# Machine Learning & Data Mining

Prova scritta 17 Giugno 2015

1) **Concept learning**

Indicare sotto quali condizioni l'algoritmo Candidate-Elimination converge all'ipotesi che descrive correttamente il concetto target e quando si dice che il concetto è appreso "esattamente".

2) **Alberi di decisione**

2.a) Descrive il metodo per decidere il best split che utilizza l'information gain, del quale si dia la definizione. Discutere quando questo metodo può essere problematico e indicare come possa essere modificato per evitare tale problema.

2.b) Cosa si intende per Search Bias in un algoritmo per l'apprendimento induttivo di un albero di decisione? Che differenza c'è con il concetto di hypothesis/language bias di Candidate-Elimination?

2.c) Descrivere i passi dell'algoritmo di base per la costruzione (attraverso learning induttivo) di un albero di decisione.

3) **Valutazione di algoritmi**

Dare le definizioni di true error, sample error e estimation bias per un'ipotesi h.

4) **Reti Neurali**

4.a) Descrivere i passi svolti dall'algoritmo di Backpropagation per ciascuna epoca

4.b) Discutere brevemente quando è preferibile usare una rete neurale piuttosto che un albero di decisione.

5) **Learning Bayesiano**

5.a) Indicare quali sono i principali vantaggi e svantaggi del classificatore bayesiano ottimale.

5.b) Descrivere l'algoritmo (classificatore) di Gibbs e dare un bound sull'errore che esso produce rispetto all'errore del classificatore bayesiano ottimale.

6) **Clustering**

6.a) Descrivere principali problematiche dell'algoritmo k-means e possibili soluzioni al problema della scelta iniziale dei centroidi

6.b) Data la seguente matrice di similarità:

| | p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|---|
| p1 | 1.00 | 0.10 | 0.41 | 0.55 | 0.35 |
| p2 | 0.10 | 1.00 | 0.64 | 0.47 | 0.98 |
| p3 | 0.41 | 0.64 | 1.00 | 0.44 | 0.85 |
| p4 | 0.55 | 0.47 | 0.44 | 1.00 | 0.76 |
| p5 | 0.35 | 0.98 | 0.85 | 0.76 | 1.00 |

Calcolare un single link e un complete link hierarchical clustering. Mostrare i risultati disegnando i rispettivi dendrogrammi.

7) **Ensemble Methods**

Descrivere la fase di training e testing dell'algoritmo bagging

8) **ROC Curves**

Dati due sistemi di classificazione binari M1 e M2, vogliamo valutarne le performance. Dato il seguente test set con 14 istanze che contiene 5 attributi binari con label da A ad E, la seguente tabella mostra le probabilità a posteriori ottenute applicando M1 ed M2 al test set (solo le probabilità a posteriori per le classi positive sono state mostrate). Considerando un problema di classificazione binario abbiamo che P("No")=1-P("Yes") e P("No"| A,...,E)=1-P("Yes"| A,...,E)

| Istanza | Classe reale | P("Yes"| A,...,E,M1) | P("Yes"| A,...,E,M2) |
|---|---|---|---|
| 1 | No | 0.962 | 1 |
| 2 | Yes | 0.929 | 1 |
| 3 | No | 0.962 | 0.75 |
| 4 | Yes | 0.929 | 1 |
| 5 | No | 0.071 | 0 |
| 6 | Yes | 0.342 | 0 |
| 7 | No | 0.5 | 0.667 |
| 8 | Yes | 0.071 | 0 |
| 9 | No | 0.341 | 0 |
| 10 | Yes | 0.5 | 0.25 |
| 11 | Yes | 0.067 | 1 |
| 12 | Yes | 0.5 | 0.333 |
| 13 | Yes | 0.964 | 1 |
| 14 | Yes | 0.659 | 1 |

Disegnare sullo stesso grafico le ROC curve di M1 e M2. Considerando un threshold t=0.5 (le istanze di test con probabilità a posteriori maggiori di t verranno considerate come esempi positivi), calcolare le confusion matrix. Quale dei due modelli si ritiene migliore? Spiegarne il motivo.

9) **SVM**

Dati i seguenti 11 elementi del nostro dataset (in $R^2$), calcolare l'equazione della retta che rappresenta la SVM lineare che separa le due classi massimizzando il margine; rappresentare sul piano $R^2$ i punti del dataset e la retta ed indicare quali sono i support vector.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **x** | 2 | 2 | 4 | 5 | 7 | 6 | 7 | 10 | 10 | 12 | 12 |
| **y** | 4 | 7 | 2 | 5 | 2 | 10 | 7 | 6 | 9 | 4 | 7 |
| **Class** | A | A | A | A | A | B | B | B | B | B | B |