دانشکده‌ی علوم ریاضی

تحلیل رگرسیون

تمرین : رگرسیون خطی

مدرّس: دکتر محمد صفدری

مهلت تحویل ۲۹ فروردین

- پاسخ‌های خود را در قالب یک فایل فشرده در سامانه‌ی درس‌افزار شریف ارسال نمایید.
- در پوشه‌ی پاسخ لازم است به ازای هر سوال یک پوشه ایجاد کنید و کدهای مربوطه را در آن قرار دهید.
- در پوشه‌ی مربوط به هر سوال پوشه‌ای حاوی توضیحات تکمیلی مختصر به همراه نمودارهای مطلوب قرار دهید.
- پیش از اجرای هر بخش، seed تصادف را برابر ۱۰۰ قرار دهید.
- سوالات خود پیرامون تمرینات را با دستیاران درس مطرح کنید.

۱ رگرسیون خطی تک متغیره

ـ مجموعه داده‌ی خانه‌ها [۱] را از سامانه دریافت کنید.

ـ آماره‌های میانه، میانگین، و انحراف از معیار را برای هریک از ستون‌ها محاسبه و چاپ کنید. بر این اساس سعی کنید داده‌های پرت را شناسایی و حذف کنید.

ـ یک متغیر توضیح دهنده‌ی مناسب برای قیمت معرفی کنید. دلیل انتخاب خود را توضیح دهید.

ـ مدل رگرسیون خطی تک متغیره برای پیش‌بینی قیمت با استفاده از متغیر پیشنهادی خود برازش کنید و نتیجه را گزارش کنید.. خط رگرسیونی را به همراه داده‌ها تصویر کنید. آماره‌ی R^2 را گزارش کنید.

ـ مانده [۲] را محاسبه نمایید. و آن را برای مقادیر مختلف متغیر توضیح دهنده رسم کنید.

ـ مدل رگرسیونی خطی تک متغیره برای پیش‌بینی مانده برازش کنید و نتیجه را گزارش کنید. معنی‌دار بودن ضرایب این مدل را با استفاده از دانش نظری خود توضیح دهید.

[۱] houses.csv

[۲] Residual

- مدل رگرسیونی خطی تک متغیره برای پیش‌بینی مانده به توان ۲ با استفاده از متغیر پیشنهادی خود تخمین بزنید و نتیجه را گزارش کنید. خط رگرسیونی را به همراه داده‌ها تصویر کنید.
- مشاهده‌ی خودتان در قسمت قبل را به طور شهودی (با توجه به موضوع مسئله) توجیه کنید. راجع به واریانس ناهمسانی [۳] جستجو کنید و نتیجه را کوتاه توضیح دهید.

۲ رگرسیون خطی چند متغیره

تمرین ۱۰ از فصل ۳ کتاب درس را انجام دهید.

۳ متغیر dummy

- مجموعه داده‌ی تحصیل [۴] را از سامانه دریافت کنید.
- متغیرهای dummy را مشخص کنید.
- مدت زمان تحصیل [۵] را در مقابل شاخص bytest رسم کنید. در این نمودار، هر گروه از متغیر نژادی رنگ پوست [۶] یک رنگ را به خود اختصاص دهد. تابعی به طور جزئی خطی [۷] برای هر گروه نژادی در این نمودار بیاورید.
- نمودار قسمت قبل را برای هر گروه از متغیر جنسی [۸] رسم کنید.
- نمودار قسمت قبل را برای هرگروه از دو متغیر تحصیلات والدین [۹] رسم کنید. دقت کنید که در این بخش، لازم است چهار گروه در یک نمودار ظاهر شوند که هریک نشانگر یک ترکیب از تحصیل یا عدم تحصیل والدین باشند.
- کلاس متغیرهای dummy را به factor تغییر دهید.
- مدل رگرسیونی چندمتغیره برای پیش‌بینی مدت زمان تحصیل با استفاده از باقی متغیرها برازش کنید.
- ضرایب معنادار را تفسیر کنید و با شهود خود مقایسه کنید. سطح اطمینان ۹۵ درصد قابل قبول است.
- با برازش مدل رگرسیونی برای پیش‌بینی قدر مطلق مانده، صحت مدل اولیه را ارزیابی نمایید.

[۳] Heteroscedasticity
[۴] edu.xls
[۵] ed
[۶] black
[۷] partially linear
[۸] female
[۹] dadcoll, momcoll

- متغیر dummy جدیدی تعریف کنید که مقدار آن ۱ است اگر دانش‌آموز حداکثر ۱۰ مایل با یک کالج فاصله داشته و صفر از اگر فاصله تا نزدیک‌ترین کالج بیش از ۱۰ مایل بوده باشد. در ستون فاصله [۱۰] به ازای هر ۱۰ مایل ۱ واحد ثبت شده است. برای اطلاعات بیشتر راجع به داده به راهنمای آن مراجعه نمایید.

- با استفاده از دستورات group_by و group_map در هر سطح از متغیر dummy جدید، یک مدل رگرسیونی برای پیش‌بینی مدت زمان تحصیل با استفاده از باقی متغیرها بجز فاصله برازش کنید. نتیجه‌ی دو برازش را مقایسه کنید و فرض یکسان بودن هریک از ضرایب در این دو گروه را در سطح اطمینان ۹۵ درصد بیازمایید. شاید لازم باشد بکارگیری دستور group_map را جستجو کنید.

۴ توسیع‌های رگرسیون

- مجموعه داده‌ی صدف‌ها [۱۱] را از سامانه‌ی دانشگاه UCI [۱۲] دریافت کنید.

- متغیر جنسیت [۱۳] را حذف کنید.

- متغیر حلقه‌ها [۱۴] را متغیر هدف و باقی متغیرها را پیش‌گو بگیرید.

- تعامل [۱۵] های دو به دوی متغیرهای پیش‌گو و همچنین توابع لگاریتم و نمایی به ازای هر متغیر پیش‌گو برای برازش یک مدل رگرسیونی بکار گرفته شوند. متغیرهای متناظر با ضرایب بی‌معنی (در سطح اطمینان ۹۵ درصد) را حذف کنید.

- مدل رگرسیونی جدیدی با متغیرهای باقیمانده برازش کنید و دوباره متغیرهای متناظر با ضرایب بی‌معنی را حذف کنید. این فرآیند را آنقدر تکرار کنید تا در مدل نهایی تمامی متغیرها ضرایب معنادار داشته باشند. پیشنهاد می‌شود که انجام این فرآیند را در قالب یک حلقه یا با استفاده از تعریف توابع انجام دهید تا کارتان تسهیل شود.

- آماره‌ی R^2 را برای مدل نهایی محاسبه کنید.

[۱۰] dist
[۱۱] abalone.data
[۱۲] UCI machine learning repository
[۱۳] Sex
[۱۴] Rings
[۱۵] interaction