IIA ML Esercitazioni – parte 2

Si leggano le note inziali nell'esercitazione 1.

Di seguito alcune note (testo degli esercizi) di alcuni esercizi svolti nelle lezioni di esercitazione – part 2:

1 – Esercizio per validation (I):

Data la tabella con i seguenti valori di accuracy (% di classificazione corretta) per un iper-parametro lambda:

λ	TR	VL	TS
0.5	75	70	70
0.1	80	75	74
0.01	90	70	72

1. In che ordine si usano le porzioni di dati per calcolare i valori in tabella?
2. Quale modello (ossia lambda) si sceglie?
3. Che fenomeni si osservano?

2 – Esercizio per validation (II), quello suggerito alla lezione su validation (slide "Exercise"):

Data la tabella con i seguenti valori di accuracy (% di classificazione corretta)

λ	TR	VL	TS
0.5	75	70	70
0.1	80	75	70
0.01	90	70	72

1. In che ordine si usano le porzioni di dati per calcolare i valori in tabella?
2. Quale modello (ossia lambda) si sceglie?
3. Che fenomeni si osservano?

3 - Descrivere cosa si intende per generalizzazione

Provare una risposta intuitiva in base a quanto presente nel corso, raccogliendo i concetti espressi in varie parti per una sintesi.

4 - Descrivere il rapporto tra generalizzazione e il fenomeno dell'overfitting

Provare una risposta intuitiva, in base alle def. di overfitting, o in accordo al VC-bound (in accordo a quanto presentato nel corso). Preferire quella di livello più alto.

5 - Quali parametri regolano il controllo di complessità nei vari modelli visti a lezione?

Provare un risposta in base a quanto presente nel corso, raccogliendo i concetti espressi in varie parti per una sintesi.

6 - Discutere vantaggi e svantaggi relativi tra coppie di modelli viste durante il corso

Pensarlo in base a criteri informatici classici e a quelli più specifici del ML. Pensare ai criteri e poi discuterli per i vari modelli a confronto. Esercizio in auto-valutazione di ausilio al ripasso ragionato della materia.

Repetita: Testo dell'esercizio 5 di esercitazione 1: – 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.

7 -Il meteo CONTINUO (da un compito passato con qualche variazione)

Sul problema del meteo definito nell'esercizio 5 dell'esercitazione 1:

a. Proporre un possibile modello per la soluzione con SVM:
    1. Scrivere la forma (in esteso) delle ipotesi nella forma $h(\boldsymbol{x})$
    2. Come aumentare la capacità espressiva di h nel caso che le relazioni lineari sulle variabili non siano sufficienti per il problema? Esprimere la nuova forma (generale) delle $h(\boldsymbol{x})$
    3. Che rapporto c'è tra il modello del punto precedente e il fenomeno dell'overfitting? E come poterlo gestire?
b. Proporre un possibile modello per la soluzione con un K-NN:
    1. Esprimere l'equazione per la previsione del gradimento
    2. Discutere di nuovo il punto sull'aumento della capacità espressiva per il caso K-NN
    3. Discutere di nuovo il punto sul rapporto c'è tra il modello e il fenomeno dell'overfitting. E come poterlo gestire?

8 --- COMPITINO anni scorsi INZIO ---

Esercizio 8.1: Il cliente del modello lineare

Il cliente della vostra start-up vi propone un problema di classificazione a 3 variabili di ingresso per stimare preferenze (positive o negative). Vi offre 1000 dati etichettati di esempio.

a. Scrivere la $h_a(\boldsymbol{x})$ di un modello lineare per il problema.
b. Dopo avere eseguito un training (a discesa di gradiente), avete ottenuto una *accuracy* di training (ossia sul training set) del 70%. Informato del risultato preliminare, il cliente chiede il modello ottenuto e poi chiede di aggiungere anche il logaritmo delle 3 variabili originarie al modello.
   Scrivere la nuova $h_b(\boldsymbol{x})$ e calcolare il gradiente (esplicitamente, mostrando i calcoli) per i pesi (ossia per uno dei pesi) di queste tre nuove variabili.
c. L'accuracy di training con il nuovo modello h_b diviene del 75%. Il cliente, preso anche il secondo modello, chiede di aggiungere anche il quadrato e il cubo delle 3 variabili iniziali.
   Scrivere la nuova $h_c(\boldsymbol{x})$ e dire cosa cambia nella regola di apprendimento del vostro algoritmo.
d. L'accuracy di training con la h_c cosi completata raggiunge l'80%. Il cliente si ritiene soddisfatto e vi chiede di dirgli cosa si deve aspettare come accuratezza per l'uso del modello: Cosa rispondete? Cosa rischia nel prendere la h_c?
e. Il cliente fa una nuova rivelazione. Oltre ai 1000 dati forniti a voi, ha preparato altri 100 dati etichettati, diversi dai primi, con cui prova i modelli ottenuti ai passi a, b, c. Poiché il miglior risultato di queste prove sui 3 modelli lo ottiene sul modello h_b con il 73% di accuracy, misurata sui questi 100 nuovi dati, sceglie definitivamente quello e vi riferisce che è soddisfatto di avere un modello che fornirà predizioni al 73% di accuracy. Che ne pensate di questa stima? E cosa suggerireste di fare al cliente?

Esercizio 8.2: Costruire DT per funzioni booleane

a. Costruire un Decision Tree con l'algoritmo ID3 e l'uso di *Gain(S, A)* che realizzi la funzione boolena x_1 AND x_2 (su variabili booleane).
   1. Mostrare i calcoli di *Gain(S, A)*
   2. Mostrare l'albero ottenuto
   3. Disegnare nello spazio delle istanze le zone di classificazione che produce l'albero ottenuto
b. Costruire un Decision Tree con l'algoritmo ID3 e l'uso di *Gain(S, A)* che realizzi la funzione espressa in tabella.
   1. Mostrare i calcoli di *Gain(S, A)*
   2. Mostrare l'albero ottenuto
   3. Disegnare nello spazio delle istanze le zone di classificazione che produce l'albero ottenuto

x_1	x_2	Target
0	0	0
0	1	0
1	0	1
1	1	1

c. Se *n* è la dimensione del vettore di ingresso, quanti nodi potrebbero esser aggiunti da ID3 nel caso pessimo e perché?
d. Sapreste fare un esempio di funzione booleana che corrisponda al caso pessimo del punto c con 2 variabili di ingresso booleane (e mostrando l'albero)?

--- COMPITINO anni scorsi FINE ---