WM&R
INTRODUCTION TO THE FINAL TEST

R. Basili, D. Croce, G. Castellucci

a.a. 2018-19

Overview

- Program Overview
- Midterm and Final Exam
- 2° Course Section:
  - Closed Question examples
  - Open Questions: examples
- Final Projects

Course Structure

- Two major Sections
  - *Machine Learning*
  - *Advanced Web Information Retrieval applications*
- Different cross-relations between the two sections:
  - Examples:
    - *Supervised Learning* (es. NB) vs. Text Classification
    - *Semi-supervised learning* (e.g. EM) vs. *Probabilistic IR*
    - *Neural Neworks* vs. *Language Modeling*
    - *Matrix decomposition based IR Models* vs. *Lexical resources*
    - Eigenvector Analysis vs. *Link analysis*

The program synthesis

Machine Learning

- 0. Basic Notions of Geometry, Algebra and Probability
  - Elements and formalisms of probability theory
  - Elements of Information Theory
  - Vector spaces, inner product, Norms, Distances and similarity functions
  - Linear Maps, Matrices; Eigenvectors and eigenvalues

Machine Learning (2)

- 1.1. Unsupervised Learning
  - Introduction to automatic clustering.
    - Agglomerative and divisive algorithms.
  - Distance and Similarity Measures
- 1.2. Supervised Learning
  - Introduction to automatic classification.
  - Decision Tree Learning.
  - Probabilistic classification: Naive Bayes
  - Geometrical models of classification:
    - K-NN,
    - Profile-based classification: the Rocchio model.
  - On-Line Learning Algorithms

Machine Learning (3)

- 1.3. Performance Evaluation in ML
  - Gold standards and benchmarking
  - Splitting: Test vs. Training sets
  - Parameter settings: Development Sets
  - Evaluation Measures
- 1.4. Learning through Generative Models.
  - Introduction to Markov models: Sequence labeling tasks.
  - Language Models.
  - Hidden Markov Models.
  - Estimation methods for Generative Models.

Machine Learning (4)

- 1.5. Statistical Learning Theory
  - Introduction to PAC learning. The VC-dimension.
  - Support Vector Machines.
  - Kernel-based learning.
  - Complex kernels
    - Latent Semantic Kernels, Strings kernels, Tree kernels
- 1.6. Neural Networks
  - Deep Learning: training, architectures and tasks
  - Deep Learning and Language Modeling
  - Deep Learning and Natural Language Processing: Recurrent Neural Networks
  - Development of NN. Keras: MINST dataset, Named Entity Recognition

Machine Learning (5)

- 6. Semi-supervised Learning.
  - Ensemble Classifiers: bagging and boosting
  - Weakly-supervised Learning: LU learning
  - Co-training

- 1.7. Singular Value Decomposition & Latent Semantic Analysis

- 1.8. Machine Learning Tools and Applications.
  - Introduction to the WEKA machine learning platform.
  - Use of KeLP,
  - Keras
  - POS tagging as a sequence labeling task.

Information Retrieval

- 2.1 Introduction to Information Retrieval
- 2.2 Information Retrieval Models.
  - Boolean, probabilistic, algebraic
  - Systems for Information Retrieval: Lucene
- 2.3 Query processing approaches in IR
  - Query Expansion
    - Rocchio, Query Expansion and Reranking
  - Thesauri in IR
    - MeSH, Wordnet
  - Automatic Thesaurus Development
    - Automatic Global and Local Analysis
    - Wordspaces for Automatic Thesaurus acquisition
- 2.4 Evaluation of IR systems
  - Quantitative measures
    - Recall, Precision and F-measures, P@k
  - User-driven metrics

Information Retrieval (2)

- 2.5 Geometric methods for IR: Latent Semantic Indexing
- 2.6 Web Retrieval
  - Introduction to Web IR
- 2.7 Web Search: Ad search, duplicate elimination
- 2.8 Web links and Social Network Analysis.
- 2.9 Opinion Mining
- 2.10 Question Answering
  - Text-based QA
  - QA architectures for Factoid questions
  - NNs for QA

Statistical NLP & Web applications

- 3.1 Text Processing & NLP for IR
    - The standard NLP cascade
    - Morphology, Syntax & Semantic Information in Texts
- 3.2 NLP use cases:
    - POS tagging
    - Parse trees and ML kernels
    - Lexical Semantics
        - Word Sense Disambiguation & wordspaces
    - Predicate Semantics
        - Semantic Role Labeling
        - Framenet
- 3.3 Advanced Sequence Labeling for NLP
    - RNN for Named Entity Recognition

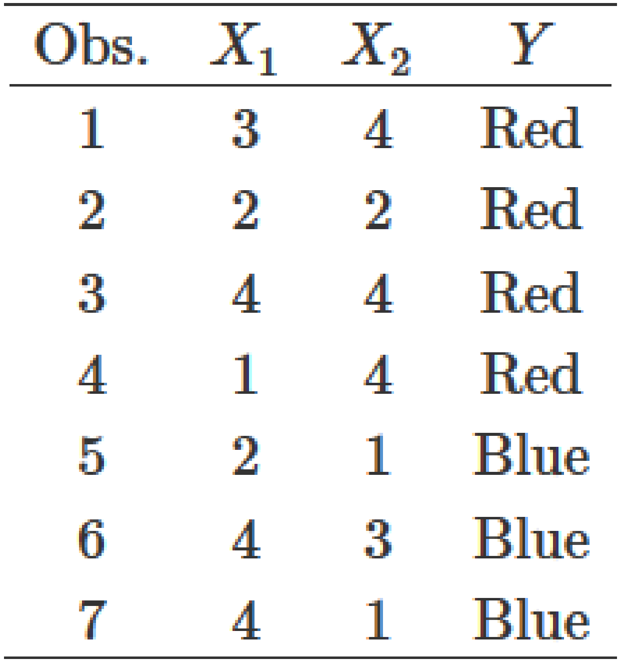
Final Test

- Second Mid Term (on the second half of the program)
  - Written test:
    - 10/12 closed questions
    - 1 open question
- First Final Written test (on the full program):
  - 15 closed questions
  - 1 open question
- Oral discussion (non mandatory for 6 CFU):
  - The final project (max 2/3 people)
  - Errors in the written tests

Topic of the second half

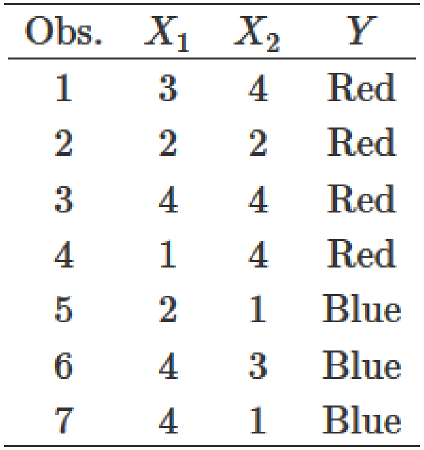
- 10/12 question in the TRM
- Targeted Topics:
  - SVM and kernels
  - Neural Networks
  - IR: models and architectures
  - IR: Query processing
  - *Distributional semantics and Matrix analysis*
  - *Distributional semantics and Neural LMs*
  - Link Analysis
  - Opinion Mining
  - Question Answering

Examples: Classification vs. Modeling



Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Questions



Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

- A. Define:
  - Equation of an hard margin linear SVM, that means estimating the *w* and *b* parameters.
  - Define the Blue vs. Red classification function
  - And compute the corresponding margin
- B. Introduce a further point P1 in the dataset that preserves the equation
- C. Introduce a point P2 for which the hard margin solution requires some changes and compute the value of the new margin.

Closed Answer Questions: examples

Questions about NNs

- What is the definition for a *convolutional neural network* and which are the main differences with *recurrent neural networks*
- What is a *non linear activation function*? What is its contribution to the training process?
- Which ones among the following techniques is specifically devoted to improve the model optimization stage of a NN (for example, by avoiding *overfitting*)?
  a) Dropout
  b) Early Stop
  c) Input normalization
  d) The Stochastic Gradient Descent
- Discuss the notion of *Loss Function* that characterizes the training of a neural network
- The backpropagation method allows to:
  a) Improve the efficiency of the learning process of a NN
  b) Maximize the Loss function over the validation data set
  c) Minimize the Loss function over the training data
  d) Maximize the Loss function over the training data

Questions (cont)

- What is a *Recurrent Neural Network*?
- How can we control the *number of epochs* required to train a NN?
- What is the advantage in adopting the *mini-batch* policy during the training stage with respect to batches based on one single instance?
- Formalize the *Stochastic Gradient Descent* algorithm and discuss its objectives.
- The *Back Propagation through Time* technique allows to maximize the Loss function in:
  a) Recurrent Neural Networks
  b) Convolutional Neural Networks
  c) Neural Networks made by a single perceptron
  d) None of the Other
- What is the *Vanishing Gradient*?

LSA (1)

- Let M= ((1 -1) (1 1) (-1 1)) be the initial co-occurrence matrix (Vocabulary V={t1,t2}). Determine the value σ_1 of the largest singular value
- R1: It is not possible: the problem is *under determined*
- R2. σ_1 =2
- R3. σ_1 =1
- R4. σ_1 =√2

Link Analysis (2)

- Let P=((0.1 0.9) (0.2 0.8)) be the matrix characterizing the graph among Web documents. Determine (with possible approximations) the vector $\underline{\pi}$ that represent the steady state of the random surfing process

- R1. The vector does not exist as the matrix does not represent an ergodic Markov process
- R2. $\underline{\pi}$ =(0.1 0.9)
- R3. $\underline{\pi}$ =(0.18, 0.82)
- R4. $\underline{\pi}$=(0.15,0.85)

Closed Questions (3)

- Determine the correct definition for the sentiment classification task amon the following:
  - (A) At document level this task corresponds to sentence classification into positive, neutral and negative polarity classes
  - (B) At sentence level the task consists into recognizing the features of objects to which the sentence sentiment makes reference
  - (C) At sentence level there are two tasks: (1) identification of subjective sentences in the input text; (2) polarity classification of individual sentences
  - (D) The task consists in the grouping of synomim expressions by which opinion holders may make reference to the object features
  - (E) None of the others corresponds to an acceptable definition

Closed Questions (4)

Signal the correct answer among the following ones:

a) Sentiment Analysis over Twitter is generally a simple task as the text corresponding to a *tweet* is limited in size.

b) User opinions in the social networks are not much interesting for the companies.

c) Sentiment Analysis is the computational study of opinions and sentiment expressed in texts.

d) Sentiment Analysis only rely on *machine learning* algorithms.

e) Sentiment Analysis is the computational study of the opinions and sentiment espressed by the topic of a text (e.g. an event or an entity)

Closed Questions (5)

Determine the correct definition for the sentiment classification task amon the following:

a) *Distributional* semantics methdos (e.g. LSA or *wordspaces*) cannot be adopted for the *relevance feedback* methods as they use vectors as representation models for terms.

b) *Distributional* semantics methdos cannot be adopted for the *relevance feedback* methods as they use lexical objects (i.e. symbols in the word dictionary) as representation models for terms and cannot be combined algebraically

c) None of the others

d) With *relevance feedback* we can impact only performances in terms of improvements in *precision*.

Closed answer Questions: solutions

Exam: Open Topics

- Part 1.
  - Generative Models
    - Modeling Sequence Labeling Tasks through generative models
    - Estimating probabilities for SLTs
  - Applications of Automatic Classification: a comparative discussion
  - Statistical Learning Theory
    - Support Vector Machines
    - Kernels
  - Latent Semantic Analysis

Exam: Open Topics (2)

- Program Second part
  - Statistical Learning Theory
    - Support Vector Machines
    - Kernels
  - Latent Semantic Analysis
  - Neural Network Learning
    - MLP, CNNs, RNNs,
    - Applications
- Parte 2.
  - IR models
    - Comparative discussion between vector space models and boolean models
    - Probabilistic reranking (EM)
  - Embedding in IR
    - LSA and its applications to IR and NLP.
    - Motivations and techniques for word embeddings
  - IR applications in the Web
    - Link analysis
    - Opinion Mining

Open Questions: examples

- Targets methods:
  1. Metodi di neural learning
  2. Automatic Global Analysis (for Query Expansion) o Word Embedding
  3. Ranking in Web search
  4. Sentiment Analysis sulle reviews

- Request:
  - Define basic methodological assumptions of the problem (model assumptions, type of obesrvations available, external resources)
  - Describe the pseudo-algorithm or the functional architecture adopted for solving the task
  - Discuss the possible evaluation metrics
  - Discuss the potential applications of the proposed solution

Open Questions

1. Discuss the main differences between Support Vector Machines and the approach of Neural Networks in supervised learning tasks
2. Discuss the main architectures for deep learning and their differences, describing also their main applications
3. Please define and discuss the notion of kernel function in the area of Statistical Learning Theory. Provide examples of their application in IR or NLP tasks
4. Discuss link analysis methods and their applications to Web search, by possibly comparing the different paradigms presented in the Course
5. Please determine the main computational challanges related to the task of Sentiment Analysis and discuss potential resources used to approach them.

Final Test: Open Question

Please discuss the application of markov modeling to *sequence labeling tasks*.
(Make use of an example through an applications, such as POS tagging of natural languae sentences)

- Request:
  - Define basic methodological assumptions of the problem
  - Define the basi cnotion of state, transition and emission
  - Define the general model equations
  - Describe the pseudo-algorithm adopted for solving the task
  - Discuss the possible evaluation metrics

Variants

- Apply the HMM modeling to the problem of URL recognition in free texts.
- Please make use of the state lables such as IOB for the start (B), inner (I) and outern (O) elements of a valid URL.
- Define the state vocabulary, the transition and emission matrices. Discuss possible parameter estimate techniques and their corresponding challenges.

WMR: Exam Dates

- Summer Session (2018-2019)

- Second Mid Term & First Final test:
  - **June 20 2019**, **Room C11**

- Second Final test:
  - **20 July 2019**, **Room C6**

- Project discussion (mandatory for 9 CFU):
  - After first written test
  - Before the end of july (approx. 28)
  - Afer the summer:
    - After the September exam dates