

Obiettivo pratico del corso:

Costruire un sistema in grado di leggere e processare le risposte degli studenti a dei questionari che vengono erogati ad ogni lezione.

MODELLO INDUTTIVO := Utilizzo i dati per generare i modelli corretti.
(MACHINE LEARNING)

MODELLO DEDUTTIVO := Definisco un modello e lo applico ai dati per vedere se è corretto.

Un primo movimento che possiamo effettuare è assegnare alle varie parole una classe, a seconda di come la parola viene utilizzata nel contesto. Questo modo prende il nome di POS-tagging.

POS-TAGGING (LIVELLO I) ~ 49:00

Prima di iniziare notiamo il fatto fondamentale che il livello di analisi del POS-tagging non esiste nei dati ma viene IMPOSTO da chi sta analizzando la frase. L'analisi effettuata dipende quindi dalla conoscenza della grammatica di chi sta analizzando la frase.

In altre parole, la produzione di frasi è NATURALE, mentre il modello grammaticale è una IMPOSTAZIONE che viene portata nel modello di produzioni delle frasi. La grammatica viene DOPO il linguaggio, e viene utilizzata per dare una struttura od ems.

ESEMPIO:

FRASE	GLI	ALGORITHMI	SONO	PRESENTI	OVUNQUE
GRAMM. SANTORO	ART	NOUN	VERB	AUX	ADV
GRAMM. ZANZÙ	ART	NOUN	VERB	ADJ	ADV

Date una frase P dobbiamo quindi trovare una interpretazione $I(P)$ e chiederci: (1:03:00 min)

i) Come sono tutte le possibili interpretazioni di P ?

ii) Qual è l'interpretazione giusta?

In particolare dobbiamo capire l'interpretazione più "COMODA" e "UTILE" rispetto ai nostri scopi.

(!) Un problema fondamentale è quindi capire cosa vuol dire che una interpretazione è UTILE nel contesto del NLP.

ESEMPIO: In alcuni modelli non c'è differenza tra CONGIUNZIONI e PROPOSIZIONI SUBORDINATIVE.

In generale chi ha più SOLDI e POTERE punta il proprio standard. Ad esempio c'è UNIVERSAL DEPENDENCIES, finanziato da Google, prima c'era PENN-TREEBANK.

La scelta dei TAGS dipende quindi dalle risorse memorie per passare da P a $I(P)$. Quelle che non si è in grado di fare al livello corrente

viene DELEGATO al livello successivo.

La necessità di ricevere dei FINANZIAMENTI per sviluppare modelli di NLP è stata anche una conseguenza del fatto che ormai vengono utilizzate delle FRASI REALI, che quindi non costano da ottenere.

INTER ANNOTATOR AGREEMENT

Se inizialmente il modello veniva dato a dei LINGUISTI, che dovevano utilizzarlo per andare da P a I(P), col passare del tempo, anche al fine di ridurre i costi, questo compito è stato anche delegato a persone prese dalla strada.

Ma è quindi il problema di capire se il MODELLO, emesso tramite delle GUIDELINES, è stato compreso correttamente da tutti gli ANNOTATORI.

Questo problema equivale al calcolo dell'IAA, noto come INTER ANNOTATOR AGREEMENT.

Se il calcolo dell'IAA viene effettuato a mani per capire se:

- i) Le GUIDELINES sono scritte male
- ii) Il MODELLO non è soddisfacente

Se è IAA e abbastanza alto, allora possiamo procedere all'instatazione del DATASET.

Date una frase f e due annotatori, abbiamo la seguente situazione:

ANNOTATORE	I	t_1^1	t_2^1	...	t_m^1
ANNOTATORE	II	t_1^2	t_2^2	...	t_m^2

Vogliamo definire un COEFFICIENTE numerico che esprime l'ACCORDO tra i due annotatori.

Un primo approccio è quello di calcolare il # di volte in cui gli annotatori effettuano la stessa scelta:

$$A_1 := \sum_{i=1}^m \frac{\mathbb{1}(t_i^1 = t_i^2)}{m}$$

Questo coefficiente però è troppo BIASED rispetto nel caso in cui gli annotatori effettuano le scelte a CASO. Risulta quindi importante tenere in considerazione l'ACCORDO CASUALE per la metrica finale, specialmente se il # di dati da scegliere è limitato.

Sia A_0 la FREQUENZA DI ACCORDO CASUALE.
Una normale metrica per l'accordo finale
è detta KAPPA COEFFICIENT ed è calcolata
come segue

$$A := \frac{A_1 - A_0}{1 - A_0}$$

Il modo in cui A_0 viene calcolato dipende
dalle assunzioni che facciamo. In particolare:

- i) Prob. uniformi per le varie classi.
- ii) Prob. per le varie classi calcolate
in modo BAYESIANO.

È poi possibile estendere il KAPPA COEF. al caso
in cui si hanno più di due annotatori.

OSS: Amazon Turks non utilizza il calcolo
dell'IAA per capire se le annotazioni
iniettano il modello definito.

CONFRONTARE MODELLI ~ (1:56:00 min)

Nell'apprendimento automatico dobbiamo stare attenti all'ADEGUAMENTO al TEST SET che noi facciamo una volta che abbiamo visto i primi risultati negativi.

Quando confrontiamo dei modelli che apprendono da dei DATASETS dobbiamo stare attenti che gli eventuali miglioramenti PERCENTUALI non siano un frutto del CASO o peggio dello specifico DATASET. (2:10:00 min)

Un notevole approccio è quindi quello di mantenere una parte del DATASET BLIND durante la fase di addestramento. Rimane comunque il problema di ottenere un sampling rappresentativo, ovvero dei dati i.i.d.

Per capire se i risultati ottenuti non sono il frutto del CASO possiamo usare la teoria della PROBABILITA' e della STATISTICA.

La PROBABILITA' ci dice come effettuare i CALCOLI con le probabilità, mentre la STATISTICA ci dice come STIMARE le probabilità e definire degli INTERVALLI DI CONFIDENZA in cui si trovano i PARAMETRI della DISTRIBUZIONE che stiamo studiando.

Poniamo quindi vedere i risultati ottenuti dai vari sistemi come dati generati da due distribuzioni di probabilità. Tramite la TEORIA DELLE DECISIONI possiamo quindi definire

$H_0 :=$ Le distribuzioni dei due sistemi non UGUANI.

A questo punto possiamo cercare di verificare se i dati SUPPORTANO H_0 con una buona probabilità. Nel caso in cui questo non fosse vero con la probabilità voluta (tipicamente intorno al 95%) dobbiamo RIGETTARE H_0 .