IstatLearning
AULA E ONLINE

Tecniche di Classificazione dei dati

- Analisi in componenti principali
- Analisi Cluster
- Random Forest

Il corso intende illustrare le tecniche di classificazione che vengono utilizzate:

1. per trovare gruppi nei dati, ovvero per prevedere le variabili categoriali target che non vengono misurate (analisi dei cluster).
2. per riepilogare i dati attraverso la riduzione delle dimensioni, eseguita utilizzando l'analisi delle componenti principali.

Supponiamo di voler valutare un tratto non misurabile, come la felicità e che le tue unità di destinazione siano regioni geografiche. La felicità può essere misurata indirettamente attraverso una serie di variabili (questionari, indici, ecc.) il cui valore può essere riassunto in un'unica misura ottenuta riducendo le dimensioni del data-base. L'analisi in componenti principali aiuta ad ottenere la misura ottimale e l'analisi dei cluster separerà le regioni in pochi (due, tre, quattro) gruppi, rispetto ai livelli di felicità. È quindi possibile pianificare diverse politiche per ogni gruppo. L'ultima parte del corso sarà dedicata ai random forest, un modello ensemble, che si avvale del bagging come metodo di ensemble e l'albero decisionale come modello individuale.

Individualmente, le previsioni fatte dagli alberi decisionali potrebbero non essere accurate, ma combinate insieme, le previsioni saranno in media più vicine al risultato.

Data: 23 e 30 marzo,
7 e 14 aprile

Orario: 9.30 – 13.00

Luogo: Aula Virtuale

Docenti:
Barbara Guardabascio, Marco Ballin

Info: Stefania Brandetti

PROGRAMMA

1° Giornata

9.30 – 13.00

Analisi in componenti Principali

- Nozioni preliminari: Autovalori e Autovettori
- Componenti principali: direzione, peso e scala
- Stima
- Interpretazione
- Scelta del numero di componenti
- Rappresentazioni Grafiche: Screen plot e Bi-plot
- *Applicazioni in R*

2° giornata

9.30 – 13.00

Analisi Cluster

- Introduzione
- Cluster non gerarchici:
  K-means e PAM
- *Applicazione in R*

3° giornata

9.30 – 13.00

Cluster gerarchico

- Dendogramma
- *Applicazione in R*
- Misure di validità del cluster
- Coefficiente di Coesione
- Coefficiente Silhouette
- *Applicazione in R*

4° giornata

9.30 – 13.00

Random Forest

- Regression and classification trees
- Random forest
- *Applicazione in R*