UNIVERSITÀ DI PISA

STATISTICA

MARCO ROMITO

Anno accademico	2019/20
CdS	ARTIFICIAL INTELLIGENCE AND DATA ENGINEERING
Codice	697AA
CFU	6

Moduli	Settore/i	Tipo	Ore	Docente/i
STATISTICA	MAT/06	LEZIONI	60	MARCO ROMITO

Obiettivi di apprendimento

Conoscenze
Al termine del corso lo studente avrà acquisito conoscenze di metodi della statistica multivariata, metodi di classificazione e clustering, analisi di serie storiche, sia da un punto di vista teorico che attraverso l'implementazione con un software statistico.

Modalità di verifica delle conoscenze
Lo studente sarà valutato riguardo la sua abilità di risolvere problemi e discutere concetti e applicazioni di statistica.

Capacità
Al termine del corso

- lo studente sarà in grado di formulare il modello statistico più opportuno per l'analisi quantitativa di un problema industriale,
- lo studente saprà implementare l'analisi formulata per mezzo di un software statistico,
- lo studente sarà in grado di trarre conclusioni e formulare previsioni sul problema industriale esaminato.

Modalità di verifica delle capacità
Analisi e implementazione di modelli statistici saranno il soggetto delle prove d'esame (scritto e orale).
Lo studente potrà preparare un progetto didattico che prevede l'analisi statistica e l'implementazione di un problema a partire da dati reali.

Comportamenti
Il corso permetterà di gestire l'analisi quantitativa di problemi industriali mediante metodi statistici.

Modalità di verifica dei comportamenti
Durante le sessioni di esame, saranno verificate le fasi di analisi statistica di un problema industriale, dal riconoscimento del modello più efficace alla sua implementazione e previsione.

Prerequisiti (conoscenze iniziali)
Ci si aspetta che lo studente conosca i concetti e le idee di base della statistica (basi elementari di probabilità, distribuzioni di probabilità principali, teoria della stima, regioni di fiducia, test statistici).

Indicazioni metodologiche

Il corso prevede lezioni frontali per la parte teorica. La parte implementativa è svolta usando i PC personali degli studenti. Il materiale della parte implementativa è reso disponibile sulla pagina del corso. Il corso prevede la possibilità di svolgere un progetto personale di analisi statistica.

Programma (contenuti dell'insegnamento)
Regressione lineare. Richiami su covarianza e coefficiente di correlazione. Regressione lineare semplice: introduzione del modello, calcolo dei coefficienti nel modello teorico, calcolo dei coefficienti nel caso campionario. Interpretazione del coefficiente di correlazione nel modello di

regressione, varianza spiegata.

Matrici di covarianza e correlazione per vettori aleatori. Simmetria e positività, teorema spettrale per la diagonalizzazione. Richiami di algebra lineare: basi ortonormali, matrici ortogonali, cambio di base. Matrici di covarianza e correlazione di una serie di dati empirici, loro simmetria e positività. Regressione lineare multipla: introduzione del modello, discussione sulla dipendenza causale, scarto quadratico medio. Discussione e descrizione del modello di regressione lineare multipla. Descrizione in termini di un modello teorico di natura probabilistica e calcolo dei coefficienti del modello di regressione nell'ambito del modello teorico. Calcolo dei coefficienti del modello di regressione a partire dai dati empirici: minimizzazione dello scarto quadratico medio. Generica unicità della soluzione trovata. Discussione dei problemi di interpretazione del modello di regressione: overfitting, variabilità statistica dei parametri ottimali, interpretazione e significato dei coefficienti, problemi derivati da differenze di scala, problemi derivati dall'allineamento di fattori, ruolo della varianza spiegata, andamento della varianza spiegata rispetto al numero di fattori e varianza spiegata corretta, p-value sui coefficienti. Discussione sull'opportunità e le modalità di riduzione del modello.

Analisi delle componenti principali. Vettori Gaussiani: vettori gaussiani standard, definizione generale, vettore delle medie, matrice di covarianza, esistenza di un vettore Gaussiano assegnati il vettore delle medie e la matrice di covarianza. Densità di vettori gaussiani non-degeneri, indipendenza e scorrelazione, vettori Gaussiani degeneri. Rappresentazione grafica e interpretazione della covarianza attraverso le curve di livello. Studio delle curve di livello di un vettore Gaussiano.

Analisi delle componenti principali: introduzione al metodo, interpretazione per mezzo di vettori Gaussiani, asse e piano principali. Proiezioni sull'asse come classificazione. Varianza lungo le componenti principali. Interpretazione della varianza delle componenti principali, proporzione di varianza spiegata, valutazione dell'efficacia dell'analisi. Matrice dei loadings.

Classificazione e clustering. Illustrazione per grandi linee dei problemi di classificazione e clustering, differenza tra i due concetti. Regressione lineare multipla applicata alla classificazione. Ponte con la regressione logistica: calcolo di una "probabilità" (tramite la funzione logistica) di classificazione. Classificazione mediante regressione logistica. Cenni a modelli lineari generalizzati e al problema di classificare con più classi. Interpretazione geometrica della classificazione: cenno grafico e concettuale.
Serie storiche. Introduzione alle serie storiche, caratteristiche essenziali della struttura di una serie storica. Funzione di autocorrelazione empirica, interpretazione delle caratteristiche strutturali (trend, stagionalità) in termini della autocorrelazione. Decomposizione di una serie storica: decomposizione additiva o moltiplicativa, medie locali e detrendizzazione, individuazione della componente stagionale, analisi dei residui. Previsione per una serie storica attraverso la decomposizione.

Metodo di smorzamento esponenziale: introduzione alla strategia del metodo, derivazione della formula per ricorrenza, ruolo del parametro, inizializzazione del metodo. Metodo di smorzamento esponenziale con trend: introduzione alla strategia del metodo, derivazione della formula per ricorrenza per intercetta e pendenza, calcolo della previsione, ruolo del parametro, inizializzazione del metodo. Metodo di Holt-Winters: smorzamento esponenziale con trend e stagionalità: introduzione alla strategia del metodo, derivazione della formula per ricorrenza per intercetta, pendenza e stagionalità, calcolo della previsione, ruolo del parametro, inizializzazione del metodo.

Regressione lineare multipla per serie storiche. Discussione delle idee di base. Implementazione elementare del modello, previsione. Funzione di cross-correlazione, fattori esogeni. Approfondimento sul ruolo dei residui: determinare parametri ottimali di un modello, confronto tra modelli, stima dell'incertezza nelle previsioni, analisi per la misura della bontà di un modello.

Parte Implementativa mediante il software R. Introduzione al software R: creazione e manipolazione di vettori, operazioni sui vettori, generazione di sequenze, creazione e manipolazione di matrici, ricerca di autovalori e autovettori, importazione di dati. Introduzione alla rappresentazione grafica dei dati empirici attraverso il software R. Diagrammi di dispersione di matrici di dati, primi comandi statistici relativi a indicatori di centralità e dispersione, e principali distribuzioni (densità, funzione cumulativa, quantili e generazione di numeri casuali). Istogrammi, confronto tra modelli teorici e dati empirici. Covarianza e correlazione.

Implementazione attraverso R di modelli di regressione, calcolo del modello, rappresentazione della retta sovrapposta al diagramma di dispersione, standardizzazione di una tabella, modelli di regressione differenziati per sottogruppi, bande empiriche di confidenza. Confronto con un campione casuale e interpretazione del risultato.
Regressione per lo studio di un indice azionario, confronto con il modello di regressione logaritmica dell'indice, studio dell'andamento della varianza spiegata al variare dei dati. Esempio di regressione polinomiale.

Esempio di regressione multipla, riduzione dei fattori e regressione ai fini della previsione per i dati nella Scheda 4 del libro di testo. Previsione attraverso la regressione multipla per l'esempio su indici azionari, autovalutazione del modello tramite il confronto tra dati noti e previsioni. Analisi di un modello con fattori fortemente allineati, sua riduzione. Esempio di regressione non-lineare.

Vettori Gaussiani nel piano: rappresentazione di vettori casuali con varianze uguali e differenti, trasformazioni (rotazioni). Vettori Gaussiani nello spazio: rappresentazione di vettori casuali con varianze uguali e differenti. Covarianza, diagonalizzazione della covarianza, calcolo di autovalori e autovettori, calcolo della radice quadrata di una matrice simmetrica e semidefinita positiva, generazione di una covarianza casuale.

Analisi delle componenti principali: esempio artificioso con 5 fattori ma di dimensione 2. Analisi delle componenti principali sull'esempio relativo a indicatori economici/sanitari. Analisi delle componenti principali sull'esempio relativo alla produzione agricola, standardizzazione della tabella, confronto delle analisi tra le tabelle standardizzate e non, classificazione attraverso l'asse principale, esplorazione dei piani principali e commenti sulla risoluzione di clustering apparenti. Analisi delle componenti principali dell'esempio relativo alle caratteristiche degli iris, confronto dei piani principali, risoluzione dei cluster.

Rappresentazione grafica di serie storiche. Analisi elementare di serie artificiose (a scopo didattico), funzione di autocorrelazione, decomposizione di serie additive e moltiplicative, funzione di autocorrelazione di campioni aleatori. Analisi delle serie storiche tratte dalle schede 11 e 12 del libro di testo: funzione di autocorrelazione, decomposizione, analisi dei residui, andamento annuale medio e sua incertezza.

Decomposizione con stagionalità non uniforme. Analisi delle serie storiche generate dalla decomposizione, analisi dei residui.

Smorzamento esponenziale di serie create ad-hoc e della serie tratta dalla scheda 11 del libro di testo. Scelta ottimale del parametro e della condizione iniziale. Previsione. Smorzamento esponenziale con trend di serie create ad-hoc e della serie tratta dalla scheda 11 del libro di testo. Scelta ottimale dei parametri e della condizione e pendenza iniziale. Previsione. Previsione mediante lo smorzamento esponenziale e lo smorzamento esponenziale con trend attraverso il trend proveniente dalla decomposizione della serie storica. Serie stagionali con lo smorzamento esponenziale e lo smorzamento esponenziale con trend. Metodo di Holt-Winters per la serie tratta dalla scheda 12 del libro di testo. Previsione con il metodo di Holt-Winters. Confronto tra previsioni. Auto-validazione del modello. Analisi dei residui. Valutazione dell'incertezza nelle previsioni.

Metodi regressivi per serie storiche. Funzione di autoregressione parziale. Riduzione del modello di autoregressione. Previsione attraverso il modello di autoregressione e confronto con la previsione di Holt-Winters. Auto-validazione del modello autoregressivo. Autoregressione con il metodo Yule-Walker per la serie tratta dalla scheda 12 del libro di testo: previsione, sua incertezza e analisi dei residui. Autoregressione con il metodo dei minimi quadrati per la serie tratta dalla scheda 11 del libro di testo: previsione, sua incertezza e analisi dei residui.

Bibliografia e materiale didattico
Note disponibili sulla pagina web del corso

Indicazioni per non frequentanti
La frequentazione del corso è altamente consigliata. Rivolgersi al docente per istruzioni in caso di impossibilità.

Modalità d'esame
L'esame prevede una prova scritta e una prova orale. La prova scritta è finalizzata alla verifica della capacità di formulazione dei modelli statistici e loro implementazione, ed ha durata di due ore. La prova orale è finalizzata alla verifica della conoscenza dei concetti di base del corso, e prevede due o più domande. In alternativa alla prova scritta, durante il corso, lo studente frequentante può svolgere un progetto autonomo in cui si cimenta in una analisi statistica basata su dati reali.
Modalità eccezionali per gli appelli estivi (Giugno e Luglio): la prova scritta viene rimpiazzata dalla preparazione di un progetto didattico di analisi e interpretazione di tabelle di dati, svolto autonomamente dallo studente, analogo alla modalità d'esame che è stata svolta nel primo semestre per gli studenti frequentanti. La consegna del progetto è fissata in corrispondenza del giorno previsto per la prova scritta. I dettagli sulla consegna del progetto sono inviati agli iscritti all'appello.

Pagina web del corso
http://people.dm.unipi.it/romito/Teaching/2020/stat2_ing

Ultimo aggiornamento 19/05/2020 16:37