پروژه اول درس یادگیری ماشین

پریا پاسه‌ورز

شماره دانشجویی: 810101393

سوال دوم

الف) تعریف L1 Regularization و L2 Regularization و تفاوتهای آنها:

Regularization تکنیکی است که از آن برای کم کردن overfitting می‌شود. برای این کار در مدلهای پیچیده، یک penalty معرفی می‌کنیم که پیچیدگی مدل را کاهش دهد و مدل general تری ارائه دهد. این کار سبب ممکن است سبب شود که کارکرد مدل در برابر دیتای train کاهش یابد، ولی در برابر دیتای test بهتر عمل می‌کند.

L1 Regularization (Lasso Regularization)

$$\hat{\theta} = \arg\min_{\theta}\left[\sum_{i=1}^{n}\left(y^{(i)} - x^{(i)}\theta\right)^2 + \lambda|\theta|\right]$$

L2 Regularization (Ridge Regression)

$$\hat{\theta} = \arg\min_{\theta}\left[\sum_{i=1}^{n}\left(y^{(i)} - \theta^T x^{(i)}\right)^2 + \lambda|\theta|^2\right]$$

- اولین تفاوتی که در این دو روش مشهود است در این است که lasso regularization از ضرایبی از مقادیر پارامترها به عنوان penalty استفاده می‌کند، در حالی که ridge regularization از ضرایب مربعات این پارامترها استفاده می‌کند.
- Sparsity به این معناست که بعضی پارامترها در مدل دقیقا صفر باشند، که این نکته سبب می‌شود تاثیر بعضی featureها در مدل از بین برود. Lasso regularization در واقع الگویی است که منجر به sparsity می‌شود. برای درک بهتر نمودار قدر مطلق را در نظر بگیرید. مقادیر حول صفر به صورت یکنواخت به صفر نزدیک نمی‌شوند و حالت الماس شکل دارند(در ابعاد بالا). این sharp بودن سبب می‌شود وقتی مقادیر پارامترها به صفر نزدیک می‌شوند، مدل آنها را صفر در نظر بگیرد و sparsity افزایش یابد. در حالی که در ridge regression یک انحنا حول نقطه صفر وجود دارد. استفاده از توان دو پارامترها باعث می‌شود مقادیر حول صفر به صفر نزدیک شوند ولی لزوما به اینکه صفر باشند force نمی‌شوند.
- Lasso regularization نسبت به داده‌های outlier حساس‌تر است، چون Penalty از جمع قدرمطلق‌ها به دست می‌آید، در حالی که در مدل ridge regularization، با توان دو داده‌ها یکنواخت‌تر حول صفر توزیع می‌شوند و تاثیر داده‌های outlier کمتر می‌شود.
- همانطور که در مورد دوم توضیح دادیم، Lasso regularization بعضی پارامترها را صفر در نظر می‌گیرد و تاثیر feature متناظر با آنها در مدل از بین می رود، پس تعدادی از featureها خودبه‌خود حذف می‌شوند، در حالی که در ridge regularization تمامی featureها تاثیر خود را حفظ می‌کنند.
- Lasso regularization حساسیت کمتری نسبت به featureهایی دارد که با هم، همبستگی دارند، چون فقط یکی از آنها را نگه می‌دارد، در حالی که ridge regularization این featureها را حذف نمی‌کند.

- کاربرد lasso regularization زمانی است که تعداد زیادی feature داشته باشیم که لزوما در مدل تاثیرگذار نباشند و بتوانیم با این روش تاثیر آنها را از بین ببریم، در حالی که کاربرد ridge regularization وقتی است که بخواهیم کارکرد مدل را بدون حذف هیچ یک از featureها بهبود ببخشیم.

ب) اگر در مسئله رگرسیون خطی زیر، λ یک ضریب ثابت مثبت باشد، فرم بسته مقدار بهینه w را در تابع هزینه محاسبه کنید.

$$\boldsymbol{L(w) = \sum_{i=1}^{n}(w^T x_i - y_i)^2 + \lambda||w||_2^2}$$

اول سعی می‌کنیم این عبارت را به فرم ماتریسی بنویسیم:

$$w^T x_i - y_i = \begin{bmatrix} w^T x_1 - y_1 \\ w^T x_2 - y_2 \\ \vdots \\ w^T x_n - y_n \end{bmatrix}$$

حال اگر تعریف کنیم:

$$X = \begin{bmatrix} {x_1}^T \\ {x_2}^T \\ \vdots \\ {x_n}^T \end{bmatrix}$$

در آن صورت می‌توان نوشت:

$$w^T x_i - y_i = (Xw - y)$$

و به طور کلی رابطه تابع هزینه خواهد بود:

$$L(w) = (Xw - y)^T (Xw - y) + \boldsymbol{\lambda w^T w}$$

حال از رابطه بالا نسبت به w مشتق می‌گیریم:

$$\frac{\partial L(w)}{\partial w} = \frac{\partial}{\partial w}((Xw)^T Xw - (Xw)^T y - y^T (Xw) + y^T y + \boldsymbol{\lambda w^T w})$$

$$= \frac{\partial}{\partial w}(w^T X^T Xw - w^T X^T y - (X^T y)^T w + \boldsymbol{\lambda w^T w})$$

از آنجایی که $\frac{\partial}{\partial X}(X^T AX) = 2AX$:

$$= 2X^T Xw + \frac{\partial}{\partial w}(-(Xw)^T y - (X^T y)^T w + \boldsymbol{\lambda w^T w})$$

$$= 2X^T Xw + \frac{\partial}{\partial w}(-y^T (Xw) - (X^T y)^T w + \boldsymbol{\lambda w^T w})$$

$$= 2X^T Xw + \frac{\partial}{\partial w}(-(X^T y)^T w - (X^T y)^T w + \boldsymbol{\lambda w^T w})$$

$$= 2X^T Xw + \frac{\partial}{\partial w}(-2(X^T y)^T w + \boldsymbol{\lambda w^T w})$$

$$= 2X^TXw + \frac{\partial}{\partial w}(-2w^T(X^Ty) + \boldsymbol{\lambda w^T w})$$

$$= 2X^TXw - 2X^Ty + \frac{\partial}{\partial w}(\boldsymbol{\lambda w^T w})$$

از طرف دیگر:

$$\lambda w^T w = \lambda \sum_{i=1}^{d} w_i^2$$

$$\frac{\partial}{\partial \mathrm{w}}\left(\lambda \sum_{\mathrm{i}=1}^{\mathrm{d}} \mathrm{w}_{\mathrm{i}}^2\right) = 2\lambda \mathrm{w}$$

$$= 2X^TXw - 2X^Ty + 2\lambda \mathrm{w} \;=\; 0$$

$$w(X^TX + \boldsymbol{\lambda}\mathrm{I}) = X^Ty$$

$$w = (X^TX + \boldsymbol{\lambda}\mathrm{I})^{-1}(X^Ty)$$