Versione: 15 settembre 2025, 20:45

Università degli Studi di Salerno
Laurea Magistrale in Scienze Statistiche per la Finanza
(curriculum Big Data Analytics)

ADVANCED STATISTICAL LEARNING I

Prof. Pietro Coretto
pcoretto@unisa.it

SYLLABUS

Sommario. Il corso si propone di introdurre i metodi di apprendimento statistico automatico (*Machine Learning*) per l'analisi dei grandi data sets (big data) che ormai dominano il settore della finanza. Lo spirito del corso è quello di illustrare algoritmi e metodi che rappresentano lo *state-of-the-art* della letteratura. Il corso si propone di affrontare tre temi principali: metodi ed algoritmi di simulazione; metodi di previsione *ensamble*; metodi di clustering model-based. Le lezioni affronteranno questioni concettuali, metodologiche e computazionali. Gli algoritmi e i metodi oggetto del corso saranno implementati ed illustrati mediante l'uso dell'ambiente di calcolo R.

Indice

1 Organizzazione della didattica

Lezioni. Le lezioni si terranno nel periodo 18 settembre – 15 dicembre 2023 per un totale di 30h corrispondenti a 5cfu. Di seguito si riporta la programmazione settimanale.

```
LUN   10:10-11:50  Laboratorio SP5   [Edificio D2]
GIO   08:30-10:10  Laboratorio SP5   [Edificio D2]
```

Avvisi e ricevimento. Per quanto riguarda gli avvisi ed il ricevimento studenti consultare l'homepage del docente. Il calendario esami sarà disponibile attraverso la bacheca appelli di ateneo.

2 Materiali didattici

Slides, software, scripts, note, esercitazioni e prove d'esame saranno distribuiti durante tutta la durata del corso. La distribuzione del materiale si basa sulla piattaforma di ateneo Google Workspace. L'accesso alle risorse condivise richiede un account *@studenti.unisa.it*.

Durante le lezioni ci sarà un continuo approfondimento delle questioni computazionali attraverso l'uso del linguaggio di programmazione `R`. Il software `R-base` e le librerie aggiuntive necessarie per il corso sono disponibili sotto licenza GPL attraverso il *Comprehensive R Archive Network* (CRAN) per tutti i principali sistemi oparativi.

Il libro di Hastie et al. (2009) è il testo di riferimento del corso. Tuttavia, sarà necessario integrare la bibliografia consigliata con i materiali forniti durante le lezioni. La valutazione del profitto riguarderà solo gli argomenti affrontati al corso, di conseguenza i testi consigliati vanno studiati solo per le parti trattate al corso.

3 Programma dettagliato

Introduzione al corso e richiami di teoria della probabilità. Ruolo dei metodi di learning nelle applicazioni finanziare – breve panoramica sui temi del corso – successioni di variabili casuali – nozioni di convergenza stocastica – relazioni tra regimi di convergenza – teoremi di Mann–Wald e Slutsky – leggi dei grandi numeri (LLN) e teorema centrale del limite (CLT) – teorema di Cramér-Wold e convergenza di vettori casuali – teorema di Glivenko-Cantelli – trasformazione integrale di probabilità (e *Inverse Transform Sampling*) – teorema del previsore ottimale – previsore ottimale nel casi di funzioni di rischio quadratiche, 0-1 loss, etc.

Metodi ed algoritmi di simulazione. Rilevanza del calcolo di momenti e probabilità nelle applicazioni (es: errore di previsioni, probabilità di default, etc.) – distinzione tra approssimazioni analitiche e non – il problema generale dell'approssimazione numerica di integrali – algoritmo "Ordinary Monte Carlo" (OMC) – esempi di utilizzo dell'algoritmo OMC – errore di approssimazione asintotico – stima efficiente del Montecarlo Standard error basato sull'OMC a due stadi – breve panoramica sui metodi di riduzione della varianza (sampling stratificato, importance sampling etc.) – brevi cenni sui metodi Markov Chains – esempi di utilizzo di metodi Monte Carlo per la simulazione di scenari complessi.

Metodi ensemble. Regressione nonparametrica – alberi di regressione (RT, regression-trees) – regressione e classificazione con RT – il problema di stima dei RT – analisi dell'algoritmo CART – adaptive pruning – kFCV per la scelta della complessità ottimale del previsore – problemi dei RT: discontinuità, instabilità e varianza di previsione – bootstrap resampling – bagging: algoritmo e sue varianti – consensus bagging e bagged posteriors – data splitting e correlazione dei previsori – random forest – gestione degli iperparametri nel random forest – CART con dati pesati e boosting – algoritmo AdaBoost – scelta della lunghezza del tree e del *base learaner* – generalizzazioni del boosting – feature importance – vantaggi e limiti dei metodi ensamble.

Model-based clustering. Richiami sul problema della ricerca dei gruppi – modelli di mistura di probabilità – partizionamento basato sull'allocazione MAP – stima ML dei pa-

rametri di un modello di mistura – algoritmo EM – cluster ellittico-simmetrici e modello Gaussiano – algoritmo EM per misture Gaussiane – problema dell'esistenza dello stimatore ML – regolarizzazione delle matrici di varianza-covarianza – parametrizzazione parsimoniose delle matrici di varianza-covarianza – scelta del numero di gruppi e selezione della parametrizzazione – criteri di selezione e validazione: AIC, BIC ed ICL.

4 Esame e valutazione del profitto

L'esame consiste in una prova scritta in laboratorio ed un colloquio orale. La prova scritta è valutata con un massimo di 30 punti allocati su diversi quesiti. Essa si considera superata con punteggio $\geq$ 18pt. Generalmente durante il colloquio orale si procede alla correzione della prova scritta e alla discussione di eventuali punti da chiarire. Tuttavia, coloro i quali conseguono punteggi $\geq$ 27pt alla prova scritta dovranno difendere il voto con un esame orale approfondito.

5 Riferimenti bibliografici

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (Second ed.). Springer Series in Statistics. Springer, New York.