UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO	Scienze Economiche, Aziendali e Statistiche		
ANNO ACCADEMICO OFFERTA	2024/2025		
ANNO ACCADEMICO EROGAZIONE	2024/2025		
CORSO DILAUREA MAGISTRALE	STATISTICA E DATA SCIENCE		
INSEGNAMENTO	MACHINE LEARNING AND STATISTICAL MODELS C.I.		
CODICE INSEGNAMENTO	23845		
MODULI	Si		
NUMERO DI MODULI	2		
SETTORI SCIENTIFICO-DISCIPLINARI	SECS-S/01		
DOCENTE RESPONSABILE	CHIODI MARCELLO	Professore Ordinario	Univ. di PALERMO
ALTRI DOCENTI	CHIODI MARCELLO SOTTILE GIANLUCA	Professore Ordinario Ricercatore a tempo determinato	Univ. di PALERMO Univ. di PALERMO
CFU	12		
PROPEDEUTICITA'			
MUTUAZIONI			
ANNO DI CORSO	1		
PERIODO DELLE LEZIONI	1° semestre		
MODALITA' DI FREQUENZA	Facoltativa		
TIPO DI VALUTAZIONE	Voto in trentesimi		
ORARIO DI RICEVIMENTO DEGLI STUDENTI	**CHIODI MARCELLO** Martedì 15:00 17:00 stanza del docente (edificio 13); eccezionalmente su teams Venerdì 12:00 13:00 stanza del docente (edificio 13); eccezionalmente su teams **SOTTILE GIANLUCA** Lunedì 10:00 12:00 Ufficio del docente Mercoledì 10:00 12:00 Ufficio del docente		

DOCENTE: Prof. MARCELLO CHIODI

PREREQUISITI	Conoscenza dei fondamenti e metodi dell'inferenza statistica classica (al livello del Corso di Inferenza Statistica STAD) e dell'inferenza sui Modelli Lineari (al livello del Corso di Modelli Lineari STAD); conoscenza dell'ambiente di programmazione statistica R a livello intermedio o altri software scientifici opensource (p.e. Python).
RISULTATI DI APPRENDIMENTO ATTESI	1. Conoscenza e capacita' di comprensione 1.1. Conoscenza dei metodi avanzati dell'inferenza statistica classica (basata sull'approccio di verosimiglianza). 1.2. Conoscenza dei metodi di base dell'inferenza Bayesiana. 1.3. Comprensione delle giustificazioni teoriche nel caso di modelli a più parametri, con approfondimenti teorici e con tecniche di simulazione 2. Capacita' di applicare conoscenza e comprensione 2.1. Capacita' di specificare il modello statistico con un approccio critico, partendo dagli obbiettivi conoscitivi/operativi dello studio esaminato. 2.2. Capacita' di usare in modo integrato le conoscenze acquisite in corsi precedenti per trattare problemi applicativi reali, inclusi problemi non-standard. 2.3. Capacita' di dimostrare risultati teorici in modo formale. 3. Autonomia di giudizio 3.1. Comprensione critica delle caratteristiche, potenzialita' e limiti di modelli statistici gia' conosciuti, e capacita' di arricchirli con estensioni e nuove caratteristiche quando necessario. 4. Abilita' comunicative 4.1. Capacita' di discutere le caratteristiche di un dato problema inferenziale, sia con statistici che con non-statistici. 4.2. Capacita' di scrivere un rapporto tecnico-scientifico, focalizzato sul modello statistico scelto e sull'interpretazione sostantiva dei risultati. 5. Capacita' d'apprendimento 5.1. Capacita' di utilizzare le nozioni e competenze acquisite in successivi corsi di Statistica e Statistica Applicata e nella tesi finale. 5.2. Capacita' di consultare e comprendere la letteratura statistica internazionale, allo scopo di aggiornare le proprie conoscenze teoriche e competenze tecniche.
VALUTAZIONE DELL'APPRENDIMENTO	Prova finale scritta e orale. La prova scritta consiste nell'analisi di un dataset reale in laboratorio informatico, con utilizzo dell'ambiente di programmazione statistica R. Il candidato ha di norma tre ore a disposizione, alla fine delle quali deve consegnare un rapporto tecnico finale. La prova scritta ha come esito solo due possibili risultati: "Ammesso alla prova orale" vs. "Non ammesso alla prova orale". La condizione necessaria per il superamento della prova scritta e' che il candidato dimostri una sufficiente capacita' di: (i) utilizzare in modo autonomo e critico i metodi statistici appresi a lezione per analizzare gli specifici problemi che caratterizzano il dataset proposto; (ii) interpretare i risultati statistici raggiunti; (iii) scrivere in modo efficace un rapporto tecnico-scientifico. La prova orale, cui sono ammessi solo gli studenti che abbiano superato la prova scritta, si articola in due fasi: (i) la discussione del rapporto tecnico finale redatto dal candidato nella prova scritta; (ii) la verifica della conoscenza e capacita' del candidato di illustrare e discutere i principali risultati teorici presentati nelle lezioni frontali. In caso di superamento, il voto finale (espresso nel campo di variazione 18/30 - 30/30, piu' l'eventuale lode) riflettera': (i) il livello mostrato dal candidato, nella prova scritta di laboratorio, di raggiungimento dei "Risultati di apprendimento attesi", con particolare riferimento alle voci sub. 2 e 4.2 fino ad un massimo di 15/30; (ii) il livello mostrato dal candidato, nella prova orale, di raggiungimento dei "Risultati di apprendimento attesi", con particolare riferimento alle voci 1.1, 1.2, 1.3, 2.3, 4.1 (fino ad un massimo di 15/30). Il voto finale sara' ottenuto per somma delle due componenti ora descritte. Per superare l'esame, e ottenere quindi un voto non inferiore a 18/30, lo studente deve dimostrare un livello sufficiente di raggiungimento dei "Risultati di apprendimento attesi" sia nella prova scritta che in quella orale. Per conseguire la valutazione di 30/30, lo studente deve dimostrare un livello ottimo di raggiungimento dei "Risultati di apprendimento attesi" sia nella prova scritta che in quella orale. La lode e' riservata agli studenti che dimostrano una padronanza eccellente dei contenuti del corso ed uno spiccato senso critico nel loro utilizzo.
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali, esercitazioni in laboratorio informatico, analisi di casi di studio reali.

MODULO
STATISTICAL MACHINE LEARNING

Prof. GIANLUCA SOTTILE

TESTI CONSIGLIATI

Lecture notes made available by the professor on the University portal. Online resources indicated by the teacher during the course.
- B. Boehmke, B.M. Greenwell (2020). Hands-On Machine Learning with R, First Edition, CRC Press Taylor & Francis Group (Chap. 1, 2, 4, 5, 8, 9, 10, 11, 12, 13, 14, 17, 20, 21)
- A. Ghatak (2017). Machine Learning with R, Springer (Chap. 1, 3, 4, 5)
- S.V. Burger (2018). Introduction to Machine Learning with R, Publisher(s): O'Reilly Media, Inc. (Chap. 2, 3, 4, 5, 6, 7, 8)
- E. Alpayidin, F. Bach (2014). Introduction to Machine Learning, Fourth Edition, The MIT Press (Chap. 1, 2, 6, 7, 9, 11, 13, 18)
- C. Lesmeister, S.K. Chinnamgari (2019). Advanced Machine Learning with R, Packt Publishing Limited (Chap. 2, 3, 5, 6, 7, 8, 9, 10)
- A. Kassambara (2017). Machine Learning Essentials: Practical Guide in R. STHDA (Chap. 3, 5, 12, 13, 14, 21, 24, 26, 27, 28, 29, 31, 32, , 33, 34, 35)

TIPO DI ATTIVITA'	B
AMBITO	70296-Formazione matematico-statistica
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	108
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	42

OBIETTIVI FORMATIVI DEL MODULO

La materia rappresenta un punto di contatto tra i modelli non lineari e i corrispondenti algoritmi di apprendimento per risolvere fondamentalmente problemi dei seguenti tipi:
- Classificazione;
- Raggruppamento; - Regressione;
- Predizione.

Ognuno di questi problemi richiede un approccio che può essere diverso dal punto di vista dell'apprendimento. Pertanto, il corso inizia rivedendo i concetti e le definizioni di base dei diversi tipi di apprendimento automatico: apprendimento supervisionato, non supervisionato, semi-supervisionato, attivo e di rinforzo. Successivamente, impareremo come valutare i risultati di questi problemi, le diverse metriche esistenti, la necessità di suddividere i set di dati per garantire prestazioni accettabili e i possibili miglioramenti che possono sorgere, come le tecniche di boosting o bagging per generare ensemble. Alla fine, gli studenti saranno in grado di sviluppare il modello di machine learning più appropriato, combinando diverse tipologie di essi e interpretare i risultati della soluzione fornita in un ambiente multidisciplinare.

PROGRAMMA

ORE	Lezioni
2	Argomenti preliminari - Definizioni: campione, modello, variabile, algoritmo, dimensionalità, ... - Standardizzazione e codifica - Selezione delle variabili - Estrazione delle variabili: scomposizione in valori singolari, analisi delle componenti principali, ...
4	Approcci all'apprendimento e problemi connessi - Apprendimento supervisionato: problemi di classificazione; e problemi di regressione e predizione - Apprendimento non supervisionato: K-Means e clustering gerarchico - Approcci semi-supervisionati: apprendimento semi-supervisionato e apprendimento attivo - Apprendimento per rinforzo: problemi di ottimizzazione
4	Valutazione del modello - Divisione del set di dati: K-fold e leave-one-out - Valutare le prestazioni dei modelli di machine learning: problemi di classificazione e regressione - Miglioramento del modello: potenziamento, ensemble e bagging
2	Support Vector Machine - Introduzione - Iperpiano di separazione ottimale - Il trucco del kernel - Supported Vector Regresso (SVR)
2	Alberi decisionali - Rappresentanza - Entropia e guadagno di informazione - Potatura - Alberi di Classificazione e Regressione (CART) - Foresta casuale (RF)
10	Tendenze attuali nell'apprendimento automatico - Gradient Boosting - K-Nearest Neighbours (KNN) - Artificial Neural Networks (ANN) - Deep Learning.

ORE	Esercitazioni

18	Implementazione dei modelli descritti nelle lezioni teoriche: - Pre-elaborazione di insiemi di dati: standardizzazione, codifica, selezione ed estrazione delle variabili; - Valutazione del modello: K-fold, Leave-one-out, ...; - Apprendimento supervisionato: problemi di classificazione e regressione (e.g., linear, logistic, support vector machine, decision tree-based, random forest, gradient boosting...); - Apprendimento non supervisionato: problemi di clustering; - Apprendimento semi-supervisionato: apprendimento attivo e di rinforzo; - Apprendimento profondo e rete neurale artificiale.

MODULO
STATISTICAL MODELS

Prof. MARCELLO CHIODI

TESTI CONSIGLIATI

a)appunti di lezione (lecture notes);
b)Agresti, A., (2015) Foundations of Linear and Generalized Linear Models- Wiley eds.
c) Mc Cullagh, Nelder, (1989) Generalized Linear Models-Chapman and Hall eds.
d) Wood, S. (2006) , Generalized Additive Models_ An Introduction with R- Chapman and Hall
e) Pawitan, Y. (2001) In All Likelihood. Oxford Science Publications, Oxford

TIPO DI ATTIVITA'	B
AMBITO	70296-Formazione matematico-statistica
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	108
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	42

OBIETTIVI FORMATIVI DEL MODULO

Questo corso mira ad arricchire il bagaglio teorico ed applicativo dello studente nella costruzione e interpretazione dei modelli statistici, approfondendo le unita' didattiche: (a) gli sviluppi in ambito di modelli di tipo regressivo non lineare (GLM ed estensioni); (b) Approfondimento di alcuni aspetti dell'inferenza

parametrica classica; (c) cenni all'inferenza Bayesiana; (d) Cenni alla selezione del modello con riferimento alle capacità descrittive e/o predittive. La parte teorica, erogata nelle lezioni frontali, sara' integrata dal punto di vista applicativo nelle esercitazioni in laboratorio, realizzate nell'ambiente statistico R. Dopo aver frequentato questo corso con successo, gli studenti preparati dovrebbero essere capaci di:
(i) specificare un modello statistico appropriato per i dati in esame (GLM o altri modelli), fare inferenza su tale modello e interpretare i risultati;
(ii) riconoscere situazioni in cui e' necessario ricorrere ad una estensione dei GLM standard, e fare inferenza su tali modelli estesi;
(iii) avere un approccio critico al processo di costruzione dei modelli; (iv) sviluppare competenze di base sull' inferenza Bayesiana

PROGRAMMA

ORE	Lezioni
8	(a) Richiami sui modelli lineari, ordinari e generali, predittori lineari e configurazione della matrice del disegno. La distribuzione normale multivariata. Teoria asintotica dell'inferenza su più parametri nel caso regolare
4	Approcci generali all'inferenza: Cenni all'inferenza Bayesiana. Distribuzioni a priori e a posteriori; il ruolo della verosimiglianza. Stima Bayesiana puntuale e intervallare.
12	I modelli lineari generalizzati. Ruoli diversi dei vari elementi: predittore lineare, funzione legame, distribuzione appartenente alla famiglia esponenziale. Metodi numerici per la stima dei parametri (IWLS). Proprietà asintotiche. Residui, strumenti di diagnostica. Confronto fra modelli. Selezione di modelli. Aspetti computazionali ORE

ORE	Esercitazioni
14	Sviluppi in tema di modelli di tipo regressivo: esercitazioni di laboratorio con R. GLM: esempi su varie distribuzioni, casi di studio, software R e package vari. Stima dei parametri, interpretazione dei risultati, confronto fra modelli, diagnostica. Qualche applicazione di tecniche di simulazione
4	Sviluppi in tema di inferenza: esercitazioni di laboratorio con R.