

گزارش کار تمرین شماره یک درس یادگیری ماشین

محمد لشکری ۴۰۰۱۱۲۰۸۷

۸ آذر ۱۴۰۰

۱ نتایج رگرسیون خطی یک متغیره

در فایل linreg.py کدها به صورت ماترسی و بدون حلقه با استفاده از تابع np.dot پیاده سازی شده است تا محاسبات سریع تر انجام شود. مقادیر مشاهده شده برای تابع هزینه با مقادیر متفاوت هایپرپارامترها به صورت زیر است:

$$\theta_0 = [15, 15]^T, \alpha = 0.1, n = 1500 \Rightarrow J(\theta) \simeq Nan \quad (1.1)$$

$$\theta_0 = [10, 10]^T, \alpha = 0.01, n = 1500 \Rightarrow J(\theta) \simeq 4/54 \quad (2.1)$$

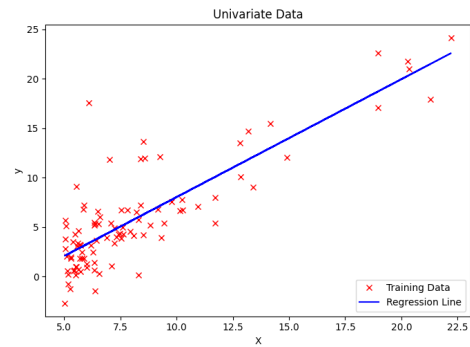
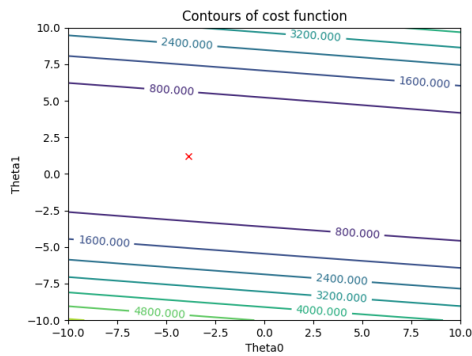
$$\theta_0 = [5, 5]^T, \alpha = 0.001, n = 1500 \Rightarrow J(\theta) \simeq 8/24 \quad (3.1)$$

$$\theta_0 = [15, 15]^T, \alpha = 0.1, n = 2000 \Rightarrow J(\theta) \simeq Nan \quad (4.1)$$

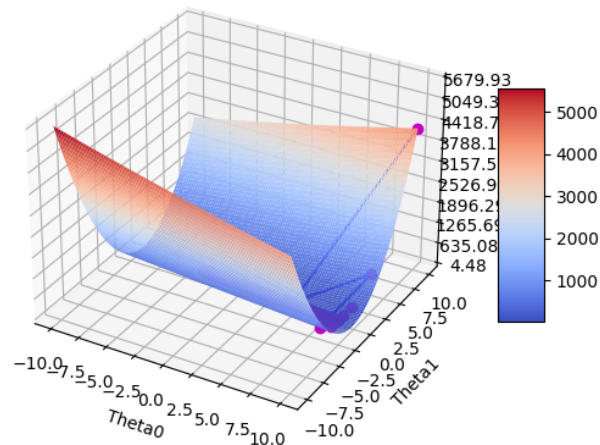
$$\theta_0 = [10, 10]^T, \alpha = 0.01, n = 2000 \Rightarrow J(\theta) \simeq 4/49 \quad (5.1)$$

$$\theta_0 = [5, 5]^T, \alpha = 0.001, n = 2000 \Rightarrow J(\theta) \simeq 7/62 \quad (6.1)$$

اگر نرخ یادگیری زیاد باشد، میزان نوسانات بالا می‌رود و ممکن است الگوریتم گرادین کاهشی همگرا نشود که این مطلب، عبارات (۱.۱) و (۴.۱) را توجیه می‌کند. از طرفی اگر نرخ یادگیری خیلی کم باشد سرعت همگرایی پایین می‌آید و باید تعداد تکرار را افزایش دهیم تا مقدار تابع هزینه کاهش یابد که این، گواهی بر نتایج (۳.۱) و (۶.۱) است. اما بهترین نتیجه با نرخ یادگیری مناسب و تعداد تکرار کافی برای $\alpha = 0.01, n = 2000$ رخ داده است که مقادیر ضرایب رگرسیون خطی برای این دو مقدار برابر $\theta \simeq [-3/67, 1/17]^T$ است که $\theta_0 = [10, 10]^T$ همچنین نمودارهای مربوط به این مدل را می‌توانید در ادامه مشاهده کنید:



Surface plot of the cost function



۲ نتایج رگرسیون خطی چند متغیره

نتایج رگرسیون خطی چند متغیره به ازای مقادیر مختلف α و n به شرح زیر است:

$$\alpha = 0.1, n = 2000 \Rightarrow J(\theta) \simeq 2043280050/60$$

$$\alpha = 0.1, n = 2000 \Rightarrow J(\theta) \simeq 2043280083/21$$

$$\alpha = 0.1, n = 3000 \Rightarrow J(\theta) \simeq 2043280050/60$$

$$\alpha = 0.1, n = 3000 \Rightarrow J(\theta) \simeq 2043280050/60$$

می‌دانیم تابع هزینه رگرسیون محدب است و همانطور که مشاهده می‌شود بعد از چند تکرار، تابع هزینه ثابت می‌ماند (به جز عبارت دوم که نیاز به تعداد تکرار بالا تری دارد) پس مینیمم تابع پیدا شده است اما مقدار آن زیاد است که این امر می‌تواند دو دلیل داشته باشد: ۱. کم بودن تعداد داده ها ۰.۲ عدم برقراری فرض خطی بودن وابستگی. همچنین بردار ضرایب که در هر ۴ بار اجرا تقریباً یکسان بود، به صورت زیر است:

$$\theta \simeq [340412/66, 109439/19, -6569/75]^T$$

۳ خطای مدل روی داده های فایل holdout.npz

با فرض اینکه متغیرهای مستقل مجموعه تست را با X_{test} نمایش دهیم، این مجموعه قبل از پیش بینی به صورت زیر استاندارد شده است:

$$X_{test} = \frac{X_{test} - \mu}{\sigma}$$

که در آن $\mu = \bar{X}$ و $\sigma = S_X^2$. در جدول ۱، خطای تست که بر اساس تابع هزینه محاسبه شده است به ازای مقادیر مختلف هایپارامترها قابل مشاهده است:

خطای تست	تعداد تکرار	نرخ یادگیری
۱۱۵۳۰۰۴۱۸۹/۳۶	۲۰۰۰	۰/۱
۱۱۵۳۰۳۰۳۳۹/۸۵	۲۰۰۰	۰/۰۱
۱۱۵۳۰۰۴۱۸۹/۳۶	۳۰۰۰	۰/۱
۱۱۵۳۰۰۴۱۸۹/۳۶	۳۰۰۰	۰/۰۱

جدول ۱: خطای داده های تست

خطای تست در هر ۴ بار اجرا به جز اجرای دوم، تقریباً عددی یکسان و بزرگ است که با توجه به بالا بودن مقدار تابع هزینه در انتهای آموزش نتیجه دور از ذهنی نیست. خطای داده های آموزش و تست از هم فاصله دارند و پیچیدگی مدل کم است پس underfit رخ داده است. بنابراین این مدل مطلوب نیست و برای بهبود آن باید تعداد داده ها زیاد شود و پیچیدگی مدل تغییر یابد.