هدف از این تمرین پیاده‌سازی یک رگرسیون خطی منظم‌شده[1] است. مجموعه داده‌های مورد استفاده در فایل data.txt موجود است. داده‌ها دارای یک ستون x و یک ستون y هستند.

بخش اول، رسم داده‌ها) مجموعه داده را به صورت نموداری که y تابعی از x باشد رسم کنید. (با توجه به یک بعدی بودن فضای ویژگی (X)، این مجموعه داده را می‌توان در یک نمودار ۲ بعدی نشان داد.)

خروجی بخش اول: نمودار مجموعه داده

بخش دوم، انتخاب تابع فرضیه) با مشاهده‌ی نمودار رسم شده در بخش اول، می‌توان مشاهده کرد که استفاده از یک خط مستقیم برای تخمین بسیار ساده‌انگارانه است. به جای استفاده از خط، می‌خواهیم از برازش یک چندجمله‌ای با درجه‌ی بالا بر روی داده‌ها استفاده کنیم. تابع فرضیه برای چندجمله‌ای درجه ۳ و چند جمله‌ای درجه ۶ را مشخص کنید.

خروجی بخش دوم: تابع فرضیه‌ی مربوط به چند جمله‌ای درجه ۳ و درجه‌ی ۶

بخش سوم، انتخاب تابع هزینه) با توجه به کم بودن تعداد داده‌ها به نسبت درجه‌ی چندجمله‌ای مورد استفاده، خطر بیش‌برازش[2] مدل بر روی داده‌ها بسیار زیاد است. برای کاهش این مشکل از روش رگرسیون منظم‌شده استفاده می‌کنیم. پس تابع هزینه که قصد کمینه کردن آن را داریم به صورت زیر در نظر می‌گیریم:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

در این رابطه، m تعداد داده‌های آموزشی، h تابع فرضیه‌ی در نظر گرفته شده، θ_j ها ضرایب چندجمله‌ای و n درجه‌ی چندجمله را نشان می‌دهند.

توضیح دهید در رابطه‌ی بالا، λ چیست و چه تاثیری دارد؟

تاثیر عدم وجود پارامتر θ_0 در قسمت Regularization عبارت بالا را بیان کنید.

خروجی بخش سوم: توضیح تاثیر پارامتر λ و علت عدم وجود θ_0

[1] Regularized linear regression
[2] Over fitting

بخش چهارم، یافتن پارامترهای بهینه با استفاده از روش معادله‌ی نرمال[3] یک روش برای یافتن پارامترهای بهینه‌ی مدل، استفاده از معادله‌ی نرمال است. جواب معادله‌ی نرمال برای رگرسیون خطی منظم‌شده، به صورت زیر است:

$$\theta = (X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix})^{-1} X^T \vec{y}$$

ماتریس موجود در این رابطه، یک ماتریس (n+1)*(n+1) بعدی است. همچنین در این رابطه، X یک ماتریس m*(n+1) بعدی است. ستون اول آن تماما ۱ است و هر یک از ستون‌ها یکی از درجات چندجمله‌ای را مشخص می‌کند. بردار y نیز دربردارنده‌ی خروجی مطلوب برای داده‌هاست.

برای هر دو حالت چند جمله‌ای درجه ۳ و درجه ۶ فعالیت‌های زیر را انجام دهید:

- از معادله‌ی بالا استفاده کرده و θ بهینه را برای سه حالت λ=0، λ=1 و λ=10 را به دست‌آورید.
- نمودار اندازه‌ی بردار θ بر حسب λ های مختلف را رسم کنید. نتیجه را تحلیل کنید.
- چندجمله‌ای به دست آمده را برای سه حالت λ=0، λ=1 و λ=10 در کنار داده‌ها رسم کنید. نتیجه را تحلیل کنید.
- نمودار خطای MSE را بر حسب λ های مختلف رسم کنید. نتیجه را تحلیل کنید.
- نتایج مربوط به حالت چند جمله‌ای درجه ۳ با درجه‌ی ۶ را با یکدیگر مقایسه کنید.

خروجی بخش چهارم: موارد بالا

بخش امتیازی، حل معادله‌ی نرمال برای رگرسیون خطی منظم‌شده) نشان دهید جواب معادله‌ی نرمال برای رگرسیون خطی منظم‌شده که با تابع هزینه‌ی بخش ۳ به دست می‌آید، به صورت رابطه‌ی بیان شده در بخش ۴ خواهد بود. به زبان ساده‌تر، رابطه‌ی بیان شده در بخش ۴ را ثابت کنید.

شیوه‌ی تحویل تمرین: تا ساعت ۲۳:۵۵ روز جمعه ۱۴ اسفند فرصت دارید تا تمرین را در مودل بارگذاری کنید. تمام فایلهای پیاده‌سازی را به همراه فایل pdf مربوط به گزارش تمرین، در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود قرار دهید. (برای مثال 93131130.rar)

در صورت وجود هر گونه سوال می‌توانید از طریق ایمیل با یکی از تدریس‌یاران درس در ارتباط باشید.

MR.Molavi@gmail.com , Marjan.Moodi@gmail.com , NavidFumani@gmail.com

[3] Normal equation