Data Mining

Competitions

Aldo Solari

Data sets

Name	Problem	n	m	p	Evaluation
`wine`	Classification	5199	1298	9	Accuracy
`ames`	Regression	1460	1470	82	RMSE(log)
`orange`	Classification	22494	27506	230	AUC

Go to http://www.bee-viva.com/competitions

Deadlines

Name	Days	Start	End
A			
`wine`	18	9/10/17	27/10/2017 h. 12:00
`ames`	18	9/10/17	9/11/2017 h. 12:00
`orange`	27	13/10/17	9/11/2017 h. 12:00
B			
`wine`	10	13/1/18	23/1/2018 h. 12:00
`ames`	10	13/1/18	23/1/2018 h. 12:00
`orange`	10	13/1/18	23/1/2018 h. 12:00
C			
`wine`	10	8/6/2018	18/6/2018 h. 12:00
`ames`	10	8/6/2018	18/6/2018 h. 12:00
`orange`	10	8/6/2018	18/6/2018 h. 12:00

Rules

- You can partecipate in A, B or C (only once)
- Score: 0-10 points
- You must submit valid predictions for all data sets, otherwise your score will be 0
- You must send the R code used for the final predictions, otherwise your score will be 0
- Your score will be valid for the any of the following exams: 13/11/17, 24/1/18, 13/2/18, 19/6/18, 4/7/18, 12/9/18
- For B and C, you must participate individually
- For A, you may participate as a team (max. team size: 3)
- You must register here https://goo.gl/forms/rKchboY4BMoTgECs2

e-mail with R code

Student with badge number 123456

- To: aldo.solari@unimib.it
- Subject: DM competitions + 123456
- Text
  - Name, Surname
  - Badge number
  - Curriculum (SPI, STAT, MAF, OTHER)
  - Exam type (Data Mining M (6CFU) F8204B014, Data Mining (module of Data Science M) F8204B014)
  - Team name (optional)
  - Session (A, B or C)
  - Methods (old and new)
- Attachments:
  - wine_123456.R
  - ames_123456.R
  - orange_123456.R

e-mail with R code

- Inviare un'unica e-mail con il codice R per le tre competizioni al termine delle competizioni (e.g. per A dopo il 9/11/17)
- Se nella previsione finale e' previsto l'utilizzo di metodi visti in corsi precedenti (e.g. principal components analysis), questi vanno specificati in Methods Old: e.g. Methods Old = pca, etc.
- Se nella previsione finale e' previsto l'utilizzo di nuovi metodi mai in corsi precedenti, questi vanno specificati in Methods New: e.g. Methods New = Nonparametric Missing Value Imputation using Random Forest (R package missForest)
- I Methods New non possono essere utilizzati come black-box: mi riservo di fare domande sulla teoria e l'implementazione in R di tali metodi (in occasione della prova d'esame), ed eventualmente di modificare il punteggio ottenuto nella competizione

Calcolo del punteggio

- Il calcolo del punteggio e' funzione dello score s_{min} ottenuto dal modello "benchmark" (identificato dal partecipante solari.aldo) e dallo score s_{max} ottenuto dal modello "best" (miglior modello in classifica finale)
- La percentuale di punti ottenuti da un generico partecipante con punteggio s nella classifica finale viene calcolato come

$$\frac{s - s_{min}}{s_{max} - s_{min}}$$

- Quindi, ad esempio, in una competizione da 4 punti con $s_{min} = 0.6$, $s_{max} = 0.8$ e $s = 0.7$ i punti ottenuti sono pari a $4 \cdot (0.7 - 0.6)/(0.8 - 0.6) = 2$
- Mi riservo la possibilita' di modificare la regola di calcolo del punteggio qualora lo ritengo opportuno