فاطمه حسینی - پروژه اول ماشین لرنینگ

مسأله رگرسیون خطی منظم‌سازی شده با نرم L2 :

← این مسئله در واقع برای حل بیش برازش معرفی شده است به این شکل که یک ترم جریمه به تابع هزینه مدل رگرسیون استاندارد اضافه می‌کنیم. ترم جریمه وزن‌های بزرگ $w_i$ در مدل را جریمه می‌کند. هدف این است که مدلی پیدا کنیم که نه تنها به خوبی داده‌های آموزشی را برازش کند بلکه وزن‌های آن نیز تا حد امکان کوچک باشند.

تابع هزینه در این مدل به صورت ⇐

$$E(w) = J(w) = \frac{1}{2}\sum_{i=1}^{N} (w^T x_i - y_i)^2$$

تعداد نمونه‌ها

اما در حالت منظم‌سازی شده داریم ⇐

$$E(w)_{Regu} = \frac{1}{2}\sum_{i=1}^{N} (w^T x_i - y_i)^2 + \alpha \cdot \text{RegularizationTerm}(w)$$

نوع نرم L2 / L1

بردار وزن‌های مدل — بردار ویژگی‌های نمونه — مقدار هدف برای نمونه i — پارامتر تنظیم برای نرم

منظم‌سازی L2 ⇐ مربع نرم L2 وزن‌ها، ترم منظم‌سازی است:

$$RT(w) = \|w\| = \left(\sqrt{\sum_{j=1}^{p} w_j^2}\right)^2 \qquad L_2 = \sqrt{\sum_i x_i^2}$$

p تعداد ویژگی‌ها است

تابع هزینه به صورت زیر بازنویسی می‌شود:

$$E(w)_{Regu} = \frac{1}{2}\sum_{i=1}^{N} (y_i - w^T x_i)^2 + \frac{\alpha}{2}\sum_{j=1}^{p} w_j^2$$

الف) با کمک تکنیک کاهش گرادیان تصادفی یک رابطه برای یادگیری یک رابطه برای یادگیری بر خط این مدل ارائه دهید. (sgd)

در این روش در هر گام وزن‌ها را در جهت منفی گرادیان تابع هزینه نسبت به وزن‌ها بروزرسانی می‌شوند. گرادیان جهت بیشترین افزایش تابع هزینه را نشان می‌دهد. بنابراین حرکت در جهت منفی آن باعث کاهش هزینه تابع می‌شود.

تکنیک کاهشی گرادیان تصادفی یک نسخه از GD است که به جای استفاده از کل مجموعه داده برای محاسبه گرادیان در هر گام تنها از یک نمونه داده تصادفی (یا یک دسته کوچک از داده‌ها) استفاده می‌کند.

تابع خطا که در مرحله قبل معرفی کردیم را در نظر می‌گیریم، سپس مشتق آن را نسبت به بردار وزن $w$ محاسبه می‌کنیم:

$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(w^T x_i - y_i)^2 + \|w\|^2 \frac{\lambda}{2}$$

حالا برای گام $t$ ام داریم:

$$E_t(w) = \frac{1}{2}(w^T x_t - y_t)^2 + \|w\|^2 \frac{\lambda}{2}$$

$$\frac{\partial}{\partial w}\left(\frac{1}{2}(w^T x_t - y_t)^2\right) = -(y_t - x_t(w^T x_t))$$

$$\frac{\partial}{\partial w}\|w\|^2\frac{\lambda}{2} = \lambda w$$

$$\Rightarrow \nabla_w E_t(w) = x_t(w^T x_t) - y_t + \lambda w$$

حالا رابطه بروزرسانی وزن ها در sgD: در این الگوریتم وزن ها در جهت منفی گرادیان بروزرسانی می‌شوند. در sgD این بروزرسانی بر اساس گرادیان محاسبه شده برای نمونه تصادفی $t$ انجام می‌شود:

بردار وزن در گام $t$

$$w^{(t+1)} = w^{(t)} - \eta \nabla_w E_t(w^{(t)})$$

$\eta$ (اتا) نرخ یادگیری ما است یا طول گام است که اندازه گام در جهت منفی گرادیان را کنترل می‌کند. انتخاب مناسب نرخ یادگیری برای همگرایی موفق الگوریتم بسیار مهم است.

حالا با جایگذاری عبارت گرادیان که بدست آوردیم، رابطه بروزرسانی وزن ها برای رگرسیون خطی منظم شده با ترم L2 با کمک sgd به صورت زیر است:

$$w^{(t+1)} = w^{(t)} + \eta(y_t - w^{(t)T} x_t)\, x_t - \eta \lambda w^{(t)}$$

ب) در تکنیک کاهشی گرادیان ، $\eta$ ضریب یادگیری نامیده می‌شود و در محیط بهینه‌سازی طول گام است . در رابطه‌ای که برای قسمت الف ارائه کرده‌اید ، طول گام بهینه را با حل معادلات ریاضی بدست آورید. منظور از طول گام بهینه ، مقداری برای ضریب یادگیری $\eta$ است که به ازای آن ، مقدار تابع هزینه در راستای بردار گرادیان به کمترین مقدار ممکن برسد. برای این کار ابتدا پارامتر وزن جدید را که برحسب $\eta$ تعریف شده است مجدد وارد معادلات مدل ، و تابع زیان کنید تا تابع زیان ، تابعی برحسب $\eta$ داده شود. حال نسبت به $\eta$ مشتق بگیرید و برابر صفر قرار دهید و $\eta$ را حل کنید.

$\eta$ : ابتدا پارامتر وزن جدید را $(\omega')$ برحسب $\eta$ در تابع هزینه جایگذاری می‌کنیم و تعریف می‌کنیم

$\omega$ وزن فعلی و $\omega'$ وزن جدید باشد:

$$\omega' = \omega + \eta\,(y_t - x_t\,\omega^T) - \eta\lambda\omega$$

$$E(\omega') = \frac{1}{2}\left(y_t - (\omega')^T x\right)^2 + \frac{\lambda}{2}\|\omega'\|^2$$ ⇐ تابع هزینه (*)

جایگذاری $\omega'$ در (*) ⇐

$$E(\eta) = \frac{1}{2}\left(y - \left(\omega + \eta\,(y - t_t\,\omega^T) - \eta\lambda\omega\right)^T x\right)^2 + \frac{\lambda}{2}\left\|(\omega + y)\eta - x\,(x\omega^T - \omega\lambda\eta)\right\|^2$$

حال برای $\eta$ بهینه از معادله بالا مشتق گرفته و برابر صفر قرار می‌دهیم:

$$\frac{\partial E(\eta)}{\partial \eta} = 0$$

$$\frac{\partial E(\omega')}{\partial \omega} = \nabla_\omega E_t(\omega') = -\left(y_t - \omega^T x_t\right) x_t + \lambda\omega$$

$$\omega' = \omega + -\eta\,\nabla_\omega E_t(\omega')$$

$$E(\eta) = \frac{1}{2}\left(y - \left(\omega - \eta\,\nabla_\omega E_t(\omega')\right)^T x\right)^2 + \frac{\lambda}{2}\left\|\omega - \eta\,\nabla_\omega E_t(\omega')\right\|^2$$

$$\frac{\partial E(\eta)}{\partial \eta} = \left(y - \left(\omega - \eta\,\nabla_\omega E_t(\omega')\right)^T x\right)\left(\nabla_\omega E_t(\omega')^T x\right) + \lambda\left(\omega - \eta\,\nabla_\omega E_t(\omega')\right)^T - \left(\nabla_\omega E_t(\omega')\right) = 0$$

F

$$y - \omega^T x + \eta (\nabla_\omega E_t(\omega'))^T x \, (\nabla_\omega E_t(\omega'))^T x - \lambda \omega^T \nabla_\omega E_t(\omega') + \lambda\eta (\nabla_\omega E_t(\omega'))^T (\nabla_\omega E_t(\omega')) = 0$$

$$\Rightarrow (y - \omega^T x)((\nabla_\omega E_t(\omega'))^T x) + \eta ((\nabla_\omega E_t(\omega'))^T x)^2 - \lambda (\omega^T)(\nabla_\omega E_t(\omega'))^T) + \lambda \eta \|\nabla_\omega E_t(\omega')\|^2 = 0$$

$$\Rightarrow \eta \left( ((\nabla_\omega E_t(\omega'))^T x)^2 + \lambda \|\nabla_\omega E_t(\omega')\|^2 \right) = -(y - \omega^T x)((\nabla_\omega E_t(\omega'))^T x) + \lambda \omega^T \nabla_\omega E_t(\omega')$$

$$\Rightarrow \eta^* = \frac{-(y-\omega^T x)(-(y-\omega^T x)x + \lambda\omega)^T x + \lambda \omega^T(-(y-\omega^T x)x + \lambda\omega)}{((-(y-\omega^T x)x + \lambda\omega)^T x)^2 + \lambda \|-(y-\omega^T x)x + \lambda\omega\|^2}$$

ج) به طور مشابه به سوالات الف و ب، در صورتی که مساله رگرسیون خطی با نرم L1 منظم سازی شده باشد، پاسخ دهید.

نرم L1:

$$\|\omega\| = L_1 = \sum_i |x_i| \quad \Leftarrow \omega = [x_1, x_2, \dots, x_n]$$

$$E(\omega) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \omega^T x_i)^2 + \lambda\|\omega\|_1$$

بردار علامت ω

$$sgd \Rightarrow \nabla_\omega E_t(\omega) = -(y_t - \omega^T x_t)\, x_t + \lambda \cdot \overrightarrow{sgn(\omega)}$$

$$sgn(\omega_j) = \begin{cases} 1 & \omega_j > 0 \\ -1 & \omega_j < 0 \\ 0 & oth \end{cases}$$

$$\omega^{(t+1)} = \omega^{(t)} - \eta \nabla_\omega E_t(\omega^{(t)})$$

$$\omega^{(t+1)} = \omega^{(t)} + \eta\,(y_t - (\omega^{(t)})^T x_t)\, x_t - \eta\lambda \cdot sgn(\omega^{(t)})$$

حالا برای پیدا کردن طول گام بهینه برای منظم سازی L1 باید مشتق بگیریم. تابع هزینه ها به دلیل وجود قدر مطلق پیچیده است و مشتق آن ممکن است پیوسته نباشد.

یک روش معمول در این حالت استفاده از روش linesearch است. اگر بخواهیم یک تقریب خوب برای طول گام بهینه در نظر بگیریم می توانیم از تقریب مرتبه دوم تابع هزینه (روش نیوتن) کمک بگیریم.