

معرفی کتاب

در این نشریه کتابی در زمینه یادگیری آماری معرفی میکنیم. یادگیری آماری، به مجموعه ای از ابزارهای مدل سازی و درک داده های با ساختار پیچیده گفته می شود. در این زمینه به تازگی توسعه یافته، ترکیبی است از روش های آماری (مانند رگرسیون لاسو، رده بندی، مدل های گرافیکی) و علوم کامپیوتر. به ویژه موضوع یادگیری ماشین، با مطرح شده موضوع کلان داده ها یادگیری آماری موضوعی پر کاربرد شده است. عنوان کتابی که به معرفی آن میپردازیم

An Introduction to Statistical Learning with applications in R است که توسط انتشارات اسپرینگر در سال ۲۰۱۳ منتشر شده است. نویسندگان کتاب گارت جیمز، دانیلا ویتن، تریور هاستی و روبرت تیشرانی هستند. هاستی و تیشرانی دو نفر از نویسندگان نخستین کتاب منتشر شده در زمینه یادگیری آماری هستند.

کتاب در ده فصل تدوین شده است. فصل اول ضمن معرفی مجموعه داده های که در کتاب مورد استفاده قرار میگیرند، برخی زمینه های کاربردی یادگیری ماشین را توضیح می دهد. در فصل دوم نخست به چسبی یادگیری ماشین و انواع آن (نظارتی و غیر نظارتی) می پردازد. سپس ضمن بیان کلیاتی در مورد مدل سازی و انتخاب مدل، محتوای فصل های مختلف کتاب را برای خواننده توضیح می دهد.

خواننده ای که با رگرسیون خطی، رگرسیون لوجستیک و کاربرد آن در رده بندی آشنایی دارد، می تواند از فصل های ۳ و ۴ به سرعت عبور کند. فصل پنجم به روش های باز نمونه گیری اختصاصی دارد. فصل ششم به روش های انتخاب مدل های رگرسیونی می پردازد. در این فصل به رگرسیون ستیجی، رگرسیون لاسو و رگرسیون مولفه های اصلی برای داده های با بعد بالا نیز پرداخته می شود. موضوع فصل هفتم، رگرسیون غیر خطی و رگرسیون چند جمله ای است. روش های درختی، مینا، مانند درخت های تصمیم، درخت های رگرسیونی، درخت های رده بندی و جنگل های تصادفی در فصل هشتم بحث می شود. بحث فصل نهم، که نیاز به آشنایی با بحث یادگیری ماشین دارد، به ماشین های بردار پشتیبان برای رده بندی می پردازد. فصل دهم، به بحث مولفه های اصلی و خوشه بندی که نوعی یادگیری غیر نظارتی هستند، پرداخته شده است.

تمرکز کتاب بر کاربرد است و کمتر خواننده را درگیر جزئیات ریاضی موضوعات می کند. پایان هر، به توضیح کاربرد روش های یادگیری آماری در تحلیل داده ها با بخش با عنوان آزمایشگاه اختصاصی دارد. پایان هر فصل تمرین هایی نیز گنجانیده شده است. انرم افزار کتاب در قالب یک درس با عنوان یادگیری آماری برای دانشجویان دوره کارشناسی ارشد و همچنین برا دانشجویان دوره کارشناسی که درس رگرسیون را گذرانیده اند، قابل تدریس است.

عنوان فصل های کتاب:

- 1 Introduction
- 2 Statistical Learning
- 3 Linear Regression
- 4 Classification
- 5 Resampling Methods
- 6 Linear Model Selection and Regularization
- 7 Moving Beyond Linearity
- 8 Tree-Based Methods
- 9 Support Vector Machines
- 10 Unsupervised Learning

چونیه انجمن آمار ایران، سال ۶۳، دیماه و آبانستان ۹۷، شماره ۹ و ۹۸

کتاب: An Introduction to Statistical Learning

Springer

پژاورد

نشریه انجمن علمی آمار دانشگاه سمنان

« صاحب امتیاز: انجمن علمی آمار دانشگاه سمنان »
 « مدیر مسئول: فاطمه رجایی نازکی »
 « سر دبیر: زهرا وفايي نژاد »
 « امور تحرير و نايبست: فاطمه رجایی نازکی »
 « گرافیک: محمّد رجایی نازکی »
 « صفحه آرایی: پاچاپ »
 « با تشكر از اساتيد محترم گروه آموزشي آمار: دكتر كريمي و دكتر حسيني »

سرمقاله:

سپاس پروردگار هستی بخش را که به لطف و کرمش، توفیقی حاصل شد تا بار دیگر نشریه دانشجویی پژاورد با همت انجمن علمی آمار دانشگاه سمنان به چاپ برسد. مؤسسان ملی بسیار مشتاق روی هم، انباشتن اعداد و ارقام هستند؛ آنها اعداد را گردآوری می کنند، با هم جمع می کنند، به توان چند می رسانند، ریشه سوم می گیرند و نمودارهای قشنگی تهیه می کنند. اما چیزی که هرگز نباید فراموش کنید آن است که هر یک از آن ارقام در وهله اول حاصل کار آن آمارگیر روستاست، که ... هر چه را دلش بخواهد ثبت می کند. (Sir Josiah Stamp)

در این شماره می خواهیم نظر همه را به موضوعی مهم جلب کنیم. آمار ثبتی یک نظام آماری وقتی به سمت پیشرفت و تکامل حرکت می کند که آمارهای که در آن تولید می شوند ماهیت ثبتی داشته باشند. امیدوارم مطالب این شماره مفید باشد و در نهایت از هیات تحریریه خاطر تلاشی های فراوان تشکر می کنم.

دکتر امید کریمی

فهرست
 صفحه
 عنوان
 آمارهای ثبتی مینا
 همیشه می توان از محاسبات آماری، نظریه های علمی استخراج کرد
 بوت استرپ
 A Brief History of Statistical Learning
 تهیه سازی و مسئله فرارونده دوره گرد
 شمر، مرد آماری
 جذبات و آمار
 معرفی کتاب

مرد آماری

مرد آماری دوان در صبح و شام
کوچه بن بست و خیابان‌ها شلوغ
مرد آماری نشان بی نشان
سوز سرما، مرد آماری روان
چون عطشان بارد ز سوی آسمان
چون بیابک خانه‌ها، کارخانه‌ها
صنعت و مسکن که باشد در فراز
وارد شدن در تراز با صادرات
از دبستان از کتب آرد خسر
لشعل رو-ای شیرین جوان
از عسل گندم، عسل‌دار چون
مرد آماری دوان و پرفشاناب
تا بساند چشودن ایران زمین
گر نباشد یک هوایی در جهان
گر نباشد هیچ آماری در میان
خلق این ایران زمین دارد پیام

خاک این ایران زمین گیرد قوام
چون دغالی از سیاهی بی فروغ
بی نشانان را بگسود یک نشان
سوی یک کارخانه‌ای او بی‌امان
در تلاش از خانه‌ای گیرد نشا
در شرف آید از این شکرانه‌ها
خلق این ایران زمین در انتظار
ملک پسر گسود نیانی با نشاط
میرد آسمانی نشان از سیفر
بعثت شوق انگیز دیرین جوان
کل‌گالایی که باشد در وطن
در میان‌سالای روان و در شتاب
تا نگردد این وطن را ظلم و کین
در دم خاموش کردیم بی‌گمان
چیلگی نابود در رود زمستان
مرد آماری تلاشست مستدام



jermy

پراورد
مجموعه نشریات تخصصی و علمی
تأسیس در سال ۱۳۸۵

آمارهای ثبتی مبنا

دکتر امید کریمی



سازمان های ملی آمار سه نوع آمار مبتنی بر ریز داده ها منتشر می کنند :
آمار های مبتنی بر آمار گیری نمونه ای
آمار های مبتنی بر سر شماری
آمار های مبتنی بر ثبت های اداری
این دوره در مورد نوع سوم یعنی آمار های مبتنی بر ثبت های اداری است که به جای جمع آوری داده ها از طریق آمار گیری های نمونه ای و سر شماری ، ثبت های اداری را با استفاده از منابع مختلف تعدیل و پردازش می کنند تا داده ها را برای اهداف آماری مناسب سازند . این نوع آمار گیری ، آمار گیری ثبتی نامیده می شوند .

آمار گیری ثبتی

آمار های ثبتی مبنا که حاصل از اجرای یک آمار گیری ثبتی هستند ، آمار هایی هستند که از پردازش (طبقه بندی ، مرتب کرد ، محاسبات ، تلخیص و ...) داده های یک ثبت حاصل می شوند .

برای مثال ، فرم ثبت نام دانشگاه شامل اقامت اطلاعاتی نظیر نام ، نام خانوادگی ، شهرستان محل اقامت ، معدل دیپلم ، محل تولد ، نام و شغل پدر و غیره می باشد . چنانچه این نوع داده ها برای همه دانشجویانی که ثبت نام نموده اند درون یک ثبت جمع آوری شوند از جمع بندی ، دسته بندی و پلايش داده های این ثبت آمار های ثبتی مبنا تولید خواهد شد .

نظام آمار های ثبتی

نظام آمار های ثبتی ، نظامی است که داده های ثبتی تولید شده در حین عملیات روزمره و جاری یک سازمان را به اطلاعات آماری مورد نیاز مصرف کنندگان تبدیل کرده و آنها را به نحو مناسب انتقال دهد .

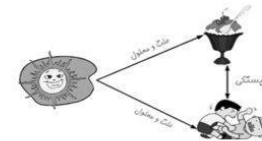
در این نظام هدف فعالیت هایی با عنوان "جمع آوری آمار" نیست ، بلکه فعالیت های جاری و روزمره سازمان می باشد .

استفاده از فعالیت های جاری سازمان برای ثبت و انتقال داده ها و تولید آمار ثبتی به منظور تأمین نیاز کاربران اطلاعات از ویژگی های مهم این نظام است . در این نظام مفهوم آمار های ثبتی با ثبت آمار یکسان نیست ، به عنوان مثال "پارنامه" برای تولید آمار و اطلاعات تکمیل نمی شود بلکه مجوزی برای بازگویی ورود ، خروج یا انتقال بار است ، همچنین قبض برق ، آب ، گاز و تلفن برای تولید آمار و اطلاعات صادر نمی شود بلکه هدف از صدور آن دریافت هزینه های مربوط به تولید و انتقال می باشد .

پراورد
مجموعه نشریات تخصصی و علمی
تأسیس در سال ۱۳۸۵

آشنایی با نظریه های نظام گیری و آمار ثبتی مبنا / دکتر امید کریمی / آبان ۱۳۸۵ / سازمان پراورد

همیشه نمی توان از مطالعات آماری ، نظریه های علمی استخراج کرد



ما معمولاً واقعیت های آماری را با نظریه های علمی اشتباه می گیریم. اینکه آمار مصرف بستنی و آمار درگیری ها همزمان افزایش و کاهش پیدا می کنند یک واقعیت آماری است. اما اینکه بین بستنی و درگیری، «رابطه علت و معلول» در نظر بگیریم، یک خطای جدی علمی است.

شاید حالا بتوان درک کرد که چگونه در حوزه های علمی مختلف، نتایج تحقیقات ضد و نقیض منتشر می شود. اینکه چرا جای و فیهو به فاصله چند روز از عامل سرطان به عامل بازدارنده سرطان (یا بالعکس) تغییر هویت می دهند. یا اینکه تحقیقات در مورد ضرر سیگار (در عین اشتراک زیان های شدید آن) تا این حد تفاوت عددی دارد و...

سالها پیش تحقیقات در آمریکا نشان داد که دقیقاً با افزایش فروش بستنی، تعداد درگیری های خیابانی افزایش پیدا می کند. با توجه به این واقعیت آماری، کدام مورد را نتیجه می گیرید؟

(الف) خوردن و فروش بستنی کار بسیار خطرناکی است و می تواند منجر به افزایش درگیری های خیابانی شود.
(ب) احتمالاً عاملی مثل گرم شدن هوا، همزمان موجب عصبی تر شدن مردم و نیز افزایش تقاضا برای بستنی و فروش آن می شود.

انسانهای سیگاری، طول عمر کمتری نسبت به انسانهای غیرسیگاری دارند. شما از این واقعیت آماری، کدام مورد زیر را نتیجه می گیرید؟

(الف) این مصرف سیگار است باعث کاهش عمر می شود.
(ب) احتمالاً عواملی مثل تنش و استرس، فشار روانی، مصرف کافئین و... وجود دارند که همزمان، ضمن ترغیب انسانها به سیگار، آسیب های فیزیولوژیک متعدد دیگری هم به بدن می زنند و حاصل ماجرا، کاهش عمر است. به عبارت دیگر، مصرف سیگار و کاهش عمر هر دو ناشی از علت دیگری به نام «تنش و استرس» هستند.

بوت استرپ

در آمار، بوت استرپ یک روش محاسباتی-آماري – کامپیوتری است برای تعیین میزان دقت برآوردگرهای حاصل داده نمونه. در این تکنیک تنها با یک روش خیلی ساده می توان تقریباً هر آماره ای از توزیع داده های نمونه را تخمین زد.

در آمار بوت استرپ یک مدل کامپیوتری است برای نسبت دادن معیار دقت به تخمین های داده نمونه. در این تکنیک تنها با یک روش خیلی ساده می توان تقریباً هر آماره ای از توزیع داده های نمونه را تخمین زد. به طور عمومی این روش از روش های باز نمونه گیری به حساب می آید. بوت استرپ در واقع تخمین ویژگی های (مثل واریانس) یک تخمین زنده است با استفاده از اندازه گیری همین ویژگی ها در یک توزیع تقریبی از کل داده های نمونه. یک انتخاب استاندارد برای توزیع تقریبی، توزیع تجربی داده های مشاهده شده است.

البته لازمه کنترل فرایند آماری، ثبت کلی اطلاعات بخشهای مختلف فرایند است که زیرساخت لازم را برای تحلیل های آماری فراهم می آورد. استفاده از تکنیک های آماری در کنترل مواد اولیه، مدل ارتباطی مواد اولیه با تولید، پیش بینی تولید، تعیین عوامل مؤثر بر تولید محصول نامرغوب از جمله موارد استفاده از تکنیک های آماری در بخش تولید است. در بخش ستادی نیز موارد متعددی از کاربرد آمار را می توان اشاره نمود. بررسی رضایت مدتی مشتریان، آگاهی از نظرات پرسنل ستادی، تولید و خدماتی، ارزیابی عملکرد کارکنان، کاهش مدت زمان فرایندهای وابسته به زمان، مانند فرایند خرید قطعات و مواد اولیه و بسیاری از موارد دیگر اشاره به این کاربردها است.

در ادامه این بخش، به ذکر یک تجربه از کاربرد آمار می پردازم که در عین سادگی در کاربرد نتیجه های جالب در برداشته است. در یکی از صنایع بزرگ تولیدی، مدیر عامل مربوطه مقرر کرده بود که تمامی کارکنان صبح و در بدو ورود به محل کار به مدت چند دقیقه به طور دسته جمعی ورزش نمایند. پس از مدتی یکی از کارمندان آشنا با آمار تعدادی از کارکنان را ب همصادف انتخاب نموده و با چند پرسش ساده نظرشان را در خصوص ورزش صبحگاهی جویا شده و نتایج را به مدیریت گزارش می دهد. نتایج برای مدیریت صنعت بسیار جالب بوده و ضمن اعمال نظر کارکنان طبق نظرسنجی، پاداش مناسبی نیز به کارمند خوش فکر اعطا شد.

در خصوص کاربرد آمار در بخش خدمات نیز مصادیق متعددی وجود دارد. شاید بتوان گفت بهترین استفاده از آمار در این بخش نظرسنجی از حسن خدمات ارائه شده توسط واحدهای خدماتی به مخاطبان خود است که اصطلاحاً رضایت مدتی مشتری می نامند. در ادامه با ذکر دو کاربرد ساده دیگر از علم آمار در این بخش بحث خاتمه داده میشود. یک فارغ التحصیل آمار را پس از چند سال دیدم و از کارش سؤال کردم، گفت که یک شرکت خدمات کامپیوتری را ایجاد نموده است. سؤال دوم برآیم بوجود آمد که از آموخته های آماری چه بهره ای در این کار برده ای؟ با یک حالت پروژهمندانه گفت که تکنیک های آماری برای من بسیار با ارزش بود. سپس توضیح داد که بازار خرید و فروش کامپیوتر را با استفاده از تکنیک های سری های زمانی مدل سازی نمودم. این کار باعث شد که اطلاعات ارزشمندی از بازار و پیشبینی آینده آن داشته باشم و در نتیجه موفقیت اقتصادی در این کار نصیب من شد.

مصادیق دیگر استفاده از تکنیک های آماری در سطح جامعه مربوط به اتحادیه و دارای مدرک کارشناسی ارشد است. به من مراجعه کرد و در خصوص انجام یک نظرسنجی در سطح شهر اصفهان درخواست همکاری نمود. طبق اطلاع واصله، جلی بین اصناف این اتحادیه و اداره بازرگانی در خصوص قیم گذاری یک محصول که در سال در یک محدوده زمانی کوتاهی تولید و مصرف می شود، در جریان بود. نیمی از دانشجویان کارشناسی آمار این نظرسنجی را انجام دادند. طبق نتایج به دست آمده تمامی افشار مردم خواستار کیفیت محصول بدون در نظر گرفتن قیمت آن بودند. نتایج برای مسئولین بازرگانی بسیار جالب بود.

در خاتمه شایان ذکر است که بیان کاربرد آمار در صنعت و خدمات خود مثنوی است که لازمه هفتاد من کاغذ است و در این مجال نمی گنجد.



[illegible]

پراورد
مجله تخصصی پژوهش‌های علمی و اجتماعی
پایه علمی و تخصصی پژوهش‌های علمی و اجتماعی
پایه علمی و تخصصی پژوهش‌های علمی و اجتماعی

[illegible]

The top section of the slide features two logos. On the left is the Bootstrap logo, which consists of a large white letter 'B' inside a black circle, with the word 'Bootstrap' in white text to its right. On the right is the Pajonk logo, which is a circular arrangement of various white icons representing different web development and design concepts, with the word 'پاژنک' in white Persian text below it.

A Brief History of Statistical Learning

Though the term statistical learning is fairly new, many of the concepts that underlie the field were developed long ago. At the beginning of the nineteenth century, Legendre and Gauss published papers on the method of least squares, which implemented the earliest form of what is now known as linear regression. The approach was first successfully applied to problems in astronomy. Linear regression is used for predicting quantitative values, such as an individual's salary. In order to predict qualitative values, such as whether a patient survives or dies, or whether the stock market increases or decreases, Fisher proposed linear discriminant analysis in 1936. In the 1940s, various authors put forth an alternative approach, logistic regression. In the early 1970s, Nelder and Wedderburn coined the term generalized linear models for an entire class of statistical learning methods that include both linear and logistic regression as special cases.

By the end of the 1970s, many more techniques for learning from data were available. However, they were almost exclusively linear methods, because fitting non-linear relationships was computationally infeasible at the time. By the 1980s, computing technology had finally improved sufficiently that non-linear methods were no longer computationally prohibitive. In mid 1980s Breiman, Friedman, Olshen and Stone introduced classification and regression trees, and were among the first to demonstrate the power of a detailed practical implementation of a method, including cross-validation for model selection. Hastie and Tibshirani coined the term generalized additive models in 1986 for a class of non-linear extensions to generalized linear models, and also provided a practical software implementation. Since that time, inspired by the advent of machine learning and other disciplines, statistical learning has emerged as a new subfield in statistics, focused on supervised and unsupervised modeling and prediction. In recent years, progress in statistical learning has been marked by the increasing availability of powerful and relatively user-friendly software, such as the popular and freely available R system. This has the potential to continue the transformation of the field from a set of techniques used and developed by statisticians and computer scientists to an essential toolkit for a much broader community.



بهبه سازی و مسئله فروشنده دوره گرد

به صورت کلی مسئله فروشنده دوره گرد دارای ۳ حالت زیر می باشد.

۱ فروشنده دوره گرد متقارن

در حالت متقارن مسئله، تعدادی شهر داریم و هزینه رفتن مستقیم از یکی به دیگری را می دانیم. مطلوب است کم هزینه ترین مسیری که از یک شهر شروع شود و از تمامی شهرها دقیقاً یکبار عبور کند و به شهر شروع بازگردد.

۲ فروشنده دوره گرد نامتقارن

مساله ی فروشنده ی دوره گرد نامتقارن، یک TSP است که فاصله بین رئوس آن، متقارن نیست. TSP بسیار مشکل تر از TSP است، در حقیقت در حالی که TSP متقارن، حتی در گراف های با چندین هزار رأس به طور پهنه قابل حل است، تنها نمونه های خاصی از TSP را که ماتریس فاصله ی آنها، تقریباً متقارن است، تنها در گراف های دارای چندین دوجین رأس، می توان به طور پهنه حل کرد. به کاربردن هوش مصنوعی برای TSP، راحت و سر راست است، چون هیچ تغییراتی در الگوریتم اصلی، لازم ندارد. پیچیدگی محاسباتی در حلقه ی الگوریتم، برنامه ی کاربردی TSP یکسان است، زیرا تنها تفاوت آنها در فاصله ها و ماتریس های ردپا است که در اینجا دیگر متقارن نیستند.

۳ فروشنده دوره گرد با پنجره های زمانی

مسئله فروشنده دوره گرد با پنجره زمانی، شامل یافتن کوتاهترین طول توری است که به وسیله یک فروشنده دوره گرد طی می شود با این شرایط که فروشنده باید هر کره را فقط یکبار ملاقات کند و در پنجره زمانی معینی به آن سرویس دهد. به این معنا که اگر فروشنده زودتر از محدوده زمانی تعیین شده به آن کره برسد باید منتظر بماند تا بازه زمانی سرویس دهی مربوط به آن کره شروع شود. همچنین اگر دیرتر از پنجره زمانی برسد ارائه سرویس به آن کره دیگر امکان پذیر نخواهد بود.

یک اکثر شناسی در مرکز فرمان یک تلسکوپ، اغلب میل دارد اندازه گیری های کوتاه مدتی را برای فهرست بلندی از ستارگانی که در پهنه آسمان پراکنده اند، انجام دهد. حرکت دادن تلسکوپ از ستاره ای به ستاره دیگر، فرایندی وقت گیر است. بنابر این اکثر شناسی مزبور سعی می کند برای پرهریز از حرکت های غیر ضروری، کاراترین دنباله از دنبادهار را انتخاب کند. این مسئله بخصوصی در مورد یک تلسکوپ فضایی که در بالای جو در حال گردش به دور زمین است، تشدید می شود. عمل نشانه روی در چنین تلسکوپی اغلب از زمین به کمک یک چت گازی کوچک تنظیم می شود و سوختی که در دسترس قرار دارد، بشدت محدود است. این مسئله صورتی از مسئله فروشنده دوره گرد است و معمولاً به عنوان مسئله جستجو برای کوتاهترین مسیر بسته بین مجموعه ای از شهر ها، بطوریکه هر شهر فقط یک بار مورد بازدید قرار گیرد، بیان می شود. این، انگوی نمونه ای از یک رده مسائل است که می توان آنها را به طور خلاصه بیان کرد اما راه حل مستقیم آنها چالش برانگیز است و تنها می توان به کمک رشته حدس های عالمانه حلشان کرد.

مسئله فروشنده دوره گرد

مسئله فروشنده دوره گرد، مسئله ای مشهور است که ابتدا در سده ۱۸ مسائل مربوط به آن توسط ویلیام همیلتون و توماس کرکمن مطرح شد و سپس در دهه ۱۹۳۰ شکل عمومی آن به وسیله ریاضیدانی مثل کارل منگر از دانشگاه هاروارد و هاسلر ویتنی از دانشگاه پرینستون مورد مطالعه قرار گرفت.

شرح مسئله بدین شکل است

تعدادی شهر داریم و هزینه رفتن مستقیم از یکی به دیگری را می دانیم. مطلوب است کم هزینه ترین مسیری که از یک شهر شروع شود و از تمامی شهر ها دقیقاً یک بار عبور کند و به شهر شروع باز گردد.

این مسئله علاوه بر جنبه نظری از جنبه عملی نیز کاربرد فراوانی دارد به عنوان مثال در مواردی مانند مسیریابی، ساخت ترانه های الکترونیکی، زمان بندی کارها و غیره مورد استفاده قرار گیرد. اما در مواجهه با چالش حل مسائل پهنه سازی، که این نوع مسائل در دنیای واقعی بسیار زیاد هستند، روش های کلاسیک اغلب با مشکل مواجه می شوند. به همین دلیل معمولاً از روشهای فرا ابتکاری همانند الگوریتم ژنتیک و سایر الگوریتم های تکاملی برای حل این نوع مسائل استفاده میشود