تمرین سری سوم – تحلیل آماری کاربردی

نکات مهم:

- برای تحویل تمارین گزارشی با استفاده از RMarkdown ایجاد نموده که در آن به صورت کامل صورت سوال، کد مربوطه، خروجی کد و توضیحات مربوطه آورده شود و سپس در سامانه آموزش مجازی بارگذاری شود. برای بارگذاری، ابتدا فایل word خروجی RMarkdown را تهیه کرده و سپس آن را pdf نموده و فایل pdf را به همراه فایل Rmd در سامانه بارگذاری کنید. صورت سوال را می توانید به صورت یک عکس در گزارش بیاورید. نام فایل ها حتما شامل شماره تمرین و شماره دانشجویی باشد (به عنوان مثال، HW2-9913418023).
- دقت کنید که حتما برای پاسخ به تمامی قسمت های سوالات کدی نوشته و پاسخ را به صورت خروجی کد بیاورید.

مساله ۱:

برای انجام این مساله مجموعه داده های برآورد، اعتبار سنجی و آزمون از یک متغیر پیش بینی کننده X و یک متغیر پاسخ Y در اختیار شما قرار خواهد گرفت. ابتدا این مجموعه داده ها را در نرم افزار R بارگذاری کرده و سپس به سوالات زیر پاسخ دهید:

- ۵ مدل چند جمله ای از مراتب ۱، ۳، ۵، ۷ و ۹ بر روی مجموعه داده برآورد برای پیش بینی Y با استفاده از متغیر X برازش دهید. برای هر مدل، مقدار RMSE مجموعه داده برآورد و اعتبار سنجی را محاسبه کنید. نتایج را با استفاده از یک نمودار که در آن مقادیر RMSE برآورد و اعتبارسنجی بر حسب مرتبه چندجمله ای آمده است گزارش کنید. بهترین مدل را انتخاب کنید. کدام یک از مدلها احتمالا دچار بیش برازش و کدام یک دچار کم برازش شده است؟
- ۵ مدل KNN با مقادیر k مساوی ۱، ۱۱، ۲۱، ۳۱ و ۴۱ بر روی مجموعه داده برآورد برای پیش بینی Y با استفاده از متغیر X برازش دهید. برای هر مدل، مقدار RMSE مجموعه داده برآورد و اعتبار سنجی را محاسبه کنید. نتایج را با استفاده از یک جدولی که توسط نرم افزار تولید شده و ستونهای آن مقادیر k و RMSE برآورد و اعتبارسنجی هستند، گزارش کنید. بهترین مدل را انتخاب کنید. کدام یک از مدلها احتمالا دچار بیش برازش و کدام یک دچار کم برازش شده است؟
- ۵ مدل درخت رگرسیون با مقادیر پارامتر cp مساوی ۰، ۰,۰۰۱، ۰,۰۱، ۰,۱ و ۱ بر روی مجموعه داده برآورد برای پیش بینی Y با استفاده از متغیر X برازش دهید. مقدار پارامتر minsplit را مساوی ۲ قرار دهید. برای هر مدل، مقدار RMSE مجموعه داده برآورد و اعتبار سنجی را محاسبه کنید. نتایج را با استفاده از یک جدولی که توسط نرم افزار تولید شده و ستونهای آن مقادیر cp و RMSE برآورد و اعتبارسنجی هستند، گزارش کنید. بهترین مدل را انتخاب کنید. کدام یک از مدلها احتمالا دچار بیش برازش و کدام یک دچار کم برازش شده است؟
- یک نمودار Y بر حسب X بر اساس مجموعه داده برآورد رسم کنید. سپس بهترین مدل انتخاب شده چند جمله ای، KNN و درخت رگرسیون بر اساس قسمت های قبل را به صورت خط یا منحنی به آن اضافه کنید. از رنگها و نوع خطوط مختلف برای هر یک از مدلها استفاده کنید. همچنین راهنمایی (legend) برای مشخص کردن هر یک از مدلها به نمودار اضافه کنید.
- از بین مدلهای برازش داده شده در قسمت های قبل بهترین مدل را انتخاب کنید و آن را بر روی مجموعه داده آموزشی (برآورد + اعتبارسنجی) برازش دهید. سپس خطای تعمیم این مدل انتخاب شده را با استفاده از مجموعه داده آزمون محاسبه کنید.

تمرین سری سوم – تحلیل آماری کاربردی

مساله ۲:

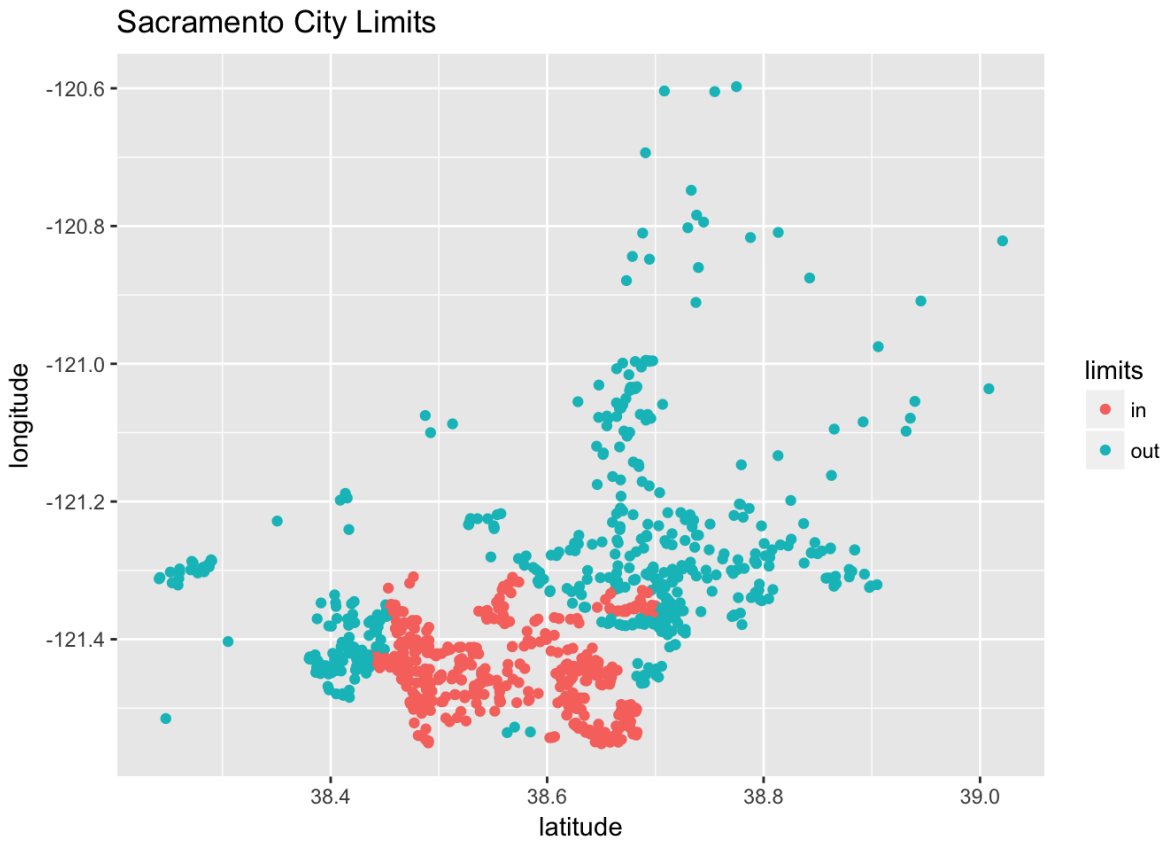
برای این مســاله، از مجموعه داده "Sacramento" از کتابخانه "caret" اســتفاده می شــود. ابتدا این مجموعه داده را بارگذاری کنید و help نرم افزار R را در رابطه با این مجموعه داده و توضــیحات مربوط به آن را بخوانید. هدف این مســاله پیش بینی متغیر قیمت (price) خانه است.

قبل از مدل سازی، آماده سازی داده ها به شرح زیر انجام می شود. به جای استفاده از متغیرهای city یا zip که در مجموعه داده وجود دارد، برای ساده سازی یک متغیر (limits) ایجاد می کنیم که نشان می دهد خانه در محدوده شهر ساکرامنتو قرار دارد یا خیر. متغیر baths را نیز از مجموعه داده حذف کنید.

```
data(Sacramento)
sac_data = Sacramento
sac_data$limits = factor(ifelse(sac_data$city == "SACRAMENTO", "in", "out"))
sac_data = subset(sac_data, select = -c(city, zip, baths))
```

نمودار طول جغرافیایی در مقابل عرض جغرافیایی به ما این حس را می دهد که محدوده شهر کجاست.

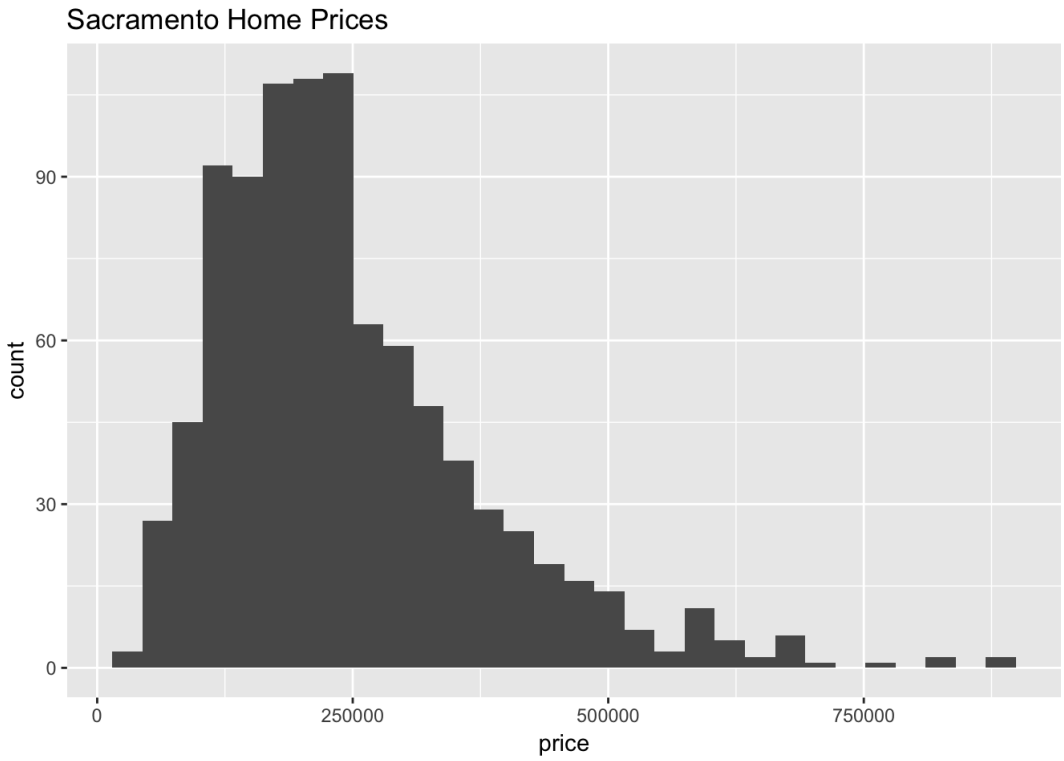
```
qplot(y = longitude, x = latitude, data = sac_data,
      col = limits, main = "Sacramento City Limits ")
```



تمرین سری سوم – تحلیل آماری کاربردی

شما می توانید، از روشهای EDA برای مصور سازی و کشف رابطه بین متغیرها استفاده کنید. فعلا اینجا فقط هیستوگرام متغیر price رسم می شود.

```
qplot(x = price, data = sac_data, main = "Sacramento Home Prices")
```



حال مجموعه داده را به صورت تصادفی (با دستور set.seed(42)) به ۳ قسمت برآورد، اعتبارسنجی و آزمون تقسیم کنید و با استفاده از آنها به سوالات زیر پاسخ دهید.

- حداقل ۴ مدل رگرسیون خطی برای پیش بینی متغیر price ایجاد کنید. بدین منظور، می توانید از مدلهایی با تمامی متغیرها، اضافه کردن اثرات متقابل، مدل های چند جمله ای و الگوریتم های انتخاب متغیر ا ستفاده کنید. جدولی ایجاد کنید که در ستونهای آن نام مدل، متغیرهای استفاده شده در مدل (با استفاده از دستورات نرم افزار)، RMSE برآورد و اعتبارسنجی باشد. بهترین مدل را انتخاب کنید. آیا متغیرهای وابسته به مکان خانه (limits, latitude, longitude) در پیش بینی متغیر price موثر به نظر می رسند؟
- یک بار با نرمالسازی متغیرهای کمی (تنظیم ۱) و یک بار دیگر بدون نرمالسازی (تنظیم ۲)، مدلهای KNN را با مقادیر مختلف k از ۱ تا ۱۰۰ برازش دهید و برای هر یک از مدلها (۲۰۰ مدل)، مقادیر RMSE اعتبارسنجی را محاسبه کنید. نتایج را در قالب یک نمودار که مقدار RMSE اعتبارسنجی را بر حسب تابعی از k نشان می دهد گزارش کنید. تنظیم ۱ و ۲ را با رنگ و نوع خط مختلف در نمودار متمایز سازید (برای نمودار legend نیز تهیه کنید). آیا نرمالسازی متغیرهای کمی، بهبودی در نتایج مدل‌ها ایجاد می کند؟ بهترین مدل را انتخاب کنید.

تمرین سری سوم – تحلیل آماری کاربردی

- ۱۰ مدل درخت رگرسیون با ۵ مقدار مختلف cp و ۲ مقدار مختلف minsplit (مقادیر ۵ و ۲۰) برازش دهید. در قالب یک جدول، مقادیر RMSE اعتبار سنجی این مدل ها را بر حسب مقدار cp و minsplit گزارش کنید. بهترین مدل را انتخاب کنید و درخت رگرسیون مربوط به آن را رسم و تفسیر کنید.
- در قالب یک نمودار ۳ قسمتی، نمودار مقادیر پیش بینی هر یک از مدلها در برابر مقادیر واقعی (Predicted vs Actual plot) مجموعه اعتبار سنجی را رسم و تفسیر کنید. برای تحلیل بهتر نمودارها خط Y=X را به هر یک از آنها اضافه کنید.
- بهترین مدل های انتخاب شده در ۳ قسمت قبل را از نظر RMSE اعتبارسنجی مقایسه کرده و بهترین مدل را انتخاب کنید. بهترین مدل انتخاب شده را ابتدا بر روی کل مجموعه آموزشی (برآورد+اعتبار سنجی) برازش داده و سپس با استفاده از آن، مقدار RMSE مجموعه آزمون را محاسبه کنید.
- **(تشویقی)** آیا با تغییر عدد کد set.seed جواب قسمت های قبل تغییر خواهد کرد؟ به عنوان مثال، عدد را از ۴۲ به ۴۳۲ تغییر دهید و نتایج را مقایسه کنید.

موفق باشید– شادمان