دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

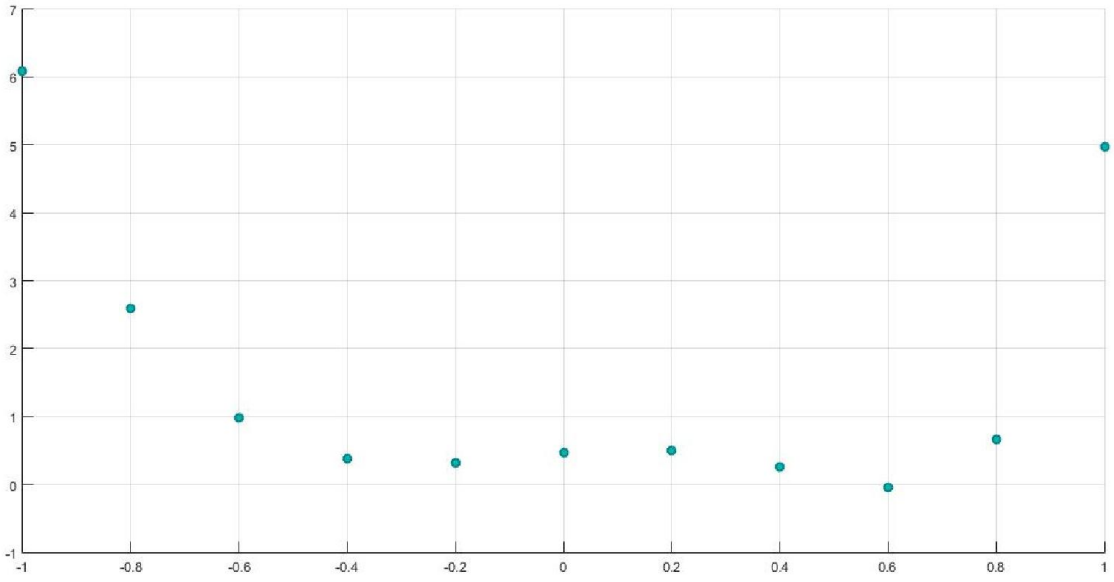
گزارش تکلیف اول درس یادگیری ماشین
پیاده‌سازی یک رگرسیون خطی منظم‌شده

دانشجو:
سید احمد نقوی نوزاد
ش-د: ۹۴۱۳۱۰۶۰

استاد:
دکتر ناظرفرد

بهار ۹۵

۱. رسم داده‌های مسئله



۱- نمودار مجموعه داده‌ها

۲. تابع فرضیه مربوط به چندجمله‌ای درجه‌ی ۳ و درجه‌ی ۶

تابع فرضیه مربوط به چندجمله‌ای درجه‌ی ۳:

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

تابع فرضیه مربوط به چندجمله‌ای درجه‌ی ۶:

$$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5 + \theta_6 x^6$$

۳. انتخاب تابع هزینه

در عبارت زیر:

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

در اینجا $\lambda\sum_{j=1}^{n}\theta_j^2$، عبارت تنظیم‌سازی (regularization term) نامیده شده و پارامتر λ نیز پارامتر تنظیم‌سازی (Regularization Parameter) می‌باشد. کاری که λ انجام می‌دهد برقراری یک tradeoff میان دو هدف متفاوت است؛ هدف اول این است که می‌خواهیم مدل ما با داده‌های آموزشی به خوبی fit شود و هدف دوم این است که می‌خواهیم پارامترهای مدل ما تا حد امکان کوچک باشند تا تخمین ما نسبتا ساده بوده و از overfitting اجتناب نمائیم و یا اثر feature های کم‌اهمیت را کم کرده و یا حتی آن‌ها را حذف کنیم. به عبارتی به ازای λهای کوچک، تمرکز بر روی کم‌کردن RSS (Residual Sums of Errors) یا همان $\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$ بوده و مدل به سمت overfitting رفته و در نتیجه دارای بایاس اندک و واریانس بالا خواهد بود؛ و نیز به

ازای λهای بزرگ، تمرکز بر روی کم‌کردن اندازه‌ی پارامترها بوده و مدل به سمت **underfitting** رفته و در نتیجه دارای بایاس بالا و واریانس اندک خواهد بود.

در اینجا ما پارامتر θ_0 را بنا به سنّت اصطلاحا جریمه ننموده و البته انجام یا عدم انجام این مسئله تغییر چندانی در نتایج نهائی ایجاد نمی‌نماید(بنا به گفته‌ی آقای **Andrew Ng** !!؟)؛ اما مسئله این است که در صورت شمول پارامتر θ_0 و البته بزرگ‌بودن λ در عبارت تنظیم‌سازی، مقدار این پارامتر نیز منقبض (**shrink**) شده و تابع تخمین نهائی ما به سمت $h(x)=0$ متمایل شده و این البته مطلوب ما نمی‌باشد. اما در صورت عدم شمول پارامتر θ_0 و نیز بزرگ‌بودن λ ، مدل ما به سمت حالتی می‌رود که مقدار تابع تخمین به ازای هر کدام از داده‌های آموزشی برابر میانگین مقادیر مطلوب (**target values mean**) می‌باشد ($h(x)=mean(\bar{y})$) و این از حالت تابع تخمین $h(x)=0$ به مراتب بهتر است.

۴. یافتن پارامترهای بهینه با استفاده از روش معادله‌ی نرمال

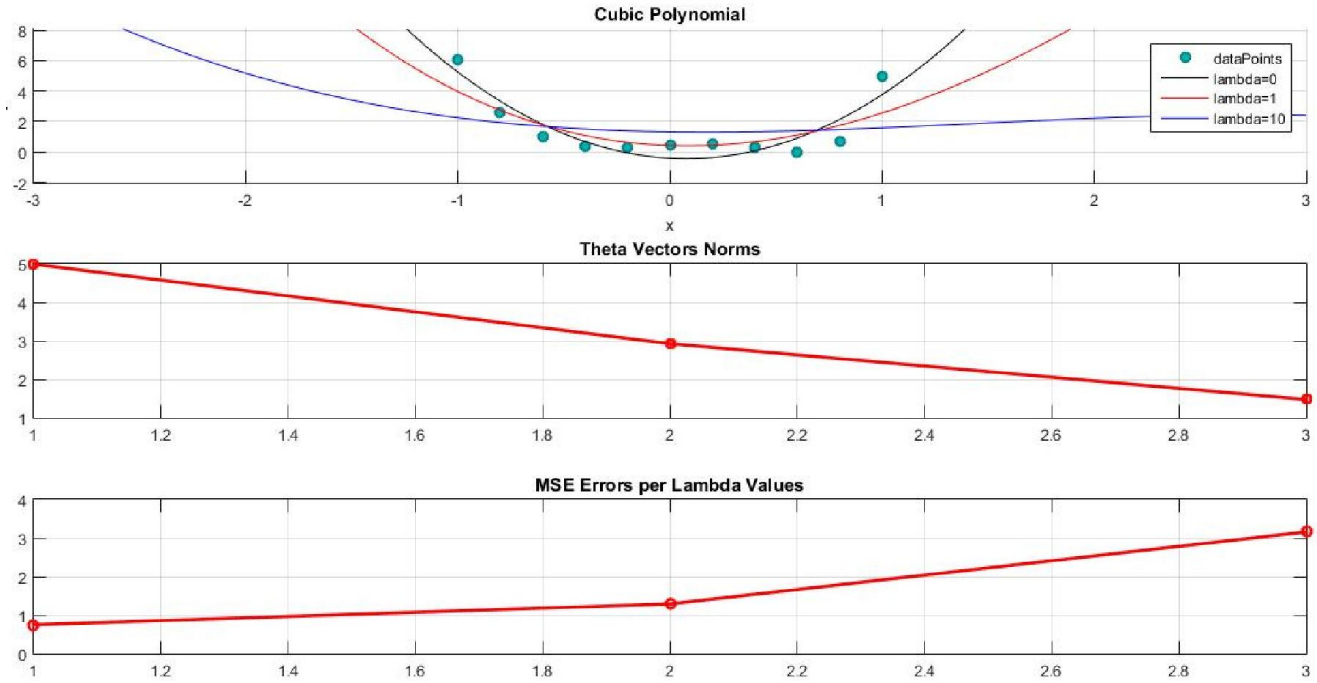
مقادیر **θ** بهینه برای سه حالت **λ=0** ، **λ=1** و **λ=10** برای چندجمله‌ای درجه‌ی ۳ به شرح زیر است:

Cubic Polynomial			
λ=0		λ=1	λ=10
θ_0	-0.4020	0.4261	1.3258
θ_1	-0.7543	-0.4701	-0.1982
θ_2	4.9125	2.8422	0.5930
θ_3	0.0130	-0.2362	-0.1353

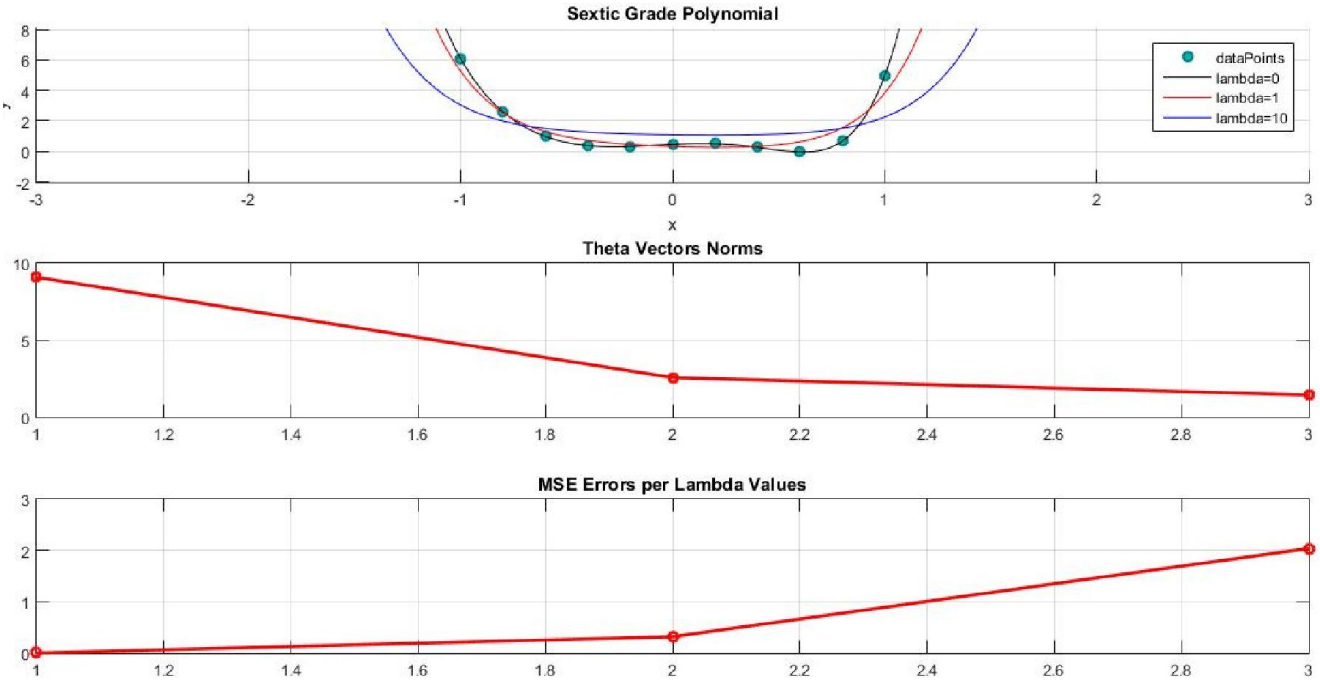
مقادیر **θ** بهینه برای سه حالت **λ=0** ، **λ=1** و **λ=10** برای چندجمله‌ای درجه‌ی ۶ نیز به شرح زیر است:

Sextic Degree Polynomial			
λ=0		λ=1	λ=10
θ_0	0.4646	0.3188	1.0909
θ_1	0.6844	-0.4652	-0.1853
θ_2	-1.3959	1.0289	0.4630
θ_3	-5.9607	-0.2256	-0.1223
θ_4	2.4826	1.4948	0.5379
θ_5	4.7160	-0.0227	-0.0861
θ_6	3.9786	1.7049	0.5601

سایر موارد خواسته‌شده در صورت مسئله (چندجمله‌ای به‌دست‌آمده به ازای λهای مختلف؛ اندازه‌ی(نُرم) بردار θ بر حسب λهای مختلف؛ نمودار خطای MSE بر حسب λهای مختلف) در نمودارهای زیر برای دو حالت چندجمله‌ای درجه‌ی ۳ و ۶ قابل مشاهده است:



۲- نمودارهای مربوط به چندجمله‌ای درجه‌ی ۳



۳- نمودارهای مربوط به چندجمله‌ای درجه‌ی ۶

همانطور که از نمودارهای بالا قابل مشاهده است با توجه به پیچیدگی و ظرافت‌های موجود در مجموعه‌ی داده‌ی فعلی، چندجمله‌ای با درجه‌ی ۳ نسبت به چندجمله‌ای با درجه‌ی ۶ کمتر نسبت به داده‌ها fit شده و حتی می‌توان دید که به ازای $\lambda = 0$ در چندجمله‌ای با درجه‌ی ۶، منحنی حاصله کاملا بر داده‌های ما منطبق شده و مدل ما اصطلاحا دچار overfitting شده است.

همینطور قابل مشاهده است که با افزایش اندازه‌ی λ از پیچیدگی‌های هر دو منحنی کاسته شده و مدل ما به سمت یک مدل ساده‌تر حرکت می‌نماید و اصطلاحا به ازای λهای کوچک مدل دچار **overfitting** شده و با افزایش اندازه‌ی λ از این وضعیت دور می‌شویم.

می‌توان دید که با افزایش اندازه‌ی λ، اندازه‌ی(نُرم) بردار **θ** در هر دو حالت چندجمله‌ای درجه‌ی ۳ و ۶ کاهش یافته و نیز میزان خطای **MSE** در هر دو مورد افزایش می‌یابد و این شاهد بر همان مطالبی است که در پاسخ سؤال سوم بیان شد که با افزایش اندازه‌ی λ تمرکز مدل بر روی کم‌کردن اندازه‌ی پارامترهای مسئله می‌باشد، که در نتیجه‌ی آن مدل ما ساده‌تر شده و کمتر نسبت به داده‌های آموزشی **fit** می‌شود و در نتیجه میزان خطای **residual** به ازای هر داده افزایش یافته و در نهایت خطای کل **MSE** بالا می‌رود.

۵. پاسخ بخش امتیازی

$$J(\bar{\theta}) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(y_i - h(x_i)\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] =$$

$$\frac{1}{2m}\left[\sum_{i=1}^{m}\left(y_i - \overline{x_i}^T\bar{\theta}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] =$$

$$\frac{1}{2m}\left[(\bar{y} - X\bar{\theta})^T(\bar{y} - X\bar{\theta}) + \lambda \|\bar{\theta}\|_2^2\right]$$

که در اینجا **X** یک ماتریس **m*(n+1)** می‌باشد که **m** تعداد داده‌های آموزشی و **n** درجه‌ی چندجمله‌ای بوده و $\overline{x_i}^T$ نیز برداری سطری با ابعاد **1*(n+1)** شامل درایه‌های هر سطر ماتریس **X** می‌باشد ؛ $\bar{y}$ نیز یک بردار **m*1** بوده و $\bar{\theta}$ هم یک بردار **1*(n+1)** می‌باشد. داریم:

$$\frac{\partial}{\partial\bar{\theta}}J(\bar{\theta}) = \frac{1}{2m}\left[-2X^T(\bar{y} - X\bar{\theta}) + 2\lambda\bar{\theta}\right] = 0 \quad \Rightarrow$$

$$-X^T\bar{y} + X^TX\bar{\theta} + \lambda\bar{\theta} = 0 \quad \Rightarrow$$

$$X^TX\bar{\theta} + \lambda\bar{\theta} = X^T\bar{y} \quad \Rightarrow$$

$$\left(X^TX + \lambda I_n\right)\bar{\theta} = X^T\bar{y} \quad \Rightarrow$$

$$\bar{\theta} = \left(X^TX + \lambda I_{n+1}\right)^{-1}X^T\bar{y} =$$

$$\left(X^TX + \lambda\begin{pmatrix}0 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ 0 & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}_{(n+1)*(n+1)}\right)^{-1}X^T\bar{y} \quad \therefore \text{ Proved!}$$