


دانشکده‌ی علوم ریاضی

تحلیل رگرسیون

تمرین : دسته‌بندی

مدرّس: دکتر محمد صفدری *مهلت تحویل ۲۶ اردیبهشت*

(مقدار seed را در همه تمرین‌ها برابر ۱۰۰ قرار دهید.)
در این سری از تمرینات می‌خواهیم تا به پیاده‌سازی روش‌های مختلف دسته‌بندی بپردازیم. داده‌ای که در این تمرین از آن استفاده می‌کنیم، داده‌ی قبولی دانش‌آموزان در یک دانشگاه خارجی است. این داده چهار ستون دارد:

- ستون gre : بیانگر نمره فرد در آزمون gre است که از ۸۰۰ محاسبه می‌شود.
- ستون gpa : بیانگر معدل شخص است که از ۴ محاسبه می‌شود.
- ستون rank : بیانگر درجه مدرسه‌ای است که شخص در دبیرستان در آن تحصیل کرده است. مقادیر این ستون یکی از اعداد یک تا چهار است که یک بیانگر مدرسه با بیشترین کیفیت و چهار بیانگر مدرسه با کم‌ترین کیفیت است.
- ستون admit : یکی از مقادیر صفر یا یک را می‌گیرد و نشان‌دهنده قبول‌شدن یا نشدن شخص در آن دانشگاه است.

هدف ما در این تمرین توسعه‌دادن مدل‌هایی است که بتوانیم با آن قبول‌شدن یا نشدن شخص در دانشگاه را با دانستن سه متغیر دیگر پیش بینی کنیم.

۱. تقسیم داده به دو بخش آموزش و آزمون

یکی از اولین کارهایی که هنگام کار روی یک مجموعه داده باید انجام دهیم، تقسیم آن به دو بخش آموزش(train) و آزمون(test) است. پس از این تقسیم‌بندی، بخش آزمون را کنار می‌گذاریم و صرفا در انتها برای سنجش نهایی از آن استفاده می‌کنیم.

حال شما نیز باید مجموعه داده را با نسبت چهار به یک، به دو بخش آموزش و آزمون تقسیم‌بندی کنید. توجه کنید که برش‌زدن از ابتدای داده، اگر داده‌ها به صورت تصادفی پخش‌نشده باشند، ممکن است موجب اریب داده‌های آموزش یا آزمون شود به طوری که توزیع آن‌ها مانند توزیع داده اصلی نباشد. لذا بهترین کار این است که این تقسیم‌بندی به صورت تصادفی روی کل مجموعه داده انجام شود؛ مثلا می‌توان ابتدا داده را بر زد، سپس عملیات برش‌زدن از ابتدا را انجام داد. یا این‌که بدون جایگذاری از داده‌ها به تعداد موردنیاز نمونه‌برداری شود.
برای پیاده‌سازی از توابع آماده R مثل تابع sample نیز می‌توانید استفاده کنید.
داده‌ها را در دو دیتافریم به نام‌های train و test ذخیره کنید.

۲. خواندن داده و بررسی اولیه

طبیعتا هنگامی که با داده‌ای برخورد می‌کنیم، اولین کار(بعد از تقسیم‌بندی) بررسی دقیق داده است تا بتوانیم نسبت به داده دیدگاه خوبی پیدا کنیم.(تحلیل اکتشافی داده)
برای این کار ابتدا با دستور summary خلاصه‌ای از داده را نمایش دهید.(توجه کنید که همه این کارها باید روی داده آموزش انجام شود.)
سپس نمودارهای زیر را رسم کنید:

- نمودار توزیع نوع مدارس
- نمودار توزیع gpa
- نمودار توزیع نمره gre
- نمودار توزیع پذیرش و عدم پذیرش
- نمودار نقطه‌ای نمرات gpa و gre (رنگ نقاط بیانگر پذیرش یا عدم پذیرش، و اندازه نقاط بیانگر نوع مدرسه باشد)

هم‌چنین میزان هم‌بستگی تک‌تک متغیرهای پیش‌بینی را با متغیر هدف به دست آورید.
پس از این مراحل، چه نکاتی درباره داده به نظرتان می‌آید که در مدل‌سازی می‌تواند کمک‌کننده باشد؟

۳. KNN

اولین مدلی که روی داده امتحان می‌کنیم، مدل KNN است. با استفاده از تابع knn پکیج class ، مدل KNN را به ازای $k = 1$ روی داده‌های آزمون اجرا کنید. سپس دقت(accuracy) مدل را با توجه به برچسب‌های حقیقی، حساب کنید.
به ازای k های مختلف(از یک تا پانزده) مدل را بسازید و هر دفعه دقت مدل را روی داده‌های آزمون حساب کنید و آن را چاپ کنید.
بیشترین دقت مدل به ازای کدام k به دست می‌آید؟
برای بهترین مدل به دست‌آمده، موارد زیر را حساب کنید(برای محاسبه این موارد مجاز نیستید از توابع آماده استفاده کنید):

- True Positives
- True Negatives
- False Positives
- False Negatives
- Precision
- Recall

هر کدام از موارد بالا، شاخصی برای سنجش میزان عملکرد مدل است.
چرا شاخص دقت(accuracy) برای سنجش مدل کافی نیست و لازم است از شاخص‌هایی مثل Precision و Recall نیز استفاده شود؟(با مثال توضیح دهید)
آزمون GRE امروزه مقیاس دیگری دارد و حداکثر نمره آن به جای ۸۰۰ برابر ۳۴۰ است. مقیاس نمرات GRE را تغییر دهید به نحوی که با مقیاس جدید تطبیق داشته باشد و مجددا مدل نهایی را روی داده‌ها اجرا کنید. آیا نتایج با مرحله قبل(قبل تغییر مقیاس) مشابه است؟ دلیل مشاهده چنین نتیجه‌ای چیست؟ آیا این ویژگی KNN برای چنین مسئله‌ای مطلوب است؟

۴. Logistic Regression

حال مدل رگرسیون لاجستیکی را روی داده‌های آموزش، برازش دهید.(همه متغیرها را به عنوان متغیرهای پیش‌بینی در نظر بگیرید)
سپس این مدل را روی داده‌های آزمون اعمال کنید تا احتمالی به هر کدام از آن‌ها نسبت داده شود. حال با در نظر گرفتن آستانه یک دوم، به داده‌ها با احتمال بیش از آستانه برچسب یک و به بقیه داده‌ها برچسب صفر دهید و دقت مدل را محاسبه کنید. این کار را به ازای آستانه‌ها یک دهم، دو دهم،.... تا نه دهم انجام دهید و دقت مدل را در هر حالت حساب کنید. بیشترین دقت به ازای کدام آستانه به دست می‌آید؟ این دفعه مدل را صرفا با یک متغیر آموزش دهید(با توجه به نتایج بخش دو، متغیر را انتخاب کنید) و به ازای بهترین آستانه به دست‌آمده در مرحله قبل، این بار هم دقت مدل را حساب کنید و آن را با حالت قبل مقایسه کنید. نتیجه به دست آمده چه معنایی دارد؟

در پناهِ لطفِ «او»، سالم باشید :)