Machine Learning Workshop

Emanuela Boros
University of La Rochelle, France

4 February 2021

Organization

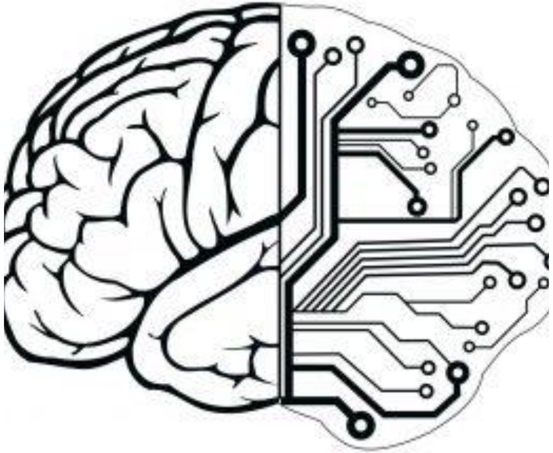
1. Text pre-processing & classification
2. Machine learning basics
3. Deep learning examples (extra)

- #ia (Discord)
- Python 3.7+
- Docker (or not, depends)
- Github (github classrooms, invitation link)
- Jupyter Notebook
- Computation libraries (numPy)
- Data libraries (pandas)
- NLP libraries (NLTK, spaCy)
- Ml libraries (scikit-learn, tensorflow, keras, etc)

Object Classification 01

- **Human brains** are wired to **recognize patterns** and classify objects for learning and making decisions
- .. they are not able to treat every object as **unique**
- .. we don't have a **lot of memory resources** to be able to process the world around us
- → our brains develop “concepts” or mental representations of “categories of objects”

- **Classification** is **fundamental** in language, prediction, inference, decision making and all kinds of environmental interactions
- **Language**: for example, how the meaning of words in a sentence can be contextualized by earlier words or concepts



QUICK COMMENT:

AI can be sexist and/or racist
Racist data? It’s the human bias that is Infecting the AI development

Object Classification
02

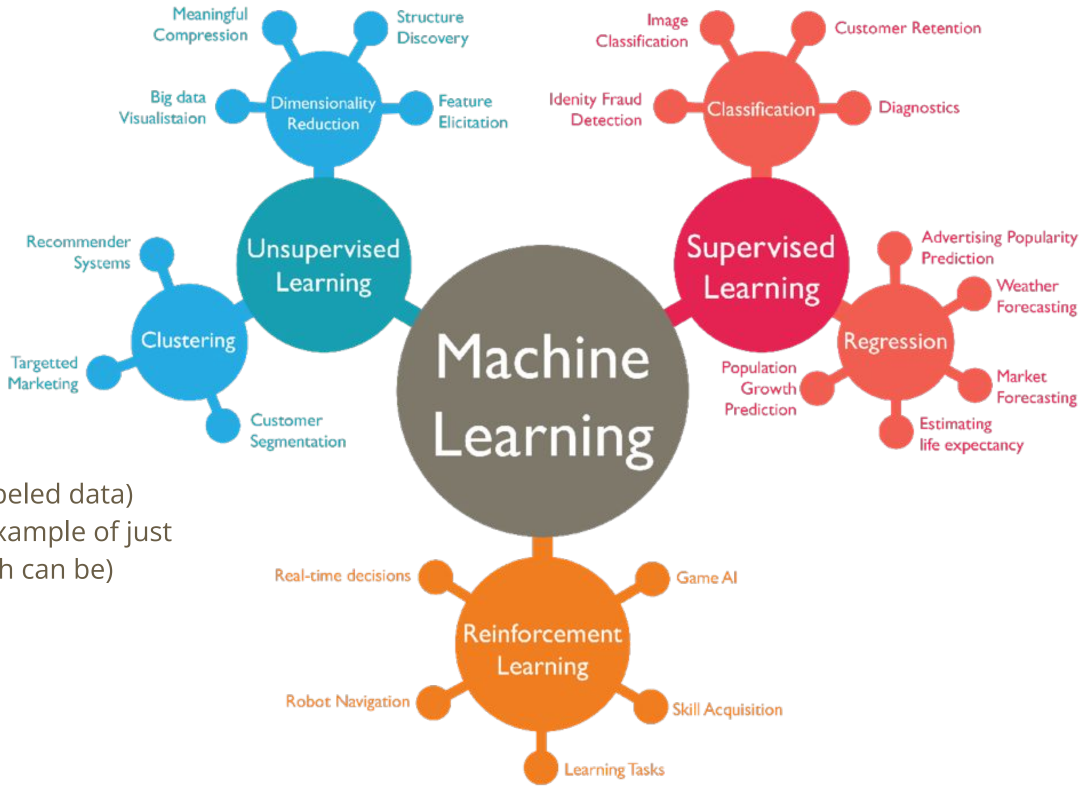
- The **classification of objects** consists of assigning a class to an object.
- These objects can be of the type:
  Text image audio video

- We do classification all the time:
  - We can **recognize the way back home** from university
  - We can r**ecognize a cat that is black** even if we have only seen white and orange cats before
  - We can even distinguish between a **Chihuahua** and a **muffin**

Concepts
01

- **Corpus/dataset** (corpora/datasets)
  - *Spam or not? Cat or dog?*
  - *Positive, negative or neutral?*
  - *House price prediction*
- **Machine/Deep learning types** of algorithms
  - **Supervised** learning (labeled data)
  - **Unsupervised** learning (unlabeled data)
  - **Semi-supervised** learning (labeled+unlabeled data)
  - **Reinforcement** learning (rules, a good example of just how broad the overall “learning” approach can be)
- **Machine/Deep learning evaluation**

  Train **development** **test**
  - Evaluation metrics



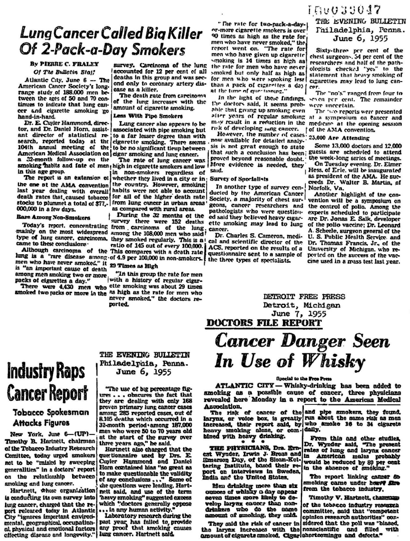
Dataset: Tobacco

The US government has sued five major US tobacco companies for raising large profits by lying about the dangers of smoking. The tobacco companies agreed in 1953 to "jointly carry out a vast public relations campaign to counter the increasingly obvious evidence of a link between tobacco consumption and serious illness".

In this trial 6,910,192 documents were collected and digitized. In order to facilitate the use of these documents by lawyers, you are in charge of setting up an automatic classification of the types of documents:

Advertisement **Email** **Form** **Letter** **Memo**

News **Note** **Report** **Resume** **Scientific**



TI0033047

THE EVENING BULLETIN
Philadelphia, Penna.
June 6, 1955

Lung Cancer Called Big Killer Of 2-Pack-a-Day Smokers

By PIERRE C. FRALEY
Of The Bulletin Staff

Atlantic City, June 6 — The American Cancer Society's long-range study of 188,000 men between the ages of 50 and 70 continues to indicate that lung cancer and cigarette smoking go hand-in-hand.

Dr. E. Cuyler Hammond, director, and Dr. Daniel Horn, assistant director of statistical research, reported today at the 104th annual meeting of the American Medical Association on a 32-month follow-up on the smoking habits and fate of men in this age group.

The report is an extension of the one at the AMA convention last year dealing with overall death rates that caused tobacco stocks to plummet a total of $77,000,000 in a few days.

Rare Among Non-Smokers

Today's report, concentrating mainly on the most widespread type of lung cancer, carcinoma, came to these conclusions:

Although carcinoma of the lung is a "rare disease among men who have never smoked," it is "an important cause of death among men smoking two or more packs of cigarettes a day."

There were 4,430 men who smoked two packs or more in the survey. Carcinoma of the lung accounted for 12 per cent of all deaths in this group and was second only to coronary artery disease as a killer.

The death rate from carcinoma of the lung increases with the amount of cigarette smoking.

Less With Pipe Smokers

Lung cancer also appears to be associated with pipe smoking but to a far lesser degree than with cigarette smoking. There seems to be no significant tieup between cigar smoking and lung cancer.

The rate of lung cancer was high in cigarette smokers and low in non-smokers regardless of whether they lived in a city or in the country. However, smoking habits were not able to account for all of the higher death rate from lung cancer in urban areas as compared with rural areas.

During the 32 months of the survey there were 152 deaths from carcinoma of the lung among the 108,000 men who said they smoked regularly. This is a ratio of 145 out of every 100,000. This compares with a death rate of 4.9 per 100,000 in non-smokers.

29 Times as High

"In this group the rate for men with a history of regular cigarette smoking was about 29 times as high as the rate for men who never smoked," the doctors reported.

"The rate for two-pack-a-day-or-more cigarette smokers is over 90 times as high as the rate for men who have never smoked," the report went on. "The rate for men who have given up cigarette smoking is 14 times as high as the rate for men who have never smoked but only half as high as for men who were smoking less than a pack of cigarettes a day at the time of questioning."

In the light of these findings, the doctors said, it seems probable that giving up smoking even after years of regular smoking may result in a reduction in the risk of developing lung cancer.

However, the number of cases now available for detailed analysis is not great enough to state that such a conclusion has been proved beyond reasonable doubt. More evidence is needed, they said.

Survey of Specialists

In another type of survey conducted by the American Cancer Society, a majority of chest surgeons, cancer researchers and pathologists who were questioned said they believed heavy cigarette smoking may lead to lung cancer.

Dr. Charles S. Cameron, medical and scientific director of the ACS, reported on the results of a questionnaire sent to a sample of the three types of specialists.

Sixty-three per cent of the chest surgeons, 54 per cent of the researchers and half of the pathologists checked "yes" to the statement that heavy smoking of cigarettes may lead to lung cancer.

The "no's" ranged from four to seven per cent. The remainder were uncertain.

The two reports were presented at a symposium on cancer and medicine at the opening session of the AMA convention.

25,000 Are Attending

Some 13,000 doctors and 12,000 guests are scheduled to attend the week-long series of meetings.

On Tuesday evening, Dr. Elmer Hess, of Erie, will be inaugurated as president of the AMA. He succeeds Dr. Walter B. Martin, of Norfolk, Va.

Another highlight of the convention will be a symposium on the control of polio. Among the experts scheduled to participate are Dr. Jonas E. Salk, developer of the polio vaccine; Dr. Leonard A. Scheele, surgeon general of the U. S. Public Health Service, and Dr. Thomas Francis, Jr., of the University of Michigan, who reported on the success of the vaccine used in a mass test last year.

THE EVENING BULLETIN
Philadelphia, Penna.
June 6, 1955

Industry Raps Cancer Report

Tobacco Spokesman Attacks Figures

New York, June 6—(UP)—Timothy B. Hartnett, chairman of the Tobacco Industry Research Comittee, today urged smokers not to be "misled by sweeping generalities" in a doctors' report on the relationship between smoking and lung cancer.

Hartnett, whose organization is conducting its own survey into lung cancer, charged that the report released today in Atlantic City "ignores important environmental, geographical, occupational, physical and emotional factors effecting disease and longevity."

"The use of big percentage figures . . . obscures the fact that they are dealing with only 168 proven primary lung cancer cases among 285 reported cases, out of 8,105 deaths which occurred in a 32-month period among 187,000 men who were 50 to 70 years old at the start of the survey over three years ago," he said.

Hartnett also charged that the questionnaire used by Drs. E. Cuyler Hammond and Daniel Horn contained bias "so great as to make questionable the validity of any conclusions . . ." Some of the questions were leading, Hartnett said, and use of the term "heavy smoking" suggested excess which "doctors generally oppose . . . in any human activity."

Laboratory research during the past year has failed to provide any proof that smoking causes lung cancer, Hartnett said.

DETROIT FREE PRESS
Detroit, Michigan
June 7, 1955

DOCTORS FILE REPORT

Cancer Danger Seen In Use of Whisky

Special to the Free Press

ATLANTIC CITY — Whisky-drinking has been added to smoking as a possible cause of cancer, three physicians revealed here Monday in a report to the American Medical Association.

The risk of cancer of the larynx, or voice box, is greatly increased, their report said, by heavy smoking alone, or combined with heavy drinking.

• • •

THE PHYSICIANS, Drs. Ernest Wynder, Irwin J. Bross and Emerson Day, of the Sloan-Kettering Institute, based their report on interviews in Sweden, India and the United States.

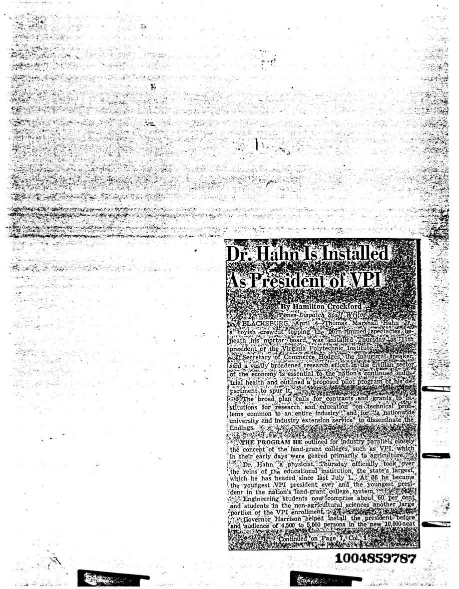
Men drinking more than six ounces of whisky a day appear seven times more likely to develop larynx cancer than nondrinkers who do the same amount of smoking, they said.

They said the risk of cancer in the larynx increases with the amount of cigarets smoked. Cigar and pipe smokers, they found, run about the same risk as men who smoke 16 to 34 cigarets daily.

From this and other studies, Dr. Wynder said, "The present rates of lung and larynx cancer in American males probably would be reduced by 80 per cent in the absence of smoking."

The report linking cancer to smoking came under heavy fire from the tobacco industry.

Timothy V. Hartnett, chairman of the tobacco industry research committee, said that "competent opinion research authorities" considered that the poll was "biased, nonscientific and filled with shortcomings and defects."



Dr. Hahn Is Installed As President of VPI

By Hamilton Crockford
Times-Dispatch Staff Writer

BLACKSBURG, April 4—Thomas Marshall Hahn Jr., a boyish crewcut topping the horn-rimmed spectacles beneath his mortar board, was installed Thursday as 11th president of the Virginia Polytechnic Institute.

Secretary of Commerce Hodges, the inaugural speaker, said a vastly broadened research effort in the civilian sector of the economy is essential to the nation's continued industrial health and outlined a proposed pilot program of his department to spur it.

The broad plan calls for contracts and grants to institutions for research and education "on technical problems common to an entire industry" and for "a nationwide university and industry extension service" to disseminate the findings.

THE PROGRAM HE outlined for industry parallels closely the concept of the land-grant colleges, such as VPI, which in their early days were geared primarily to agriculture.

Dr. Hahn, a physicist, Thursday officially took over the reins of the educational institution, the state's largest, which he has headed since last July 1. At 36 he became the youngest VPI president ever and the youngest president in the nation's land-grant college system.

Engineering students now comprise about 60 per cent and students in the non-agricultural sciences another large portion of the VPI enrollment.

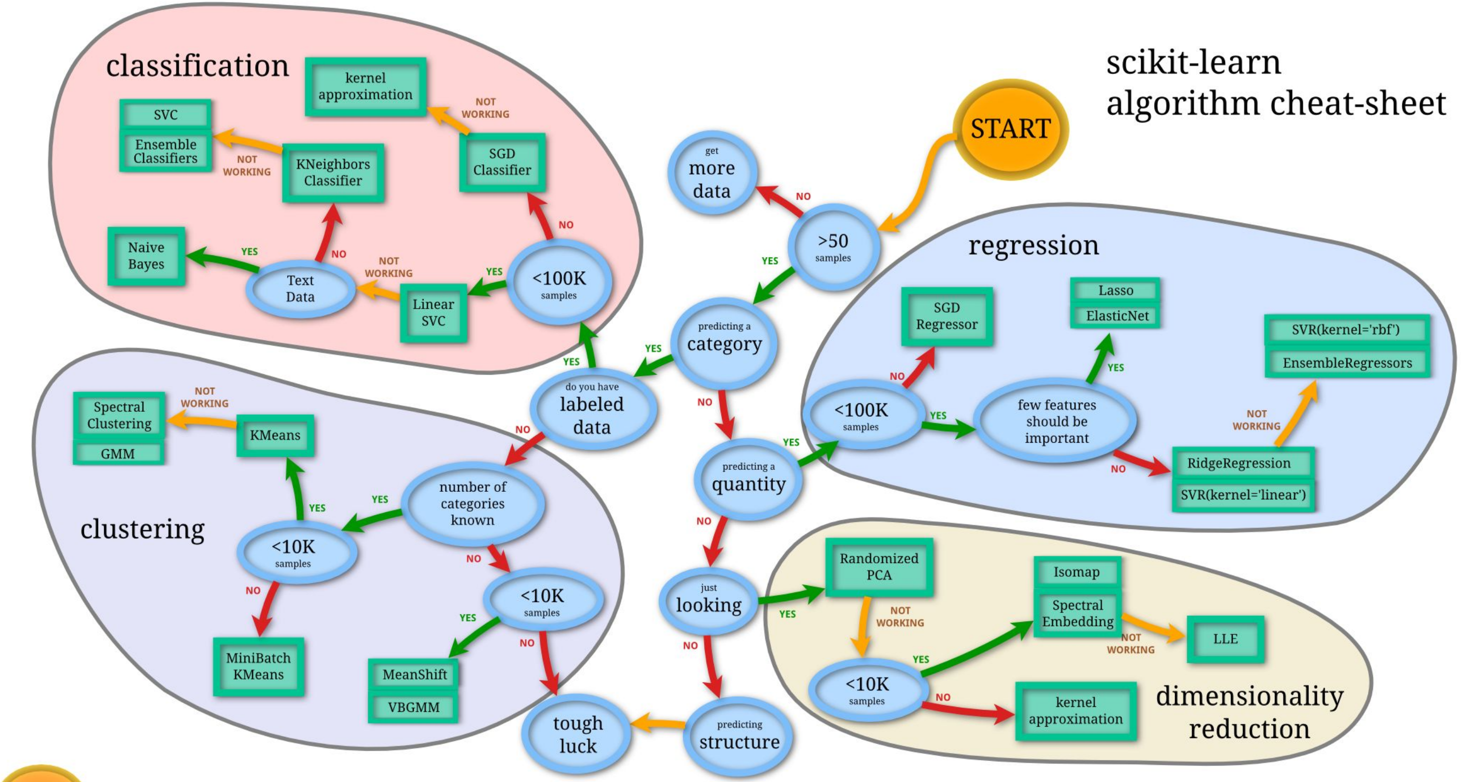
Governor Harrison helped install the president before and audience of 4,500 to 5,000 persons in the new 10,000-seat

Continued on Page 7, Col. 1

1004859787

http://tc11.cvc.uab.es/datasets/Tobacco800_1
David Doermann, Tobacco 800 Dataset (Tobacco800)

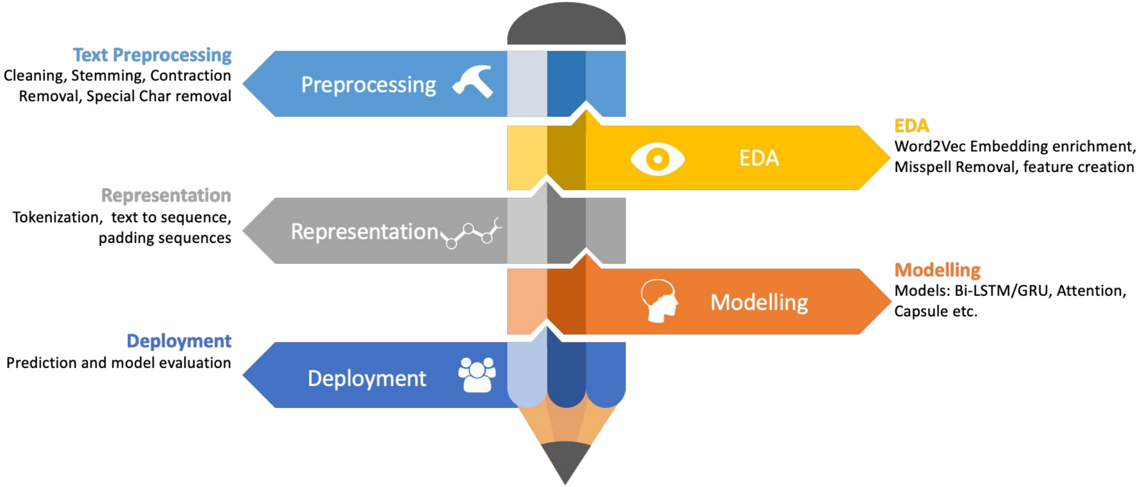
Scikit-learn



Back

Overview

1. Recovery of the text corpus (* .csv, * .txt, * .json, etc.)
2. Text pre-processing: tokenization etc
3. Exploration of the corpus (EDA, exploratory data analysis): frequency analysis
4. Word representations (bag of words, TF-IDF, word embeddings, word embeddings)
5. Machine learning and deep learning methods
6. Error analysis (evaluation, etc.)



Text Preprocessing
Cleaning, Stemming, Contraction Removal, Special Char removal

EDA
Word2Vec Embedding enrichment, Misspell Removal, feature creation

Representation
Tokenization, text to sequence, padding sequences

Modelling
Models: Bi-LSTM/GRU, Attention, Capsule etc.

Deployment
Prediction and model evaluation

Links

https://www.kaggle.com/competitions

Machine learning Coursera famous courses, Andrew Ng, https://www.coursera.org/learn/machine-learning

Machine learning Coursera (on youtube), Andrew Ng, https://www.youtube.com/watch?v=PPLop4L2eGk

The most famous book on deep learning: https://www.deeplearningbook.org/ (Ian Goodfellow, Yoshua Bengio and Aaron Courville)

ML and DL People

Andrew Ng, Founder and CEO of Landing AI, Founder of deeplearning.ai.
Fei-Fei Li, Professor of Computer Science at Stanford University.
Andrej Karpathy, Senior Director of Artificial Intelligence at Tesla.
Demis Hassabis, Founder and CEO of DeepMind.
Ian Goodfellow, Director of Machine Learning at Apple.
Yann LeCun, Vice President and Chief AI Scientist at Facebook.
Jeremy P. Howard, Founding Researcher at fast.ai, Distinguished Research Scientist at the University of San Francisco.
Ruslan Salakhutdinov, Associate Professor at Carnegie Mellon University, Director of AI Research at Apple.
Geoffrey Hinton, Professor of Computer Science at the University of Toronto, VP and Engineering Fellow at Google
Rana el Kaliouby, CEO and Co-Founder of Affectiva.
Daphne Koller, Founder and CEO of insitro, Co-Founder of Coursera, Adjunct Professor of Computer Science and Pathology at Stanford.
Alex Smola, Director, Amazon Web Services.