IIA ML Esercitazioni – parte 1 (appunti in sintesi di esercizi svolti in aula)

***WARNING*: E' estremamente utile per voi stessi provare autonomamente a risolvere gli esercizi (nei file testi-domande) PRIMA di vedere queste soluzioni**

Suggerimenti generali:

- Seguire equazioni e nozioni, specialmente quelle in accordo all'etichetta "def", ma ovviamente non solo quelle!, distribuite nelle note del corso (specie per le domande).
- Seguire tutti gli "esercizi" distribuiti nelle note del corso.
- Non ci si basa su esercizi con parti "meccaniche" in cui esercitarsi ma spesso si richiede di collegare dei concetti per esprimere una risposta ragionata a situazioni proposte d'uso del ML, che esercitano la comprensione degli argomenti di lezione.
- Si preferiscono risposte espresse in modo matematico (là dove è sensato, per chiarezza e sintesi), poi commentate a parole se occorre. *Nota: per i quiz su elearning verranno considerate forme di soluzioni adeguate ai mezzi, cioè non toglie al valore di questi esercizi preparatori qui presentati e poi discussi.*
- Non è ammessa la consultazione di materiale.
- Fare attenzione a indicare i vettori nel caso di scrittura a mano (**x** grassetto o x segnato da freccia etc. purché sia indicato chiaramente).

Di seguito alcune note (testo degli esercizi e cenni in sintesi delle soluzioni) di alcuni esercizi svolti nelle lezioni di esercitazione:

1 – Piani separatori nel piano 2D

a) Disegnare il piano separatore nello spazio delle istanze (2D) per un classificatore lineare

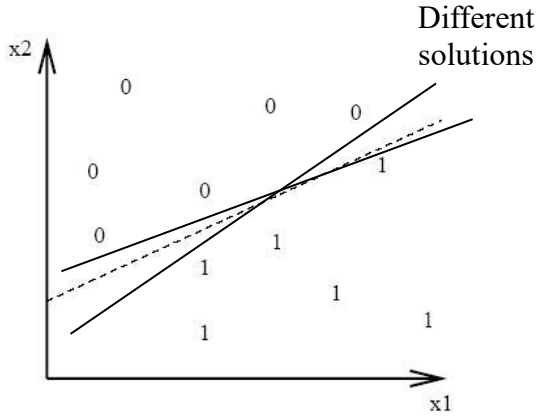
- Esercizio: Disegnare diversi *decision boundary* al variare del valore dei *w*

$$\mathbf{w}^T\mathbf{x} = w_1x_1 + w_2x_2 + w_0 = 0$$

Si risolve impostando: $x_2 = - x_1 w_1/w_2 - w_0/w_2$

Provare in particolare a disegnare (soluzione a esercitazione ed esempi analoghi o uguali sotto)

1. $x_1 + x_2 - 1 = 0$ (ossia $w_1=1$, $w_2=1$, $w_0=-1$)
2. $x_1 + x_2 - 0.5 = 0$ (ossia $w_1=1$, $w_2=1$, $w_0=-0.5$)



- Che succede al *decision boundary* moltiplicando i *w* per una costante ?
  Notare la proprietà di *Free scaling* : moltiplicando i *w* per una costante il decision boundary non cambia (si vede in $x_2 = - x_1 kw_1/kw_2 - kw_0/kw_2 = - x_1 w_1/w_2 - w_0/w_2$)

b) Conjunctions (AND) con modello lineare, come a lezione:

Realizzare $x_1 \wedge x_2 \wedge x_4$ ←→ *y con modello lineare*

$$h(\mathbf{x}) = \begin{cases} 1 & \_if \quad \mathbf{wx} + w_0 \geq 0 \\ 0 & \_\_\_\_\_otherwise \end{cases}$$

Soluzione: $1\,x_1 + 1\,x_2 + 0\,x_3 + 1\,x_4 \geq 2.5$

Ossia $h(\boldsymbol{x}) = \begin{cases} 1 & if \quad 1x_1 + 1x_2 + 1x_4 - 2.5 \geq 0 \\ 0 & otherwise \end{cases}$

Nota su uso di > o ≥: vedi sotto

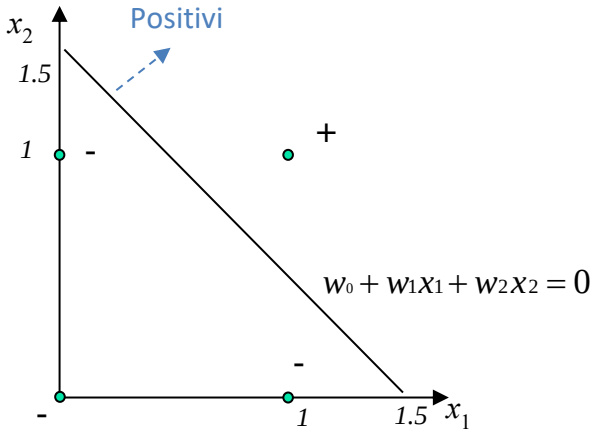
Nota generale: E' un esempio di soluzione, la soluzione NON è unica (altri valori di ***w*** soddisfano la disequazione nello spazio delle istanze). Ad esempio è valida anche con w_1=0.9, etc.

c) AND a 2 variabili:

$x_1 \wedge x_2$

$w_1=1$, $w_2=1$, $w_0=-1.5$

*h(**x**)*: scriverla come sopra con questi valori dei *w*



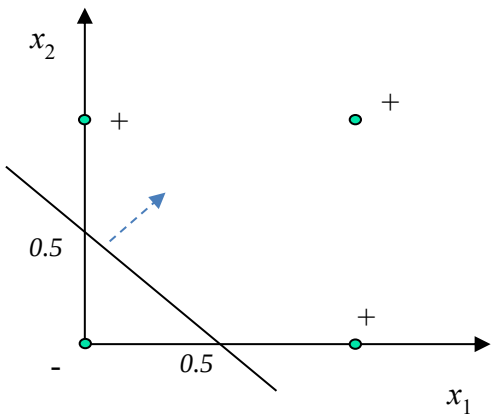
NOTA: per codifica dell'output 0/1 se si usa come sopra il maggiore o uguale (≥) in *h(**x**)*, w_0=-1.5 risolve.
Oppure si può definire *h(**x**)* con il >0 e così anche w_0=-1 risolve.
In alternativa ancora, una codifica -1/+1 (in input e output) porta ad esprimere *h(**x**)* con la funzione *sign()* e una soluzione può essere con w_0=-1 (con gli altri w invariati):
cambia ovviamente il plot nel piano cartesiano (provare il disegno).

Notare le 3 forme per lo stesso concetto: logica,
espressione lineare con la *h(**x**)*, grafica
(quest'ultima intuitiva ma non estendibile oltre le 3D)

d) OR a 2 variabili:

x_1 v x_2

$w_1=1$, $w_2=1$, $w_0=-0.5$, il resto come sopra



e) Invertire il segno (zona positiva con zona negativa del piano separatore)

Cambiando il segno ai *w* non cambia la linea del decision boundary [K=-1 punto a)] ma cambia il segno dove l'iperpiano è positivo o negativo. E.g. $x_1+x_2 \geq 0.5$ (operatore OR) versus $-x_1-x_2 \geq -0.5$, ossia $x_1+x_2 \leq 0.5$

Questo ultimo implementa un operatore logico NOR (vero solo se $x_1=x_2=0$ per input 0/1).

f) Tre punti con tutte le possibili etichette

1. Quanti modi esistono di etichettare 3 punti nel piano 2D con label di target 0/1?

2. Disegnare 3 punti (come vertici di un triangolo) e valutare se esistono delle configurazioni di assegnamento label tali che il problema risulti linearmente non separabile.

3. Esistono delle configurazione dei 3 punti per cui un assegnamento delle label risulti non linearmente separabile?

Soluzioni:

1. $2^3 = 8$

2. Per via grafica trovare (disegnare) un decision boundary lineare per tutte le 8 configurazioni

3. Si, per via grafica disegnare tre punti allineati con label al centro opposta alle label sui punti esterni

2 - Esercizio: derivare un modello lineare estremamente ridotto

Si vuol sviluppare un semplice sistema di previsioni per regressione lineare basato sul metodo Least Square e nessuna variabile di ingresso (polinomio di grado 0).

a) Scrivere l'equazione della ipotesi *h(**x**)*
b) Mostrare come risolvere per trovare il valore del parametro libero con *1* esempio di training.
c) Mostrare come risolvere per trovare il valore del parametro libero con *l (elle)* esempi di training.
d) Discutere il significato del risultato.
e) In caso di classificazione che ipotesi possiamo scrivere e che significato assume?

Soluzioni

a) $h(\mathbf{x}) = w_0$

b) Si considera un dato di ingresso con valore di target *y* e una soluzione Least Square per "apprendere" il valore di w_0.

$$\frac{\partial E(\mathbf{w})}{\partial w_0} = \frac{\partial (y - h_{\mathbf{w}}(x))^2}{\partial w_0} =$$

$$= 2(y - h_{\mathbf{w}}(x)) \frac{\partial (y - h_{\mathbf{w}}(x))}{\partial w_0} = 2(y - w_0) \frac{\partial (y - (w_0))}{\partial w_0}$$

$$\frac{\partial E(\mathbf{w})}{\partial w_0} = -2(y - w_0)$$

Imponendo $\frac{\partial E(\mathbf{w})}{\partial w_0} = 0$ si può in questo semplice caso ricavare direttamente $w_0 = y$ (ossia il target del dato disponibile).

c) Si hanno *l* dati. Non avendo variabili di input si considerano quindi solo i valori y_p, *p=1..l*

$$\frac{\partial E(\mathbf{w})}{\partial w_0} = \sum_{p=1}^{l} \frac{\partial (y_p - h_{\mathbf{w}}(x))^2}{\partial w_0} =$$

$$= \sum_{p=1}^{l} 2(y_p - h_{\mathbf{w}}(x)) \frac{\partial (y_p - h_{\mathbf{w}}(x))}{\partial w_0} = \sum_{p=1}^{l} 2(y_p - w_0) \frac{\partial (y_p - (w_0))}{\partial w_0} = \sum_{p=1}^{l} -2(y_p - w_0) = -2\left(\sum_{p=1}^{l} y_p\right) - l w_0)$$

Imponendo $\frac{\partial E(\mathbf{w})}{\partial w_0} = 0$ si può in questo semplice caso ricavare direttamente $w_0 = \frac{1}{l} \sum_{p=1}^{l} y_p$

d) Il learner emette un modello (ipotesi) che per ogni input restituisce la media dei valori dei target del training set.

e) $h(\mathbf{x}) = sign(w_0) = sign(\frac{1}{l} \sum_{p=1}^{l} y_p)$ che corrisponde ad un "voto di maggioranza" tra i dati postivi e negativi del training set.

3 - Ricavare Delta$_w$ con la *loss* della *ridge regression* (esercizio proposto a lezione sui linear model, qui per regressione)

Soluzione:

$$Loss(h_w)=\sum_{p=1}^{l}(y_p-h_{\boldsymbol{w}}(\boldsymbol{x}_p))^2 + \lambda \|\boldsymbol{w}\|^2$$

Svolgete (come suggerito nelle slide) i calcoli delle derivate rispetto a singoli w_i (del vettore $\boldsymbol{w}$) della nuova *Loss* con il termine di penalty separando il calcolo per i due termini sui dati ($E(\boldsymbol{w})$, da moltiplicare per η, ossia eta) e di penalty ($\lambda \Sigma_i w_i^2$), prendendo il negato del gradiente per ciascuno dei termini, ossia componete poi la regola di apprendimento (sempre con il – gradiente) e otterrete:

$$w_{i\,(new)} = w_i - \left(\eta \frac{\partial E(\boldsymbol{w})}{\partial w_i} + \lambda \frac{\partial \sum_i {w_i}^2}{\partial w_i}\right) = w_i - \left(\eta \frac{\partial E(\boldsymbol{w})}{\partial w_i} + 2\lambda w_i\right)$$
$$= w_i - \eta \frac{\partial E(\boldsymbol{w})}{\partial w_i} - 2\lambda w_i$$

Quindi (vettoriale) $\boldsymbol{w}_{\text{new}} = \boldsymbol{w} + \eta\, \Delta\boldsymbol{w} \; - 2\,\lambda\, \boldsymbol{w}$

Con il Delta$_w$ sul "data term" $E(\boldsymbol{w})$ che è ancora $\Delta w_i = -\frac{\partial E(\mathbf{w})}{\partial w_i} = \sum_{p=1}^{l}(y_p - \mathbf{x}_p{}^T \mathbf{w}) \cdot x_{p,i}$

4- Discutere il bias induttivo del modello lineare semplice (lineare rispetto agli input), con il suo algoritmo di apprendimento visto a lezione

Soluzione

Il *bias di linguaggio* è dovuto alla scelta di H come insieme di funzioni lineari (non permettendo di risolvere problemi non-lineari) (Nota: questo bias si può attenuare tramite la *linear basis expansion* permettendo modelli non-lineari rispetto alle variabili di input. Resta una assunzione di linearità sui *w)*.

Il *bias di ricerca* è dovuto all'ordine di ricerca imposti dalla Least Squares minimization con discesa di gradiente. Non è una strategia di ricerca completa (si ricordi che è una ricerca locale). Inoltre, nelle forme più articolate possiamo scegliere un metodo diverso, come ad esempio imporre nella Loss restrizioni sul valore dei pesi, che portano a diverse soluzioni con altre proprietà sul controllo della complessità (ad esempio con la ridge regression, Tikhonov) .

Nota fuori soluzione: Si noti come anche un modello così semplice comporti numerosi fattori di scelta del bias induttivo. La teoria del ML cerca di rispondere in modo generale a queste problematiche.

5 – Definire un task (il meteo)

Definire un task e un sistema di apprendimento per previsioni di gradimento sulle condizioni meteo, basato su variabili che misurano l'umidità (in percentuale) e la temperatura (in gradi Celsius) di oggi, che stimi se per voi sia o no una giornata piacevole. Si dispone di una serie di misurazioni delle variabili di interesse nel passato.

a) Definire l'input e il tipo delle variabili
b) Definire l'output e il target e il loro tipo
c) Definire il data set
d) Definire il tipo di task
e) Proporre un possibile modello per la soluzione
f) Descrivere cosa significa fare apprendimento e training per il problema.

Nota: Definire un compito (task) di apprendimento significa definire le risposte alle domande a, b, c, d.

Soluzioni

a) Variabili continue u e t; $u \in IR$, $t \in IR$ (a valori reali)
b) Output/target y= 1 se la giornata è piacevole, 0 altrimenti.
   y (e *target)* sono variabili discrete/categoriche binarie/Boolean oppure 0/1 oppure -1/+1.
   Si può anche sintetizzare con: funzione target: $R^n \rightarrow \{0/1\}$ etc., (ove qui n=2)
c) Data set: un insieme di l esempi del tipo coppie $([u_p, t_p], y_p)$ $p=1..l$
   oppure, definendo $\boldsymbol{x}_p=[u_p, t_p]$, un insieme di coppie $(\boldsymbol{x}_p, y_p)$ $p=1..l$
d) Si tratta di un task supervisionato di classificazione binaria (o concept learning)
e) Se si è scelto ad esempio la codifica del target e uscite in -1/+1,
   un possibile modello per la soluzione: $h(\boldsymbol{x})= sign(w_1 u + w_2 t + w_0)$
f) Abbiamo molte formulazioni del problema nel corso. Ad esempio possiamo dire che:
   Fare <u>*apprendimento*</u> significa trovare un'ipotesi del tipo $h(\boldsymbol{x})$ che sia una buona approssimazione della funzione target (che è nota solo sugli esempi), ossia che minimizzi l'errore vero, cioè l'errore su tutto l'insieme dei dati possibili (R in SLT). Nel caso dell'esercizio si tratta in particolare di inferire una funziona Booleana da un insieme di esempi positivi e negativi.
   Il <u>*training*</u> (in accordo al principio induttivo della minimizzazione del rischio empirico) si effettua utilizzando i dati/esempi disponibili (ad esempio gli l dati del data set o una loro porzione) per cercare i valori dei parametri liberi w che minimizzano l'errore/Rischio empirico $E(\boldsymbol{w})$:

$$E(\mathbf{w}) = \sum_{p=1}^{l} (y_p - \mathbf{x}_p{}^T \mathbf{w})^2 = \| \mathbf{y} - \mathbf{X}\mathbf{w} \|^2$$

   (ossia, nell'approccio LMS, cercare i valori di w tali che sia minimo l'errore quadratico (medio) su l'insieme di dati di training [ma è preferibile forniate la versione con l'equazione!])

6 - Find-S e Candidate Elimination (esercizio proposto alla lezione su concept learning)

Provare Find-S e Candidate Elimination sul seguente instance space, usando x_i per i letterali (o l_i), aggiungendo "not(x_i)" allo spazio delle "simple conjunctive rules", risolvendolo nel formalismo "alla Mitchell" e notando poi che rispetto a quello sia ha 0 → not(x_i), 1 → x_i, ? → no literal

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Soluzioni

Find-S: Risoluzione nello linguaggio "alla Mitchell"

h_0 =<Ø, Ø, Ø, Ø, Ø> = h_1 = h_2 (in Find-S ci sono modifiche solo per esempi positivi, qui come 1)

h_3= <0,0,1,1>

h_4= <?,0,?,1> = h_5 = h_7

STOP

Soluzione finale: h(***x***) = not(x_2) and x_4 (verificatela sulla tabella sopra)

Candidate Elimination:

S_0: {< Ø, Ø, Ø, Ø >} G_0: {< ?, ?, ?, ? >}

$\xrightarrow{Ese\ 1}$ S_1: {< Ø, Ø, Ø, Ø >} G_1: {< 1, ?, ?, ?, ? >, < ?, 1, ?, ? >, < ?, ?, 0, ? >, < ?, ?, ?, 1 >} (perché non anche <?,0,?,1> ?)

$\xrightarrow{Ese\ 2}$ S_2: {< Ø, Ø, Ø, Ø >} G_2: {< 1, ?, ?, ? >, <?, 1, 1, ?>, >, <?, 0, 0, ?>, < ?, ?, ?, 1 >} (le due centrali di G_1 darebbero 1 su Ese 2, per quella in seconda posizione trova la specializzazione <?, 1, 1, ?> , per la terza <?, 0, 0, ?>; altre varianti sono meno generali di ipotesi già presenti in G o inconsistenti, etc.)

$\xrightarrow{Ese\ 3}$ S_3: {< 0, 0, 1, 1 >} G_3: {< ?, ?, ?, 1 >} (stesso passo visto in Find-S e le prime di G_2 darebbero 0 su Ese 3, inconsistenti)

$\xrightarrow{Ese\ 4}$ S_4: {< ?, 0, ?, 1 >} G_4: {< ?, ?, ?, 1 >} (stesso passo visto in Find-S)

$\xrightarrow{Ese\ 5\ e\ 6}$ S_6: {< ?, 0, ?, 1 >} G_6: {< ?, ?, ?, 1 >} (specializzare G_6 ma non in posizione 1 o 3, altrimenti quella h non sarebbe più generale di S, i.e. alcuni membri di S devono essere più specifici di quella h)

$\xrightarrow{Ese\ 7}$ $S_7 = G_7$: {< ?, 0, ?, 1 >}

STOP

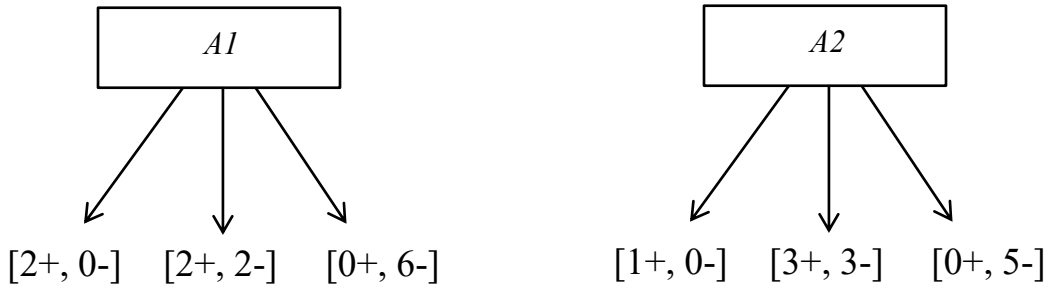
Nota che converge a una singola ipotesi (ammettendo in fondo in questo caso h_7=S=G)

Soluzione finale: h(***x***) = not(x_2) and x_4

7 - Esercizio: Decision Tree costruction

Nella costruzione di un Decision Tree,

a) scegliere tra i 2 nodi con attributi candidati descritti in figura e motivare la scelta in base al calcolo dell'Information Gain ($log_2 1/2=-1$ e assumere $0 log_2 0=0$). Mostrare i calcoli.



b) E' possibile o frutto di un errore considerare un attributo A3 che dia luogo alla ripartizione seguente? [2+, 0-] [3+, 3-] [0+, 4-]

Soluzioni

a) Ricordando le definizioni

Entropy (S) $\equiv -p_+ log_2 p_+ - p_- log_2 p_-$ *[assume 0 $log_2 0$ = 0]*

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv)$$

si può svolgere il seguente calcolo:

Gain per A1: **G(S,A1)**= E(S)-(2/12(-2/2 $log_2$2/2-0/2 $log_2$0/2)+4/12(-2/4$log_2$2/4-2/4 $log_2$2/4)
+6/12(-0/6 $log_2$0/6-6/6 $log_2$6/6))= E(S)- (2/12 (0) +4/12 (1) + 6/12 (0))= E(S)-4/12=**E(S)-1/3**

Analogo per A2: **G(S,A2)**= E(S)-1/12 (0) -6/12(1) – 5/12 (0)= E(S)-6/12= E(S)-6/12=**E(S)-1/2**

Si noti che E(S) è lo stesso per entrambi. Il maggiore è quindi G(S,A1), che vince.

Si sceglie A1.

b) E' frutto di un errore, A3 non può ripartire in quel modo poiché il totale dei positivi (5) e negativi (7) non è lo stesso degli altri due che erano 4 positivi e 8 negativi.