UNIVERSITÀ DI PISA

STATISTICA II

MARCO ROMITO

Anno accademico	2023/24
CdS	INGEGNERIA GESTIONALE
Codice	750AA
CFU	6

Moduli	Settore/i	Tipo	Ore	Docente/i
STATISTICA II	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 nell'ottica dell'interpretazione di analisi statistica sui dati.

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, formulare previsioni e stimare incertezze sul problema industriale esaminato.

Modalità di verifica delle capacità
Le capacità saranno verificate attraverso l'analisi statistica e implementazione di modelli statistici per lo studio di un problema a partire da dati reali.

Comportamenti
Il corso è finalizzato a fornire gli strumenti per sviluppare l'analisi quantitativa di problemi industriali mediante metodi statistici.

Modalità di verifica dei comportamenti
Durante le sessioni di esame, sarà verificata la padronanza delle diverse fasi dell'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, quali quelle contenute nel corso di Statistica I, e dell'algebra lineare.

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.

Programma (contenuti dell'insegnamento)
Elementi di teoria
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.

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.

Discussione sulla utilità dei residui: identificazione dei parametri ottimali per un modello, confronto di modelli diversi attraverso la varianza spiegata. Analisi quantitativa della varianza e identificazione di eventuale struttura residua. Il metodo della cross-validation per giudicare le capacità predittive dei modelli. Stima dell'incertezza nelle previsioni mediante metodi di bootstrap.

Analisi delle componenti principali
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.
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 loading e interpretazione delle componenti principali, rotazioni delle componenti. Analisi fattoriale: cenni.
Classificazione e clustering
Introduzione al problema della classificazione. Descrizione dell'approccio Bayesiano per la classificazione, regola di appartenenza per mezzo delle probabilità a-posteriori, punto cdi vista geometrico e formulazione della regola di appartenenza in termini geomtetrici. Regressione lineare multipla applicata alla classificazione. Regressione logistica: introduzione al modello, formulazione del problema di verosimiglianza, interpretazione dei risultati. Cenni a modelli lineari generalizzati e al problema di classificare con più classi. Interpretazione geometrica della classificazione: cenno grafico e concettuale. Introduzione alla analisi discriminante. Regola di appartenenza mediante le probabilità a-posteriori e corrispondenti insiemi geometrici. Analisi discriminante quadratica e analisi discriminante lineare.

Introduzione al problema del clustering: ruolo della distanza, similarità. Metodi di clustering per punti prototipo: discussione preliminare sulla scelta del numero di classi, strategia generale basata su centroidi, algoritmo di ottimizzazione per i centroidi. Metodo k-means, discussione sulla convergenza dell'algoritmo iterativo per l'ottimizzazione di k-means. Metodo partition around medoids. Metodi gerarchici: costruzione del dendrogramma, distanze tra cluster: single linkage, complete linkage, average linkage. Confronto con i metodi a punti prototipo. Analisi dell'output di un clustering: silhouette, calcolo della silhouette.

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.
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.
Metodi autoregressivi per serie storiche. Discussione delle idee di base. Implementazione elementare del modello, previsione. Funzione di autocorrelazione parziale, metodi AR (Yule-Walker, minimi quadrati). Funzione di cross-correlazione, fattori esogeni, uso dei fattori esogeni a fini interpretativi.

Parte Implementativa mediante il software R

Introduzione al software R. Strutture e operazioni elementari, importazione di dati. Introduzione alla rappresentazione grafica dei dati empirici. Primi comandi statistici e principali distribuzioni. Campioni casuali.
Implementazione attraverso R di modelli di regressione semplice. Intervalli di confidenza e bande empiriche di confidenza. Predizione, intervalli di predizione. Elementi implementativi relativi all'analisi dei residui. Regressione nonlineare. Esempi di regressione multipla implementati in R, previsione, intervalli di confidenza e predizione. Autovalutazione del modello tramite il confronto tra dati noti e previsioni. Riduzione dei fattori in regressione multipla.
Vettori Gaussiani nel piano e nello spazio. Autovalori, autovettori e radice quadrata della matrice di covarianza. Analisi delle componenti principali, interpretazione, rotazioni.
Classificazione mediante regressione multivariata. Accuratezza, matrice di confusione, robustezza del modello rispetto a informazioni non corrette. Classificazione mediante regressione logistica, problemi di convergenza. Curva ROC, area sotto il grafico (AUC). Analisi discriminante lineare e quadratica,
confronto mediante le curve ROC, l'area sotto la curva ROC e l'autovalidazione. Classificazione multiclasse attraverso l'analisi discriminante lineare e la regressione logistica.
Clustering con i metodi k-means e pam, instabilità di k-means. Valutazione del numero di cluster mediante la varianza entro ogni classe e

l'analisi della silhouette. Clustering gerarchico.
Rappresentazione grafica di serie storiche. Funzione di autocorrelazione. Decomposizione additiva di serie storiche caratterizzate principalmente da trend, da stagionalità. Decomposizione moltiplicativa. Confronto tra decomposizioni attraverso i residui. Decomposizione con stagionalità non uniforme.
Smorzamento esponenziale di serie, ruolo del parametro e sua scelta ottimale, previsione. Smorzamento esponenziale con trend, scelta ottimale dei parametri e delle condizioni iniziali, previsione. Metodo di Holt-Winters, previsione e stima dell'incertezza parametrica e non parametrica. Auto-validazione del modello. Analisi dei residui. Valutazione dell'incertezza nelle previsioni. Metodo di Holt-Winters con trend e stagionalità additiva e moltiplicativa, confronto tra i residui e validazione dei metodi.
Metodi regressivi per serie storiche. Funzione di autocorrelazione parziale. Autocorrelazione e autocorrelazione parziale per serie tipicamente autoregressive. Previsione attraverso il modello di autoregressione e confronto con la previsione di Holt-Winters attraverso analisi dei residui e autovalidazione. Autoregressione con il metodo Yule-Walker, previsione, stima delle incertezza e confronto con Holt-Winters mediante autovalutazione. Autoregressione con il metodo dei minimi quadrati, previsione e confronto con Holt-Winters. Funzione di cross-correlazione. Previsione con fattori esogeni.

Bibliografia e materiale didattico
Note disponibili sulla pagina web del corso

Indicazioni per non frequentanti
La modalità d'esame attraverso la realizzazione di un progetto personale di analisi dei dati è riservata agli studenti frequentanti.

Modalità d'esame
L'esame prevede una prova scritta e una prova orale. La prova scritta consiste nell'analisi di tabelle di dati attraverso gli strumenti appresi durante il corso. 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 lo studente frequentante può svolgere un progetto autonomo di analisi statistica basata su dati reali.

Pagina web del corso
http://people.dm.unipi.it/romito/Teaching/2024/s2

Note
Il corso è mutuato anche dalla LM in Artificial Intelligence and data engineering. Per le studentesse e gli studenti di tale laurea valgono tutte le indicazioni fornite per la LM in Ingegneria Gestionale.
Il corso è mutuato anche dalla LM in Matematica. Per le studentesse e gli studenti di tale laurea valgono tutte le indicazioni fornite per la LM in Ingegneria Gestionale, con l'eccezione delle modalità d'esame. Per tale laurea la prova orale è rimpiazzata dalla presentazione (di circa 30') di un argomento affine/complementare assegnato dal docente.

Ultimo aggiornamento 22/08/2023 12:10