انجمن علمی آمار
دانشگاه اراک
۱۳۹۷

نشریه علمی آنالیز

شماره اول- آذرماه ۱۴۰۰- سال اول
نشریه علمی دانشجویی انجمن علمی آمار دانشگاه اراک

در این شماره می‌خوانیم:

سخن سردبیر

دفتر را با نام او می‌گشایم که هر امر مبهمی بی‌یاد او بی‌حاصل است.

شماره اول نشریه انیگما را در آذر ماه ۱۴۰۰ کلید می‌زنیم این نشریه زیر نظر انجمن علمی آمار دانشگاه اراک منتشر می‌شود. به همین بهانه چند سطری را با شما مخاطبان نشریه در رابطه با انتخاب نام این مجله در میان می‌گذارم.

همه‌ی ما با گستره علم آمار در علوم مختلف آشنا هستیم که بیان آن تکرار مکررات است. هر آماردان، آمارخوان یا تحلیل‌گرداده امروزی نیازمند آشنایی با علوم‌کامپیوتری به روز است چرا که در عصر اطلاعات و ارتباطات دیجیتال این عنوان‌ها را کسب می‌کند. به همین علت گذری بر گذشته برنامه‌نویسی، الگوریتم‌نویسی و هوش‌مصنوعی زدیم و نام انیگما را برگزیدیم.

انیگما اولین دستگاه رمزگذاری و کدنگاری بود که از هوش مصنوعی برای محافظت از ارتباطات تجاری، دیپلماتیک و نظامی به طور گسترده در جنگ جهانی دوم توسط ارتش آلمان نازی به کار گرفته می‌شد.

آلن تورینگ، پدر علم کامپیوتر و هوش مصنوعی، به کمک ماشین تورینگ، فرمولاسیون موثری برای روش‌های الگوریتم‌نویسی و محاسبه تهیه کرد که موفق به شکست کدنگاری‌های ماشین انیگما شد. به همین جهت نام انیگما را برای این نشریه انتخاب کردیم تا اهمیت علم آمار و نقش داده‌ها را در هوش مصنوعی شفاف‌تر کنیم.

امید است قدم کوچکی در جهت گسترش و معرفی علم آمار برداریم.

داده را از کجا پیدا کنیم؟

یکی از عمده مشکلاتی که نیازمندان تحلیل‌های آماری با آن سرو کار دارند یافتن یک مجموعه داده معتبر برای انجام تحقیقات اولیه و گزارش نویسی دانشگاهی است. در این بخش به معرفی وبسایت‌هایی می‌پردازیم که مجموعه داده‌های شناسه داری را رایگان در اختیار همگان قرارمی‌دهند:

سایت مرکز ملی آمار ایران

در این سایت انواع داده‌های سرشماری‌های کشوری، مطالعات و تحقیقاتی که زیر نظر مرکز ملی آمار ایران انجام شده است در اختیار همگان قرار دارد و منبع اصلی دسترسی به داده‌های ملی است.

سایت data.world

در Data world داده‌ها را کشف کنید و به اشتراک بگذارید. با مردم ارتباط برقراری کنید و برای حل مشکلات با یکدیگر همکاری کنید! این سایت داده‌های واقعی با قابلیت تجزیه و تحلیل و استفاده مجدد و قابل توسعه را در اختیار کاربران خود قرار می‌دهد. با یک کلیک موضوع خود را پیدا و داده دریافت و تحلیل کنید.

سایت kaggle

اگر علاقه مند به حوزه پردازش و تحلیل داده هستید سری به مجموعه دادهای این وبسایت بزنید. این سایت مجموعه داده‌هایی را در اختیار کاربران می‌گذارد و مسائلی در حوزه تحلیل داده راجع به آن داده‌ها را مطرح می‌کند.

به علاوه با برگزاری مسابقات پردازش داده‌ها کمی چاشنی رقابت بین تحلیلگران هم اضافه کرده است.

هر تیم شرکت کننده در بازه زمانی معین باید برنامه یا الگوریتم بهینه مورد نظر را برای رسیدن به هدف تعیین شده ارایه کند. و گاهاً جایزه‌های چندین دلاری و استخدام تیم‌های برتر را به دنبال دارد.

البته این سایت برای پروکسی‌های ایران در دسترس نیست و نیازمند فیلترشکن می‌باشد.

سایت machine learning uci repository

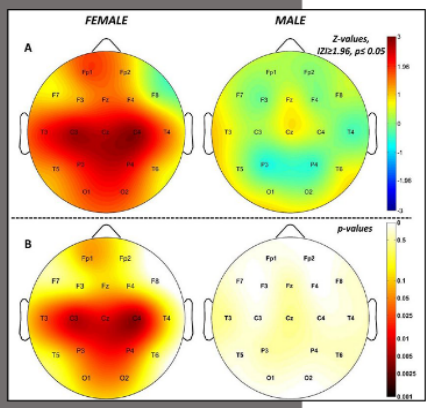
این سایت همانطور که از نامش پیداست یک مخزن دیتاست برای یادگیری ماشین می‌باشد که داده را رایگان در اختیار همگان با شناسه تحقیقاتی قرار می‌دهد.

این سایت جز برترین سایت‌های دسترسی به داده‌ها معرفی شده‌اند اما بسیاری سایت دیگر با اهداف مختلف برای دسترسی به داده وجود دارد کافیست هدف خود را گوگل کنید...

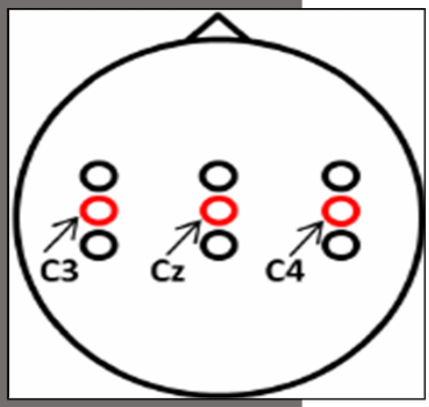
با کلیک روی نام سایت‌ها وارد سایت مربوطه شوید.

مقالک

در هر نشریه قصد داریم با ارایه مقالاتی کاربرد علم آمار را در علوم مختلف بیان کنیم. برای اولین نسخه کاربرد علم آمار در علوم اعصاب و علوم شناختی را در نظر گرفتیم:
عنوان مقاله: پردازش غیر خطی در آنالیز آماری سیگنالهای EEG در بررسی خویشاوندی و علاقه‌مندی به انتخاب یک کالا
نویسندگان: سعیده رییس دانا، سمانه صفری
در این مقاله یک طرح بازاریابی عصبی توسط پردازش سیگنالهای EEG انجام شده است که در آن میزان علاقه‌مندی افراد جامعه به خرید یک کالای تزیینی نسبتا لوکس ارزیابی می‌شود.



استخراج احساسات و ترجیحات افراد از ناخودآگاه آنها که موضوع بازاریابی عصبی است می‌تواند به عنوان یک روش قابل اطمینان برای ارزیابی میزان علاقه‌مندی افراد به کالا و خدمات مورد بررسی قرار گیرد. نورو مارکتینگ یک علم میان رشته‌ای در حال ظهور در مرز بین علوم اعصاب، روانشناسی و بازاریابی است که به لزوم درک بهتر عملکرد مغز و بررسی رفتار مصرف کننده از دیدگاه مغز تاکید دارد. تحقیقات در حوزه بازاریابی عصبی با تجزیه و تحلیل سیگنال‌های بیولوژیکی از جمله EMG , EOG,EEG و تصاویر FMRIو ردیابی چشم، آنالیز حالت چهره و ... انجام می‌گیرد.
در این مقاله روش‌های‌آماری مورد استفاده به ترتیب:
۱.ثبت دادگان
۲. ابزار پردازش و روش‌های محاسباتی
۳. آنالیز مولفه‌های مستقل با روش‌های محاسباتی و آماری برای جداسازی مولفه‌های مستقل، مجموعه‌ای از سیگنال‌ها، اندازه گیری، متغیرهای چندگانه است.

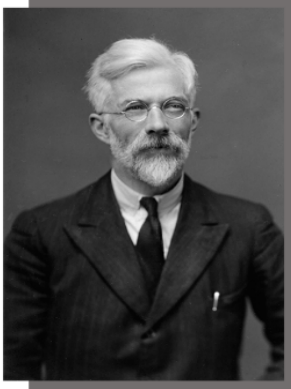


۴. شبکه نورو فازی: حل مسائل پیچیده بازشناسی الگوهای عصبی با کمک یادگیری ماشین و استنتاج مدل فازی در چارچوب شبکه عصبی چندگانه.
۵. آنالیزهای‌آماری : آزمون t ، آنالیز واریانس anova
۶. پردازش تصویری داده ها و پردازش ICA
۷. استخراج ویژگی و طبقه بندی سیگنالها
تحلیل آماری مورد استفاده جهت اعتبارسنجی نتایج این تحقیق به کمک ابزارهای استاندارد آنالیزآماری، تجزیه و تحلیل شده اند. داده‌ها توسط پرسشنامه فراهم شده‌اند. از نرم افزار آماری SPSS هم برای آزمون‌های رگرسیونی و آنالیز دادهای ناخود آگاه استفاده شده است.
در پایان هم نتایج این مقاله با مقالات مشابه مقایسه شده و نتایج را اعلام نموده‌اند. برای اطلاع بیشتر لینک مقاله به آدرس زیر است:

https://www.sid.ir/fa/journal/ViewPaper.aspx?ID=497594

با کلیک روی لینک سایت یا بارکد می‌توانید وارد سایت مقاله شوید.

فیشر



فیشر فرزند کِیتی هیث (وکیل) و جرج فیشر (صاحب یک شرکت فروش فوق‌العاده درخیابان کینگ لندن) بود. فیشر سه خواهر و یک برادر بزرگتر داشت. وقتی ۱۴ ساله بود مادرش از دنیا رفت و تنها ۱۸ ماه بعد پدرش بر اثر شکست قابل ملاحظه در معامله شغلش را از دست داد. رونالد در طول دوران تحصیل، استعداد درخشانی در ریاضیات از خود نشان داد. علی‌رغم از دست دادن مادرش در چهارده سالگی وی در همان سال، در رقابت ریاضیاتی که بین دانش‌آموزان تمام مدارس، در مدرسه «هارو» برگزار شد، موفق به کسب مدال شد. اگرچه رونالد فیشر بینایی ضعیفی داشت اما یک دانش آموز باهوش و پیشرو بود و در مسابقات ریاضی مدرسه هارو در ۱۶ سالگی مدال گرفت. به دلیل بینایی ضعیفش برای آموزش خانگی ریاضی از کاغذ و قلم استفاده نمی‌کرد، طوری که توانایی او در تجسم موارد هندسی نیز گسترش یافت. او گسترش قابل توجهی در علم بیولوژی داد و سیر تکاملی خاصی در آن ایجاد کرد.

در سال ۱۹۰۹ او موفق به دریافت بورس کالج گانویل دانشگاه کمبریج لندن شد. در کمبریج فیشر به انجام مطالعاتی جدید درباره تئوری‌های مربوط به ژنیتیک مندل دست زد و همچنین نظری بر بیومتری(کاربرد آمار درمبحث زیست شناسی) داشت و با تشکیل مجموعه‌ای ازدست نوشته‌ها در مورد روش‌های آماری موجب شد که روشی بالقوه در تطبیق با طبیعت وارث مندل حاصل شود که در آن از واریانس‌های متوالی که رتبه تغییرات نهادین بواسطه اثر تغییرات چند ژن و محیط اطراف و تکامل تدریجی آن مشخصه بوجود آورده است استفاده می‌شود.

وی پس از دریافت جایزه ولستون به مطالعاتش بر روی تئوری خطاها در دانشگاه کیمبریج ادامه داد. در واقع این علاقه فیشر بود که آن‌ها را به سمت سرمایه‌گذاری بر روی مطالعات حوزه آمار سوق می‌داد. بعد از فارغ‌التحصیل شدن از کمبریج، او دیگر از هیچ‌گونه پشتیبانی مالی برخوردار نبود و برای چند ماه در شرکتی در کانادا مشغول به کار شد. سپس به لندن بازگشت و به عنوان یک متخصص علم آمار در یک شرکت بزرگ استخدام شد. علاقه‌اش به اصلاح نژاد و گیاهان و تجربیاتش در کمبریج، فیشر را علاقه‌مند کرد به اینکه برای خود مزرعه‌ای داشته باشد.

«کارل پیرسون» به او پیشنهاد کرد تا به عنوان کارشناس ارشد آمار در لابراتوار گالتون مشغول به کار شود. همچنین مسئولیت اجرایی مشابهی در ایستگاه آزمایش‌های کشاورزی«روتهامستد» به وی پیشنهاد شد که یکی از قدیمی‌ترین مؤسسات تحقیقات کشاورزی در انگلستان بود و در سال ۱۸۳۷ به منظور مطالعه بر روی اثرات تغذیه خاک و انواع مختلف خاک بر روی باروری گیاهان تأسیس شده بود. علاقه فیشر به کشاورزی سبب شد تا مسئولیت اجرایی پیشنهادی روتهامستد را بپذیرد. جایی که او به واسطه ارائه روش‌های آنالیز و تجزیه و تحلیل نتایج آزمایش‌ها، خدمات زیادی هم به علم آمار و هم به علم ژنتیک کرد. در آنجا بود که او برروی طراحی آزمایش‌هایی بوسیله معرفی مفهوم انتخاب تصادفی و آنالیز واریانس مطالعاتی انجام داد، روش‌هایی که هم‌اکنون نیز در تمام دنیا مورد استفاده قرار می‌گیرند. وی در سال ۱۹۲۱، مفهوم احتمال را معرفی کرد. احتمال یک پارامتر، متناسب است با امکان وجود داده‌ها و تابعی ارائه می‌دهد که معمولاً دارای یک واحد ارزش ماکزیمم است، که ماکزیمم احتمال نامیده می‌شود. فیشر در سال ۱۹۲۲، تعریف جدیدی از آمار ارائه کرد. هدف او این بود که ادعا کند، کاهش (کمبود) داده وجود دارد و سه مشکل اساسی را معرفی کرد:

۱.تنوع جمعیتی که داده‌ها از آنها استخراج می‌شوند

۲. تخمین

۳. توزیع

با کلیک روی Qrکد
پادکست فیشر را بشنوید.

نرم افزار R

نرم‌افزار Rیک نرم‌افزار محاسباتی آماری و یک زبان برنامه‌نویسی نیز می‌باشدکه در زمانی کوتاهی مورد استقبال گسترده‌ای قرار گرفته است. با اطمینان می‌توان گفت که R هر روز از روز پیش کامل‌تر است، زیرا بیش‌تر کاربران آن با نوشتن برنامه‌هایی برای حل مساله‌های خود و به اشتراک گذاشتن آنها در پیشرفت نرم‌افزار نقشی اساسی دارند. امروزه ۱۸۲۷۶ پکیج مختلف در نرم افزار وجود دارد که قابل دانلود برای همگان است. این نرم‌افزار یک زبان برنامه‌نویسی و محیطی یک‌پارچه را در اختیار کاربران قرار می‌دهد تا به کمک آن بتوانند کارهای گوناگونی از جمله موارد زیر را انجام دهند:

* وارد کردن، ویرایش، پیکره بندی و ذخیره‌ی انواع داده‌ها در قالب‌های مناسب.
* اجرای روش‌ها و تحلیل‌های رایج آماری با به‌کاربستن تابع‌های موجود در R
* اجرای روش‌ها و تحلیل‌های جدید آماری با به‌کارگیری بسته‌های نرم‌افزاری موجود بر پایگاه‌های نرم‌افزاری مانندCRAN, Github , BioConductor
* اجرای محاسبات عددی مانند اعمال جبری روی بردارها، ماتریس‌ها و آرایه‌ها، حل معادله‌های خطی و ناخطی، یافتن ریشه‌ی چندجمله‌ای‌ها، بهینه‌سازی، تقریب انتگرال، مشتق گیری و...
* رسم انواع نمودارهای آماری با انعطاف پذیری بالا
* به‌کاربستن یک زبان برنامه‌نویسی ساده و کارا برای پیاده‌سازی الگوریتم‌ها و هر نوع تحلیل آماری و ریاضی دلخواه.

تاریخچه:

در اواسط قرن بیستم که انواع روش‌های اماری برای تحلیل داده‌ها توسط آماردانان و پژوهشگران دیگر رشته‌ها معرفی و در سطح گسترده‌ای از علوم و رشته‌های مهندسی فراگیر شده بودند، برای اجرای محاسبات آماری، بیش‌تر زبان‌های برنامه نویسی رایج آن دوران، به ویژه فورترن، به کار گرفته می‌شد. به همین خاطر انجام محاسبات مربوط به تحلیل داده‌ها به نوشتن برنامه‌های طولانی و مهارت بالای برنامه نویسی نیاز داشت.

درسال‌های ۱۹۷۶-۱۹۷۵ به نظر انجام ساده‌تر و سریع‌تر محاسبات آماری و عددی در مجموعه‌ی مشهور و با سابقه‌ی آزمایشگاه‌های بِل نسخه‌های آغازین زبان برنامه نویسیS توسط جان چَمبِرز و با همکاری ریک بِکِر و آلن ویک شکل گرفت. درسال ۱۹۸۸ تغییرات زیادی در زبانS ایجاد و ویژگی‌های جدیدی مانند به‌کارگیری تابع نویسی به آن افزوده شد و نسخه جدید NEW S نام گرفت. سپس ویژگی‌های برنامه نویسی شیءگرا مانند به کارگیری کلاس‌ها و مِتُدها نیز به S افزوده و نسخه‌ی جدید S۴ نامیده شد. در این میان آمارگیری به زبانS محدود به آزمایشگاه‌های بل نشد و این زبان در اختیار دیگر افراد به ویژه پژوهشگران در دانشگاه‌ها و مراکز علمی قرار گرفت. سادگی برنامه‌نویسی با S ، امکان اجرای روش‌های آماری رایج با دستورهای ساده و کوتاه، رسم انواع نمودارهای گرافیکی و انتشار چندکتاب درمورد شیوه بکارگیری S توسط تیم توسعه دهنده‌ی آن باعث شد آن زبان مورد توجه قرار گیرد و گسترش یابد. به دلیل استقبال از S ، یک پیاده‌سازی تجاری از آن با عنوانS_PLUSتوسط یک شرکت تجاری منتشر و برای فروش به بازار عرضه شد. درسال ۱۹۹۳ راس ایهاکا و رابرت جِنتِلمَن از دانشگاه اوکلَند نیوزیلند با افزدون ویژگی‌هایی از زبان برنامه‌نویسی Scheme به S، پروژه‌ی توسعه‌ی یک پیاده‌سازی غیرتجاری و متن باز از S را آغاز کردند و آن را R نامیدند. این اسم، حرف اول اسم هردوی آنان بوده و حرف قبل از S در الفبای انگلیسی است. پس از آن متخصصان دیگری از سراسر جهان، از جمله جان چمبرز خالق S ، به آن‌ها پیوستند و تیم هسته‌ی توسعه‌ی را R تشکیل دادند که تا به امروز مسؤلیت توسعه و به‌روز رسانی Rرا به عهده دارد.

برای دانلود آخرین نسخه نرم افزار R و R studio به لینک‌های زیر مراجعه کنید:

https://cran.um.ac.ir/bin/windows/base/R-4.1.1-win.exe

https://download1.rstudio.org/desktop/windows/RStudio-1021.09.0-351.exe

چرا نرم افزار R؟

چرا R برای Data Science مهم است؟

چون می‌توانید کد خود را بدون هیچ کامپایلری اجرا کنید؛ یعنی R یک زبان تفسیر شده است. از این رو می‌توان بدون هیچ کامپایلری کد را اجرا کرد . R کد را تفسیر می‌کند و توسعه کد را آسان تر می‌کند. از طرفی دیگر بسیاری از محاسبات با بردارها انجام می شود؛ یعنی R یک زبان برداری است. بنابراین هر کسی می تواند توابع را به یک بردار اختصاص دهد، بدون اینکه حلقه ایجاد کند. از این رو، R قدرتمندتر و سریع‌تر از زبان‌های دیگر است.

چرا R برای تجارت(Business) خوب است؟

دلیل اصلی این است که R یک برنامه open-source است؛ بنابراین می‌تواند براساس نیاز کاربر اصلاح و توزیع شود. برای عملیات مصورسازی داده‌ها (visualization) عالی است و در مقایسه با سایر ابزارها قابلیت های بسیار بیشتری دارد. برای مشاغل داده محور، عدم وجود دانشمند داده(data scientist) یک نگرانی بزرگ است. شرکت‌ها از برنامه نویسی R به عنوان پلتفرم اصلی خود استفاده می‌کنند و برنامه نویسان آموزش دیده R را استخدام می‌کنند.

زبان R یک شغل پردرآمد

زبان R به طور گسترده ای در data science مورد استفاده قرار می‌گیرد که برخی از پردرآمدترین مشاغل روز جهان را در بردارد. اگر می‌خواهید وارد حوزه data science شوید و درآمد بالایی بدست آورید، قطعاً باید R را یاد بگیرید.

منبع باز open-source

زبان R یک زبان منبع باز است که توسط جامعه ای از کاربران فعال توسعه داده می‌شود و به کار خود ادامه می دهد؛ از همین رو می‌توانید از R به صورت رایگان بهره‌مند شوید. شما می‌توانید توابع مختلف را در R اصلاح کرده و بسته های نرم افزاری خود را بسازید. از آنجا که R تحت مجوز عمومی(GNU)صادر شده است، هیچ محدودیتی در استفاده از آن وجود ندارد.

محبوبیت

زبان R به یکی از محبوب ترین زبان های برنامه نویسی در صنایع تبدیل شده است. به طور معمول، R بیشتر در دانشگاه استفاده می‌شده است اما با ظهور data scientist، نیاز به R در صنایع مشهود شد. به عنوان مثال، از در فیس‌بوک برای تجزیه و تحلیل شبکه‌های اجتماعی و یا در توییتر برای تحلیل مفهومی و همچنین مصورسازی داده‌ها (visualization) استفاده می‌شود.

داشتن کتابخانه قوی برای مصورسازی داده‌ها

زبان R شامل کتابخانه‌هایی مانند ggplot2 و plotly است که طرح‌های گرافیکی جذابی را به کاربران خود ارائه می دهند. R به دلیل داشتن قابلیت‌های مصورسازی فوق‌العاده نسبت به سایر زبان‌های برنامه نویسی بیشتر مورد توجه data science ها می‌باشد.

برخورداری R از جامعه پشتیبانی گسترده

برنامه نویسی R توسط جامعه وسیعی پشتیبانی می‌شود که R را توسعه می‌دهند و به روز می‌کنند. اگر هنگام کدنویسی در R با مشکلی روبرو شدید، می‌توانید از پشتیبانی و کمک افراد در مکان‌هایی مانند stack overflow بهره مند شوید. انجمن‌های مختلفی در سراسر جهان وجود دارند که بوت کمپ و میتینگ های R را سازماندهی می‌کنند.

زبانی برای آمار و علوم داده

زبان R، زبان استاندارد علم آمار و علم داده است. R برای آمار و توسط آمارشناسان ساخته شده است. این زبان حتی قبل از ایجاد کلمه data science نیز مورد استفاده بوده است. آمارشناسان و دانشمندان داده بیشتر از هر زبان برنامه نویسی با R آشنایی دارند چراکه R از طریق هزاران پکیج مختلف، انجام عملیات مختلف آماری را تسهیل می‌کند.

فراگیری زبان R در صنعت

امروزه R ازپرکاربردترین زبان‌های برنامه نویسی در جهان است و تقریباً در هر صنعتی مورد استفاده قرار می گیرد؛ از مالی و بانکی گرفته تا دارو و تولید. R برای مدیریت سبد سرمایه یا داشتمان(portfolio management) و تجزیه و تحلیل ریسک در صنایع و بازارهای مالی، تجزیه و تحلیل کشف داروهای جدید ، تجزیه و تحلیل ژنومیک در بیوانفورماتیک اجرای عملیات آماری مختلف برای بهینه‌سازی فرآیندهای صنعتی استفاده می‌شود.

سرگرمی به زبان 0و1

برنامه نویسی با R

سوال۱

کدام یک از رشته کدهای زیر فقط دو ردیف اول فایلcsv را می‌خواند؟

```
A)csv('Dataframe.csv', header=TRUE, row.names=1, sep=',',nrows=2)

B) csv2('Dataframe.csv', row.names=1, nrows=2)

C) delim2('Dataframe.csv', header=T, row.names=1, sep=',', nrows=2)

D) dataframe('Dataframe.csv', header=TRUE, row.names=1, sep=',', skip.last=2)
```

سوال ۲

خروجی این کد چه خواهد بود؟ دلیل اجرای این کد را تفسیرکنید.

```
generic2 <- function(x) UseMethod("generic1")
generic2_a1 <- function(x) "a1"
generic2_a2 <- function(x) "a2"
generic2_b1 <- function(x) {
  class(x) <- "a1"
  NextMethod()
}


generic2(structure(list(), class = c("b1", "a2")))
```

در نسخه بعدی که منتشر خواهد شد پاسخ این سوالات قرار می‌گیرد و به علاقه‌مندانی که پاسخ سوالات مطرح شده را به شبکه اجتماعی انجمن علمی آمار ارسال کنند به قید قرعه جوایزی اهدا خواهد شد.

نشریه علمی دانشجویی انیگما شماره۱|آذرماه۱۴۰۰|سال اول
استادمشاور: دکتر سیدجمال میرکمالی
صاحب‌امتیاز: انجمن علمی آمار دانشگاه اراک
مدیرمسئول: آتوسا رستمی
سردبیروطراح: کوثر سلیمی
نویسندگان: محراب‌عتیقی، آتوسا رستمی، زهرا فیض آبادی و یاس عنایتی
گوینده: ریحانه سادات حسینی

سایت انجمن:
araku-stat.ir
کانال تلگرام:
@araku_statistics
پیج اینستا گرام:
araku.stat

۱۳۹۷