Fondamenti concettuali del Machine Learning

Giorgio Gambosi

Università di Roma Tor Vergata & IASI - CNR

18 Maggio 2020

ML: cos'è?

- Progettazione di metodi algoritmici tali da fornire risultati sulla base dell'esame di dati o di esperienza passata
- Le azioni da intraprendere non sono codificate a priori, ma sono apprese: approccio induttivo
- Utile in tutti i casi in cui la descrizione formale di un metodo di risoluzione di problemi (eventualmente osservato in natura) risulta impossibile o proibitiva

ML è CS

Computer Science è ...

Risolvere problema mediante algoritmi, codificati come programmi

- Analisi del problema
- Modellazione formale
- Progetto di un algoritmo (operante sul modello formale)
- Implementazione in un linguaggio di programmazione
- Verifica di correttezza ed efficienza

Apprendimento

“Processo di acquisizione e di modificazione di capacità e abilità comportamentali degli organismi viventi animali e umani, nel corso delle esperienze nell’ambiente”
– *Enciclopedia Treccani*

“In psicologia cognitiva l’apprendimento consiste nell’acquisizione o nella modifica di conoscenze, comportamenti, abilità , valori o preferenze e può riguardare la sintesi di diversi tipi di informazione. Possiedono questa capacità gli esseri umani, gli animali, le piante e alcune macchine.”
– *Wikipedia*

Apprendimento

Processo attraverso il quale l'esperienza viene convertita in conoscenza

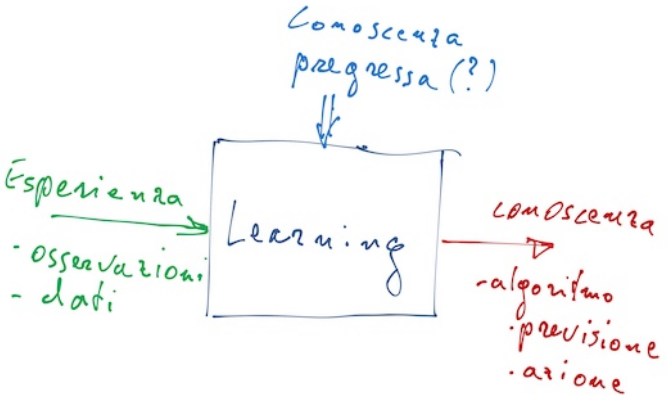
Ma che definizione di conoscenza adottiamo?

Definizione "operativa" (scientifica): capacità di

1. effettuare predizioni
2. agire (prevedere esiti)

in modo accurato/corretto

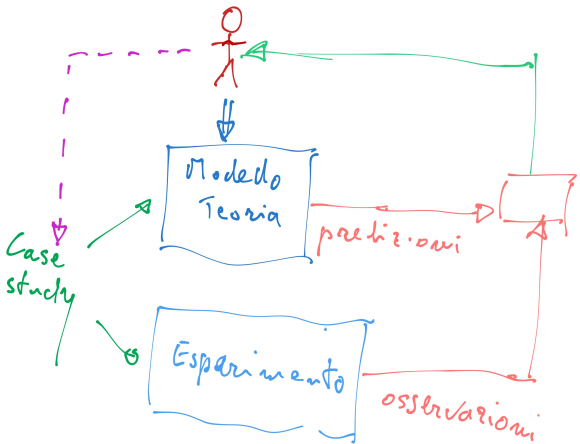
Apprendimento



Metodo scientifico

- Derivare conoscenza da esperienza è tipico della scienza e obiettivo dell'applicazione del metodo scientifico
- Il modello di riferimento è però molto diverso
- Diverso utilizzo dei dati

Metodo scientifico



Metodo scientifico vs ML

- Nel processo di ricerca scientifica, i dati vengono utilizzati prioritariamente per verificare l'affidabilità di un modello proposto dallo "scienziato" (ed eventualmente per suggerire, opportunamente analizzati, possibili modifiche o nuove soluzioni)
- In ambito ML, l'intero processo di selezione del "miglior" modello (all'interno di una classe predefinita) è guidato dai dati e si svolge in modo algoritmico, almeno entro certi limiti

Metodo scientifico

Questioni critiche:

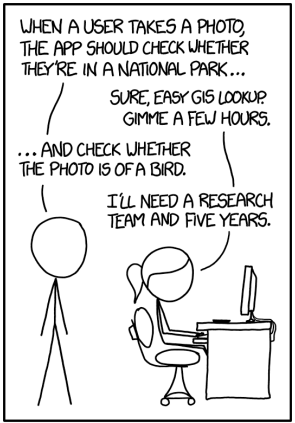
- come viene definito il modello?
- come viene modificato in presenza di incongruenze?

Necessità di intervento umano (knowledge, skill, ingenuity)

Il problema deve essere sufficientemente “semplice” da poter essere almeno affrontato e formalizzato da un soggetto umano

Apprendimento

- In molti casi, risulta difficile anche solo “pensare” un punto di partenza.
- o formalizzarlo
- no expertise disponibile
- expertise disponibile ma non traducibile in modello/algoritmo



Pensare un modello/algoritmo

Algoritmo: sequenza di passi elementari che risolvono un determinato problema in generale.

Quante “i” compaiono in questo testo?

When Mr. Bilbo Baggins of Bag End announced that he would shortly be celebrating his eleventy-first birthday with a party of special magnificence, there was much talk and excitement in Hobbiton. Bilbo was very rich and very peculiar, and had been the wonder of the Shire for sixty years, ever since his remarkable disappearance and unexpected return. The riches he had brought back from his travels had now become a local legend, and it was popularly believed, whatever the old folk might say, that the Hill at Bag End was full of tunnels stuffed with treasure. And if that was not enough for fame, there was also his prolonged vigour to marvel at. Time wore on, but it seemed to have little effect on Mr. Baggins. At ninety he was much the same as at fifty. At ninety-nine they began to call him *well-preserved*, but *unchanged* would have been nearer the mark. There were some that shook their heads and thought this was too much of a good thing; it seemed unfair that anyone should possess (apparently) perpetual youth as well as (reputedly) inexhaustible wealth. ’It will have to be paid for,’ they said. ’It isn’t natural, and trouble will come of it!’ But so far trouble had not come; and as Mr. Baggins was generous with his money, most people were willing to forgive him his oddities and his good fortune.

Un modello di “i”

- è necessario definire un modello di “i” che consenta di distinguerlo da altri caratteri
- dipende dal modello di calcolo assunto
  - un segno in un testo dattiloscritto i
  - una sequenza di bit in una codifica binaria (ad esempio 1101001 in ASCII)
- in entrambi i casi, la definizione è semplice, univoca e si assume verificabile da un agente di calcolo

Ma non è sempre così semplice

Quante volte compare Emma Stone, qui sotto?



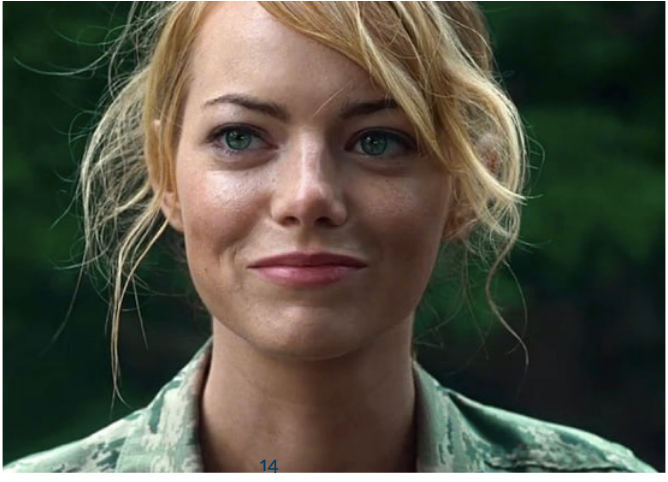







Per chi non la conoscesse

Emma Stone

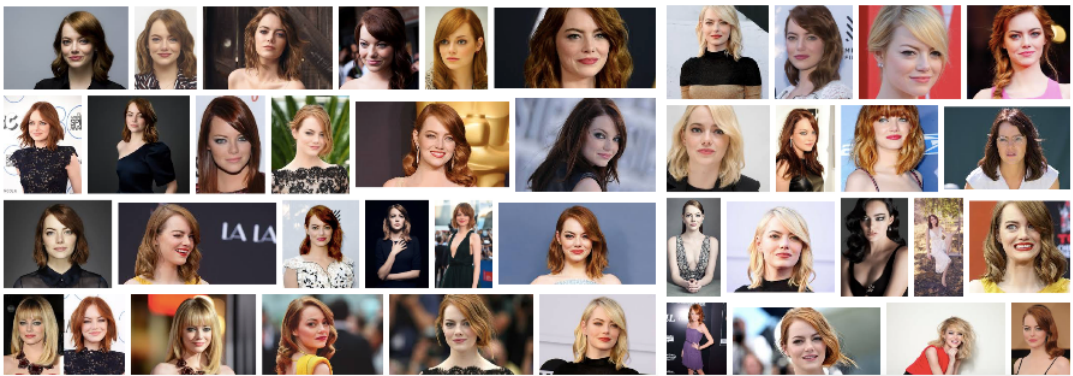


Un modello di Emma Stone

- difficile da definire come insieme di proprietà verificabili
- approccio deduttivo: da universale a particolare. Poco applicabile
- meglio seguire un approccio induttivo: da particolare a universale?

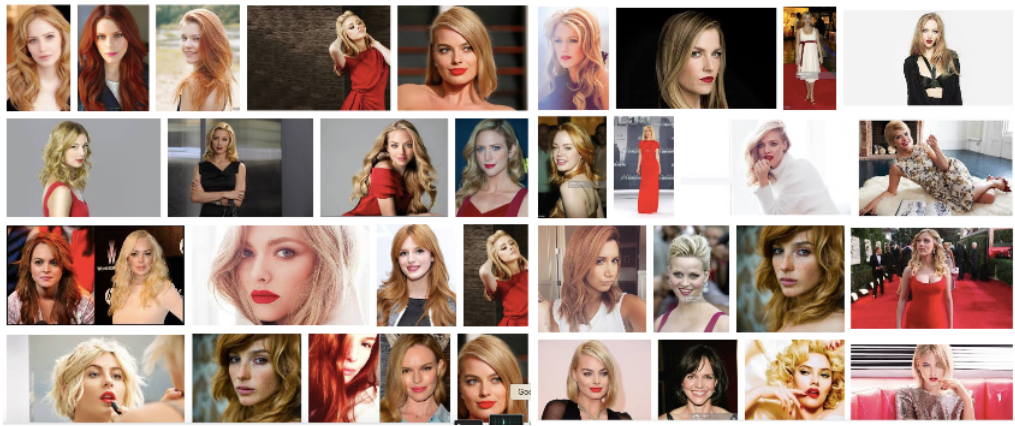
Una definizione “induttiva” di Emma Stone

Per esempi



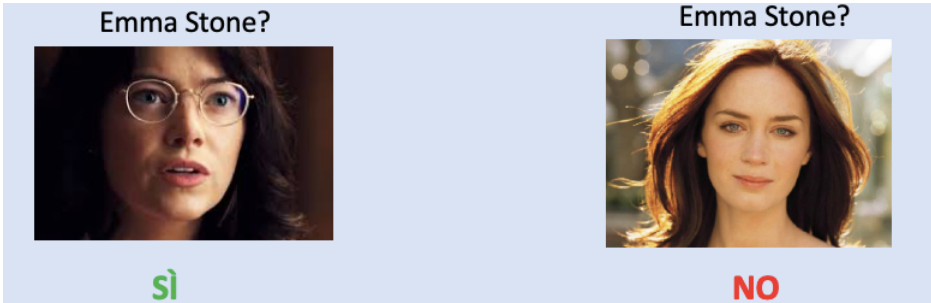
Una definizione “induttiva” di Emma Stone

Per esempi negativi



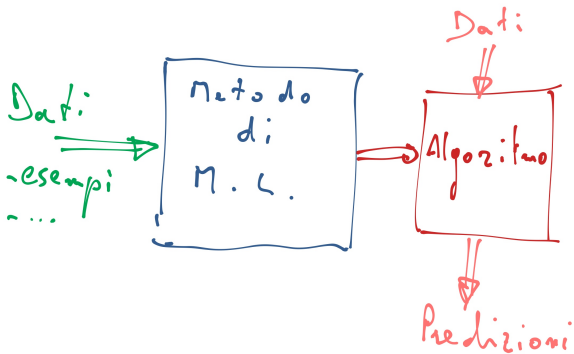
Apprendere a riconoscere Emma Stone

- Non posso spiegare come riconoscere Emma Stone e codificarlo in un algoritmo
- Posso solo mostrare tanti esempi di Emma Stone (e di non Emma Stone)
- Ho bisogno di un algoritmo che a partire dagli esempi “impari” a riconoscere Emma Stone, in modo che . . .

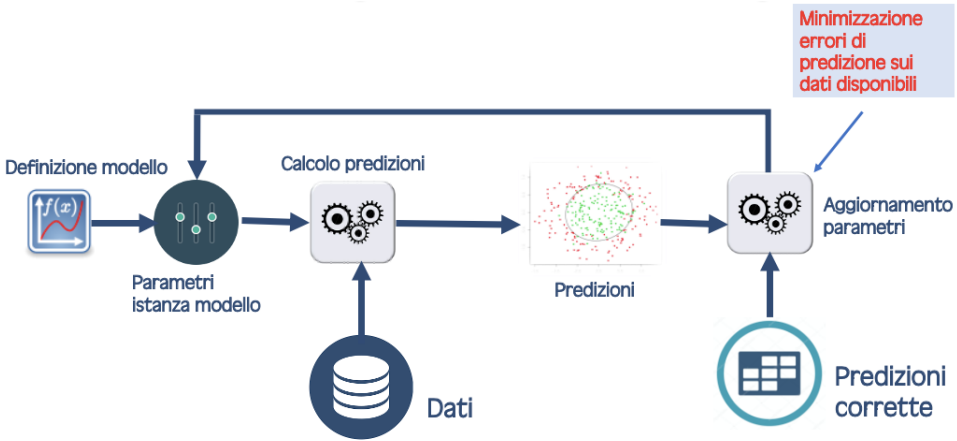


Cosa mi serve?

In realtà , ho bisogno di un algoritmo che, a partire da un insieme di esempi (ES sì, ES no), produca un ulteriore algoritmo, da utilizzare per effettuare predizioni.



Schema di apprendimento



Problemi “difficili”

Qualche esempio di problema difficile da risolvere “top-down”

- classificazione di testi (spam, sentiment analysis, hate speech)
- traduttori automatici, question answering
- assistenti virtuali, speech recognition
- classificazione di immagini (face recognition, immagini mediche)
- individuazione di correlazioni in dataset complessi
- sistemi di recommendation, profiling
- anomaly detection
- estrazione di informazione latente
- guida in processi di ricerca in spazi complessi: giochi, chimica computazionale, biologia computazionale, drug design
- robotica, self operating devices

Tipi diversi di apprendimento

Quel che possiamo apprendere dipende dai dati a disposizione

- Dati rappresentati da un insieme di coppie (x, t), esempi di applicazione di una funzione sconosciuta f ($t = f(x)$, o più in generale, $t = f(x) + \varepsilon$): apprendimento supervisionato
  - t valori reali: regressione
  - t valori binari: classificazione binaria
  - t valori finiti: classificazione multiclasse
  - t valori (reali, binari) su coppie: ranking

Tipi diversi di apprendimento

Quel che possiamo apprendere dipende dai dati a disposizione

- Dati rappresentati da un insieme di elementi x, rispetto ai quali vogliamo estrarre informazione sintetica mediante individuazione di pattern: apprendimento non supervisionato
  - partizioni in insiemi di elementi “simili”: clustering
  - individuazione variabili (o trasformazioni di variabili) più significative: dimensionality reduction
  - probability estimation and anomaly detection

Tipi diversi di apprendimento

Quel che possiamo apprendere dipende dai dati a disposizione

- Dati rappresentati da una sequenza di azioni eseguite dinamicamente da un agente x che transita da stato a stato. Per ogni azione, reward assegnato da funzione predefinita e sconosciuta, si vuole apprendere strategia per massimizzare il reward totale: apprendimento con rinforzo
  - strategie di esplorazione efficiente di spazi di grande dimensione (giochi, dispositivi "intelligenti", drug design, protein folding)

Framework generale di learning (supervisionato)

Input per l'apprendimento

Dominio degli elementi $\mathcal{X}$: insieme degli elementi rispetto ai quali si vuole effettuare la predizione

Dominio delle predizioni $\mathcal{Y}$: dominio dei valori (label) che possono essere predetti

Training set $\mathcal{S} = \{(x_i, t_i) : i = 1, \ldots, n, x_i \in \mathcal{X}, t_i \in \mathcal{Y}\}$: insieme di esempi di predizione. Si assume che i t_i derivino dall'osservazione "imprecisa" dei risultati dell'applicazione di una funzione sconosciuta (concetto) $f : \mathcal{X} \mapsto \mathcal{Y}$ che si vuole "apprendere": questa situazione è modellata attraverso una distribuzione congiunta $p(x, t)$

Framework generale di learning (supervisionato)

Modello probabilistico di riferimento

Natura Distribuzione sconosciuta $p(x, t)$ su $x \in \mathcal{X}, t \in \mathcal{Y}$ di cui assumiamo il training set sia un campione. Generazione:

1. sampling da $p(x)$ di un elemento
2. sampling da $p(t|x)$ della label associata.

Probabilità $p(x)p(t|x) = p(x, t)$

Active learner Il training set viene generato su richiesta dal learner: specifica x e riceve $t = f(x) + \varepsilon$

Framework generale di learning (supervisionato)

Output dall'apprendimento

Regola di predizione funzione $h : \mathcal{X} \mapsto \mathcal{Y}$ o, in alternativa, distribuzione predittiva $p(t|x)$
per ogni $x \in \mathcal{X}, t \in \mathcal{Y}$

Misura di qualità

Errore di un classificatore.

Rischio Frazione attesa del numero di elementi mal classificati da h, dove il valore è atteso rispetto alla distribuzione $p(x, t)$

$$\mathcal{R}(h) = \mathbb{E}_p[\mathbf{1}[h(x) \neq t]] = \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbf{1}[h(x) \neq t] p(x, t) dx dt$$

in modo del tutto equivalente, probabilità che un elemento estratto casualmente secondo la distribuzione $p(x, t)$ venga mal classificato da h

$$\mathcal{R}(h) = P_{(x,t)\sim p}[h(x) \neq t]$$

Rischio

Il rischio di dice quindi quanto ci aspettiamo che ci costi prevedere $h(x)$ invece del valore corretto t, assumendo che:

1. x sia estratto a caso dalla distribuzione marginale

$$p(x) = \int_{\mathcal{Y}} p(x, t)dt$$

2. il relativo valore corretto da predire sia estratto a caso dalla distribuzione condizionata

$$p(t \mid x) = \frac{p(x, t)}{p(x)}$$

3. il costo dell'errore sia rappresentato dalla funzione $[\mathbf{1}[h(x) \neq t]]$, sia cioè pari a 1 se la predizione effettuata da h è corretta e 0 altrimenti

Importante caso particolare

Assumiamo che la relazione funzionale sconosciuta $f : \mathcal{X} \mapsto \mathcal{Y}$ sia direttamente osservabile: allora, $p(x, f(x)) = p(x)$ e $p(x, y) = 0$ altrimenti. Ne deriva che, data la distribuzione $p'(x)$, il rischio sarà definito come

$$\mathcal{R}(h) = \mathbb{E}_{p'}[\mathbf{1}[h(x) \neq f(x)]] = \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbf{1}[h(x) \neq f(x)] p'(x) dx = P_{x \sim p'}[h(x) \neq f(x)]$$

Un approccio generale al learning

Un possibile approccio, molto diffuso, consisterebbe nell'individuare la funzione di predizione h^* che minimizza il rischio

$$h^* = \underset{h}{\operatorname{argmin}}\ \mathcal{R}(h)$$

Il rischio non è però valutabile, in quanto la distribuzione p è sconosciuta.

Rischio empirico

- Una approssimazione utile, in presenza di un data set $(X, t) = \{(x_i, t_i), i = 1, \ldots, n\}$ è rappresentata dal rischio empirico, definito come frazione di elementi mal classificati all'interno del training set

$$\overline{\mathcal{R}}(h; X, t) = \frac{1}{n} \sum_{i=1}^{n} [\mathbf{1}[h(x_i) \neq t_i]]$$

- Ciò evidentemente corrisponde a sostituire il valore atteso sull'intera popolazione di una variabile di Bernoulli con una media aritmetica su un campione, che ne rappresenta uno stimatore "unbiased", con varianza che tende a 0 al crescere di n

Rischio empirico

Un approccio praticabile per l'apprendimento viene ad essere quello di selezionare la funzione di predizione $\overline{h}^*$ che minimizza il rischio empirico

$$\overline{h}^* = \underset{h}{\operatorname{argmin}}\ \overline{\mathcal{R}}(h; X, t)$$

Rischio vs. rischio empirico

Una serie di osservazioni possono essere effettuate a partire dallo schema di apprendimento definito sopra: ne elenchiamo alcune.

- il training set è un campione casuale dell'universo dei possibili elementi: che garanzia abbiamo che minimizzare il rischio empirico su un training set casuale ci fornisca, almeno con alta probabilità, un classificatore efficace (ad esempio con probabilità di errore minore di ε su elementi estratti a caso da $p(x, t)$)?
- in un insieme infinito, per un qualunque training set esiste una funzione che fornisce errore 0 "specializzandosi" sul training set stesso: questo cosa comporta in termini di generalizzazione, e quindi di rischio?

Overfitting

- l'ottimizzazione dell'errore con riferimento al training set è finalizzata a minimizzare gli errori che un classificatore effettua sugli elementi disponibili: cosa possiamo dire riguardo all'efficacia nel classificare nuovi elementi (generalizzazione)? Il classificatore potrebbe comportarsi benissimo nel training set ma molto male, in generale, nelle predizioni relative a nuovi elementi (overfitting). Questo pone delle ulteriori questioni:
  - è possibile evitare (o meglio limitare) l'overfitting indipendentemente dall'insieme $\mathcal{H}$ delle ipotesi che si considerano?
  - in questo caso, esiste una indicazione di quanto deve essere numeroso un training set, a tal fine?

PAC learning

- PAC (probably approximately correct) learning: un insieme $\mathcal{H}$ di possibili ipotesi è (agnostic) PAC learnable se esiste un algoritmo di apprendimento $A_{\mathcal{H}}$ tale che, per ogni coppia ε, δ, in presenza di un training set X, Y composto un numero adeguato $m(\varepsilon, \delta)$ di elementi (campionati i.i.d. dalla distribuzione sconosciuta $p(x, t)$) fornisce con probabilità almeno $1 - \delta$ una ipotesi $\hat{h}$ con rischio al più ε maggiore rispetto al minimo possibile tra tutte le ipotesi in $\mathcal{H}$

$$\mathcal{R}(\hat{h}) \leq \min_{h \in \mathcal{H}} \mathcal{R}(h) + \varepsilon$$

Pura induzione

Considerare la minimizzazione in un insieme infinito di funzioni (ipotesi) presenta importanti problematiche per quanto riguarda la generalizzazione

- no free lunch: per ogni algoritmo di apprendimento $\mathcal{A}$ (che, per ogni training set, fornisce una funzione di predizione h), esistono un concetto f e una distribuzione $p(x, t)$ degli elementi tali che
  - f descrive perfettamente i dati osservabili, nel senso che il suo rischio è nullo

$$\mathcal{R}(f) = P_{(x,t)\sim p}[f(x) \neq t] = 0$$

  - per una frazione significativa dei possibili training set X, Y, la funzione di predizione h derivata da $\mathcal{A}$ a partire da X, Y, ha rischio che non può essere troppo piccolo
- una classe infinita di ipotesi non è, in generale, PAC learnable

Conoscenza pregressa (inductive bias)

La situazione cambia se assumiamo qualche limitazione nella scelta di h

- tutte le classi $\mathcal{H}$ finite sono PAC-learnable: è quindi possibile limitare (in probabilità) l'overfitting utilizzando training set sufficientemente grandi
- in realtà, anche classi $\mathcal{H}$ infinite sono PAC-learnable, se la corrispondente dimensione di Vapnik-Chervonenkis $VC(\mathcal{H})$ è finita: in questo caso, il rischio di h, derivata da un opportuno algoritmo A a partire da un training set X, Y, è $\leq \varepsilon$ con probabilità $1 - \delta$ se X, Y contiene almeno $\frac{VC(\mathcal{H})}{\varepsilon^2} \log \frac{1}{\delta}$ elementi

Quindi…

- l'overfitting è limitabile aumentando opportunamente la dimensione del training set e limitando la possibilità di scelta complessità del classificatore all'interno di un insieme opportuno: questo equivale ad assumere un qualche tipo di conoscenza pregressa (non tutte le ipotesi sono accettabili, ma solo alcune)
- in modo equivalente, si possono introdurre meccanismi di penalizzazione opportuni (regolarizzazione), che limitino in modo "soft" l'insieme delle possibili ipotesi
- una eccessiva limitazione può però determinare il fenomeno opposto, di underfitting: il classificatore derivato si comporta male sia sul training set che su altri elementi

Come risolvere il problema della conoscenza pregressa?

- conoscenza del dominio del problema
- model selection, cross validation
- spesso, trovare una “giusta” rappresentazione dei dati permette di adattare uno stesso metodo a contesti diversi: è il caso delle reti neurali (o, in misura minore, dei kernel)

Minimizzazione

- l'apprendimento è stato ridotto a un problema di ottimizzazione: come è possibile definire delle condizioni in cui tale operazione è effettuabile in modo efficace?
- l'ottimizzazione è stata definita in relazione alla funzione $\mathbf{1}[h(x_i) \neq t_i]$, che presenta una serie di inconvenienti
  1. non si applica su problemi diversi, come la regressione
  2. nel caso considerato, è un problema NP-hard
- la minimizzazione su un insieme generale di funzioni $\mathcal{H}$ richiederebbe l'uso di metodi di analisi funzionale

Minimizzazione

Per rendere il problema più semplice da affrontare tipicamente:

- $\mathcal{H}$ è definito come un insieme parametrico di funzioni: l'ottimizzazione è effettuata in uno spazio di coefficienti
- in questo modo, un problema di apprendimento è ridotto a un problema di ottimizzazione in uno spazio di possibile elevata dimensionalità
- difficile utilizzare metodi analitici: tipicamente necessità di metodi iterativi (del primo o del secondo ordine)

Minimizzazione

- si utilizzano funzioni di costo *loss* diverse e più adatte al contesto, che consentano l'uso efficace di metodi iterativi (smooth, convesse se possibile)
  - smooth: gradiente ovunque
  - convex: un solo minimo

Funzioni di costo

Mean Squared Error $(h(x) - t)^2$

Mean Absolute Error $| h(x) - t |$

Cross entropy $-t \log h(x) - (1 - t) \log(1 - h(x))$

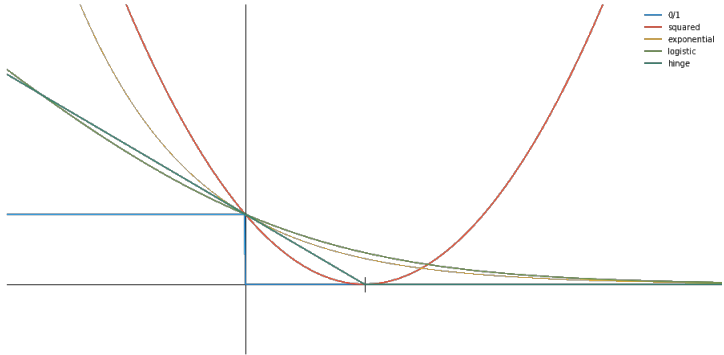
Softmax $\frac{e^{h_i(x)}}{\sum_k e^{h_k(x)}}$

Hinge $\max(0, 1 - h(x)t)$

Logistic $\log(1 + e^{-yt})$

Exponential e^{-yt}

Funzioni di costo



Funzioni di costo diverse vengono utilizzate in problemi di predizione diversi (cross entropy logistic regression, softmax softamx regression, hinge loss “traslata” perceptron, hinge loss con regolarizzazione SVM, mean squared regressione lineare, etc.)

Modellazione dei coefficienti: approccio deterministico

- i coefficienti caratterizzano una funzione loss predefinita che, dato un training set, associa un valore (il rischio empirico) ad ogni ipotesi
- si ricerca il valore dei coefficienti che minimizza la funzione

Modellazione dei coefficienti: approccio probabilistico

- i coefficienti caratterizzano modello probabilistico generativo dei dati osservati: nel caso supervisionato, il modello fornisce una distribuzione di probabilità dei valori target t_i dati i valori osservabili x_i e i coefficienti del modello
- l'apprendimento si riduce all'inferenza del miglior modello (nella classe considerata) che rappresenta i dati del training set

Modellazione dei coefficienti: approccio probabilistico

Approccio frequentista. I coefficienti sono parametri, il cui valore sconosciuto va stimato:

- utilizzo della verosimiglianza dei valori osservati, dati i coefficienti del modello

$$p(X, Y|w) = p(Y|X, w)p(X|w)$$

  dove per l'ipotesi i.i.d. e assumendo che il training set sia indipendente dal modello $p(X|w)$ è una costante, funzione soltanto della dimensione del training set
- verosimiglianza interpretabile come probabilità, per ogni valore di w, che i dati generati dal modello corrispondano a X, Y
- massimizzazione della verosimiglianza al variare di w

Modellazione dei coefficienti: approccio probabilistico

Approccio bayesiano. I coefficienti sono variabili casuali, caratterizzate da distribuzioni di probabilità:

- la regola di Bayes consente, a partire da distribuzioni di probabilità a priori dei parametri e delle osservazioni nel training set, di derivare distribuzioni di probabilità a posteriori dei parametri, e quindi dei modelli stessi
- utilizzo della distribuzione a posteriori delle probabilità dei coefficienti, a posteriori della conoscenza dei valori osservati, dati i coefficienti del modello

$$p(w|X, Y) \propto p(X, Y|w)p(w)$$

  dove il primo termine a destra dell'equazione è ancora la verosimiglianza
- massimizzazione della probabilità a posteriori al variare di w

Bayesian learning

La distribuzione a posteriori non viene semplicemente massimizzata, ma utilizzata per il calcolo di un valore atteso di predizione

- effettuazione di predizioni non semplicemente utilizzando un particolare modello h precedentemente selezionato per ottimizzazione, ma componendo le predizioni di tutti i modelli in $\mathcal{H}$, opportunamente pesate sulla base della relativa probabilità a posteriori dei modelli stessi

Bayesian learning

$$p(t|x) = \int_w p(t|x, w)p(w|X, Y)dw$$

Valore atteso

- delle predizioni effettuate dai singoli modelli $pt|x, w$, dato l'elemento x rispetto al quale effettuare la predizione
- pesati dalle probabilità a posteriori dei modelli, dato il training set $p(w|X, Y)$

Il risultato è una distribuzione di probabilità del valore predetto: informazione più ricca rispetto ai casi precedenti