WM&R
PREPARAZIONE TEST FINALE

R. Basili, D. Croce, G.Castellucci

a.a. 2015-16

Overview

- Overview del programma
- Struttura dell'Esame Finale
- Struttura del secondo esonero
  - Esempi di domande Chiuse
  - Esempi di domande Aperte
- Proposte di Progetti e Relazioni Esame finale

Struttura del Corso

- 2 Sezioni Fondamentali del Programma (Section 1 e 3 da programma)
  - *Machine Learning*
  - *Information Retrieval*
- Sono numerose le correlazioni tra gli argomenti trattati nelle diverse sezioni
  - Esempi
    - Supervised Learning (es. NB) vs. Text Classification
    - Embedding models vs. Lexical vectors
    - Analisi agli autovalori vs. link analysis

Machine Learning

- 1. Nozioni Preliminari di Geometria, Algebra e Probabilità
  - Elementi e notazioni di teoria della probabilità
  - Elementi di teoria dell'Informazione
  - Spazi vettoriali, prodotto interno, Norme e funzioni di similarità
  - Trasformazioni Lineari, Matrici e Autovettori

Machine Learning (2)

- 2. Unsupervised Learning
  - Introduction to automatic clustering.
    - Agglomerative and divisive algorithms.
  - Distance and Similarity Measures
- 3. 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)

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

Machine Learning (4)

- 6. Statistical Learning Theory
  - Introduction to PAC learning.
  - Introduction to the VC-dimension.
  - Support Vector Machines.
  - Kernel-based learning.
  - Complex kernels
    - Latent Semantic Kernels
    - Strings kernels
    - Tree Kernels

Primo Esonero

- 7. Deep Learning via Neural Networks
  - from perceptron to multi-layered neural networks
  - how to optimize neural networks
    - Stochastic Gradient Descent

Secondo Esonero

Machine Learning (5)

- 8. Semi-supervised Learning.
  - Ensemble Classifiers: bagging and boosting
  - Weakly-supervised Learning: LU learning
  - Co-training
- 9. Singular Value Decomposition and Latent Semantic Analysis
  - see later in “Question Processing”
- 10. Machine Learning Tools and Applications.
  - Introduction to The WEKA machine learning platform.
  - Use of KeLP

Information Retrieval

- 2.1 Introduzione all'Information Retrieval
- 2.2 Modelli di Information Retrieval.
  - Boolean, probabilistic, algebraic
  - Sistemi di Information Retrieval: Lucene
- 2.3 Metodi di query processing per l'IR
  - Query Expansion
    - Rocchio, Expansion and Reranking
  - Thesauri IN IR
    - Wordnet
  - Automatic Thesaurus Development
    - Automatic Global and Local Analysis
  - Latent Semantic Analysis
    - Wordspaces/Word Embeddings for Automatic Thesaurus Population

Information Retrieval (2)

- 2.4 La valutazione dei sistemi di IR
  - Misure Oggettive
    - Recall, Precision and F-measures, MAP, NDCG
  - Misure basate sull'utente
- 2.5 Web Retrieval and Ranking
  - Introduzione all'IR nel Web
    - Ad-words
    - Duplicate Removal
  - Pagerank
  - HITS

Information Retrieval (3)

- 2.7 Learn to Rank
- 2.8 Opinion Mining
- 2.9 Web links and Social Network Analysis

Struttura dell'esame finale

- Secondo Esonero (Seconda parte del programma)
  - 8 domande chiuse
  - 1 domanda aperta
- Esame finale (per chi non ha superato il primo esonero)
  - 12 domande chiuse
  - 1 domanda aperta
  - Scope: intero programma (prima e seconda parte del programma)

- Esame sulla terza parte del programma (9 CFU)
  - Prova orale
  - Discussione a scelta su:
    - un progetto sperimentale (eseguito da 2/3 persone)
    - approfondimento teorico (1 persona) (vd bibliografia delle lezioni)

- Alla verbalizzazione verrà chiesto di visionare l'esito degli esercizi visti a lezione

ESEMPI DI DOMANDE

LSA (1)

- Sia M= ((1 -1) (1 1) (-1 1)) la matrice di co-occorrenza inziale (vocabolario V={t1,t2}). Determinare il valore σ_1 del piu' grande dei valori singolari
- R1: Non è possibile: il problema è sottodeterminato
- R2. σ_1 =2
- R3. σ_1 =1
- R4. σ_1 =√2

Link Analysis (2)

- Sia P=((0.1 0.9) (0.2 0.8)) la matrice che caratterizza il grafo tra documenti Web. Determinare (con eventuali approssimazioni) il vettore $\underline{\pi}$ che rappresenta lo stato stazionario del processo di navigazione casuale
- R1. Non esiste poiche' la matrice non rappresenta un processo ergodico
- R2. $\underline{\pi}$ =(0.1 0.9)
- R3. $\underline{\pi}$ =(0.18, 0.82)
- R4. $\underline{\pi}$=(0.15,0.85)

Dom Chiuse (3)

3. Determinare tra le seguenti la definizione corretta per il task di *sentiment classification*.
(**A**) A livello di documento questo task coincide con la classificazione delle singole frasi in positive, neutre o negative. [+0]
(**B**) A livello di frase il task consiste nel riconoscere le feature di un oggetti a cui la frase fa riferimento. [+0]
(**C**) A livello di frasi esistono due sottotask: (1) identificazione delle frasi soggettive di un testo e (2) classificazione delle frasi individuali. [+0]
(**D**) Il task consiste nel raggruppamento delle espressioni sinonime con cui l'opinion holder fa riferimento alle *features* del prodotto. [+0]
(**E**) Nessuna delle alternative costituisce una definizione accettabile. [+0]

Dom Chiuse (4)

Segnalare **la** risposta corretta tra le seguenti

a) La Sentiment Analysis su Twitter è generalmente un task semplice in quanto il testo di un tweet è limitato in lunghezza.
b) Le opinioni degli utenti in rete sono di scarso interesse per le aziende.
c) La Sentiment Analysis è lo studio computazionale delle opinioni e del sentimento espresso nei testi.
d) Nella Sentiment Analysis si fa uso esclusivamente di algoritmi di machine learning.
e) La Sentiment Analysis è lo studio computazione delle opinioni e del sentimento espresso nei testi, ma necessita il riconoscimento dei topic espressi nei testi

Dom Chiuse (5)

Segnalare **la** risposta corretta tra le seguenti.

a) I metodi di semantica distributional (ad es. LSA o wordspaces) non possono essere adottati per i metodi di relevance feedback perché non usano vettori come modelli di rappresentazione .

b) I metodi di distributional semantics non possono essere usati per task di relevance feedback perché usano oggetti lessicali (cioè simboli discreti del dizionario, parole) come modelli di rappresentazione non consentendone alcuna combinazione algebrica.

c) Nessuna delle altre

d) Con il relevance feedback si possono migliorare solo le prestazioni in termini di aumento della precision.

Dom Chiuse (6)

Riguardo al meccanismo delle *ad-words* .

a) Non può usare meccanismi di *machine learning* poiché si applica a simboli discreti del dizionario cioè parole individuali
b) Lo score di rilevanza con cui il termine *t* contribuisce al ranking di un advertiser *a* e' limitata superiormente dal *bid* di *a* su *t*.
c) Nessuna delle altre
d) Lo score di rilevanza con cui il termine *t* contribuisce al ranking di un advertiser *a* e' maggiore del *bid* di *a* su *t*.
e) Lo score di rilevanza con cui il termine *t* contribuisce al ranking di un advertiser *a* dipende unicamente dal *click-through* rate di *a* rispetto a *t*.

Dom Chiuse (7)

Quali valori può assumere la funzione di attivazione sigmoid g(z) = 1/(1+e^-z))?

a) (0,1)
b) [0,1]
c) [-1,1]
d) (-1,1)

Dom Chiuse (8)

Dati i punti $x=(x_1,x_2;y)$: $x_1=(2,2;\ 1)$, $x_2=(2,4;\ 1)$, $x_3=(6,3;\ 0)$, $x_4=(7,5;\ 0)$, calcolare i valori di θ e b della funzione $h(x)=g(\theta^T x+b)$ e g(z) = 1/(1+e^-z)) dopo aver osservato il punto x_1 con l'algoritmo di SGD con learning rate $\alpha=0.1$ e con valori iniziali $\theta=(0.4,\ 0)$, $b=1$

a) $\theta_1= 0.4074$; $\theta_2= -0.013$; $b=1.007$
b) $\theta_1= -1.2768$; $\theta_2= 0.87680$; $b=-1.3584$
c) $\theta_1= 0.4074$; $\theta_2= 0.013$; $b=-1.007$
d) $\theta_1= 0.2768$; $\theta_2= 1.87680$; $b=1(0,1)$

Esempio Domanda Aperta

- Il candidato descriva l'algoritmo di Page Rank per la stima di rilevanza della pagine web nella rete. Si discuta tale algoritmo, ponendo attenzione a elencare le assunzioni di base e si descriva la forma generale delle equazioni.
- Si descriva inoltre la possibile applicazione di tale algoritmo ad un problema diverso rispetto al problema di ordinamento delle pagine web.

Esempio Domanda Aperta

- Il candidato descriva un algoritmo per il *learning to rank* utilizzabile per migliorare la qualità di un sistema di ricerca documentale. Si discuta tale metodo descrivendo come esso possa essere applicato in maniera efficace all'ordinamento di documenti.
- Inoltre si descriva la possibile architettura di un sistema che utilizzi la metodologia adottata, descrivendo ponendo enfasi sulle componenti software necessarie per la costruzione del materiale di addestramento.
- Si espongano inoltre alcuni assunzioni di tali algoritmi che possono essere troppo semplificative in scenari di retrieval e fornire possibili soluzioni.

Date esami

- Sessione estiva (2015-2016)
  - Secondo Esonero : Prova Scritta. **6 Giugno 2016**, h. 14:00-16:00 nell'aula del corso
  - Discussione Orale: nei giorni immediatamente successivi alla (2°) prova scritta (a meno di richieste esplicite).

  Gli studenti che superano la prima prova scritta potranno chiedere di anticipare la discussione orale ad una data precedente.

Progetto Finale:
Sentiment Analysis in Twitter in italiano

Il candidato deve definire e sviluppare un sistema per il riconoscimento automatico dei sentimenti espressi in messaggi provenienti da Twitter in lingua italiana. E' quindi richiesto di riconoscere la polarità delle opinioni espresse nei messaggi.

Per ciascun messaggio, occorre associare una delle seguenti classi di polarità, tra ***positive***, ***negative***, ***neutral*** (non è espressa nessuna opinione).

Il candidato dovrà definire e applicare un metodo di classificazione per il riconoscimento della polarità basato su uno degli algoritmi visti a lezione. Al candidato verrà fornito un dataset annotato secondo lo schema sopra.

Opzionalmente il candidato potrà partecipare alla competizione SENTIPOLC 2016 confrontando il proprio sistema con sistemi provenienti da altre università italiane (e non solo): `http://www.di.unito.it/~tutreeb/sentipolc-evalita16/index.html`

Qualora vengano rispettate le date della competizione, sarà possibile pubblicare un articolo scientifico che descriva il sistema realizzato negli atti della conferenza Evalita 2016: `http://www.evalita.it/2016/`

Progetto Finale: *Human Robot Interaction*

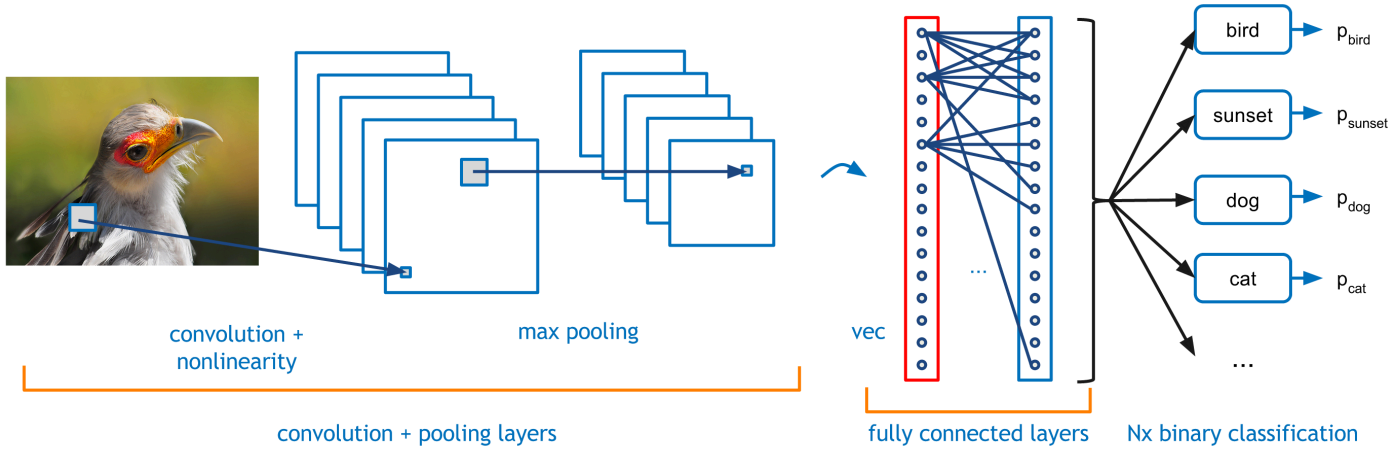
- Il candidato deve definire e sviluppare un sistema per l'interpretazione semantica di comandi per robot utilizzabile per realizzare interfacce robotiche basate sull'uso del linguaggio naturale.
- Ad esempio nella frase:

"*Prendi il libro sul tavolo*"

- Il sistema dovrà produrre una struttura dati simile alla seguente
  - $Prendi_{\text{TAKING}}$ $[il\ libro]_{\text{THEME}}[sul\ tavolo]_{\text{SOURCE}}$
- che rappresenta il comando (TAKING) l'oggetto da prendere (il THEME) e la sorgente da cui prendere l'oggetto (il SOURCE)
- **Al candidato verranno forniti: (1) la descrizione della teoria alla base della costruzione delle strutture dati che ospitano le interpretazioni del comando (2) un dataset annotato in italiano (da validare); (3) un sistema realizzato dal gruppo SAG per l'interpretazione di comandi in inglese basato su metodi di apprendimento automatico (da adattare alla nuova lingua)**.
- Una volta che il candidato avrà validato i testi dal dataset, esso verrà utilizzato per addestrare/validare il sistema.

Progetto Finale
Deep Learning for Computer Vision

- Al candidato è richiesta la progettazione e sviluppo di un sistema basato su metodi di *Deep Learning* per la annotazione semantica di immagini.



- Al candidato verrà fornita una selezione di articoli scientifici che descrivano il task e alcuni metodi stato dell'arte. Al candidato verrà fornito un dataset annotato.

Progetto Finale: Survey sulle tecniche di *Deep Learning*

Al candidato è richiesto una analisi della letteratura scientifica relativa al tema del *Deep Learning.* Al candidato verrà fornita una selezione di articoli scientifici e verrà richiesto di presentare al docente il tema affrontato. Non è preclusa la possibilità di approfondire ulteriormente il tema attraverso la lettura di altri articoli.

- Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing, Antoine Bordes, Xavier Glorot, Jason Weston and Yoshua Bengio (2012), in: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)
- Deep Learning for Efficient Discriminative Parsing. R. Collobert., In AISTATS, 2011.
- Parsing Natural Scenes and Natural Language with Recursive Neural Networks, Richard Socher, Cliff Lin, Andrew Y. Ng, and Christopher D. Manning. The 28th International Conference on Machine Learning (**ICML 2011**)
- Efficient Estimation of Word Representations in Vector Space. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. In Proceedings of Workshop at ICLR, 2013.