ESE 502

Introduction to Spatial Analysis

Description

The course is designed to introduce students to modern statistical methods for analyzing spatial data. These methods include nearest-neighbor analyses of spatial point patterns, variogram and kriging analyses of continuous spatial data, and autoregression analyses of area data. The underlying statistical theory of each method is developed and illustrated in terms of selected GIS applications. Students are also given some experience with ARCMAP, JMPIN, and MATLAB software.

STAT 503/770

Data Analytics and Statistical Computing

Description

This course will introduce a high-level programming language, called R, that is widely used for statistical data analysis. Using R, we will study and practice the following methodologies: data cleaning, feature extraction; web scrubbing, text analysis; data visualization; fitting statistical models; simulation of probability distributions and statistical models; statistical inference methods that use simulations (bootstrap, permutation tests).

EDUC 691

Core Methods in Educational Data Mining

Description

Students will learn how to execute core educational data mining methods in standard software packages, the limitations of existing implementations of these methods, and when and why to use these methods. The course will also cover how EDM differs from more traditional statistical and psychometric approaches.

ESE 545

Data Mining: Learning from Massive Datasets

Description

Many scientific and commercial applications require us to obtain insights from massive, high-dimensional data sets. In this graduate-level course, students will learn to apply, analyze and evaluate principled, state-of-the-art techniques from statistics, algorithms and discrete and convex optimization for learning from such large data sets. The course both covers theoretical foundations and practical applications.

STAT 571

Modern Data Mining

Description

Modern Data Mining: Statistics or Data Science has been evolving rapidly to keep up with the modern world. While classical multiple regression and logistic regression technique continue to be the major tools we go beyond to include methods built on top of linear models such as LASSO and Ridge regression. Contemporary methods such as KNN (K nearest neighbor), Random Forest, Support Vector Machines, Principal Component Analyses (PCA), the bootstrap and others are also covered. Text mining especially through PCA is another topic of the course. While learning all the techniques, we keep in mind that our goal is to tackle real problems. Not only do we go through a large collection of interesting, challenging real-life data sets but we also learn how to use the free, powerful software "R" in connection with each of the methods exposed in the class.

CPLN 505

Planning by Numbers

Description

This class emphasizes the theory, practice, and use of statistics as applied to planning and policy problems and data. Starting with a review of basic descriptive statistics and measures of association, this course will introduce students to the regression techniques, including multiple regression analysis and logistical and probabilistic models for categorical data; data mining techniques, measures of spatial autocorrelation, and time-series modeling; and causal inference techniques, including structural equation modeling (SEM). A basic familiarity with descriptive and inferential statistics at the upper-division undergraduate level is expected at the beginning of the class. This course uses the popular, free, and open source statistical software R. Meets methods breadth requirement.

CIS 519

Applied Machine Learning

Description

Machine learning has been essential to the success of many recent technologies, including autonomous vehicles, search engines, genomics, automated medical diagnosis, image recognition, and social network analysis, among many others. This course will introduce the fundamental concepts and algorithms that enable computers to learn from experience, with an emphasis on their practical application to real problems. This course will introduce supervised learning (decision trees, logistic regression, support vector machines, Bayesian methods, neural networks and deep learning), unsupervised learning (clustering, dimensionality reduction), and reinforcement learning. Additionally, the course will discuss evaluation methodology and recent applications of machine learning, including large scale learning for big data and network analysis.

CIS 520

Machine Learning

Description

This course covers the foundations of statistical machine learning. The focus is on probabilistic and statistical methods for prediction and clustering in high dimensions. Topics covered include linear and logistic regression, SVMs, PCA and dimensionality reduction, EM and HMMs, and deep learning.

ESE 542

Statistics for Data Science

Description

The course covers the methodological foundations of data science, emphasizing basic concepts in statistics and learning theory, but also modern methodologies. Learning of distributions and their parameters. Testing of multiple hypotheses. Linear and nonlinear regression and prediction. Classification. Uncertainty quantification. Model validation. Clustering. Dimensionality reduction. Probably approximately correct (PAC) learning. Such theoretical concepts are further complemented by exempla r applications, case studies (datasets), and programming exercises (in Python) drawn from electrical engineering, computer science, the life sciences, finance, and social networks.

EPID 600

Data Science for Biomedical Informatics

Description

Data science refers broadly to using statistics and informatics techniques to gain insights from large datasets. Biomedical informatics refers to a range of disciplines that use computational approaches to analyze biomedical data to answer pre-specified questions as well as to discover novel hypotheses. In this course, we will use R and other freely available software to learn fundamental data science applied to a range of biomedical informatics topics, including those making use of health and genomic data. After completing this course, students will be able to retrieve and clean data; perform exploratory analyses, build models to answer scientific questions, and present visually appealing results to accompany data analyses; be familiar with various biomedical data types and resources related to them; and know how to create reproducible and easily shareable results with R and github.

HPR 608

Applied Regression Analysis for Health Policy Research

Description

This course deals with the work-horse of quantitative research in health policy research--the single outcome, multiple predictor regression model. Students will learn how to 1) select an appropriate regression model for a given set of research questions/hypotheses, 2) assess how adequately a given model fits a particular set of observed data, and 3) how to correctly interpret the results from the model fitting procedure. After a brief review of fundamental statistical concepts, we will cover analysis of variance, ordinary least squares, and regression models for categorical outcomes, time to event data, longitudinal and clustered data. We will also introduce the concepts of mediation, interaction, confounding and causal inference.

MSSP 710

Democratizing Data: Analytics for Social Change

Description

With the advent of digital technologies nad the increasing power of computational analytics, the proliferation and ubiquity of data production has increased at exponential rates enabling new possibilites for social analysis. This course will examine the emergence of democratizing data -- the movement to make government and other data more widely or publicly available and its potential enabling for democratic possibilities. The types of data being made available, through various analytic systems, and the ways in which their accessibility and inaccessibility is contributing to reconfigured power relations, will be described. The paradigmatic tensions and shifts that have emerged in the debates on "Big Data," such as deductive versus inductive reasoning and the challenges posed to statistical sampling theory, will be interrogated. The appropriation of machine learning and predictive analytic algorithms for social analysis will be critically explored. Issues related to the ethical and legal use of administrative data, particularly data related to patient, client, student, and taxpayer information will be considered, as well as from internet-based sources including social media. Potential solutions to data security challenges will be additionally considered. Methods for web-scraping of data, analysis of web traffic data, and the use of social networking data in the modeling of social phenomena and public opinion will be examined. Students will learn how to make results accessible to non-technical audiences via data visualization tools, such as web-based data dashboards and web-based maps. These topics will be discussed for the analysis of health, education, and social policy as well as their implications for questions pertaining to race, gender, class, sexuality, dis/abilities, age and youth culture. This course will develop students' knowledge of computational and data analytics and its applications for social policy analysis.

CPLN 691

Data Wrangling and Visualization

Description

The purpose of this course is to familiarize students with the "pipeline" approach to data science. This involves the process of gathering data; sorting the data; analyzing the data and visualizing the data such that non-technical managers can make use of it for decision making. The first part of the course teaches students how to gather data by way of scraping, APIs, Google Big Query, Twitter and other unstructured sources. The second part of this course, teaches students how to store and retrieve these data in a database. The third part of the class teaches some more esoteric machine driven analytics. The fourth and final component of the class is data visualization both in static and dynamic (web-based) form. The students will be expected to replicate this pipeline on a dataset of their own choosing for their final project.

CIS 545

Big Data Analytics

(No Course Description)

BEPP 780

APPLIED DATA ANALYSIS

Description

This course will examine how and when data can be used specifically to infer whether there is a causal relationship between two variables. We will emphasize (a) the critical role of an underlying economic theory of behavior in interpreting data and guiding analysis, as wel as (b) a range of advanced techniques for inferring causality from data, such as randomized controlled trials, regression discontinuity, difference-in-difference, audit study (mystery shopping) approaches and stock-market event studies. The issue of causality, and the relevance of thinking about models and methods for inferring causality, is just as central and important for "Big Data" as it is when working with traditional data sets in business and public policy. The emphasis will not be on proofs and derivations but rather on understanding the underlying concepts, the practical use, implications and limitations of techniques. Students will work intensively with data, drawing from examples in business and public policy, to develop the skills to use data analysis to make better decisions. All analysis will be conducted using R. The goals of the course are for students to become expert consumers able to interpret and evaluate empirical studies as well as expert producers of convincing empirical analysis themselves.

BDS 516

Data Science and Quantitative Modeling

Description

(This course fulfills the MBDS program's quantitative course requirement.) Increasingly, decision-makers and systems rely on intelligent technology to analyze data systematically to improve decision-making. Data science is opening new pathways to improve decision-making in private and public organizations. Through lectures and real-world examples, this course will present a practical understanding of the fundamental methods used by data scientists including data management techniques, quantitative modeling, and data visualization. The primary emphasis is on understanding the fundamental concepts and applications of data science in the context of behavioral and decision sciences. We will cover several algorithms though this is not an algorithms course. We will examine real-world examples and cases to place data science techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. Non-MBDS students may request a permit to register at <https://www.sas.upenn.edu/lps/graduate/mbds/permit-request>.

BSTA 632

Statistical Methods for Categorical and Survival Data

Description

This is the second half of the methods sequence, where the focus shifts to methods for categorical and survival data. Topics in categorical include defining rates; incidence and prevalence; the chi-squared test; Fisher's exact test and its extension; relative risk and odds-ratio; sensitivity; specificity; predictive values; logistic regression with goodness of fit tests; ROC curves; the Mantel-Haenszel test; McNemar's test; the Poisson model; and the Kappa statistic. Survival analysis will include defining the survival curve, censoring, and the hazard function; the Kaplan-Meier estimate, Greenwood's formula and confidence bands; the log rank test; and Cox's proportional hazards regression model. Examples of medical and biologic data will be used throughout the course, and use of computer software demonstrated.

BSTA 656

Longitudinal Data Analysis

Description

This course covers both the applied aspects and methods developments in longitudinal data analysis. In the first part, we review the properties of the multivariate normal distribution and cover basic methods in longitudinal data analysis, such as exploratory data analysis, two-stage analysis and mixed-effects models. Focus is on the linear mixed-effects models, where we cover restricted maximum likelihood estimation, estimation and inference for fixed and random effects and models for serial correlations. We will also cover Bayesian inference for linear mixed-effects models. The second part covers advanced topics, including nonlinear mixed-effects models, GEE, generalized linear mixed-effects models, nonparametric longitudinal models, functional mixed-effects models, and joint modeling of longitudinal data and the dropout mechanism.

ECON 705

Econometrics I: Fundamentals

Description

Violations of classical linear regresson assumptions, nonlinear regression models (including logit, probit, etc.), diagnostic testing, distributed lag models, panel data models, identification, linear simultaneous-equations model.

STAT 501

Introduction to Nonparametric Methods and Log-linear Models

Description

An applied graduate level course for students who have completed an undergraduate course in basic statistical methods. Covers two unrelated topics: loglinear and logit models for discrete data and nonparametric methods for nonnormal data. Emphasis is on practical methods of data analysis and their interpretation. Primarily for doctoral students in the managerial, behavioral, social and health sciences. May be taken before [STAT 500](https://catalog.upenn.edu/search/?p=STAT%20500) with permission of instructor.

STAT 511

Statistical Inference

Description

Graphical displays; one- and two-sample confidence intervals; one- and two-sample hypothesis tests; one- and two-way ANOVA; simple and multiple linear least-squares regression; nonlinear regression; variable selection; logistic regression; categorical data analysis; goodness-of-fit tests. A methodology course.

STAT 590

Causal Inference

Description

Questions about cause are at the heart of many everyday decisions and public policies. Does eating an egg every day cause people to live longer or shorter or have no effect? Do gun control laws cause more or less murders or have no effect? Causal inference is the subfield of statistics that considers how we should make inferences about such questions. This course will cover the key concepts and methods of causal inference rigorously.

STAT 724

Text Analytics

Description

This course introduces methods for the analysis of unstructured data, focusing on statistical models for text. Techniques include those for sentiment analysis, topic models, and predictive analytics. Course includes topics from natural language processing (NLP), such as identifying parts of speech, parsing sentences (e.g., subject and predicate), and named entity recognition (people and places). Unsupervised techniques suited to feature creation provide variables suited to traditional statistical models (regression) and more recent approaches (regression trees). Examples that span the course illustrate the success of text analytics. Hierarchical generating models often associated with nonparametric Bayesian analysis supply theoretical foundations.