

Università degli Studi di Salerno
Laurea in Statistica per i Big Data

STATISTICAL LEARNING
Prof. Pietro Coretto
pcoretto@unisa.it

SYLLABUS

Sommario. Il corso si propone di introdurre i metodi di apprendimento statistico automatico (*Machine Learning*) con particolare enfasi sulle applicazioni per i grandi data sets moderni (big data). Oggetto di studio sono i principali metodi ed algoritmi di apprendimento supervisionato e non supervisionato (*supervised e unsupervised learning*). 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	1
2	Materiali didattici	2
3	Programma dettagliato	2
4	Esame e valutazione del profitto	4
5	Riferimenti bibliografici	4

1 Organizzazione della didattica

Lezioni. Le lezioni si terranno nel periodo 18 settembre – 15 dicembre 2023 per un totale di 60h corrispondenti a 10cfu. Di seguito si riporta la programmazione settimanale.

MAR	08:30-10:10	Aula 4	[Edificio C]
GIO	11:50-13:30	Aula Informatica	[Edificio D1]
VEN	11:50-13:30	Aula 5	[Edificio C2]

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 operativi.

Il libro di [James et al. \(2021\)](#) è da considerarsi come testo principale del corso. In molti casi faremo anche riferimento al testo di [Hastie et al. \(2009\)](#). 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. È necessario integrare la bibliografia consigliata con i materiali forniti durante le lezioni.

3 Programma dettagliato

Presentazione del corso e nozioni introduttive. Learning – supervised vs. unsupervised learning – batch vs online learning – learning ed ottimizzazione – metodi ed algoritmi – nozione intuitiva di complessità – accuratezza vs interpretabilità dei modelli – training, testing e validation – rappresentazione vettoriale del data set – dati osservati e popolazione – richiami di teoria della probabilità: variabili casuali 0-1, funzione indicatrice, valori attesi di funzioni indicatrici, legge dei grandi numeri per variabili casuali 0-1, valore atteso condizionato e sue proprietà.

Unsupervised learning e classificazione. Problema di classificazione – 0-1 loss, error rate, error rate condizionato – problema di classificazione 2-class: classificatore ottimale, derivazione del classificatore ottimale Bayesiano derivazione e calcolo del Bayes Error – problema multi-class: classificatore Bayesiano ottimale, error rate ed error rate condizionato calcolo del Bayes Error.

Metodi di classificazione basati su modelli di regressione. Modello di regressione logistica per problemi 2-class – classificatore logistico – stima di massima verosimiglianza (ML) – algoritmi ed implementazioni software – parametri e rilevanza delle features – linearità della decision boundary – separabilità delle classi – error rate e limiti della stima in-sample – regressione logistica per problemi multi-class – metodo “*one-vs-all*” – modello multinomial-logit – classificazione multinomial-logit.

Elementi di geometria e probabilità sullo spazio \mathbb{R}^p . Nozione intuitiva di distanza in \mathbb{R} e \mathbb{R}^2 – distanza e metriche in \mathbb{R}^p – distanza di Minkowski e casi particolari (Manhattan, Euclidea, Chebyshev) – intorno e prossimità geometrica di un punto in \mathbb{R}^p – matrice di distanze e sua visualizzazione – momenti di variabili casuali a valori in \mathbb{R}^p – matrice di varianza-covarianza – proprietà della matrice di varianza-covarianza – matrice di correlazione – modelli di covarianza: diagonale, sferico e pieno – stimatori empirici dei momenti – proprietà asintotiche degli stimatori empirici – dispersione lungo una proiezione – misure di dispersione multivariata – variabile casuale Normale multivariata – cenni alle

variabili casuali ellittico-simmetriche – scatter non-sferici e misure di distanza – distanza di Mahalanobis – centralità multivariata nel caso sferico e non sferico.

Metodi di classificazione basati sulle densità di classe. Prior probability e class conditional density – classificatore ottimale Bayesiano – metodi parametrici *vs.* nonparametrici – panoramica sui principali metodi (LDA, QDA, metodi Kernel, e naive Bayes) – differenza concettuale tra metodi logistici e metodi density based – modello con classi Gaussiane omoschedastiche – LDA: classificatore ottimale Bayesiano, linearità della decision boundary, stime ML e loro proprietà asintotiche, garanzie teoriche e performance sperimentali – modello con classi Gaussiane eteroschedastiche – QDA: classificatore ottimale Bayesiano, forma quadratica della decision boundary, stime ML e loro proprietà asintotiche, garanzie teoriche e performance sperimentali – Utilizzo dei metodi LDA e QDA in contesti non Gaussiani – k -Nearest Neighbors (kNN): classificatore kNN, ruolo dell'iperparametro k , non-linearità della decision boundary, garanzie teoriche ed evidenza sperimentale.

Metriche di performance e metodi thresholding. Matrice di confusione nei problemi 2-class – paradosso dell'accuratezza – metriche di performance basate sulla matrice di confusione: FPR , specificity e sensitivity (TPR), FDR , precision, F-score, coefficiente di Matthew – trade-off tra FPR e TPR – thresholding e controllo del FPR - vs - TPR – curve ROC e AUC.

Model selection and model validation. Modelli, complessità, flessibilità ed iperparametri – test error (generalization error), expected prediction error, e training error – decomposizione dell'errore e *bias-variance trade-off* – ottimismo del training error – validazione e selezione del modello – bias e varianza nei tasks di selezione e validazione – cross-validation – Monte Carlo Cross-Validation (MCCV) – k -fold Cross-Validation (kFCV) – Leave-One-Out Cross-Validation (LooCV) – repeated k -fold Cross-Validation (RkFCV) – standard errors (molto) approssimati per la CV – pseudo-intervalli di confidenza approssimati per la CV – trade-off nella scelta di k nella kFCV – garanzie teoriche e performance sperimentale della CV – CV per la selezione del modello – regola di selezione *one-standard-error* – CV per la validazione del modello – selezione e validazione nel caso in cui si dispone di un *external validation set* (EVS) – selezione e validazione mediante nested-CV nel caso in cui non si dispone di un EVS.

Unsupervised learning, riduzione della dimensionalità e data compression. Il problema della riduzione della dimensionalità – differenza tra riduzione e compressione – direzioni e proiezioni nello scatter – definizione delle componenti principali (PC) – PC come soluzione di un problema di ottimizzazione sequenziale – geometria delle PC – decomposizione spettrale della matrice di covarianza – calcolo delle PC basato sullo spettro della matrice di covarianza – proprietà statistiche delle componenti principali – cenni all'analisi delle componenti principali (PCA): correlation circle plots e biplots – Problema generale della compressione dei dati – differenza tra i problemi di compressione e codifica dei dati – compressione *lossy* vs compressione *lossless* – reversibilità della trasformata PC – reconstruction error (Frobenius loss) – *low-rank reconstruction* e teorema di Eckart-Young-Mirsky – immagine (raster) come struttura dati: matrice raster, grayscale e RGB color mode – compressioni delle immagini basata sulle PC

Unsupervised learning e clustering. Il problema generale del clustering – criterio di omogeneità, funzione obiettivo ed algoritmi di partizionamento – affinità/prossimità tra oggetti Euclidei e non-Euclidei – funzioni di similarità e dissimilarità e loro proprietà – dis-

similarità ottenuta da una distanza – altre forme di dissimilarità – matrice di dissimilarità e sua visualizzazione – decomposizione del *total point scatter*. – formulazione generale del problema di partizionamento – k-Means: funzione obiettivo e problema di ottimizzazione, relazione tra funzione obiettivo e centralità/scatter dei dati, soluzioni approssimate, algoritmo di Lloyd e sue proprietà, inizializzazione. – cluster prototypes: centroidi e medoids – k-Medoids: funzione obiettivo e problema di ottimizzazione, soluzioni approssimate, algoritmo PAM e sue proprietà, inizializzazione. – metodi gerarchici agglomerativi e divisivi – dendrogramma e grafo di un algoritmo agglomerativo – taglio ed interpretazione del dendrogramma – linkage e dissimilarità tra gruppi – single, complete, ed average linkage e loro proprietà – effetti *chaining* e *crowding*

Cluster validation. Validazione di una partizione – metriche di confronto tra partizioni: Jaccard index, Rand index, Adjusted Rand Index – silhouette width, average silhouette width (ASW) e silhouette plot – criterio di Caliński-Harabasz – considerazioni finali sulla scelta della partizione

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 ≥ 18 pt. 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 ≥ 27 pt 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.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). *An introduction to statistical learning with applications in R* (Second ed.). Springer Texts in Statistics. Springer, New York.