DEEP LEARNING: INTRODUCTION TO THE MIDTERM

A.A. 2023-24

Roberto Basili

Università di Roma, Tor Vergata

1° MIDTERM TOPICS

- **Introduction to ML**: supervised, unsupervised and reinforcement learning and applications.
  - Machine Learning: target problems and major paradigms.
  - Automatic Classification: classical ML approaches
    - K-NN
    - Decision Trees
    - The Rocchio model
- **Evaluation of Machine Learning algorithms**
- **Language Modeling – HMM for Sequence Labeling**
  - Visible and Hidden Markov processes
  - Fundamental Tasks for HMMs
  - Example: POS tagging
- **Statistical Learning Theory**:
  - PAC-learning, VC dimension
  - The Perceptron
- **Support Vector Machines and Kernels**
  - SVM: hard margin and soft margin SVM
  - Kernel Machines
- **Neural Networks**
  - Multilayer Perceptrons: architectures, training and application
  - Convolutional Neural Networks: architecture, training and image classification

- **Lesson 0**: Deep Learning - a.a. 2023-24: Introduction: Course Organization and Exam Modalities.
  Short history of Large Language Models: perspectives for business processes.
- **Lesson 1**: Introduction to Web Mining & Retrieval.
  - Some slides of Lesson 1 refer to the discussion of the link: **"A visual introduction to ML"**, slide 24).
  - **Lesson 1.1: Machine Learning: target problems and major paradigms**.
- **Lesson 2**: Machine Learning Metrics and Evaluation (part I: metrics for Text Classification).
- NEW! **Lesson 3: Language Modeling - an Introduction to Hidden Markov Models for Sequence Labeling**.

  Complementary Materials (Non mandatory):
  - **Lesson 3a**. Parameter Estimation for Language Modeling: the Baum-Welch algorithm.
  - **Lesson 3b**. Parameter Estimation and Rare Phenomena in Language Modeling.
- **Lesson 4**: (A gentle) Introduction to PAC learning and VC dimension.
  The slides used for the Course have been postedited from a kindly published version by Ethem Alpaydin, that you can find HERE.
  - **C. Burges's Tutorial on SVM and VC dimension.**
  - **D. Haussler discussion of PAC Learning**, 1999.
  - **Valiant L. G. *A Theory of the Learnable***, Communications of the ACM, Volume 27 Issue 11, Nov. 1984 Pages 1134-1142 .
- NEW! **Lesson 5 and 6**: Support Vector Machines and Kernels (Full package).
  - **An animated Perceptron**.
  - **Dan Klein's tutorial** on **Lagrange methods for the SVM optimization problem**.

- **Section II - Introduction to Neural Networks and Deep Learning Architectures**

NEW! **Lesson 7 An Introduction to Neural Learning.** The MultiLayer Perceptron: defining and training MLPs.
- **Lab 1 - Introduction to Keras:** the XOR example.
- **Lab 2a** - A Linear classifier and a MLP for image classification over the MNIST dataset in Keras.
- **Lab 2b** - A Linear classifier and a MLP for image classification over the MNIST dataset in Pytorch.


MidTerm open questions: some examples

TEMI D'ESAME: DOMANDA APERTA

Discutere la applicazione di una modellazione markoviana ai task di tipo *sequence labeling*.
(E' utile nella discussione presentare un esempio di applicazione, come ad esempio i processi di *Part-Of-Speech tagging* di frasi in linguaggio naturale)

- Definire le assunzioni di base,
- La nozione di stato, transizione ed emissione
- Le equazioni generali del modello
- I metodi di soluzione
- Possibili misure di valutazione

VARIANTE

- Utilizzare una tecnica di tipo HMM per il problema della *tokenizzazione* di un testo libero.
  - Si usino come etichette di stato le etichette IOB che stabiliscono l'inizio (B), l'interno (I) e la uscita (O) da un *token*.
  - Si definiscano **l'alfabeto degli stati** e quello delle **osservazioni**, le **matrici di transizione** e di **emissione**.
  - Si discuta infine la possibile tecnica di **stima dei parametri** applicabile al task, e gli eventuali problemi ad essa connessi.

MIDTERM TOPICS: OPEN QUESTION

Discuss the application of a Markov model to a *sequence labeling* task.
(Please use a concrete task, such as POS tagging, as an example)

Define and discuss:

- The basic assumptions of the method,
- The notion of state transition and emission
- The General Equations of the method
- Algorithmic Methods to automate the inference
- Available Evaluation Metrics

VARIANT

- Use an HMM for solving the problem of text *tokenization*.
- (Make use of the IOB state labels that determine the beginning (B), the inner (I) and the outing (O) of a *token*.
- Define the state and observation vocabularies as well as the transition and emission matrices. Finally, discuss the main challenges and solution methods of the parameter estimation problem.

TEMI D'ESAME: DOMANDA APERTA (3)

- Discutere la differenza tra un modello multivariato (binomiale) ed un modello multinomiale nei processi di classificazione bayesiana.
- (E' utile nella discussione presentare un esempio di applicazione, come ad esempio i processi di classificazione di documenti)
  - Definire le assunzioni di base,
  - La nozione di evento, spazio campione e caso possibile
  - Le equazioni generali del modello
  - I metodi di soluzione
  - Possibili misure di valutazione

MIDTERM TOPICS: OPEN QUESTION (3)

- Discuss the difference between the binomial multivariate model (Bernoulli) and a multinomial univariate model in Bayesian text classification tasks.

- Define and discuss
  - Basic Assumptions
  - The notion of stochastic e[illegible], mple space and possible case
  - General Equations of t[illegible]nethod
  - Algorithmic Solutions
  - Performance Metrics

TEMI D' ESAME: DOMANDE APERTE (4)

- Discutere il processo di analisi linguistica di un testo in processi di Information Extraction
- Definire in particolare le differenze tra le diverse informazioni estratte e la nozione di semantica lessicale di un testo libero.

- Discutere un approccio logico alla analisi semantica dei testi (testo, grammatica e forma logica). Esempio su una certa frase fornita in ingresso.

- Definire le diverse potenziali applicazioni della analisi linguistica quali:
  - Analisi semantica dei testi: ad es. Framenet labeling
  - Word Sense Disambiguation nei testi Web

MIDTERM TOPICS: OPEN QUESTION (4)

- Define a process of **Information Extraction** and discuss the **cascade of linguistic tasks** that are applied to source texts
  - Discuss in particular, the main differences between the semantic information extracted from a text and the notion of lexical semantics of a word.

- Discuss the notion of sem[illegible]ntics that characterizes a logical approach to the se[illegible]c analysis of texts (from texts to grammatical analysis to th[illegible]cal form). Provide an example on a given sentence use[illegible]s a[illegible]xample.

- Discuss different tasks and applications of linguistic analysis in current Web applications:
  - Semantic Role Labeling: such as Framenet labeling
  - Text classification or sentiment analysis

DOMANDA APERTA

- Definire un modello markoviano che esprima un modello probabilistico del linguaggio:

  $a^n b^m c^k$ con $n,m,k>0$

- che esprime stringhe del tipo

  abc, aaabcc, abbbbc, aabbccc,

- e non stringhe del tipo

  cbba, cbbc, aaacc, bba, ...

- Si definiscano i parametri del modello in modo tale che valga $p(abcc \mid \lambda) > 2p(abbc \mid \lambda)$

OPEN QUESTION

- Define a Markov model of the language:

  $a^n b^m c^k$ with $n,m,k>0$

- The language includes strings such as:

  abc, aaabcc, abbbbc, aabbccc,

- and excludes strings such as:

  cbba, cbbc, aaacc, bba, ...

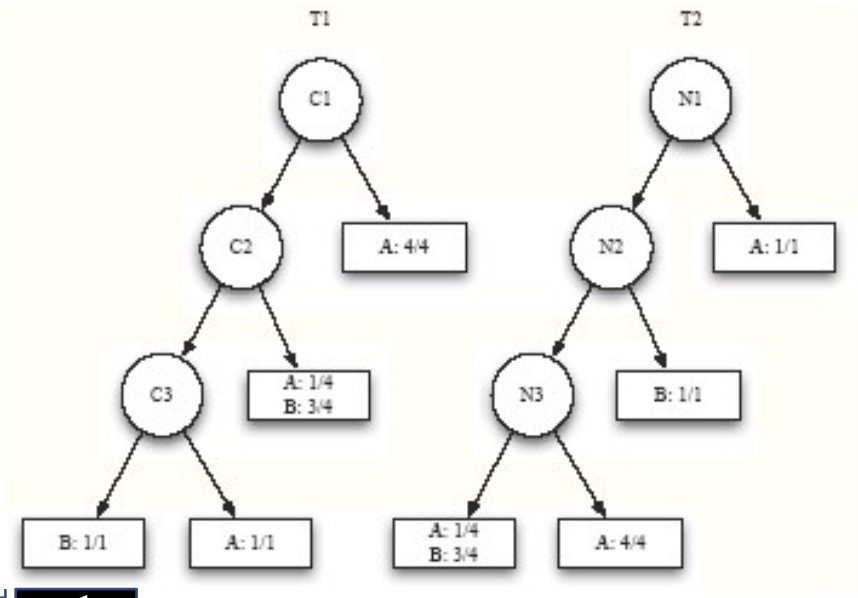
- Define also the model parameters λ such that:

  $p(abcc \mid \lambda) > 2p(abbc \mid \lambda)$

MULTIPLE ANSWER QUESTIONS: EXAMPLES AND SOLUTIONS

QUESTIONS WITH SOME COMPUTATION (1)

5. Dati gli alberi in figura scegliere **le affermazioni più corrette**



solo

solo

(A) La probabilità del nodo C1 di individuare la classe A correttamente è maggiore del nodo N1 [ ]

(B) La probabilià del nodo C1 di individuare la classe A correttamente è uguale del nodo N1 [ ]

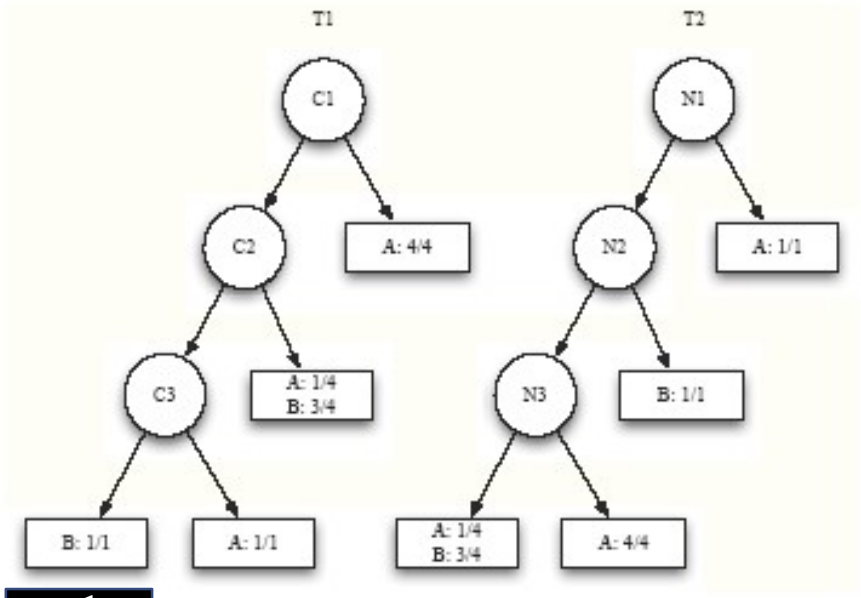
(C) La classe A viene sempre correttamente riconosciuta in entrambi gli alberi [ ]

(D) L'Information Gain del nodo C2 è maggiore del nodo N2 [ ]

Mulquestion: è possibile marcare più risposte

QUESTIONS WITH SOME COMPUTATION (1)

5. Dati gli alberi in figura scegliere **le affermazioni più corrette**



only

only

(A) La probabilità del nodo C1 di individuare la classe A correttamente è maggiore del nodo N1 [ ]

(B) La probabilià del nodo C1 di individuare la classe A correttamente è uguale del nodo N1 [ ]

(C) La classe A viene sempre correttamente riconosciuta in entrambi gli alberi [ ]

(D) L'Information Gain del nodo C2 è maggiore del nodo N2 [ ]

Mulquestion: è possibile marcare più risposte

MULTIPLE ANSWER QUESTIONS (2)

- What is the difference between a *clustering* algorithm and a classification one?
  - None: the K-mean algorithm applies to both tasks
  - Clustering is a supervised process/task while classification is not
  - *None of the other answers*
  - No difference, as both algorithms are based onto a numerical definition (metrics) of the similarity between istances
  - They are different as clustering generates one or more classes while classification generates example instances

MULTIPLE ANSWER QUESTIONS (3)

- Rocchio

26. Given a class C_i and the following (Rocchio) classifier,
$(\sum_{\vec{d} \in C_i} \frac{\beta}{|C_i|} \vec{d} - \sum_{\vec{d} \notin C_i} \frac{\gamma}{|\overline{C_i}|} \vec{d}) \cdot \vec{x} - \tau > 0$, with threshold $\tau > 0$
mark the correct statement among the following ones:
(A) The corresponding separating function is a polynomial of degree $n > 2$.
(B) The corresponding separating function is an hyperplane characterized by the maximal margin among all the separating hyperplanes.
(C) The corresponding separating function is an hyperplane whose gradient is the difference between the average of positive examples and the average of the negative examples.

(D) The corresponding separating function is the vector sum of all vectors representing positive documents.

MULTIPLE ANSWER QUESTIONS (4)

- Performance Evaluation

30. What does *n-fold cross validation* mean?
(A) Given the training and testing examples, models are acquired from training and testing, and then are tested on the test set.
(B) Given the training and testing examples, models are acquired from the testing data, and then are tested on the training set.
(C) The corpus is partitionaed in equal n parts; at each step, one partition is used a testing set and the remaining $n-1$ are used for the training.
(D) The training set is divided in n parts e the classifier is trained n times; at every time, the performance is measured over the test-set.

MULTIPLE ANSWER QUESTIONS (5)

- SVM

59. If $\vec{x}_i$ is a *support vector* obtained through the hard-margin SVM algorithm which among the following statement is $false$?

(A) $y_i(\vec{w} \cdot \vec{x}_i + b) - 1 = 0$.

(B) The associated Lagrange multiplier is null, i.e. $\alpha_i = 0$.

(C) If $\vec{x}_j$ is another different *support vector* with $y_j = -y_i$ then $b = -\frac{\vec{w}\cdot\vec{x}_i + \vec{w}\cdot\vec{x}_j}{2}$

(D) The geometric margin of the training set is $y_i(\vec{w} \cdot \vec{x}_i + b)$

ESEMPI

- SVM

51. Se $\vec{x}_i$ è un support vector ottenuto con l'algoritmo delle hard-margin SVMs quale affermazione risulta falsa?

(A) $y_i(\vec{w} \cdot \vec{x_i} + b) - 1 < 0$.

(B) Il moltiplicatore di Lagrange associato $\alpha_i \neq 0$.

(C) Se $\vec{x}_j$ è un'altro support vector con $y_j = -y_i$ allora $b = -\frac{\vec{w} \cdot \vec{x}_i + \vec{w} \cdot \vec{x}_j}{2}$

(D) Il margine geometrico del training set è $y_i(\vec{w} \cdot \vec{x_i} + b)$

MULTIPLE ANSWER QUESTIONS (6)

- Soft Margin SVM

61. Select the wrong statement with respect to the following system:

$$\begin{cases} min \quad ||\vec{w}|| + C\sum_{i=1}^{m}\xi_i^2 \\ y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1-\xi_i, \quad \forall i = 1,..,m \\ \xi_i \geq 0, \quad i = 1,..,m \end{cases}$$

(A) If parameter C goes to infinity the constraints $y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1-\xi_i$ tend to be equivalent to $y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1$
(B) $\sum_{i=1}^{m}\xi_i^2$ does not count exactly the number of errors of the hyperplene.
(C) if an $\xi_i > 0$ exists, the point $\vec{x_i}$ is wrongly classified, so the separating hyperplae does not exists.
(D) $\sum_{i=1}^{m}\xi_i$ is an alternative measure of the error.

ESEMPI

- Soft margin SVM

57. Individuare l'affermazione *errata* rispetto al seguente sistema:

$$\begin{cases} min \quad ||\vec{w}|| + C\sum_{i=1}^{m} \xi_i^2 \\ y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i, \quad \forall i = 1, .., m \\ \xi_i \geq 0, \quad i = 1, .., m \end{cases}$$

(A) Se il parametro C tende a 0 i vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i$ tendono ad essere equivalenti ai vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$

(B) $\sum_{i=1}^{m} \xi_i^2$ non conta esattamente il numero degli errori commessi dal iperpiano di separazione.

(C) Se esiste $\xi_i > 1$ il punto $\vec{x_i}$ non è classificato correttamente.

(D) $\sum_{i=1}^{m} \xi_i$ è una misura alternative dell'errore.

OTHER TOPICS

- Overfitting
  - Nature of the problem
  - Definition
  - Techniques used in NN learning for counteractions
    - Regularization (early stopping, Drop-out, Ridge Regularization, …)
- Text representation models
  - One-hot representations,
  - *bag-of-word* vectors
  - *(not explained)* document/sentence embeddings
  - Syntactic representations (bigrams, dependency graphs, trees)

TOPICS NOT COVERED IN THIS SLIDES

- Neural networks
  - Perceptrons, Multilayer Perceptrons
  - Terminology (layers, parameters, bias, loss, …)
- Training of NNs
  - Activation functions
  - SGD
  - Backpropagation algorithms
  - Problem settings: multiclass problems, multiple labeling problems
- Convolutional Neural Networks
  - Convolution filters
  - Training of CNNs
  - Applications to image processing