Lezione 4

Estensioni: Regole di classificazione

Giovedì, 1 Febbraio 2007

Giuseppe Manco

ICAR CNR
Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Classificatori Rule-Based

- **Si classifica utilizzando una collezione di regole "if…then…"**
- **Regola: (*Condition*) $\rightarrow$ *y***
  - ***Condition* è una congiunzione di attributi**
  - ***y* è l'etichetta di classe**
  - ***LHS*: antecedente**
  - ***RHS*: conseguente**
  - **Esempi:**
    - **(Blood Type=Warm) $\wedge$ (Lay Eggs=Yes) $\rightarrow$ Birds**
    - **(Taxable Income < 50K) $\wedge$ (Refund=Yes) $\rightarrow$ Evade=No**

Un esempio

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds
R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes
R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals
R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles
R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Applicazione di un classificatore rule-based

- **Una regola *r* copre un'istanza x se l'attributo dell'istanza soddisfa la condizione della regola**

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

R1 copre la prima tupla (classificazione: Birds)

R3 Copre la seconda tupla (classificaizone Mammal)

Copertura, accuratezza

- **Copertura di una regola:**
  - **Frazione delle istanze che soddisfano l'antecedente**
- **Accuratezza di una regola:**
  - **Frazione delle istanze che soddisfano antecedente e conseguente**

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	**Single**	125K	**No**
2	No	Married	100K	**No**
3	No	**Single**	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	**Single**	85K	**Yes**
9	No	Married	75K	**No**
10	No	**Single**	90K	**Yes**

(Status=Single) $\rightarrow$ No

Copertura= 40%, Accuratezza = 50%

Utilzzo di un classificatore a regole

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

La prima tupla è coperta da R3, e quindi classificata come mammal

La seconda tupla è coperta da R4 e R5

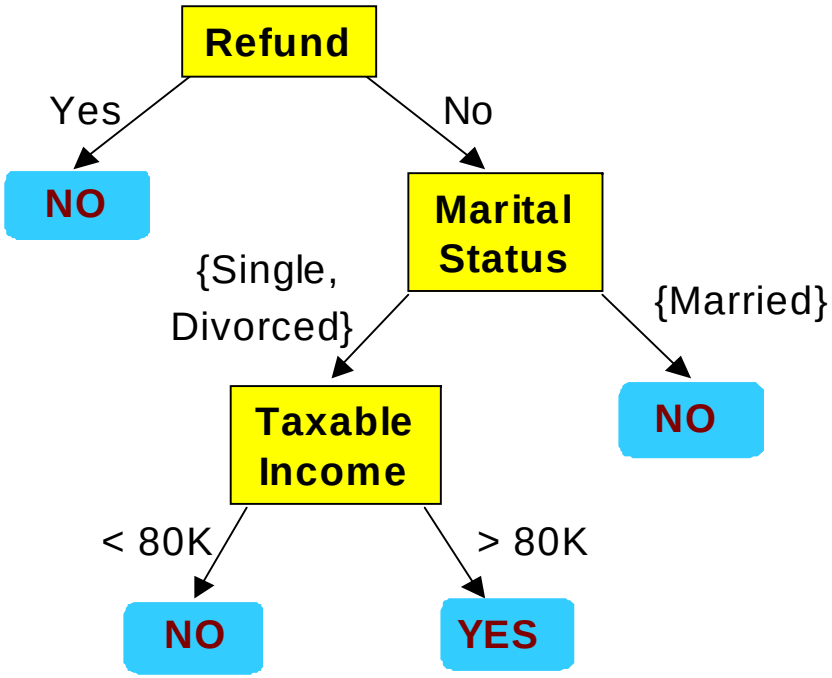
La terza da nessuna

Proprietà di un Rule-Based Classifier

- **Regole mutuamente esclusive**
  - **Le regole sono indipendenti**
  - **Ogni istanza è coperta da almeno una regola**

- **Regole esaustive**
  - **Ogni possibile combinazione di attributi è contemplata**
  - **Ogni istanza è coperta da almeno una regola**

Dagli alberi alle regole



Classification Rules

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

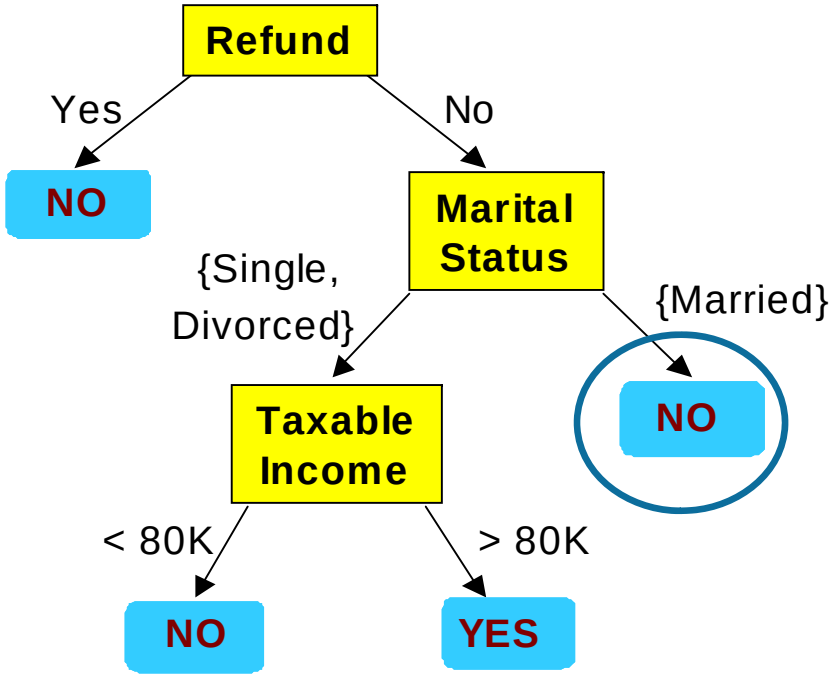
(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

Regole esclusive ed esaustive

Equivalenza

Semplificazione



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Regola iniziale: $(\text{Refund=No}) \wedge (\text{Status=Married}) \rightarrow \text{No}$

Regola semplificata: $(\text{Status=Married}) \rightarrow \text{No}$

Effetti della semplificazione

- **Le regole non sono più esclusive**
  - **Una tupla può essere coperta da più regole**
    - **Ordinamento**
    - **Schema di votazione**
- **Le regole non sono più esaustive**
  - **Una tupla può non essere coperta**
    - **Classe di default**
    - **Non sempre è un problema!**

Schemi di ordinamento

- **Rule-based ordering**
  - **Sulla base della qualità**
- **Class-based ordering**
  - **Sulla base della classe**

Rule-based Ordering

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

Class-based Ordering

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Married}) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

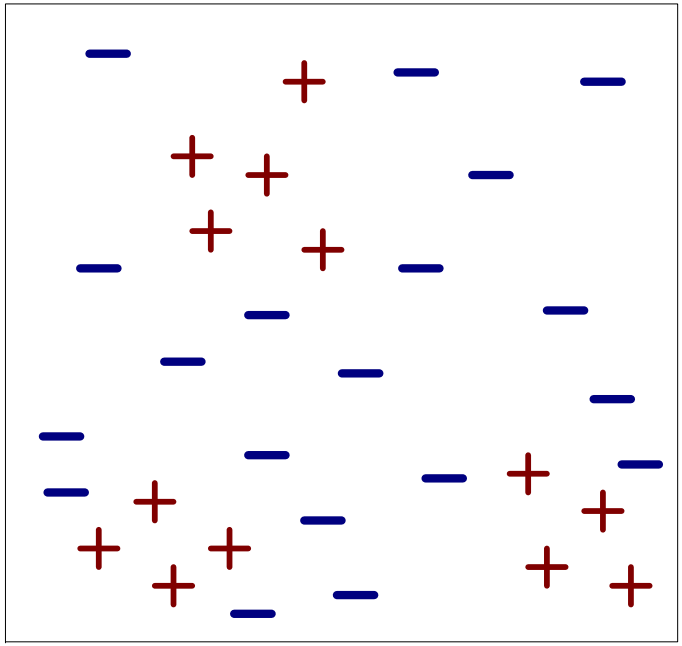
Generazione di classificatori a regole

- **Metodo diretto:**
  - **Estrazione diretta dai dati**
  - **RIPPER, CN2, …**
- **Metodo indiretto:**
  - **Estrazione da altri modelli (Alberi di decisione, reti neurali, …)**
  - **C4.5rules**

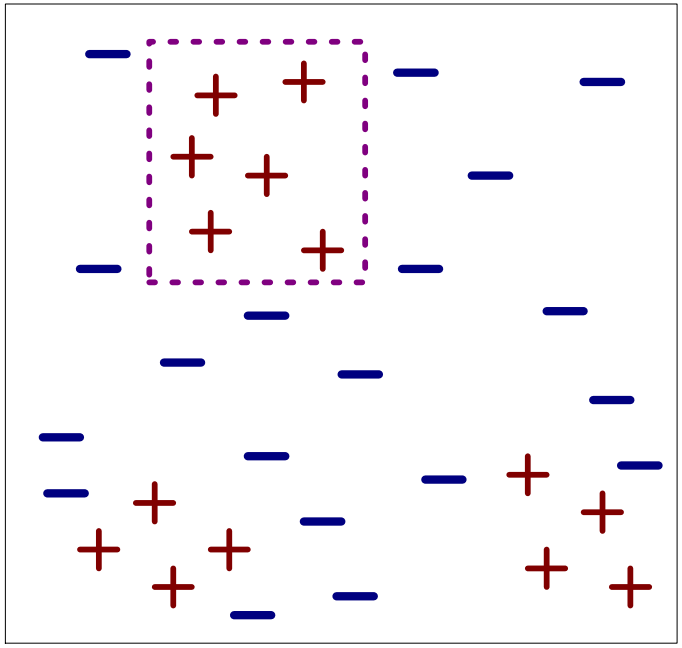
Metodo Diretto: Sequential Covering

1. Si comincia con un insieme vuoto
2. Si genera una regola
3. Si rimuovono le istanze coperte dalla regola
4. Si ripetono i passi (2) e (3) fino a quando il criterio di stop non è soddisfatto

Sequential Covering: Esempio

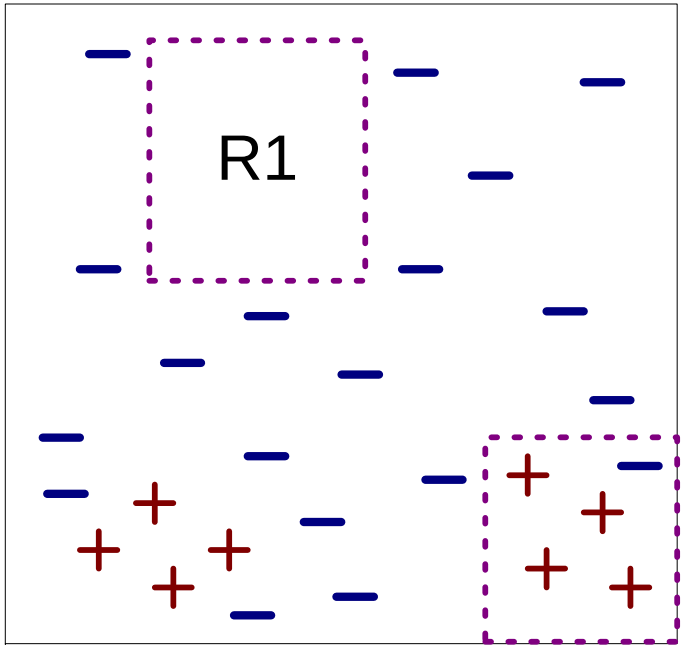


(i) Original Data

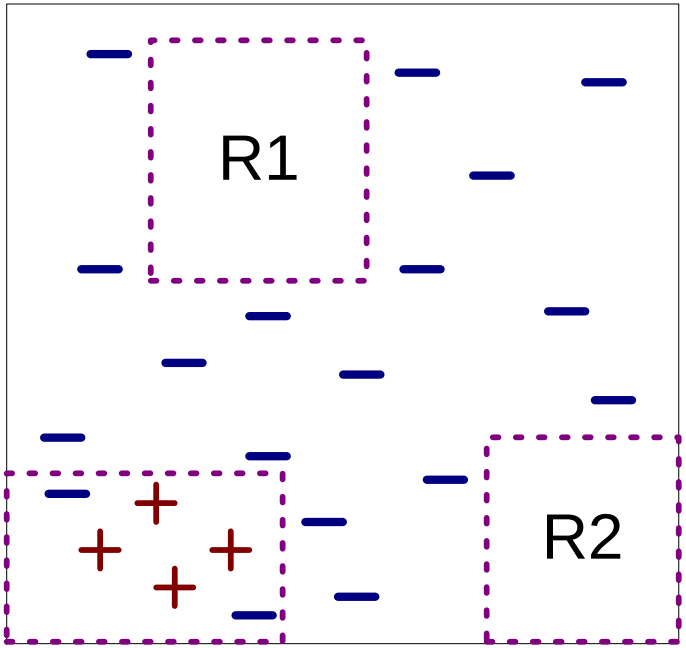


(ii) Step 1

Sequential Covering: Esempio



(iii) Step 2



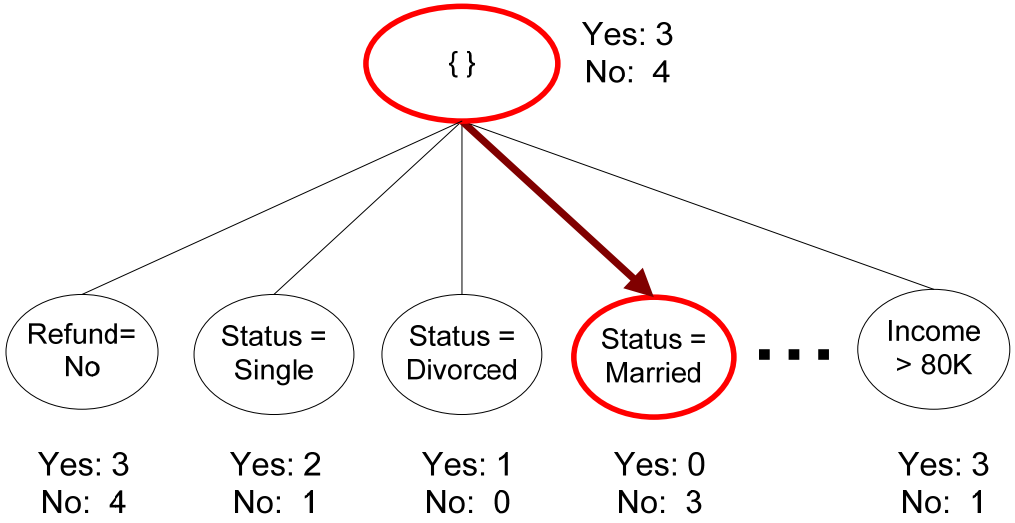
(iv) Step 3

Caratteristiche

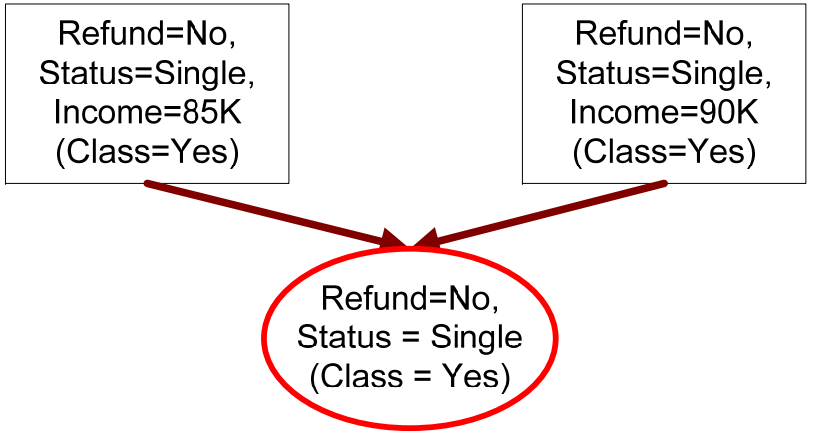
- Generazione di una regola
- Eliminazione delle istanze
- Valutazione della regola
- Criterio di stop
- Pruning

Generazione di una regola

- **Due strategie**



(a) General-to-specific



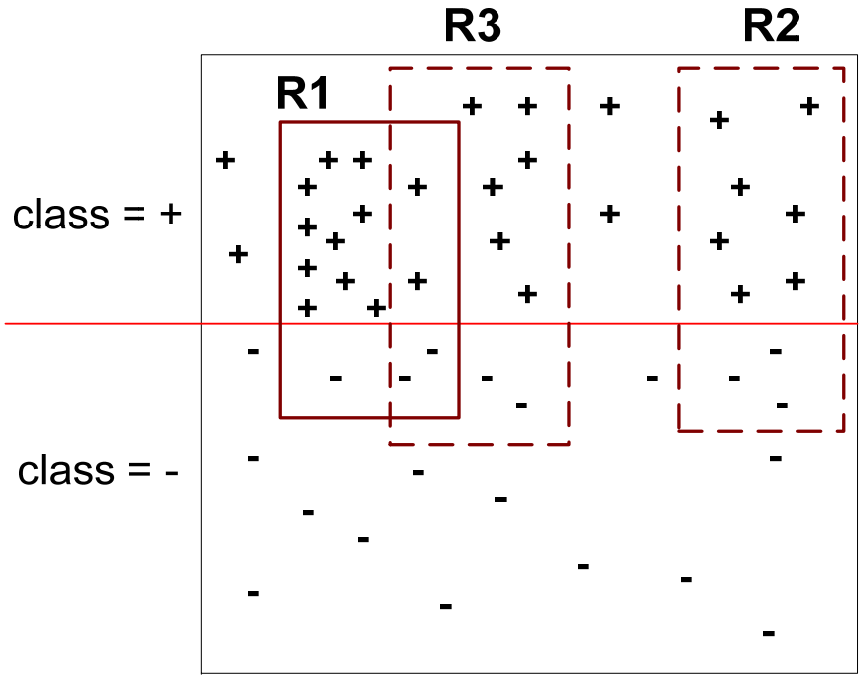
(b) Specific-to-general

generazione (Esempi)

- **CN2:**
  - **Antecedente inizialmente vuoto: {}**
  - **Aggiungi la condizione che minimizza l'entropia: {A}, {A,B}, …**
  - **Il conseguente è la classe di maggioranza delle istanze coperte**
- **RIPPER:**
  - **Regola iniziale: {} => class**
  - **Aggiungi la condizione che massimizza l'information gain:**
    - **R0: {} => class (1)**
    - **R1: {A} => class (2)**
    - **Gain(R0, R1) = t [log (p1/(p1+n1)) – log (p0/(p0 + n0))]**
    - **Dove:**
      - t: numero delle istanze positive coperte da R0 e R1
      - p0: numero di istanze positive coperte da R0
      - n0: numero di istanze negative coperte da R0
      - p1: numero di istanze positive coperte da R1
      - n1: numero di istanze negative coperte da R1

Eliminazione delle istanze

- **Perché l'eliminazione?**
  - **Evitare di generare la stessa regola**
  - **In particolare, rimuoviamo le istanze positive**
- **Vanno rimosse anche le istanze negative?**
  - **SI: evitiamo di sottostimare l'accuratezza**
  - **Confrontate R2 e R3 nel diagramma**



Valutazione di una regola

- **Metriche:**
  - **Accuratezza** $= \frac{n_c}{n}$
  - **Laplace** $= \frac{n_c + 1}{n + k}$
  - **M-estimate** $= \frac{n_c + kp}{n + k}$

n : Numero di istanze coperte dalla regola

n_c *:* Numero di istanze correttamente coperte dalla regola

k : Numero di classi

p : Prior probability

Criteri di stop e di pruning

- **Stop**
  - **Calcoliamo il guadagno**
  - **Se non è significante, eliminiamo la regola**

- **Rule Pruning**
  - **Simile al post-pruning degli alberi di decisione**
  - **Reduced Error Pruning:**
    - **Rimuovi uno dei componenti l'antecedente**
    - **Confronta l'errore con e senza la rimozione**
      - Se l'errore migliora, accetta il taglio