IIA ML Esercitazioni IV (testo con cenni di soluzioni)

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

Suggerimenti generali:

- Si vedano le note iniziali negli altri file
- *Nota: Questi esercizi a seguire sono in forma più di domande, con risposte multiple, V/F o libere, **ben adeguate anche alle prove e quiz su elearning**, soprattutto per gli esercizi **1, 3, 4 ,5, 6 e 7** (anche se alcuni erano già presenti nei file di esercitazioni precedenti, qui si vogliono fornire altri esempi in forma di quiz)*

1 – Confrontare la complessità (flessibilità in fitting) delle seguenti classi di modelli , ove x>y esprime che x ha complessità (potenzialmente) maggiore di y:

a. modello lineare (negli input) con 3 variabili di input
b. modello lineare (negli input) con 5 variabili di input
c. K-nn con 3 variabili di input e k=1
d. K-nn con 3 variabili di input e k=5
e. modello con 3 variabili di input da algoritmo Find-S
f. decision tree con 3 variabili di input

In particolare, queste affermazioni sono corrette? (esprimere tutte quelle corrette ma affermazioni indicate come corrette che non lo sono incidono negativamente sul punteggio dell'esercizio).

1. a>b (**NO**, con 5 variabili si hanno potenzialmente più parametri liberi)
2. f>e (**SI**, DT è più espressivo, e flessibile, dei soli "and" di variabili)
3. d>c (**NO**, K = 1 max. flessibilità)
4. a>c (**NO**, 1-nn più flessibile, anche decision boundary non lineari)
5. c> e (**SI**, Find-s più rigido, bias di linguaggio più forte)
6. f> a (**SI**, e.g. per risolvere lo xor, possibile con DT, non con lineare: è una domanda meno ovvia, ma si è visto un esempio nell'esercitazione 2, Esercizio 8.2.d)

2 - Rispondere ai seguenti quesiti sul Machine Learning.
Nota: risposte valide se contengono la risposta all'affermazione/domanda (Si/No o dire in che casi sia Si o No) e la motivazione [con testo libero, sintetico].

a. Si deve scegliere tra un modello Decision Tree e uno lineare (nelle variabili di ingresso) per risolvere un problema in cui sapete che la funzione target prevede un "and" tra le variabili (booleane) di ingresso. Un collega vi suggerisce il Decision Tree ravvisando il potenziale limite del modello lineare. Siete d'accordo?
   NO, un modello lineare può implementare un AND. (Fuori soluzione: si veda lezione linear model e la esercitazione I)

b. La complessità della SVM è legata al margine?
   SI, la VC-dimension della SVM decresce con alto valore del margine

c. Vi chiedono di giudicare una procedura per stabilire quando terminare il training (assunto iterativo) di un modello. Si propone in particolare di utilizzare un data set separato dal training set per decidere la terminazione in base al minimo errore raggiunto su quel data set e si suggerisce che per ottenere un modello con la miglior capacità di generalizzazione sarebbe utile valutare in base alla partizione dei dati del test set. Che cosa rispondete in merito al criterio di terminazione sulla base dell'errore e in merito all'intero processo cosi come proposto? Siete cioè d'accordo?
   Si per il criterio (è un criterio valido) se applicato con un validation set.
   No, se fatto sul test set, che non va utilizzato per decidere elementi del modello o del training ma solo per la stima di comportamento su dati nuovi (model assessment).

d. Dopo che avete calcolato la soluzione (valori alfa) di una SVM *hard-margin*, vengono aggiunti dei punti nel training set. Cambia la soluzione? Rispondere Si o No o dire in che casi sia vero o falso (e come cambia o perché non cambia).
   Dipende: Si, vero se erano punti dentro il margine (avremo una nuova soluzione con margine più stretto). Cambia anche se invece diventa un task non linearmente separabile (esempio punti negativi tra i positivi non separabili da un iperpiano): in quel caso SVM fallisce a trovare una soluzione che soddisfi i vincoli.
   Altrimenti è Falso (non cambia se sono inseriti dal lato giusto dei già correttamente classificati, esempio un positivo tra i positivi) *.
   [Nota: E' un esercizio più avanzato degli altri].
   (*) Si ricorda infatti che "The hyperplane depends only from support vectors", che sono calcolati con gli alfa diversi da zero. Se i punti non diventano vettori di supporto (entrando nel margine calcolato sinora) la soluzione non cambia.

3- La ridge regression (Tikhonov) e la LBE polinomiale (di grado M). Quali tra queste affermazioni sono vere?

Nota: vanno indicate tutte quelle vere ma affermazioni indicate come vere che non lo sono incidono negativamente sul punteggio dell'esercizio.

1. In una LBE polinomiale, il grado M regola la complessità del modello (**V**)
2. Lambda nella ridge regression non regola la complessità del modello (**F**)
3. Posso mettere M alto e poi usare il valore di lambda per controllare l'over-fitting (**V**)
4. Posso mettere M alto e poi usare il valore di lambda per controllare l'under-fitting (**V**, perché lambda alto può comunque portare in under-fitting)
5. Se ho M alto il termine con lambda non serve (**F**)
6. Se ho M basso il termine con lambda può non servire (**V, si noti il "può"**)
7. Il parametro lambda influenza il livello di regolarizzazione (**V**)
8. Più alto è il grado M del polinomio più è bene utilizzare lambda bassi (**F**)
9. Scegliere sia M che lambda non permette di trovare un bilanciamento tra under- e over-fitting (**F**)
10. Posso usare la ridge regression anche per LBE di tipo diverso da quelle polinomiali (**V**)
11. La tecnica di Tikhonov con il lambda si usa solo per problemi di regressione (**F**, vale anche per classificatori)
12. C'è un limite al grado M del polinomio da usare (**F**, dipende dal task)
13. Con pochi dati usare un polinomio di grado M elevato può richiedere un valore di lambda maggiore (**V**, vedere il bound SLT)

4- Discutere le seguenti affermazioni riguardanti possibili scelte di design di modelli di ML, ossia dire se considerarle Vere o False e motivare la risposta* (molto brevemente in una o due righe), alla luce dei principi alla base del ML [con testo libero, sintetico].

(*nei compiti con risposte libere le risposte non motivate non sono considerate)

1. Una SVM non può risolvere problemi non linearmente separabili. **[Falso,** la SVM generale, ossia se uso i Kernel, può risolverli**]**
2. In modello lineare soggetto a una *linear basis expansion* si aumenta la flessibilità del modello.
   [Vero, aumenta la complessità, ossia la VC dim. del modello; accettabile anche dire almeno che introduce non linearità rispetto agli input**]**
3. In un DT limitare il numero dei nodi a un massimo fissato garantisce un buon apprendimento.
   [Falso, dipende dal valore massimo (e dal task) per cui è falso che fissato un valore si possa garantire un buon apprendimento, e.g. con un valore basso si rischia l'underfitting, causa Remp alto**]**
4. In una discesa di gradiente non fermarsi presto assicura un buon learning.
   [Falso, rischio overfitting continuando con il fitting curando solo Remp e non la VC-confidence o si pensi a un modello dopo LBE con i *w* che crescono molto e VC-dim aumentata. In alternativa si poteva almeno dire che dovremmo comunque osservare l'errore sul VL set**]**
5. In un DT il numero di nodi non deve essere inferiore alla metà dei dati di training.
   [Falso, è una regola senza fondamento, potrei avere ottimi alberi piccoli se ho pochi attributi ben discriminanti, mentre avere molti nodi aumenta la VC-dim e con pochi dati rischio overfitting**]**
6. Per stimare la (futura) capacità predittiva di un modello è bene considerare accuratamente il risultato della model selection senza guardare al risultato in training.
   [Falso, nel senso che non basta, serve poi anche un Test Set per la stima della capacità predittiva**]**

5 - Giudicate quali tra le seguenti affermazioni sono Vere o False (SVM)

1. I vettori di supporto sono dati del problema
   (**V,** sono gli esempi nel training set che determinano piano separatore e margine)

2. I vettori di supporto sono rette nell'iperpiano usate per separare i punti
   (**F**, sono punti, ossia dati, ed è invece il decision boundary che separa i punti)

3. I vettori di supporto sono indicati dal designer dell'applicazione
   (**F**, si calcola quali siano, vedi la risoluzione del duale)

4. Il numero di vettori di supporto può eccedere il numero di dati di test
   (**V**, sono (una) parte dei dati di TR e il numero di dati di TS non è pertinente, potrebbe anche essere inferiore)

6 - Scegliere tra le seguenti le opzioni appropriate in merito a underfitting e overfitting.

Nota: Vanno indicate tra le seguenti **tutte** e solo le opzioni appropriate; le opzioni indicate appropriate che non lo sono incidono con un punteggio negativo o più sul totale dell'esercizio.

1. Il linear model tende all'overfitting se il numero di variabili in input è molto alto [**V**, aumentano i parametri liberi *w*]
2. Il linear model tende all'underfitting se la LBE prevede molti termini [**F**]
1. Il linear model può tendere all'overfitting se si usa regolarizzazione di Tikhonov con un lambda molto alto [**F,** se aumenta lambda si regolarizza molto]
2. La SVM può tendere all'overfitting con C alto [**V,** anche se qui non abbiamo specificato un kernel la tendenza è appropriata]
3. La SVM tende all'overfitting con kernel polinomiale a grado basso e C basso [**F,** e qui il kernel è chiaro]
4. La SVM tende all'overfitting con un margine ampio [**F,** max. margine porta a diminuire la VC-dimension]
5. Un Decision Tree tende all'underfitting con troppi nodi [**F**, anzi tende all'overfitting potenzialmente]
6. Un Decision Tree tende all'overfitting con pochi dati in test [**F**, non ha senso, il test set non decide come costruire il modello]
7. Il K-NN tende all'overfitting con K troppo bassi [**V**]

7. SLT e Polinomio

In una LBE polinomiale di grado "g", cosa vi aspettate con maggiore probabilità in base alla disequazione della Statistical Learning Theory?

Nota: Vanno indicate **tutte** e solo le opzioni appropriate, ma opzioni indicate come appropriate che non lo sono incidono con un punteggio negativo sul totale dell'esercizio.

1. L'aumento di g tende a far aumentare la VC-confidence (a pari valore degli altri termini e valori presenti) [**V**, aumenta la VC-dim]
2. La diminuzione di g tende a ridurre la VC-dim [**V,** giacché misura la flessibilità del modell**o**]
3. La diminuzione di g fa diminuire il valore di *l* (elle) **[F**, g non incide sul numero dati]
4. Aumentare g tende a ridurre il R_{emp} [**V**]
5. La somma dei due termini R_{emp} e VC-confidence aumenta sempre con l'aumentare di g [**F,** non è detto**,** dipende quanto aumentano o diminuiscono i due addendi R_{emp} e VC-confidence]
6. Si può generalizzare bene quando il valore di g permette un bilanciamento tra R_{emp} e VC-confidence [**V**, è la *structural risk minimization*]