:** SVM y Regresión
 - **Clase 9:** Aspectos avanzados de MD
 - **Clases 10 y 11:** Arquitecturas de redes neuronales

Organización del Curso

12 clases:

- **Clase 1:** Aspectos Generales
 - **Clase 2:** Aprendizaje Automático Supervisado (AAS)
 - **Clase 3:** AAS (continuación) - Material Complementario - Aplicaciones (Python y Weka)
 - **Clase 4:** Aprendizaje de Árboles de Decisión (AAD)
 - **Clase 5:** AAD (continuación) - Clustering
 - **Clase 6:** Clustering (continuación)
 - **Clases 7 y 8:** SVM y Regresión
 - **Clase 9:** Aspectos avanzados de MD
 - **Clases 10 y 11:** Arquitecturas de redes neuronales
 - **Clase 12:** Aplicaciones, casos de estudio y recursos

Evaluación del Curso

La nota final del curso resulta de:

- ① Notas obtenidas en las **actividades** evaluadas en las distintas Unidades de este programa.
- ② Desarrollo de un **Proyecto** (informe), de entre 5 y 12 páginas sobre alguno de los temas abordados.

Evaluación del Curso

La nota final del curso resulta de:

- ① Notas obtenidas en las **actividades** evaluadas en las distintas Unidades de este programa.
- ② Desarrollo de un **Proyecto** (informe), de entre 5 y 12 páginas sobre alguno de los temas abordados.

Alternativas:

- **Alternativa 1:** revisión e investigación de bibliografía actualizada sobre un tema específico de actualidad (a elección del asistente)

Evaluación del Curso

La nota final del curso resulta de:

- ① Notas obtenidas en las **actividades** evaluadas en las distintas Unidades de este programa.
- ② Desarrollo de un **Proyecto** (informe), de entre 5 y 12 páginas sobre alguno de los temas abordados.

Alternativas:

- **Alternativa 1:** revisión e investigación de bibliografía actualizada sobre un tema específico de actualidad (a elección del asistente)
- **Alternativa 2:** implementación y/o experimentación de algunas de la técnicas utilizadas

Evaluación del Curso

La nota final del curso resulta de:

- ① Notas obtenidas en las **actividades** evaluadas en las distintas Unidades de este programa.
- ② Desarrollo de un **Proyecto** (informe), de entre 5 y 12 páginas sobre alguno de los temas abordados.

Alternativas:

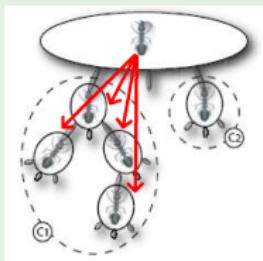
- **Alternativa 1:** revisión e investigación de bibliografía actualizada sobre un tema específico de actualidad (a elección del asistente)
- **Alternativa 2:** implementación y/o experimentación de algunas de la técnicas utilizadas
- **Alternativa 3:** uso y evaluación de una o más de las plataformas disponibles en el área

¿En qué temas trabajamos/investigamos?

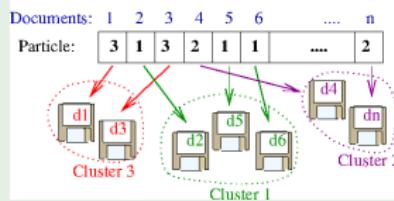
- Agrupamiento de textos cortos
- Calidad de información en la Web (esp. en [Wikipedia](#))
- Perfil del autor en medios sociales (edad, género, rasgos de personalidad, etc)
- Detección temprana de riesgos

(Agrupamiento de textos cortos - técnicas bio-inspiradas

Técnicas basadas en hormigas



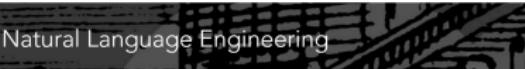
Técnicas basadas en PSO



Publicaciones recientes

 Information Sciences
Volume 265, 1 May 2014, Pages 36–49

An efficient Particle Swarm Optimization approach to cluster short texts
Leticia Cagnina^a, Marcelo Erracalde^a, Diego Ingaramo^a, Paolo Rossi^b

 Natural Language Engineering
Article
Get access
Natural Language Engineering, Volume 22, Issue 5
September 2016, pp. 687-726
Silhouette + attraction: A simple and effective method for text clustering¹

Áreas de Investigación

Calidad de Información en Wikipedia

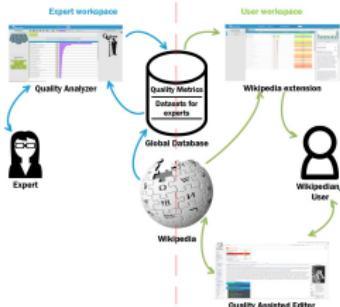
Artículos Destacados

The screenshot shows the English Wikipedia page for "United States Bicentennial coinage". The page title is "United States Bicentennial coinage". Below the title, there's a section titled "Background" with a link to "View history". The main content discusses the bicentennial coinage, mentioning the quarter, half dollar, and dollar coins. It notes that while no coins were issued in 1975, Congress passed legislation in 1976 authorizing them. The dollar coin was initially proposed by Senator Barry Goldwater but later became the "Liberty dollar". The page also mentions the "quarter dollar" and "half dollar". There are two images of coins: one showing a profile of George Washington and another showing a profile of Franklin D. Roosevelt.

Detección de Fallas

The screenshot shows the Murphaze tool interface. It displays several error notifications for different Wikipedia pages. One prominent error message is: "This article has multiple issues. Please help improve it or discuss these issues on the talk page." Another message states: "Murphaze is a tool that finds common mistakes made in articles on Wikipedia. These are some mistakes we found in this article. If you see any other mistakes, please let us know. We will fix them as soon as possible. If you think this article is complete and needs no changes, then just ignore these messages." Other messages mention missing references, self-references, and potential advertisements.

Visualización



Métricas de Calidad

The screenshot shows the Wikilyzer tool interface. It features a dashboard with various quality metrics and data visualizations. At the top, there's a search bar and a navigation menu. Below the menu, there are sections for "Quality Metrics" (with a formula $\sqrt{((A+B+C)+D)}$), "Quality Measures" (listing categories like "Article Quality", "Page Quality", "Category Quality", "User Quality", and "File Quality"), and "Recent Changes" (showing a timeline of changes). On the right side, there's a detailed view of Albert Einstein's page, including his portrait and a graph of recent edits.

Áreas de Investigación

Perfil de autor



Tareas: determinar

- ① edad
- ② género
- ③ personalidad
- ④ orientación política
- ⑤ ...

Detección de pedófilos en la Web

- Datos de entrenamiento en
www.perverted-justice.com
- Competencias recientes
pan.webis.de/clef12/pan12-web

Ejemplo:

Example 1: what nationality are u?

Exchange of information

Example 2: what r u wearing?

Grooming

Example 3: would u let me?

Example 4: thing it is me feeling u

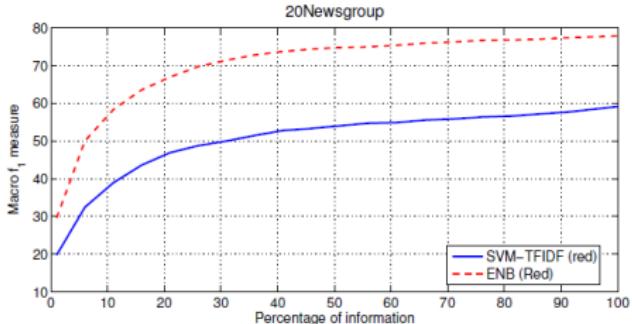
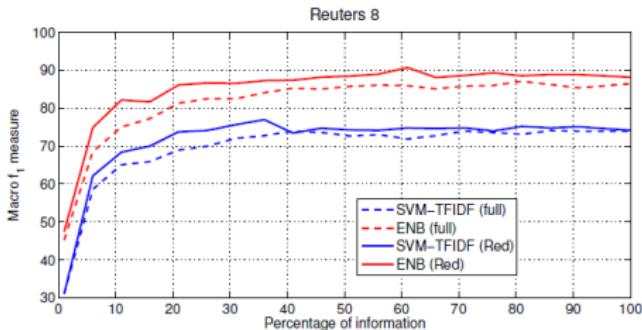
Example 5: what's your address?

Example 6: can I stay at your house overnight if i go?

Approach

Clasificación temprana / detección temprana de riesgos

- Entrenar con información secuencial completa.
- Luego clasificar, tan pronto como sea posible



La tarea ERISK 2017



Erisk 2017 - CLEF 2017 Workshop
Conference and Labs of the Evaluation Forum

Conjuntos de Datos

Entrenamiento

486 users

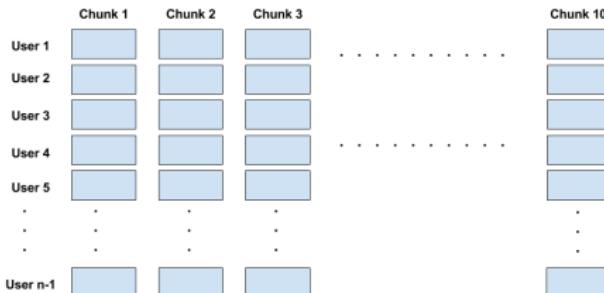
- 83 positivos (deprimidos)
- 403 negativos (no-dep.)

Testeo

401 users

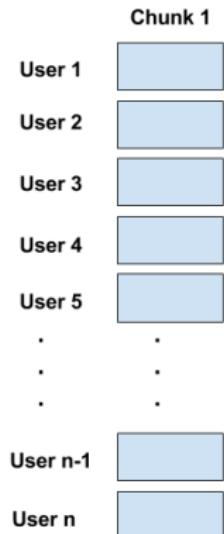
- 52 positivos (deprimidos)
- 349 negative (no-dep.)

Divididos en 10 chunks (ordenados cronológicamente):



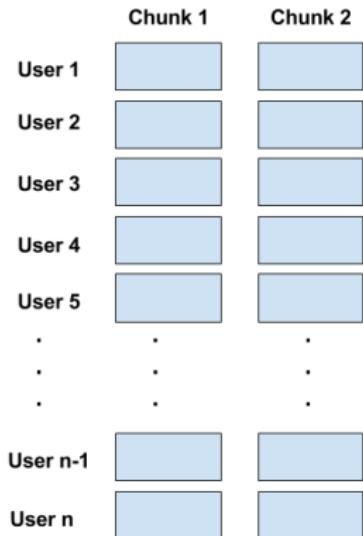
Áreas de Investigación

Datos de testeo, provistos chunk x chunk ...



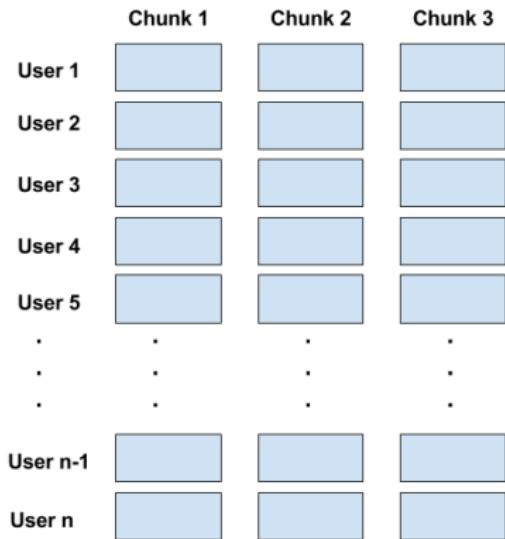
Áreas de Investigación

Datos de testeo, provistos chunk x chunk ...



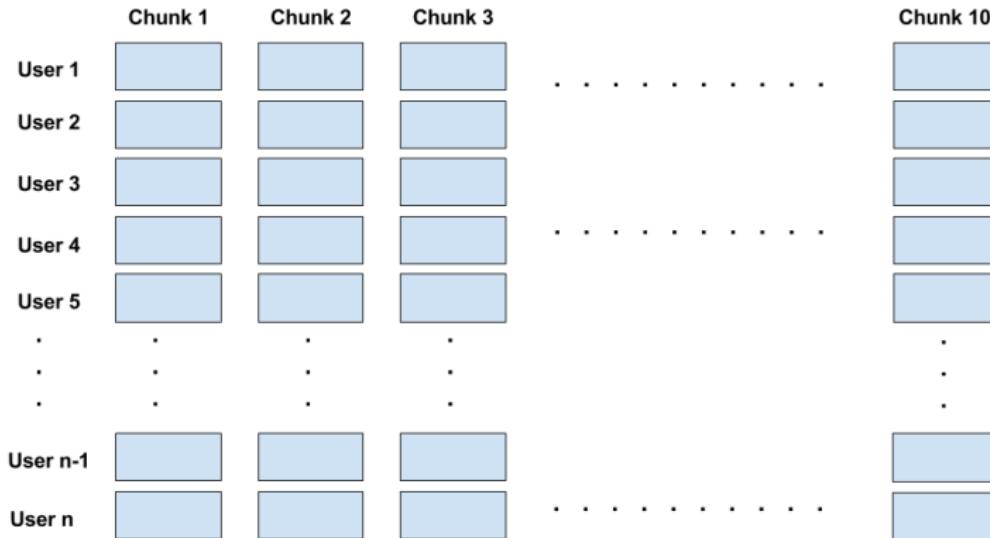
Áreas de Investigación

Datos de testeo, provistos chunk x chunk ...



Áreas de Investigación

Datos de testeo, provistos chunk x chunk ...



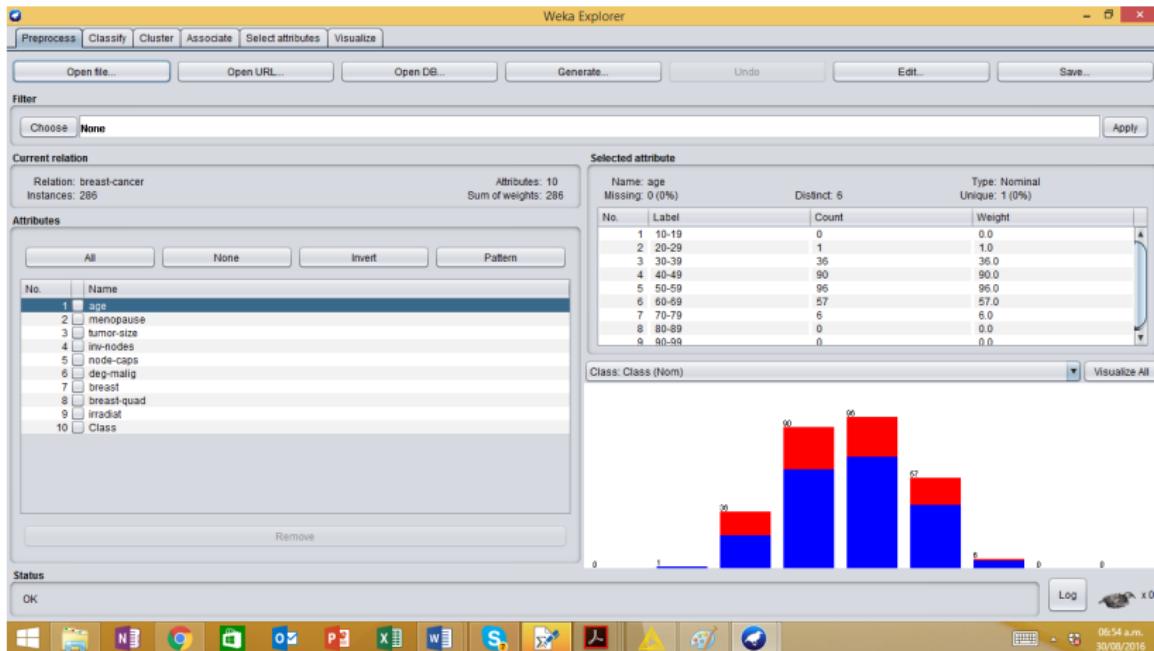
Herramientas

Algunas herramientas que soportan AA/MD

- R
- Weka
- RapidMiner
- KNIME
- TensorFlow
- SciPy y scikit-learn
- MALLET (orientado a NLP)
- MATLAB
- lush
- Julia

Herramientas

Weka



Herramientas

KNIME Analytics Platform

KNIME Analytics Platform

The screenshot shows the KNIME Analytics Platform interface with the following components:

- File Explorer:** Shows examples and local workspaces.
- Workflow Coach:** Displays recommended nodes and community activity.
- Node Repository:** Lists available nodes categorized by type (e.g., Read, Write, Other).
- Workflow Area:** Displays a workflow titled "Example Workflow" with the following steps:
 - File Reader (Read Ints.csv)
 - Color Manager
 - Partitioning
 - Statistics
 - Assign colors
 - Partitioning (Split data 60/40)
 - Decision Tree Learner
 - Train model
 - Decision Predict
 - Interactive Table
 - Explore test data
 A tooltip for the "Partitioning" node states: "Calculates statistic measures". A tooltip for the "Interactive Table" node states: "Explore test data".
- Node Description:** Provides details for the "File Reader" node, mentioning it can read ASCII files or URLs and supports various formats.
- Outline:** Shows the hierarchical structure of the workflow.
- Console:** Displays the KNIME Console output.
- Taskbar:** Shows the Windows taskbar with various application icons and the date/time (08:59 a.m., 30/08/2016).

Herramientas

TensorFlow



About TensorFlow

TensorFlow™ is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture allows you to deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile



Herramientas

scikit-learn

The screenshot shows the official scikit-learn website. At the top, there's a navigation bar with links for Home, Installation, Documentation, Examples, Google Custom Search, and a search icon. A GitHub fork button is also visible. The main content area features a large image of handwritten digits used for classification, followed by a brief introduction and a bulleted list of features. Below this, there are six categories: Classification, Regression, Clustering, Dimensionality reduction, Model selection, and Preprocessing, each with its own description, applications, algorithms, and examples links.

scikit-learn
Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ...

— Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ...

— Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ...

— Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization.

— Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics.

— Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.

— Examples

Material de Estudio

Varios de los **ejemplos en Python**, fueron tomados del libro:

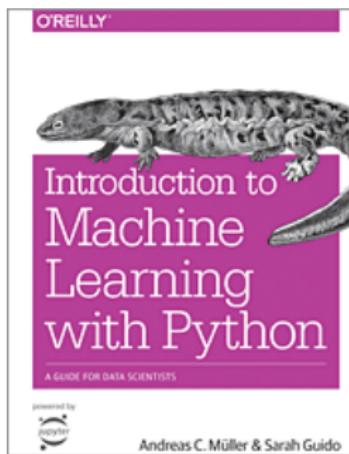
Python Data Science Handbook: Essential tools for working with data. Jake VanderPlas (1ra. Edición). 2017



Material de Estudio

Este libro se dedica específicamente al **aprendizaje automático en Python**:

Introduction to Machine Learning with Python. Andreas C. Müller and Sarah Guido. (1ra. Edición). 2016

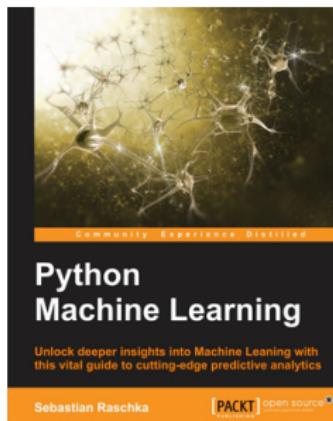


Material de Estudio

Material de Estudio

... este también, pero con bastante énfasis en redes neuronales:

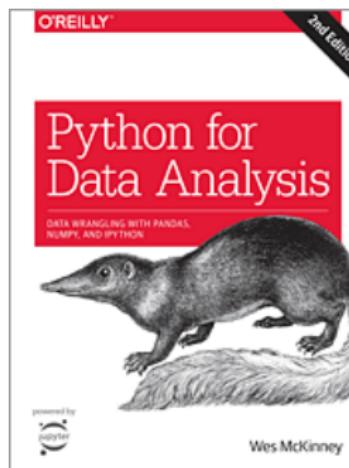
Python Machine Learning. Sebastian Raschka. 2015



Material de Estudio

Material complementario sobre **análisis de datos** (en general):

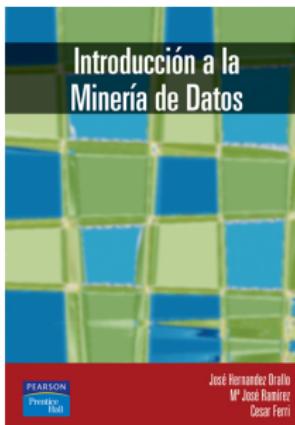
*Python for Data Analysis. Wes McKinney. (2da. Edición).
2016*



Material de Estudio

Varios ejemplos y figuras de la **parte introductoria**, están tomadas del libro:

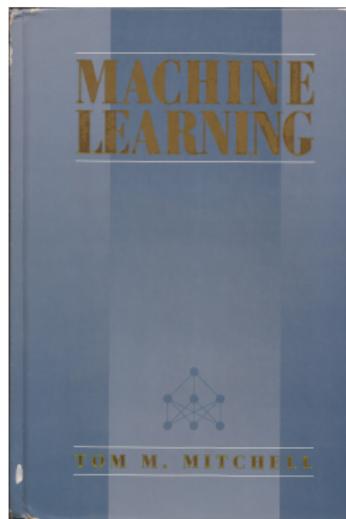
*Introducción a la Minería de Datos. J. Hernández Orallo,
M.J. Ramírez Quintana, C. Ferri Ramírez. 2004*



Material de Estudio

Un clásico del **aprendizaje automático**:

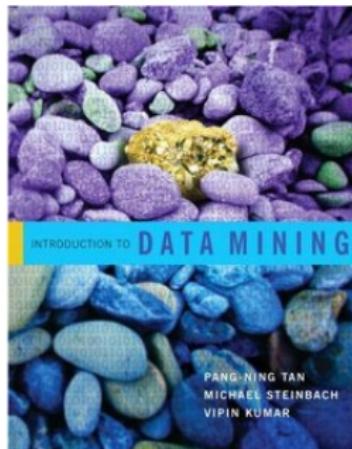
Machine Learning. Tom Mitchell. 1997



Material de Estudio

Libro interesante, buena combinación de aspectos teóricos y prácticos:

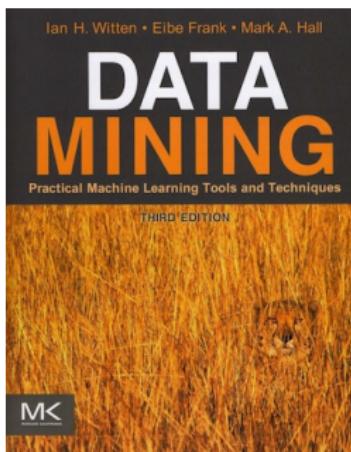
Introduction to Data Mining. P. Tan, M. Steinbach y V. Kumar. 2005



Material de Estudio

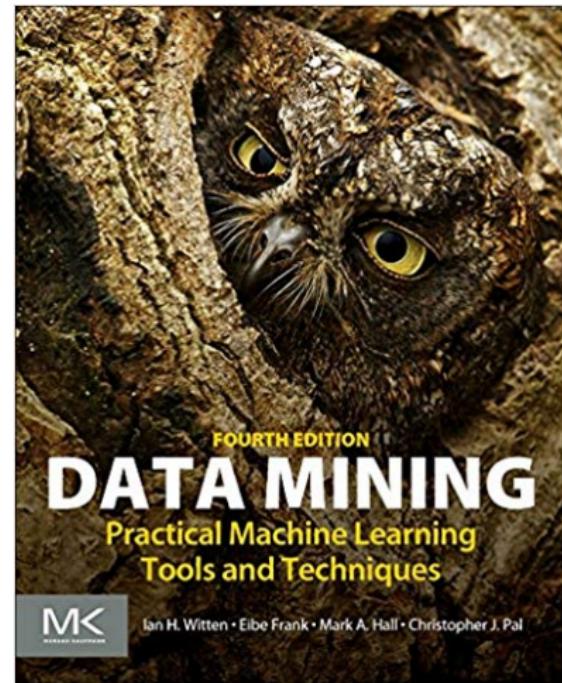
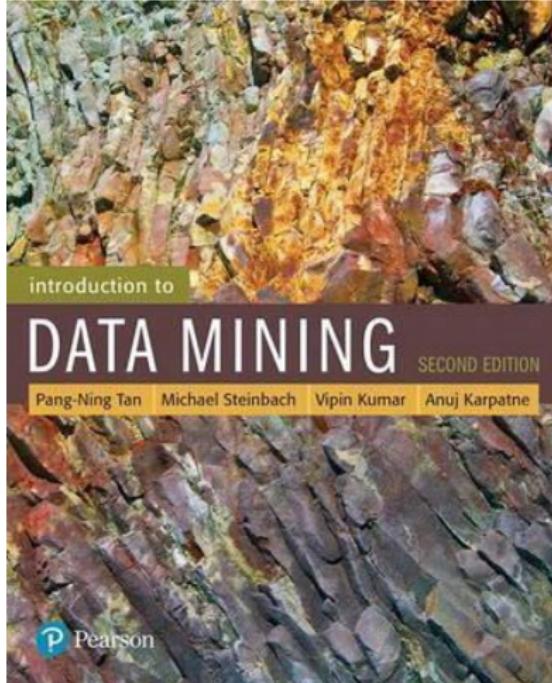
Una descripción sencilla de temas generales de minería de datos en conexión con **Weka** se puede encontrar en el libro:

Data Mining - Practical Machine Learning Tools and Techniques. I. Witten y E. Frank. 3rd. Ed. 2011



Material de Estudio

Para estos dos últimos libros existen nuevas ediciones:



Software

Nos centraremos principalmente en dos plataformas:

scikit-learn:

<http://scikit-learn.org/stable/>

Software

Nos centraremos principalmente en dos plataformas:

scikit-learn:

<http://scikit-learn.org/stable/>

Weka:

Machine Learning Group at University of Waikato

<http://www.cs.waikato.ac.nz/ml/weka/>