تمرین سری دوم درس یادگیری ماشین

تاریخ ارسال: ۱۵ آذر ۹۶

تاریخ تحویل: ۲۶ آذر ۹۶

نمره: ۲ تا ۲,۵ نمره از ۲۰ نمره پایان ترم

فولدری به نام ML_Assignment02_name1_name2 بسازید که name1 و name2 نام افراد گروه هستند. در این فولدر سه فولدر به نام های ML_Problem11_name1_name2 و ML_Problem12_name1_name2 و ML_Problem13_name1_name2 بسازید و پاسخ سه مسئله را به ترتیب در این فولدرها قرار دهید. دقت کنید در تمام فایل‌های کد و گزارش ها اسامی افراد تیم بصورت کامنت در بالای پاسخ‌ها عنوان شود. در نهایت فولدر اصلی را فشرده کنید و به آدرس taherian.khu@gmail.com ارسال کنید. فراموش نکنید که عبارت ML_Assignment02_name1_name2 (با جایگزینی نام افراد گروه) را در subject ایمیل بیاورید.

سوال:

در سال ۲۰۰۶، سه پژوهشگر تحقیقاتی را در ایالت میشیگان امریکا انجام دادند تا این فرضیه را بررسی کنند که یکی از دلایلی که باعث رای دادن افراد در انتخاب می‌شود، عوامل اجتماعی و فشارهای بیرونی است. محققان ۳۴۴۰۰۰ رای را به تصادف در گروه‌های مختلف قرار دادند. حدود ۱۹۱۰۰۰ رای در گروه کنترل قرار گرفتند و بقیه رای‌ها در ۴ دسته مختلف قرار گرفتند. این ۵ دسته متناظر با ۵ متغیر باینری در مجموعه داده‌ها هستند.

۱. گروه وظیفه شهروندی (متغیر civicduty) : به افراد این دسته تنها یک نامه ارسال شد با این متن که (وظیفه شهروندی خود را انجان دهید. رای دهید!)
۲. گروه دوم (متغیر howthorne) : به افراد این گروه، نامه وظیفه شهروندی ارسال شد بعلاوه یک پیام اضافه با این محتوا که (شما مورد مطالعه قرار دارید) و به آن‌ها اطلاع داده شد که رای یا عدم رای آن‌ها ثبت عمومی خواهد شد.
۳. گروه سوم (متغیر self): پیام وظیفه شهروندی را دریافت کردند بعلاوه سابقه رای دادن اخیر افراد آن خانه و این پیام که بعد از انتخابات پیام دیگری به همراه سوابق به روز رسانی شده آن‌ها برایشان ارسال خواهد شد.
۴. گروه چهارم (متغیر neighbors) : به افراد این گروه هر آنچه را که به افراد گروه ۳ فرستاده شده بود، ارسال شد. با این تفاوت که علاوه بر سوابق رای افراد آن خانه، سوابق رای همسایگان هم ارسال شد – بیشینه کردن فشار اجتماعی.
۵. گروه کنترل (متغیر control) : به افراد این گروه هیچ چیزی ارسال نشد و در واقع افراد این دسته، نماینده وضعیت معمول رای دادن بودند.

سایر متغیرها عبارتند از:

جنسیت (sex) : ۰ برای مرد و ۱ برای زن
سال تولد (yof: year of birth)
متغیر هدف رای دادن (voting) : ۱ برای رای دادن و ۰ برای عدم رای دادن

۱. مجموعه داده gerber.csv را بخوانید. چند درصد از افراد در ای مجموعه داده رای داده‌اند؟ کدام یک از چهار دسته بیشترین درصد افراد رای دهنده را دارند؟

۲. از تمام داده‌ها (داده ها را به آموزشی و تست تقسیم نکنید) و ۴ متغیر civicduty, howthorne, self,neighbors به عنوان متغیرهای مستقل استفاده کنید و مدل logistic regression را بسازید. کدام خصیصه‌ها از چهار مهم تشخیص داده شده‌اند؟ از حد آستانه ۰,۳ استفاده کنید (یعنی اگر احتمال حاصل از مدل بالاتر از ۰,۳ بود، پیش بینی شود که فرد رای می‌دهد). دقت (Accuracy) مدل چقدر خواهد بود؟ اگر از حد آستانه ۰,۵ استفاده کنید دقت چقدر خواهد بود؟ دو مدل پایه (همه افراد رای میدهند) و مدل پایه (همه افراد رای نمی‌دهند) را در نظر بگیرید. دقت (Accuracy) این مدل‌ها چیست؟ دقت آن‌ها را با دقت مدل‌های قبل مقایسه کنید.ROC-AUC مدل رگرسیون و مدل‌های پایه را مقایسه کنید. کدام‌یک بهتر هستند؟

۳. از تمام داده‌ها (داده ها را به آموزشی و تست تقسیم نکنید) و ۴ متغیر civicduty, howthorne, self,neighbors به عنوان متغیرهای مستقل استفاده کنید و یک درخت تصمیم CART بسازید (با استفاده از rpart). از "method="class استفاده نکنید -در واقع می‌خواهیک یک درخت رگرسیون بسازیم. می‌خواهیم درخت بسازیم تا درصد افرادی را که رای می‌دهند (احتمال رای دادن) پیدا کنیم. انتظار داریم که در صورتی که گروه‌ها احتمالات رای دادن متفاوت دارند، CART آن‌ها را از هم تفکیک کند. اگر از "method="class استفاده می‌کردیم، CART تنها در صورتی منشعب می‌شد که یکی از گروه‌ها احتمال رای دادن بالای ۵۰٪ می‌داشت. اما در درخت رگرسیون، حتی اگر تمام گروه‌ها احتمال کمتر از ۵۰٪ داشته باشند، انشعاب صورت می‌گیرد. درخت را رسم کنید.

۴. از پارامتر cp=0.0 در ساخت درخت استفاده کنید تا درخت بطور کامل رسم شود. ترتیب انشعاب‌ها به چه صورت است؟

۵. تنها با استفاده از درخت CART مشخص کنید که چند درصد از گروه civic duty رای دادند؟

۶. متغیر sex را به خصیصه‌ها اضافه کنید. قرار دهید cp = 0.0. به موقعیت و اهمیت این خصیصه در درخت توجه کنید. در گروه کنترل، احتمال رای دادن مردان بیشتر است یا زنان؟ در گروه civic duty چطور؟

۷. تنها روی گروه کنترل تمرکز کنید. با استفاده از تنها یک خصیصه (control) یک درخت رگرسیون بسازید. سپس یک درخت دیگر با استفاده از دو خصیصه control, sex بسازید. در ساخت هر دو درخت قرار دهید cp = 0.0. در درخت اول، قدرمطلق تفاضل احتمال رای دادن بین گروه کنترل و افراد خارج گروه چیست؟ (تا حداقل ۶ رقم اعشار دقت محاسبه کنید).

abs(control prediction – non-control prediction)

اکنون با استفاده از درخت دوم، مشخص کنید که چه جنسیتی بیشتر تحت تاثیر عدم عضویت در گروه کنترل بوده است؟

۸. اکنون به logistic regression بازگردید. یک مدل با استفاده از دو خصیصه control, sex بسازید. ضریب sex چیست؟ این ضریب چه معنایی دارید؟ همانطور که دیدید درخت رگرسیون احتمالات را برای هر یک از ۴ وضعیت زیر محاسبه کرد: (Man, Not Control), (Man, Control), (Woman, Not Control), (Woman, Control) اما مدل رگرسیون نمی‌تواند رخداد همزمان (Woman, Control) را در نظر بگیرد. قدر مطلق تفاضل بین خروجی مدل رگرسیون و درخت تصمیم برای (Woman, Control) چیست؟ (تا ۵ رقم اعشار). همانطور که می‌بینید این تفاضل برای این مجموعه داده زیاد نیست اما به هر حال وجود دارد. یک متغیر جدید (ترکیب متغیرهای sex, control) به logistic regression اضافه کنید بطوریکه ۱ باشد اگر فرد زن و در گروه کنترل باشد.

```
LogModel2 = glm(voting ~ sex + control + sex:control, data=gerber, family="binomial")
```

ضریب متغیر جدید چگونه به خروجی ارتباط دارد؟ در این حالت قدر مطلق تفاضل بین خروجی مدل رگرسیون و درخت تصمیم برای (Woman, Control) چیست؟ (تا ۵ رقم اعشار).