WM&R
PREPARAZIONE TEST FINALE

R. Basili, D. Croce, G.Castellucci

a.a. 2014-15

Overview

- Overview del programma
- Struttura dell'Esame Finale
- 2° Parte del Corso:
  - Esempi di domande Chiuse
  - Esempi di domande Aperte
- Proposte di Progetti e Relazioni Esame finale

Struttura del Corso

- 2 Sezioni Fondamentali del Programma
  - *Machine Learning*
  - *Information Retrieval*
- Sono numerose le correlazioni tra gli argomenti trattati nelle diverse sezioni
  - Esempi
    - Supervised Learning (es. NB) vs. Text Classification
    - Semi-supervised learning (e.g. EM) vs. Probabilistic IR
    - Matrix 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.
  - Estimation methods for Generative Models.

Machine Learning (4)

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

Machine Learning (5)

- 7. Semi-supervised Learning.
  - Ensemble Classifiers: bagging and boosting
  - Weakly-supervised Learning: LU learning
  - Co-training
- 8. Singular Value Decomposition and Latent Semantic Analysis
- 9. Machine Learning Tools and Applications.
  - Introduction to The WEKA machine learning platform.
  - Use of SVMlight and KeLP
  - POS tagging as a sequence labeling task.

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
    - Wordspaces for Automatic Thesaurus Population
- 2.4 La valutazione dei sistemi di IR
  - Misure Oggettive
    - Recall, Precision and F-measures
  - Misure basate sull’utente

Information Retrieval (2)

- 2.5 Tecniche geometriche per l'IR:
- 2.6 Web Retrieval
  - Introduzione all'IR nel Web
  - Web applications, Spidering and Search engines
- 2.7 Web Crawling and Clustering
  - Web crawling
- 2.8 Web links and Social Network Analysis.
- 2.9 Opinion Mining

Struttura dell'esame finale

- Prova scritta:
  - 15 domande chiuse
  - 1 domanda aperta
- Per coloro che hanno superato l'esonero
  - Prova scritta:
    - 10/12 domande
    - 1 domanda aperta sulla seconda parte del programma
- Prova Orale:
  - Una domanda sul programma
  - Discussione a scelta su:
    - un progetto (2/3 persone)
    - approfondimento teorico (1 persona) (vd bibliografia delle lezioni)

Domande d’Esame

- A Risposte Chiuse (2° parte)
- Temi
  - IR: modelli e architetture
  - IR: Query processing
  - Link Analysis
  - Opinion Mining
  - Modelli di *distributional semantics*

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. $\sigma_1 = 2$
- R3. $\sigma_1 = 1$
- R4. $\sigma_1 = \sqrt{2}$

LSA (1): soluzione

- I valori singolari sono le radici degli autovalori di M^TM
- Poiche'

  $M^T = ((1\ -1)\ (1\ 1)\ (-1\ 1))^T = ((1\ 1\ -1)\ (-1\ 1\ 1))$

  e $M^TM = ((3\ -1)\ (-1\ 3))$
- Gli autovalori λ soddisfano l'equazione:

$$\det(M^TM - \lambda I) = 0$$

cioe' $\lambda^2 - 6\lambda + 8 = 0$

- Essi sono: $\lambda_1 = 4$ e $\lambda = 2$
- Da cui: $\sigma_1 = 2$

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)

Soluzione (2)

- Per P=((0.1 0.9) (0.2 0.8))
- Sia $\underline{x}$=(0.5, 0.5) il vettore inziale
- Allora $\underline{x}P$=(0.15 0.85)
- e $\underline{x}P^2$=(0.19 0.82)
- e $\underline{x}P^3$=(0.18 0.82)
- ed infine $\underline{x}P^4$=(0.18 0.82) <== convergenza
- Da cui $\underline{\pi}$=(0.18, 0.82)

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 (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. [−1]
(**B**) A livello di frase il task consiste nel riconoscere le feature di un oggetti a cui la frase fa riferimento. [−1]
(**C**) A livello di frasi esistono due sottotask: (1) identificazione delle frasi soggettive di un testo e (2) classificazione delle frasi individuali. [+4]
(**D**) Il task consiste nel raggruppamento delle espressioni sinonime con cui l'opinion holder fa riferimento alle *features* del prodotto. [−2]
(**E**) Nessuna delle alternative costituisce una definizione accettabile. [−1]

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 (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.(-1)
b) Le opinioni degli utenti in rete sono di scarso interesse per le aziende.(-1)
c) La Sentiment Analysis è lo studio computazionale delle opinioni e del sentimento espresso nei testi.(+2)
d) Nella Sentiment Analysis si fa uso esclusivamente di algoritmi di machine learning.(-1)
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 (-1)

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 (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 (-1).

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 (-1).

c) Nessuna delle altre (3)

d) Con il relevance feedback si possono migliorare solo le prestazioni in termini di aumento della precision (-2).

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 (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 (-2).
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*. (2)
c) Nessuna delle altre (-1)
d) Lo score di rilevanza con cui il termine *t* contribuisce al ranking di un advertiser *a* e' minore del *bid* di *a* su *t*. (-1).
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*. (-1)

Domande d'Esame

- Domande aperte
- Parte 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

Domande d'Esame (2)

- Domande aperte
- Parte 2.
  - IR models
    - Comparative discussion between vector space models and probabilstic models
    - Probabilistic reranking (EM)
  - Embedding in IR
    - LSA e sue applicazioni.
    - Motivations and techniques for embeddings
  - IR applications in the Web
    - Link analysis
    - Opinion Mining

Esempio Domanda Aperta

Dato il task di *Word Sense Disambiguation* (*WSD*) (cioè la scelta del senso appropriato di una parola in una frase secondo un dizionario di sensi chiuso) discutere la sua mappatura in un task di classificazione:

In particolare è richiesto al candidato di:

- Definire le assunzioni di base del problema (forma del dizionario, osservazioni disponibili, risorse esterne)
- Sviluppare uno pseudo-algoritmo che descriva l'approccio proposto per il task
- Discutere possibili misure di valutazione
- Discutere possibili applicazioni che possono beneficiare dalla soluzione del task di WSD

Date esami

- Sessione estiva (2014-2015)
  - Primo Appello: Prova Scritta. **6 Luglio 2015**, h. 10:00-13:30 (aula da confermare)
  - Secondo Appello: Prova Scritta. **22 Luglio 2015**, h. 10:00-13:30 (aula da confermare)
  - 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:
Aspect Based Opinion Mining

Il candidato deve definire e sviluppare un sistema per il riconoscimento e la caratterizzazione delle opinioni nei testi secondo l'approccio **Aspect Based Opinion Mining**. Non è quindi richiesto di riconoscere solamente la polarità delle opinioni espresse nei testi, ma anche gli argomenti (*topic*) riguardo cui vengono espresse tali opinioni. Ad esempio nella frase:

"*The restaurant was too expensive*"

lo scrivente esprime una opinione negativa riguardo il ***prezzo*** del ristorante. Il dizionario dei topic è ristretto ai seguenti argomenti: `food`, `service`, `price`, `ambience`, `anecdotes/miscellaneous`.

Per ciascuna opinione, occorre associare una delle seguenti classi di polarità, tra ***positive***, ***negative***, ***neutral*** (non è espressa nessuna opinione) e ***conflict*** (nella frase è espressa sia una opinione positiva che negativa).

Il candidato dovrà definire e applicare un metodo di classificazione per il riconoscimento dei topic e delle classi di polarità basato su uno degli algoritmi visti a lezione.

Al candidato verrà fornito un dataset annotato secondo lo schema sopra indicato e un dataset di test privo delle annotazioni.

Una volta che il candidato avrà annotato i testi dal dataset di test, la prestazione del sistema verrà misurata dal docente.

Progetto Finale: *Human Robot Interaction*

- Il candidato deve definire e sviluppare un sistema di re-ranking applicato allo speech recognition usato in interfacce robotiche
- Ad esempio nella frase:

  "*Take the can in the trash bin*"

- Il candidato agirà su tutte le diverse trascrizioni di tale frase ottenuto da software di mercato (ad es. Google ASR APIs) e addestrerà il sistema affinché venga tra di esse rilevata la trascrizione corretta (o la più corretta): il meccanismo atteso è quindi quello di un *learning to rank*.

- Il candidato dovrà **definire e applicare un metodo di apprendimento automatico per il learning to rank tra le trascrizioni**
- **Al candidato verranno forniti: (1) un dataset annotato; (2) risorse di base (ad esempio un lessico ed una mappa semantica dell'ambiente dove il robot si muove) e (3) un dataset di test privo delle annotazioni**.
- Una volta che il candidato avrà annotato i testi dal dataset di test, la prestazione del sistema verrà misurata dal docente.
- La presentazione esaurisce chiude la sessione orale dell'esame

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.