پروژه درس یادگیری ماشین آماری

هدف این پروژه آشنایی با مدل‌های رگرسیون آماری است. برای این منظور از یک مجموعه داده مصنوعی و یک مجموعه داده واقعی استفاده می‌شود. آزمایش‌های خواسته شده را به دقت انجام داده و نتایج هر آزمایش را به همراه نتیجه‌گیری و استدلال خود در رابطه با نتایج در گزارش پروژه ارائه کنید.

بخش اول، مجموعه داده مصنوعی:

مجموعه داده Dataset1.csv

این مجموعه داده متشکل از ۵۰۰ نمونه است که هر یک از آنها با ۹ ویژگی مختلف مشخص شده‌اند. از میان این ویژگی‌ها، هشت ویژگی اول متغیرهای مستقل و ویژگی نهم متغیر وابسته (هدف) است. از ۴۰۰ داده ابتدایی برای آموزش و از ۱۰۰ داده بعدی برای اعتبارسنجی مدل استفاده کنید.

الف) نمودار نقطه‌ای (scatter plot) مربوط به هر یک از ویژگی‌های موجود در مجموعه داده به همراه متغیر هدف را رسم کنید. با توجه به نمودار رسم شده، ارتباط هر کدام از ویژگی‌ها با متغیر هدف را مورد بررسی قرار دهید.

ب) به ازای هر کدام از ویژگی‌های موجود، مدل رگرسیون خطی ساده‌ای برای پیش‌بینی متغیر هدف ارایه دهید. پارامترهای β_0 و β_1 مدل با استفاده از کمترین مربعات (Least Squares) تخمین زده و مقادیر حاصل را ذکر کنید. به ازای هر یک از مدل‌های به دست آمده، ابتدا خط پیش‌بینی شده را به همراه داده‌های موجود رسم کنید و سپس به ازای مجموعه داده‌های آموزشی و آزمایشی معیارهای RSS و ضریب تشخیص (R^2) را محاسبه کنید. هم‌چنین برای هرکدام از پارامترهای تخمین زده شده، انحراف معیار متناظر را نیز تخمین زده و ثبت کنید. مقدار تخمین زده شده برای σ^2 را نیز در هر کدام از مدل‌های به دست آمده، محاسبه کنید.

ج) در بخش قبل به ازای هر کدام از متغیرهای مستقل و ویژگی هدف مدل رگرسیون خطی برای پیش‌بینی متغیر هدف به دست آمد. کدام ویژگی بهترین گزینه برای پیش‌بینی متغیر هدف است؟ چرا؟ پس از انتخاب یکی از ویژگی‌ها به عنوان بهترین ویژگی، در یک فرایند رو به جلو ویژگی دوم را به ویژگی انتخابی اول اضافه کنید. در تمامی ۷ حالت به دست آمده معیار AIC را محاسبه کنید. با بررسی تغییر حاصل در معیار AIC ویژگی دوم انتخابی را مشخص کرده و به مدل اضافه کنید. پس از افزودن ویژگی دوم، معیارهای RSS و R^2 را محاسبه کنید.

د) همانند موارد ذکر شده در بخش ج، سایر ویژگی‌های موجود را به توجه به بهبود معیار AIC، به مدل اضافه کنید. در هر کدام از مراحل ویژگی افزوده شده و معیارهای RSS و R^2 را محاسبه کنید. نمودار معیار RSS را در حین افزودن ویژگی‌ها رسم کنید. چه تغییر در این معیار رخ می‌دهد؟

ه) دو بخش ج و د را با معیار BIC تکرار کنید. نتایج حاصل از دو معیار مورد استفاده برای انتخاب مدل را مقایسه کنید.

و) در این مرحله قصد داریم تا با استفاده از تمامی ویژگی‌ها به تخمین هدف بپردازیم. حال در این راستا از Least Square برای تخمین پارامترهای مدل رگرسیون خطی استفاده کرده و مدل را تشکیل دهید. ماتریس واریانس–کوواریانس پارامترهای β را به دست آورید. هم‌چنین انحراف معیار تخمینی برای پارامترهای β را نیز به دست آورید. تخمین غیربایاس‌شده σ^2 را به دست آورده و ثبت کنید. برای مدل رگرسیون خطی ارایه شده، معیار خطا را با استفاده از Leave one out cross validation به دست آورید. برای محاسبه این معیار از دو روش ذکر شده در کتاب (n بار آموزش مدل و یک بار آموزش مدل) استفاده کنید.

ز) با توجه به بخش قبل، مدلی متشکل از ۸ ویژگی برای پیش‌بینی متغیر هدف داریم. حال در یک فرایند رو به عقب در هر مرحله یک ویژگی را با استفاده از معیار Leave one out cross validation حذف کنید، تا جایی که تنها یک متغیر باقی بماند. در هنگام حذف اولین ویژگی، معیار Leave one out cross validation را در ازای حالت‌های ممکن ذکر کرده و دلیل انتخاب ویژگی نهایی را ذکر کنید. در روند حذف متغیرها تا رسیدن به تک متغیر معیار RSS را محاسبه کرده و نمودار آن را رسم کنید. این معیار در حین حذف چه تغییری دارد؟

ح) بهترین مدل به دست آمده از بخش قبل را در نظر بگیرید. با تغییر درصد داده‌های آموزش و آزمایش، پارامترهای مدل را مجددا آموزش دهید. سپس خطای RSS حاصل از مدل بر روی داده‌های آموزش و آزمایش را به دست آورده و نمودار مربوط را رسم کنید. تغییرات RSS را تحلیل کنید.

بخش دوم، مجموعه داده نمرات:

مجموعه داده dataset2.csv

این مجموعه داده متشکل از ۶ ویژگی و متغیر هدف است. ۲۰۰ داده ابتدایی را به عنوان داده آموزشی و ۴۰ داده انتهایی را به عنوان داده آزمایشی در نظر بگیرید.

الف) نمودار نقطه‌ای مربوط به هر یک از ویژگی‌های موجود در مجموعه داده را به همراه متغیر هدف رسم کنید. با توجه به نمودار رسم شده، ارتباط هر کدام از ویژگی‌ها با متغیر هدف را مورد بررسی قرار دهید.

ب) این مجموعه داده دارای مقادیر نامشخص است. این مقادیر با عدد صفر مقداردهی شده‌اند. روشی برای پر کردن مقادیر نامشخص ارایه کنید. پس از پر کردن مقادیر نامخشص مجددا نمودار نقطه‌ای را همانند بخش الف رسم کرده و تغییرات حاصل را تحلیل کنید.

ج) روش lasso را بر روی مجموعه‌ی داده اجرا نمایید. (نیازی به پیاده‌سازی این روش نبوده و می‌توانید از توابع آماده موجود استفاده کنید.)

د) با تغییر پارامتر λ مدل‌های مختلف را آموزش دهید. 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 داده آموزشی و آزمایشی محاسبه کنید.

ه) با استفاده از بهترین مدل به دست آمده، مقدار متغیر هدف را برای مجموعه داده‌ی بدون برچسب dataset2_Unlabeled ارائه شده به دست آورید. فایل خروجی مربوطه را نیز در فایل نهایی تحویل دهید.

و) مجموعه داده‌ی dataset2_extended با اندکی تغییر در داده‌های قبلی و افزودن ستونی جدید به ابتدای داده‌ها به دست آمده است. روش lasso را با پارامتر λ برابر با ۰,۰۰۱ بر روی این مجموعه داده اجرا کنید. مقادیر β حاصل را مورد بررسی قرار دهید.

نکته : کلیه روابط ریاضی استفاده شده در هر یک از بخش‌های مختلف پروژه را به شکل دقیق ذکر نمایید.

قالب گزارش:

گزارش پروژه بایستی به زبان فارسی و در قالب فایل PDF باشد. در گزارش تحلیل و نتیجه‌گیری خود در رابطه با هر بخش را بیان کنید.

فایل گزارش خود را به شکل Project1_StdNum.pdf نامگذاری کنید.

فرمت کدها:

از یکی از محیط‌های برنامه‌نویسی Matlab، R و یا Python برای پیاده‌سازی پروژه استفاده کنید.

کدهای خود را به تفکیک بخش‌های مختلف آزمایش بنویسید و در کلیه بخش‌ها کامنت کافی قرار دهید.