

Classificazione con Weka

Testi degli esercizi

A decorative horizontal bar consisting of a series of colored rectangular segments in shades of blue, teal, yellow, and black, arranged in a slightly wavy pattern across the width of the slide.

Prof. Matteo Golfarelli

Alma Mater Studiorum - Università di Bologna

Pre processing bank data

- Il date set bank-data.csv

- ✓ 600 istanze
- ✓ Nessun dato missing

Attributo	Descrizione
Id	Identificatore unico
Age	età del cliente in anni (numeric)
Sex	MALE / FEMALE
Region	inner_city/rural/suburban/town
Income	reddito del cliente (numeric)
children	sposato? (YES/NO)
car	possiede un'automobile? (YES/NO)
save_acct	ha un conto di risparmio? (YES/NO)
current_acct	ha un conto corrente? (YES/NO)
Mortgage	ha un mutuo (YES/NO)
pep	ha acquistato un PEP (Personal Equity Plan) dopo l'ultimo invio postale? (YES/NO)



Pre processing bank data

1. Caricare il file e salvarlo in formato ARFF con il nome “bank data.arff”
2. Effettuare un’analisi manuale dei dati mediante visualizzazione
 - ✓ Istogrammi attributo – attributo (da pagina Visualize)
 - ✓ Distribuzioni attributo – attributo – classe (da pagina Visualize)
 - Commentare la distribuzione Income –Children - PEP
3. Eliminare l’attributo ID e salvare nuovamente con il nome “bank data.arff”
4. Discretizzare l’attributo AGE (Equal- frequency 10 bin) e Children (Manualmente) e salvare il dataset con il nome “bank data1.arff”



Classificazione bank data

- Utilizzare i seguenti algoritmi per eseguire la classificazione e commentare il risultato
 - ✓ Valutare il risultato con Cross-validation 10 folds / Use training set / Percentage split
 - ✓ Utilizzare sia il data set discretizzato (bank data1.arff), sia quello non discretizzato (bank data.arff)
 1. J48
 - Rappresentare e discutere i decision boundary
 2. J48 senza post-pruning (unpruned = True)
 - Visualizzare l'albero di decisione
 3. JRib
 4. IBk con k=1 e k=5 sul data set non discretizzato (BankData.arff)
 - Dopo aver normalizzato i rimanenti attributi numerici
 5. Ripetere la classificazione utilizzando BankData1.arff dopo aver discretizzato anche l'attributo income (equal frequency 10 bins)
 - Salvare il training set nel file BankData2.arff

Dati di censimento

- Il date set bank-data.csv riporta dati del censimento USA (<http://cps.ipums.org/>)
 - ✓ 1000 istanze per il training
 - ✓ 31561 istanze per la validazione

Attributo	Descrizione
age	Età in anni
workclass	Classe di lavoro
fnlwtg	“Final sampling weight” peso dell’istanza (campione) rispetto alla popolazione
education	Titolo ottenuto
education-num	Numero di anni di studio
marital-status	Stato civile
occupation	Occupazione
relationship	Tipo di relazione con il capo famiglia
race	Razza
sex	Sesso
capital-gain	Utili da capitali (plus valenza)
capital-loss	Perdite da capitali (minus valenza)
hours-per-week	Ore di lavoro settimanali
native-country	Nazionalità
Total Income	L’individuo guadagna più o meno di 50K\$



Dati di censimento

- Obiettivo dello studio è trovare un modello che permetta di predire quali persone guadagnano più di 50K€
 - ✓ Ricerca di frodi fiscali
- Si proceda utilizzando la metodologia CRISP-DM
 1. Comprensione del dominio applicativo
 2. Comprensione dei dati
 3. Preparazione dei dati
 4. Creazione del modello
 5. Valutazione del modello e dei risultati
 6. *Deployment*

Applicazione: marketing lift

- *Una ditta di collocamento possiede una banca dati contenente le informazioni (classe esclusa) relative a 50000 infermiere. Si vogliono contattare tramite posta tutte le infermiere appartenenti alla classe “**priority**”. Tali infermiere saranno tutte inviate ad un insieme di ospedali che hanno fatto richiesta di nuove infermiere.*
- *L’operazione di invio delle lettere ha un costo fisso di 10,000 € e un costo individuale (per ogni infermiera contattata) pari a 5 €.*
- *Gli ospedali prenderanno in prova le infermiere selezionate dalla ditta e alla fine del periodo di prova pagheranno alla ditta 10 € per ogni infermiera che risulterà essere effettivamente una infermiera classificabile come appartenente alla classe “priority”. Decidere quale modello, tra quelli studiati permette di ottenere potenzialmente il profitto maggiore e quante infermiere (in percentuale) devono essere contattate.*
- Verificare quale delle tecniche di classificazione studiate fornisce il modello che fornisce il **lift** maggiore
 - ✓ Dataset di training: [nurseryTrain.arff](#)
 - ✓ Tecnica di validazione 10 folds cross-validation

Applicazione: marketing lift

- Il **lift** è un indicatore che permette di comparare più modelli di classificazione favorendo quelli che permettono di individuare un campione distorto della popolazione che massimizzi la probabilità di trovare istanze della classe desiderata C_i

$$\text{Lift}(\text{ModelloX}) = \frac{P(C_i | \text{Campione})}{P(C_i | \text{Popolazione})}$$

- Molto utilizzato in ambito marketing per selezionare i clienti su cui operare (campagne focalizzate).
 - ✓ La classe desiderata è quella degli utenti che risponderanno positivamente all'attività di marketing
- Calcolare il lift e il guadagno effettivo per J48 JRIB e IB1