به نام او..

پروژه اول درس یادگیری ماشین آماری رگرسیون خطی

شرح مجموعه‌ی دادگان:

برای این پروژه دو مجموعه داده پیوست شده‌است.

مجموعه داده‌ی اول به صورت مصنوعی تولید شده‌است. این مجموعه داده شامل ۵۰۰ داده است که هر داده دارای ۸ ویژگی و یک متغیرِ هدف است. متغیر هدف با یک ترکیب خطی از ۸ ویژگی به همراه نویز گاوسی تولید شده‌است. ۴۰۰ داده‌ی ابتدایی را به عنوان داده‌ی آموزشی و ۱۰۰ داده‌ی انتهایی را به عنوان داده‌ی آزمایشی در نظر بگیرید.

مجموعه داده‌ی دوم از نمرات ۲۴۰ دانشجو در ۷ درس تشکیل شده‌است. هدف استفاده از این مجموعه داده، آموزش یک مدل رگرسیون به منظور تخمین نمره‌ی درس هفتم هر فرد با استفاده از نمرات ۶ درس دیگرش است. ۲۰۰ داده‌ی اول را به عنوان داده‌ی آموزشی و ۴۰ داده‌ی انتهایی را به عنوان داده‌ی آزمایشی در نظر بگیرید.

بخش اول، مجموعه داده‌ی مصنوعی:

الف) نمودار Scatter مربوط به هر یک از ۸ ویژگی در کنار متغیر هدف را رسم کنید. به طور شهودی در مورد نحوه‌ی ارتباط هر کدام از ویژگی‌ها با متغیر هدف بحث کنید.

ب) به ازای هر یک از ویژگی‌ها، مدلی رگرسیون خطی ساده‌ای ارائه دهید که متغیر هدف را پیش‌بینی کند. برای تخمین پارامترهای β_0 و β_1 مدل از تخمینگر Least Square استفاده کنید. خطوط تخمین زده شده را در کنار داده‌ها رسم کنید. برای هر ۸ مدل طراحی شده، تخمین پارامترهای β_0 و β_1 را به همراه خطای استاندارد[1]آنها در جدولی گزارش کنید. برای هر یک از مدل‌ها تخمینی از σ^2 (واریانس ϵ) به دست آوردید. معیار RSS و R^2 را برای مدل‌های طراحی شده به ازای داده‌های آموزشی و آزمایشی گزارش کنید.

ج) پس از انتخاب بهترین ویژگی در قسمت ب، در یک روند رو به جلو[2]برای ۷ مدلی که با افزودن ویژگی دوم به ویژگی انتخاب شده ساخته می‌شوند، معیار AIC را به دست آورید. در صورت بهبود معیار AIC نسبت به حالت اولیه، متغیر دوم را به مدل اضافه کرده و معیارهای RSS و R^2 را گزارش کنید.

[1] Standard Error
[2] Forward

د) فرآیند رو به جلوی قسمت ج را تا زمانی که افزودن ویژگی به مدل باعث بهبود معیار AIC می‌شود ادامه دهید. معیارهای RSS و R^2 را برای مدل‌های ساخته شده گزارش کنید.

ه) با استفاده از تخمینگر Least Square مدل رگرسیون خطی ارائه دهید که از تمامی متغیرها استفاده می‌کند. معیار خطای Leave one out cross validation را محاسبه کنید. (برای محاسبه‌ی این معیار از هر دو روش بیان شده در کتاب مرجع درس استفاده کنید. منظور محاسبه‌ی معیار با n بار آموزش مدل و محاسبه‌ی معیار با ۱ بار آموزش مدل است.)

و) در یک فرآیند رو به عقب[3]معیار خطای Leave one out cross validation را برای ۸ مدلی که با حذف هر یک از ویژگی‌ها به دست می‌آیند گزارش کنید. روند رو به عقب را تا حذف ۷ ویژگی ادامه دهید و نمودار خطا بر حسب تعداد ویژگی‌ها را رسم کنید. بهترین مدل را مشخص کنید. ضمنا نمودار خطای RSS داده‌های آزمایشی را بر حسب تعداد ویژگی رسم کنید.

ز) بهترین مدل به دست آمده در قسمت و را در نظر بگیرید. هدف از این بخش، مشاهده‌ی تاثیر تعداد داده‌ی آموزشی بر دقت مدل است. برای این منظور، بر حسب تعداد مختلف داده‌های آموزشی، مجددا پارامترهای مدل را آموزش دهید و نمودار معیار RSS برای داده‌ها آموزشی و آزمایشی را بر حسب تعداد داده‌های آموزشی به دست آورید. نتیجه را تحلیل کنید.

بخش دوم، مجموعه داده‌ی نمرات:

الف) نمودار Scatter مربوط به هر یک از ۶ ویژگی در کنار متغیر هدف را رسم کنید. به طور شهودی در مورد نحوه‌ی ارتباط هر کدام از ویژگی‌ها با متغیر هدف بحث کنید.

ب) این مجموعه داده دارای مقادیر نامشخص[4]است که با محتوای صفر پر شده‌اند. با استفاده از یک روش انتخابی، مقادیر نامشخص را تکمیل کنید. روش خود را در گزارش بیان نمایید.

ج) روش Lasso امر تخمین پارامترها و انتخاب مدل را به طور همزمان انجام می‌دهد. این روش را بر روی مجموعه داده‌ی دوم اجرا نمایید. (نیازی به پیاده‌سازی روش نیست و می‌توانید از پیاده‌سازی‌ها و توابع موجود استفاده کنید.)

[3] Backward
[4] Missing value

د) با تغییر پارامتر λ که جریمه‌ی بزرگی ضرایب مدل را مشخص می‌کند، مدلهای مختلف را آموزش دهید و نمودار معیار Lasso را بر حسب پارامتر λ رسم کنید. بهترین مدل را مشخص کرده و برای آن مدل معیارهای RSS و R^2 را به ازای مجموعه داده‌ی آموزشی و آزمایشی گزارش کنید.

ه) با استفاده از بهترین مدل به دست آمده، مقدار متغیر هدف را برای مجموعه داده‌ی بدون برچسب ارائه شده به دست آورید. فایل خروجی مربوطه را نیز در فایل نهایی تحویل دهید.

فرمت گزارش:

گزارش بایستی به زبان فارسی و در قالب فایل PDF باشد. در گزارش تحلیل و نتیجه‌گیری خود را در رابطه با هر بخش به شکل مختصر بیان فرمایید. (در حد یک پاراگراف)

فایل گزارش خود را به شکل « Project1_StdNum.pdf» نامگذاری کنید. (مانند Project1_8931064.pdf)

فرمت کدها:

برای پروژه بایستی پیاده‌سازی در یکی از محیطهای MATLAB، R یا Python تهیه شود.

تمامی کدهای پیاده‌سازی پروژه باید دارای شرح کامل درون کد باشد. (Well Commented)

نحوه تحویل:

فایلهای کد و گزارش خود را که طبق فرمتهای فوق تهیه شده‌اند، در قالب یک فایل فشرده در سایت درس بارگذاری نمایید.

فایل فشرده را به شکل «P1_StdNum» نامگذاری کنید. (مانند P1_8931064)

مهلت ارسال تمرین ساعت ۲۳:۵۵ دقیقه‌ی روز چهارشنبه مورخ ۲۰ بهمن ماه می‌باشد.

پروژه شامل تحویل حضوری نیز خواهد بود.

هر گونه سوال در مورد پروژه را می‌توانید از طریق ایمیل به آدرس AUT.SMLf16@gmail.com بیان نمایید.