Official Statistics

Efrat Avraham and Moshe Pollak, Central Bureau of Statistics

בסיס נתונים מרחבי בנושא שימושי קרקע 2014

מידע על מערך שימושי הקרקע חיוני לצורך תכנון וקביעת מדיניות במגוון תחומים. בשנת 2014 יוצר בתחום ממ"ג-גאוגרפיה בלשכה המרכזית לסטטיסטיקה בסיס נתונים מרחבי בנושא שימושי קרקע. נבנתה רשת ארצית רציפה של תאי שטח שגודלם 100×100 מטרים עם הערך של שימוש הקרקע העיקרי בכל תא שטח. קיימים בשכבה 37 שימושי קרקע וכיסויי קרקע.

מטרת הפרויקט: השלמת חסרים בסטטיסטיקה הרשמית של מדינת ישראל בנושא קרקע, על ידי יצירת סדרה עיתית סטטיסטית בנושא שימושי הקרקע.

הפרויקט מספק מידע עדכני ורציף על שימושי הקרקע במדינה וכן על שימושי הקרקע בעיריות ובמועצות המקומיות ביהודה והשומרון ומאפשר הפקת נתונים עבור כל תיחום גאוגרפי.

שימושי הקרקע מופו בהתאם לתקן מערך קודים ארצי אחיד למיפוי תכסיות ושימושי קרקע, שהוכן במסגרת הוועדה הבין-משרדית למערכות מידע גאוגרפיות, בהתאם לתקן של האו"ם לסיווג תכסית.

המידע מעודכן לסוף שנת 2013.

מקורות המידע: מידע ממרשמים מנהליים מעוגנים (מרשם דירות ומבנים, מרשם עסקים ומרשם התושבים), שכבות מידע גאוגרפיות, שהתקבלו ממשרדים ממשלתיים, מארגונים לא ממשלתיים ומגופים פרטיים ומידע מאורתופוטו לשם טיוב שכבת היערות.

שיטת העבודה:

1. סיווג שכבת המבנים הארצית לפי שימושי קרקע, על סמך המרשמים המנהליים המעוגנים ועל סמך מגוון שכבות מידע גאוגרפיות.
2. יצירת שכבה עבור כל אחד מ-37 שימושי הקרקע על סמך שכבת המבנים המסווגת, בתוספת מעטפת ברוחב 20 מטרים ועל סמך שכבות מידע גאוגרפיות נוספות.
3. שילוב השכבות לשכבה אחת לפי סדר היררכי, שנקבע בין שימושי הקרקע.
4. טיוב שכבת היערות באמצעות מיון מבוקר של אורתופוטו בשיטה מתחום חישה מרחוק.
5. התמרה של השכבה הוקטורית שנוצרה לשכבה רסטרית 0.5 מטרים לפיקסל.
6. יצירת רשת תאים וקטורית (fishnet) בגודל תא 100 על 100 מטרים.
7. מתן ערך של שימוש הקרקע העיקרי לכל תא לפי השכבה הרסטרית.

התוצר הסופי כולל שכבה רסטרית של תאים ושכבה וקטורית נקודתית.

בהרצאה יוצגו בסיס הנתונים והחידושים בפרויקט

Prof. Avi Simhon, The National Economic Council

No title & no abstract

Dr. Hagit Glickman, The National Authority for Measurement and Evaluation in Education (RAMA)

No title & no abstract

Dr. Besora Regev, Israel Police

מה בין מקום מגורי העבריין לבין מקום ביצוע העבירה?

הניתוח נשען על נתוני סטטיסטיקה פלילית בין השנים 2014-18 כאשר מטרתו לבחון את הקשר בין מקום מגורי העבריין לטווח הגיאוגרפי של מקום ביצוע עבירות של פשיעת רכוש. הגדרת אשכולות של עבריינים בהתאם למקום ביצוע העבירה. פלטפורמה לקבלת החלטות.

Theoretical Data Science

Yuval Benjamini, The Hebrew University

The Accuracy of Multi-class Classification

The difficulty of multi-class classification generally increases with the number of classes. This raises a natural question: Using data from a subset of the classes, can we predict how well a classifier will scale as the number of classes increases? In other words, how should we extrapolate the accuracy for small pilot studies to larger problems?

In this talk, I will present a framework that allows us to analyze this question. Assuming classes are sampled from a population (and some assumptions about the classifiers), we can identify how expected classification accuracy depends on the number of classes (k) via a specific cumulative distribution function. I will present a non-parametric method for estimating this function, which allows extrapolation to K>k. I will show relations with the ROC Finally, I hope to discuss why the extrapolation problem may be important for neuroscientists, who are increasingly using mutliclass extrapolation accuracy as a proxy for richness of representation.

This is joint work with Charles Zheng and Rakesh Achanta

David Azriel, Technion

Semi-supervised linear regression

We study a regression problem where for some part of the data we observe both the label variable Y and the predictors X, while for other part of the data only the predictors are given. Such a problem arises, for example, when observations of the label variable are costly and may require a skilled human agent. If the conditional expectation E[Y |X] is exactly linear in X then typically the additional observations of the X's do not contain useful information, but otherwise the unlabeled data can be informative. In this case, our aim is at constructing the best linear predictor. We suggest improved alternative estimates to the naive standard procedures that depend only on the labeled data. Our estimation method can be easily implemented and has simply described asymptotic properties. The new estimates asymptotically dominate the usual standard procedures under certain non-linearity condition of E[Y | X]; otherwise, they are asymptotically equivalent. The performance of the new estimator for small sample size is investigated in an extensive simulation study. A real data example of inferring homeless population is used to illustrate the new methodology.

Joint work with Larry Brown, Michael Sklar, Richard Berk, Andreas Buja and Linda Zhao

Daniel Nevo, Tel-Aviv University

LAGO: The adaptive Learn-As-you-GO design for multi-stage intervention studies

In large-scale public-health intervention studies, the intervention is a package consisting of multiple components. The intervention package is chosen in a small pilot study and then implemented in large-scale setup. However, for various reasons I will discuss, this approach can lead to an implementation failure.

In this talk, I will present a new design, called the learn-as-you-go (LAGO) adaptive design. In the LAGO design, the intervention package is adapted in stages during the study based on past outcomes. Typically, an effective intervention package is sought, while minimizing cost. The main complication when analyzing data from LAGO is that interventions in later stages depend upon the outcomes in the previous stages. I will present asymptotic theory for LAGO studies and tools that can be used by researchers in practice. The LAGO design will be illustrated via application to the BetterBirth Study, which aimed to improve maternal and neonatal outcomes in India.

Meng Xu, University of Haifa

Generalized test-retest reliability based on distances

The intraclass correlation coefficient (ICC) is a classical measure of test-retest reliability. With the advent of new and complex types of data for which the ICC is not defined, there is a need for new ways to assess measurement reliability. To meet this need, we propose a new distance-based intraclass correlation coefficient (dbICC), defined in terms of arbitrary distances among observations. The Spearman-Brown formula, which shows how more intensive measurement increases reliability, is extended to encompass the dbICC. We introduce a bias correction to improve the coverage of bootstrap confidence intervals for the dbICC, and demonstrate its efficacy via simulation. As an illustration, we analyze the reliability of brain connectivity networks derived from the Human Connectome Project database.

Applied Data Science

Yedid Hoshen, Facebook research

Image and language translation without the Rosetta stone

It has been assumed that translating between languages as well as matching between images and text requires some grounding, typically in the form of supervised pairs. In this talk, we will discuss two recent works (joint with Lior Wolf), which suggest that supervision is not required for such translation tasks. We will first show how intuition from point cloud matching gives rise to a novel algorithm able to translate words across languages without the use of supervision, adversarial training or deep networks. A new unsupervised form of CCA, trained using adversarial method will then be presented. The new method, Unsupervised Correlation Analysis achieves a surprising degree of success on the challenging task of unsupervised matching between sentences and images.

Alex Zhicharevich, Intuit

Structuring financial knowledge bases with deep NLP methods

Intuit Inc. is a large software company that develops financial, accounting and tax preparation software for small businesses, accountants, and individuals. Due to the complexity of the tax laws, Intuit develops and maintains a Q&A platform for supporting its users. As the popularity of the products grow, these platforms evolved to big knowledge repositories with large amount of textual data. In this talk I will describe two projects aim to improve the effectiveness of these Q&A platforms. The first project uses deep learning methods for detecting semantically similar or duplicate questions. The second project is a system for key-phrase extraction from questions. In the talk I'll focus on the approaches we used to leverage deep learning without having large labeled datasets by exploiting additional data like search logs. I'll also discuss the advantages that DL present over traditional NLP methods for these two problems.

Asaf Noy, Alibaba

AutoML - Towards "CV as a Service"

In this talk I will review the field of AutoML, with emphasis on recent Neural Architecture Search (NAS) methods which applied continuous relaxations over the architecture space, enabling efficient search with gradient-based optimization algorithms.
Then I will present an online optimization method for differential architecture representations, based on prediction with expert advice theory, aimed to minimize the regret caused by suboptimal operations selection. The derived algorithm, Exponential Gradient Wipeout (EGW), is designed to dynamically enhance superior architectures and wipe-out inferior ones, thus smoothing the final, harsh, architecture prune step common to previous relaxation methods, while reducing both network complexity and run-time.
I will conclude by describing our goal: 'Computer-Vision as a Service'.

Shlomo Ahal, Istra research

Solving assignment problem in real time

Optimal ordering of tasks whose value is rank dependent is an assignment problem for which many solvers exist. However, in real time applications such as high frequency trading, one may need to solve this problem under real time constraints. Furthermore, in most real-life scenarios, simplifying assumptions on the data can be made. We will show a fast approximation algorithm that works for wide set of value functions and can be used in real time application.

Privacy

Yosi Rinott, The Hebrew University
Privacy in data dissemination, general and some technical background

I will demonstrate privacy issues that arise when an agency such as a bureau of statistics or a hospital disseminates data such as a sample from some population to the public or to other agencies. Various methods used by statisticians to assess the disclosure risk, and to decrease it will be reviewed (e.g., Dalenius 1977).
In general, such methods depend on scenarios regarding potential intruders, such as the intruder's prior knowledge about the sample or the population. Differential Privacy (Dwork, McSherry, Nissim and Smith2006) is an approach that avoids the need to consider such scenarios, and guarantees a well-defined notion of privacy by adding noise to released data. I will describe some basic results on differential privacy with some discussion of its application to the release of contingency tables (Dwork and Roth 2014, Rinott, O'Keefe, Shlomo, and Skinner 2018). This talk will serve as background to the two following talks in the session

Uri Stemmer, Ben-Gurion University
Local Differential Privacy

In this talk I will present the local model of differential privacy (Dwork et al. 2006, and Kasiviswanathan et al. 2008). In this model there are n users and an untrusted server. Each user is holding a private input item, and the serve's goal is to compute some function of the inputs. However, in this model, the users do not send their data as is to the server. Instead, every user randomizes her data locally, and sends a noisy report to the server, who aggregates all the reports. This is one of the models used in practice by Apple, Google, and Microsoft to ensure that private data never reaches their servers in the clear.

Moshe Shenfeld, The Hebrew University
Differential Privacy as Stability

In this talk I will present a different and somewhat surprising usage of Differential Privacy as a stability notion. A series of papers released in the last 10-15 years, raised the concern that adaptive selection of computations is eroding statistical validity of scientific findings, since the classical concentration bounds hold only for functions which were chosen independently of the sample set. As it turns out, ensuring the privacy of the sample set (regardless of the sensitivity of the information it contains), can guarantee generalization - prevents over-fitting.

Statistical Analysis of Big Data

Tamir Hazan, Technion

Learning with Perturbation Models

Predictions in modern statistical inference problems can be increasingly understood in terms of discrete structures such as shortest paths in sequential decision making, trajectories in reinforcement learning or parses in natural language processing. In a fully probabilistic treatment, all possible alternative assignments are considered, thus requiring to estimate exponentially many structures with their respective weights. However, sampling from traditional structured probabilistic models such as the Gibbs distribution is computationally expensive for many artificial intelligence applications.

Machine learning algorithms often randomly perturb the learned parameters to account for uncertainties in the prediction process. This gives rise to new probability models, which we call perturbation models, that allow for efficient reasoning in various machine learning applications. In this talk I will present their statistical properties and their application.

Ofer Lavi, IBM

Methods for evaluating model performance on future data under potential drift

Consider a situation where data arrive in sequence. We have some initial labeled dataset, called a baseline, on which we train some model. Unlabeled data ("production") that arrives in the future, may drift in distribution relative to the baseline data the model was trained on, which may cause the model's performance to differ from what was expected given only the baseline. We wish to detect changes in the production data relative to the baseline, while making few parametric assumptions about the data and maintaining statistical guarantees on detection error rates. We evaluate our method's success in detection of drift by simulating data with different types of drift, such as including classes not seen in the baseline, noise, or unbalanced class distribution. Alternatively, we may want to use outputs of the model itself, such as prediction confidence, to identify possible drift, or predict the model's performance (e.g., MSE, classification accuracy) on production data

Rami Yaari, University of Haifa and Gertner Institute

Scalable detection of rare classes in big data

We discuss the problem of generating the minority-class rules from imbalanced data, a scenario that appears in many real-life domains such as medical applications, failure prediction, network and cyber security, and maintenance. We present the Minority-Report Algorithm that uses multitude-targeted mining for boosting performance. We provide complexity analysis of the Minority-Report Algorithm that corresponds to the statistical characteristics of the data, and demonstrate its performance gain using simulations and real data. The Minority-Report uses the GFP-growth (Guided FP-growth) algorithm, a novel method also developed as part of this work, for multitude-targeted mining: finding the count of a given large list of itemsets in large data. The GFP-growth algorithm is designed to focus on the specific multitude itemsets of interest and optimizes the time and memory costs.

Joint work with Lior Shabtay and Itai Dattner

Ron Sarafian, Ben-Gurion University

Gaussian Markov Random Fields for big-scale spatio-temporal data

One of the most fundamental statistical model for space-time process over continuous domains is the Gaussian Random Field (GRF). Although it has good analytic properties, GRF is computationally hard to fit, hence, infeasible for many real-world datasets. Recent advances in the spatio-temporal statistical literature propose to alleviate the computation burden of GRFs by approximating them with Gaussian Markov Random Fields (GMRFs). GMRF is a powerful approach for learning big-scale spatio-temporal data. It allows fitting a GRF with a continuously and smoothly decaying covariance function, while computations are performed with the sparse precision matrix of a Markovian process

Prediction and uncertainty quantification - In collaboration with EMR-IBS

Giles Hooker, Cornell University

Decision Trees and CLT's: Inference and Machine Learning

This talk develops methods of statistical inference based around the popular machine learning methods of bagging and Random Forests. We show that when the bootstrap procedure in ensemble methods is replaced by sub-sampling, predictions from these methods can be analyzed using the theory of U-statistics which have a limiting normal distribution. Moreover, the limiting variance that can be estimated within the sub-sampling structure. Using this result, we can compare the predictions made by a model learned with a feature of interest, to those made by a model learned without it and ask whether the differences between these could have arisen by chance.

By evaluating the model at a structured set of points we can also ask whether it differs significantly from an additive model. We demonstrate these results in an application to citizen-science data collected by Cornell's Laboratory of Ornithology. Given time, extensions to gradient boosting will be discussed.

Or Zuk, The Hebrew University

Deep learning methods for predicting gene regulation in multiple species

Prediction of gene function and activity from DNA sequence are fundamental problems in biological gene regulation and our performance on on this problems represent our understanding of gene regulatory networks. We developed and studied novel deep-learning based methods for predicting enhancers from DNA sequence in multiple related species. Enhancer regulate gene expression from afar by providing a binding platform for transcription factors, often in a tissue-specific or context-specific manner. Despite their importance, our understanding of these DNA sequences, and their regulatory grammar, is limited. We trained deep Convolutional Neural Networks (CNNs) to identify enhancer from DNA sequences in multiple species, using in vivo binding data of single transcription factors and genome-wide chromatin maps of active enhancers in 17 mammalian species.
We obtained high classification accuracy by training enhancers vs. both non-enhancer genomic background sequences, and adversarial k-order random shuffles of enhancer sequences. By interpretation of the learned parameters vs. learned parameters in a randomized background distribution, the combined training strategy also allowed our networks to identify biologically meaningful motifs, unique to enhancers.
In addition, our learned networks were transferable between different species, showing a shared mammalian regulatory architecture.

Joint work with Dikla Cohn and Tommy Kaplan

Jonathan Yefenof, The Hebrew University

Confidence intervals for the test error in a general kernel machine classification

In statistical learning, the successes of classifiers is commonly measured by the test error which is misclassification probability. Therefore, it is of an interest to construct a high quality estimator for the test error. In this paper we consider the test error of general kernel-machine classifiers. Inference for kernel-machine classifiers, and more specifically, for the test error of these classifiers, is difficult since even the rate of convergence might be unclear. We propose confidence intervals which are asymptotically correct. The proposed confidence intervals are constructed by two different approaches. The first approach is based on converging to a normal distribution approximation and the second is based on empirical bootstrap.

Assaf Rabinowicz, Tel-Aviv University

Cross-validation for Correlated Data

K-fold cross-validation (CV) with squared error loss is widely used for evaluating predictive models, especially when strong distributional data assumptions cannot be taken. However, CV with squared error loss is not free from distributional assumptions, especially in cases involving non-i.i.d data. This talk analyzes CV for correlated data. We present a criterion for suitability of CV, and introduce a bias corrected cross-validation prediction error estimator, CVc, which is suitable in many settings involving correlated data, where CV is invalid. Our theoretical results are also demonstrated numerically.