Rule-based classifier

R: ($condition_i$) --> y
$condition_i$ = antecedente, precondizione
y = conseguente = classe

Esercizio 1

Si consideri un problema di classificazione binaria con il seguente set di attributi e di valori:

Air conditioner = {working, broken}
Engine = {good, bad}
Mileage = {high, medium, low}
Rust = {yes, no}

Si supponga che un classificatore rule-based abbia prodotto il seguente insieme di regole:

R1: mileage = high --> value = low
R2: mileage = low --> value = high
R3: air conditioner = working, engine = good --> value = high
R4: air conditioner = working, engine = bad --> value = low
R5: air conditioner = broken --> value = low

a) Le regole sono mutuamente esclusivo?

No, esistono almeno due regole che sono attivate dalle stesse istanze. Infatti, una stessa istanza può attivare sia la regola R1 che la regola R5, producendo tra l'altro due classificazioni diverse. Per risolvere questo problema sarebbe necessario imporre un ordinamento sulle regole, in modo che appena una regola nell'ordine è attivata dall'istanza, non si procede oltre valutando regole successive.

b) L'insieme regole è esaustivo?

Si, perchè esiste almeno una regola per ogni possibile istanza. Se l'insieme non fosse stato esaustivo sarebbe stato necessario predisporre una regola di default: r: {} --> classe_default da attivare se nessuna delle altre regole fosse attivabile da un'istanza.

Esercizio 2

Si supponga di avere due regole R1 ed R2 tali che:
R1 è coperta da p1 = 350 istanze positive e n1 = 150 istanze negative
R2 è coperta da p2 = 300 istanze positive e n2 = 50 istanze negative

a) Calcolare l'indice FOIL information gain di R2 rispetto a R1.

FOIL $= p_2 \times \left(\log_2 \frac{p_2}{p_2+n_2} - \log_2 \frac{p_1}{p_1+n_1}\right)$

FOIL di R2 rispetto a R1 = 300 * (log2 300/350 - log2 350/500) = 87.65

b) Calcolare l'accuratezza di R1 e R2

Acc(R) = $\frac{|A \cap y|}{|A|}$

Dove |A| è il numero di istanze che attivano R e |y| il numero di istanze che hanno come classe y.

Acc(R1) = 350/500 = 0.7 = 70%
Acc(R2) = 300/350 = 0.86 = 86%

c) Calcolare la copertura (coverage di R1 e R2), considerando che il numero totale di istanze nel training set sia 500.

Coverage(R) = $\frac{|A|}{|D|}$

Dove |A| è il numero di istanze che attivano R e |D| è il numero di istanze totali contenute nel training set.

Coverage(R1) = 1
Coverage(R2) = 350/500 = 0.7 = 70%

Esercizio 3

Si consideri un training set composto da 100 istanze positive e 400 istanze negative. Per ciascuna delle seguenti regole candidate:

R1: A → + (copre 4 istanze positive e 1 istanza negativa)
R2: B → + (copre 30 istanze positive e 10 istanze negative)
R3: C → + (copre 100 istanze positive e 90 istanze negative)

Calcolare quale è la regola migliore e peggiore rispetto a (a) il criterio di accuratezza e (b) il FOIL information gain.

(a) Acc(R) = $\frac{|A \cap y|}{|A|}$

Acc(R1) = 4/5 = 0.8 = 80%
Acc(R2) = 30/40 = 0.75%
Acc(R3) = 100/190 = 0.526 = 52.6%

R1 è la regola migliore rispetto alla metrica dell'accuratezza, mentre R3 è la regola peggiore.

(b) Si assuma che la regola iniziale (regola di default) sia R: ∅ --> + che copre p0 = 100 istanze positive e n0 = 400 istanze negative.

FOIL(R1) = $p_1 \times \left(\log_2 \frac{p_1}{p_1+n_1} - \log_2 \frac{p_0}{p_0+n_0}\right)$ = 4 * (log2 4/5 - log2 100/500) = 8
FOIL(R2) = 30 * (log2 30/40 - log2 100/500) = 57.2
FOIL(R3) = 100 * (log2 100/190 - log2 100/500) = 139.6

R3 è la regola migliore rispetto alla metrica del FOIL information gain, mentre R1 è la regola peggiore.

Nearest Neighbor (k-nn)

Esercizio 4

Si consideri la seguente tabella di istanze mono-dimensionali

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

(a) Si classifichi l'istanza x = 5.0 rispetto alla tecnica k-nn (majority vote) con i seguenti valori di k = 1-, 3-, 5- e 9-.

Calcoliamo la distanza di x rispetto alle istanze nel training set

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-
d	4.5	2.0	0.5	0.4	0.1	0.2	0.3	0.5	2.0	4.5

K = 1: l'istanza più vicina è 4.9 (d=0.1) → la classe di x è +
K = 3: k-nn = {4.9 (+), 5.2 (-), 5.3 (-)} → la classe di x è –
K = 5: k-nn = {4.9 (+), 5.2 (-), 5.3 (-), 4.5 (+), 5.5 (+)} → la classe di x è +
K = 9: k-nn = {4.9 (+), 5.2 (-), 5.3 (-), 4.5 (+), 5.5 (+), 4.5 (+), 3.0 (-), 7.0 (-), 0.5 (-)} → la classe di x è -

(b) Si classifichi l'istanza x = 5.0 rispetto alla tecnica k-nn con voto pesato rispetto alla distanza (distance weighted voting). $w = 1 / d(x', x_i)2$

K = 1: l'istanza più vicina è 4.9 (d=0.1) → la classe di x è +

K = 3: k-nn = {4.9 (+, w = 1/0.12 = 100), 5.2 (-, w=1/0.22 = 25), 5.3 (-, w = 1/0.32 = 11)} → la classe di x è +

K = 5: k-nn = {4.9 (+, w = 1/0.12 = 100), 5.2 (-, w=1/0.22 = 25), 5.3 (-, w = 1/0.32 = 11), 4.6 (+, w=1/0.42 = 6.25), 4.5 (+, w=1/0.52 = 4)} → la classe di x è +

K = 9: k-nn = {4.9 (+, w = 1/0.12 = 100), 5.2 (-, w=1/0.22 = 25), 5.3 (-, w = 1/0.32 = 11), 4.6 (+, w=1/0.42 = 6.25), 4.5 (+, w=1/0.52 = 4), 5.5 (+, w=1/0.52 =4), 3.0 (-, w=1/22 =0.5), 7.0 (- w=1/22 =0.5), 0.5 (-, w=1/4.52 = 0.049)} → la classe di x è +

Naive Bayes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

La probabilità che l'istanza abbia la classe $Y=y_i$ condizionata al fatto che i suoi attributi siano $X = x_1 \ldots x_n$

Naive bayes perchè per il calcolo di P(X) si assume l'indipendenza degli attributi; quindi, si possono moltiplicare le probabilità dei singoli attributi nel vettore X.

Esercizio 5

Si supponga che la frazione di studenti non laureati (undergraduated) che fuma sia il 15% e che la frazione di studenti laureati (graduated) che fuma sia il 23%. Se 1/5 degli studenti sono laureati e gli altri sono non-laureati.

S = smoke
UG = undergraduated
G = graduated

P(S|UG) = 0.15
P(S|G) = 0.23
P(G) = 1/5 = 0.2
P(UG) = 4/5 = 0.8

(a) Qual è la probabilità che uno studente che fuma sia uno studente laureato?

P(G|S) = qual è la probabilità che l'etichetta sia "graduated" se lo studente è fumatore.

P(G|S) = $\frac{0.23 \times 0.2}{0.15 \times 0.8 + 0.23 \times 0.2}$= 0.277

(b) Dato uno studente preso casualmente è più probabile che sia uno studente laureato o non laureato?

Uno studente non laureato: P(UG) > P(G)

(c) Si supponga che il 30% degli studenti laureati viva in un dormitorio, ma che solo il 10% degli studenti non laureati viva in un dormitorio. Se uno studente fuma e vive in un dormitorio, è più probabile che sia uno studente laureato o non laureato? Si assuma l'indipendenza tra il fatto che uno studente viva in un dormitorio ed il fatto che fumi

P(D) = probabilità che viva in un dormitorio
P(DS) = probabilità che viva in un dormitorio e che fumi

P(D|UG) = 0.1
P(D|G) = 0.3
P(D) = P(UG) * P(D|UG) + P(G) * P(D|G) = 0.8 * 0.1 + 0.2 * 0.3 = 0.14
P(S) = P(UG) * P(S|UG) + P(G) * P(S|G) = 0.8 * 0.15 + 0.2 * 0.2 = 0.166
P(DS|G) = P(D|G) * P(S|G) = 0.3 * 0.23 = 0.069 (ipotesi indipendenza!)
P(DS|UG) = P(D|UG) * P(S|UG) = 0.1 * 0.15 = 0.015 (ipotesi indipendenza!)

P(G|DS) = $\frac{0.069 \times 0.2}{P(DS)} = \frac{0.0138}{P(DS)}$
P(UG|DS) = $\frac{0.015 \times 0.8}{P(DS)} = \frac{0.012}{P(DS)}$

P(G|DS) > P(UG|DS) pertanto è più probabile che lo studente sia laureato (= che la sua classe sia G) quando gli attributi D = yes e S = yes.

Esercizio 6

Si consideri la seguente tabella

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

(a) Si stimi la probabilità condizionata per P(A|+), P(B|+), P(C|+), P(A|-), P(B|-), P(C|-)

P(A=0|+) = 2/5 = 0.4	P(A=0|-) = 3/5 = 0.6	P(A=1|+) = 3/5 = 0.6	P(A=1|-) = 2/5 = 0.4
P(B=0|+) = 4/5 = 0.8	P(B=0|-) = 3/5 = 0.6	P(B=1|+) = 1/5 = 0.2	P(B=1|-) = 2/5 = 0.4
P(C=0|+) = 1/5 = 0.2	P(C=0|-) = 0	P(C=1|+) = 4/5 = 0.8	P(C=1|-) = 1

(b) Usare le probabilità condizionate calcolate per predire la classe dell'istanza (A=0, B=1, C=0) usando l'approccio bayesiano naïve.

Sia P(A=0, B=1, C=0) = K

P(+ | A=0, B=1, C=0) = $\frac{P(A=0, B=1, C=0|+) \times P(+)}{K}$

= $\frac{P(A=0|+) \times P(B=1|+) \times P(C=0|+) \times P(+)}{K}$

= 0.4 * 0.2 * 0.2 * 0.5 / K = 0.008/K

P(- | A=0, B=1, C=0) = P(A=0, B=1, C=0 | -) * P(-) / K
= P(A=0|-) * P(B=1|-) * P(C=0|-) * P(-) / K =
= 0.6 * 0.4 * 0 * 0.5 / K = 0/K
Pertanto la classe è +.

Confronto tra modelli

Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1 = 2RP/(R+P)

Esercizio 7

Si consideri un modello M1 che produce la seguente matrice di confusione.

Predicted		+	-
Actual	+	3	2
	-	1	4

Ed un modello M2 che produce la seguente matrice di confusione.

Predicted		+	-
Actual	+	1	4
	-	1	4

Confrontare i due modelli utilizzando le metriche di precision, recall e F1.

Precision(M1) = 3/(3+1) = 75%
Recall (M1) = 3/(3+2) = 60%
F1(M1) = (2*0.75*0.60)/(0.75+0.60) = 0.667

Precision(M2) = 1/(1+1) = 50%
Recall (M2) = 1/(1+4) = 20%
F1(M2) = (2*0.50*0.20)/(0.50+0.20) = 0.2857

Il modello M1 risulta migliore del modello M2.

Esercizio 8

Si richiede di valutare la performance di due modelli di classificazione M1 e M2. Le istanze considerate contendono 26 attributi binari, etichettati da A a Z.
La tabella seguente riporta la probabilità a posteriori (per la classe +) ottenuta applicando i due modelli alle istanze nel test set.

Istanza	True class	P(+ | A, ..., Z, M1)		P(+ | A, ..., Z, M2)
1	+	10	0.73	0.61
2	+	9	0.69	0.03
3	-	4	0.44	0.68
4	-	7	0.55	0.31
5	+	8	0.67	0.45
6	+	6	0.47	0.09
7	-	1	0.08	0.38
8	-	2	0.15	0.05
9	+	5	0.45	0.01
10	-	3	0.35	0.04

(a) Stampare la curva ROC per entrambi i modelli. Quale dei due modelli è il migliore?

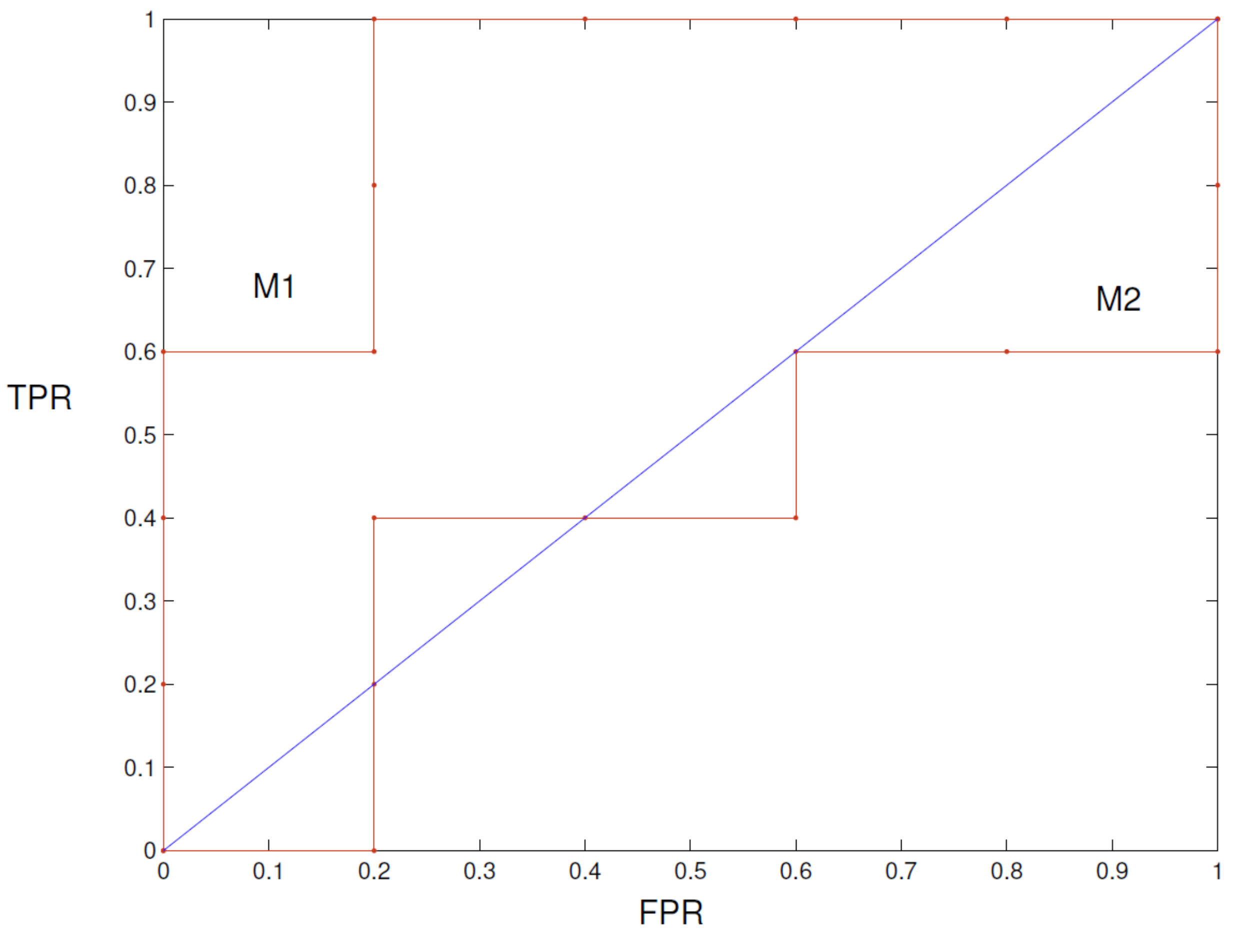
Per stampare la curva ROC bisogna prima ordinare i valori di probabilità ritornati dal modello dal più piccolo al più grande (vedi numero in grigio al fianco del valore), ed usare tali valori come threshold nella seguente tabella, dove si calcolano i valori di TP, FP, TN e FN assumendo che tutte le istanze con valore di probabilità >= threshold siano classificate come +, le altre come -.

La riga "class" indica la classe reale associata al valore di probabilità usato come threshold.

Tabella per M1

Class	-	-	-	-	+	+	-	+	+	+
Th >=	0.08	0.15	0.35	0.44	0.45	0.47	0.55	0.67	0.69	0.73
TP	5	5	5	5	5	4	3	3	2	1
FP	5	4	3	2	1	1	1	0	0	0
TN	0	1	2	3	4	4	4	5	5	5
FN	0	0	0	0	0	1	2	2	3	4
TPR	1	5/5=1	5/5=1	5/5=1	5/5=1	4/5=0.8	3/5=0.6	3/5=0.6	2/5=0.4	1/5=0.2
FPR	1	4/5=0.8	3/5=0.6	2/5=0.4	1/5=0.2	1/5=0.2	1/5=0.2	0	0	0

Ripetere la costruzione della corrispondente tabella anche per M2.



Il modello M1 risulta migliore del modello M2 perché l'area sottesa dalla curva ROC di M1 è maggiore rispetto a quella di M2. Alternativamente: la curva M1 è più vicina all'angolo in alto a sinistra.