علی آراسته – ۹۶۱۰۱۱۶۵

بهینه‌سازی محدب – تمرین کامپیوتری اول

سوال ۱:

a) ابتدا تابع هزینه را تعریف می‌کنیم:

$$f(l,w) = 2lw + (2l + \pi w)$$

این تابع به خودی خود محدب نیست؛ اما تابعی posinomial است. لذا از تغییر متغیرهای زیر استفاده می‌کنیم.

$$x = \log(l)\,, y = \log(w)$$

$$g(x,y) = 2e^{x+y} + (2e^{x} + \pi e^{y})$$

$$\min_{l,w} f(l,w) = \min_{x,y} g(x,y) = \min_{x,y}\left(2e^{x+y} + (2e^{x} + \pi e^{y})\right)$$

سپس قیود مسئله را مشخص می‌کنیم:

$$lw \geq 300\text{cm}^2 \to 300 - lw \leq 0 \to \log(300) - (x + y) \leq 0$$

$$1 \leq \frac{l}{w} \leq 2 \to w - l \leq 0, l - 2w \leq 0 \to y - x \leq 0, x - y - \log(2) \leq 0$$

$$10\text{cm} \leq w \leq 20\text{cm} \to 10 - w \leq 0, w - 20 \leq 0 \to \log(10) - y \leq 0, y - \log(20) \leq 0$$

$$20\text{cm} \leq l \leq 30\text{cm} \to 20 - l \leq 0, l - 30 \leq 0 \to \log(20) - x \leq 0, x - \log(30) \leq 0$$

b) مقادیر بهینه به صورت زیر بدست می‌آیند:

$$l \cong 21.7, w \cong 13.8$$

سوال ۲:

a) از صورت سوال داریم:

$$y_i = \phi(z_i) = \phi\left(a_i^T x + v_i\right) \quad i = 1, \dots, m \to z_i = \phi^{-1}(y_i) = a_i^T x + v_i$$

تابع ϕ^{-1} ناشناخته است؛ اما مشتق‌های آن بین $\frac{1}{\beta}$ و $\frac{1}{\alpha}$ است. در نتیجه با فرض مرتب شدن مقادیر y_i به صورت افزایشی، مقادیر z_i و y_i باید در نامساوی‌های زیر صدق کنند:

$$\frac{y_{i+1} - y_i}{\beta} \leq z_{i+1} - z_i \leq \frac{y_{i+1} - y_i}{\alpha} \quad i = 1, \dots, m-1$$

از طرفی تابع log-likelihood به صورت زیر خواهد بود:

$$l(z,x) = -\frac{1}{2\sigma^2}\Sigma_{i=1}^{m}\left(z_i - a_i^T x\right)^2 - m\log\left(\sigma\sqrt{2\pi}\right)$$

در نتیجه برای بدست آوردن مقادیر بهینه می‌توان از مسئله بهینه‌سازی زیر استفاده کرد:

$$\min_{z,x} \Sigma_{i=1}^{m}\left(z_i - a_i^T x\right)^2$$

$$\text{subject to} \quad \frac{y_{i+1}-y_i}{\beta} \le z_{i+1} - z_i \le \frac{y_{i+1}-y_i}{\alpha} \quad i = 1, \dots, m-1$$

b) مسئله بهینه‌سازی مورد نظر با استفاده از کامپیوتر حل شده است.

سوال ۳:

a) اگر $u > 0$ باشد، $1(u > 0) = 1$ و $(1+\lambda u)_+ \ge 1$ است و اگر $u \le 0$ باشد، $1(u > 0) = 0$ و $(1+\lambda u)_+ \ge 0$ خواهد بود. در نتیجه برای هر $u \in R$، $(1+\lambda u)_+ \ge 1(u > 0)$ است. با اعمال این موضوع برای $u = f_i(x)$ داریم:

$$1(f_i(x) > 0) \le \left(1 + \lambda f_i(x)\right)_+$$

$$\Sigma_{i=1}^{m} 1(f_i(x) > 0) \le \Sigma_{i=1}^{m}\left(1 + \lambda f_i(x)\right)_+$$

حال با در نظر گرفتن شرط $\Sigma_{i=1}^{m}\left(1 + \lambda f_i(x)\right)_+ \le m - k$ داریم:

$$\Sigma_{i=1}^{m} 1(f_i(x) > 0) \le m - k$$

در نتیجه $f_i(x) > 0$ حداکثر برای $m - k$ مقدار i برقرار است؛ یعنی $f_i(x) \le 0$ حداقل برای k مقدار i برقرار خواهد بود.

b) اگر $\lambda > 0$ باشد، $(\lambda u)_+ = \lambda(u)_+$ خواهد بود. در نتیجه داریم:

$$\left(1 + \lambda f_i(x)\right)_+ = \lambda\left(\frac{1}{\lambda} + f_i(x)\right)_+$$

بنابراین شرط مسئله به صورت زیر تغییرپذیر است:

$$\Sigma_{i=1}^{m}\left(\frac{1}{\lambda} + f_i(x)\right)_+ \le \frac{1}{\lambda}(m-k)$$

$$\Sigma_{i=1}^{m}\left(\mu + f_i(x)\right)_+ \le \mu(m-k), \mu > 0$$

از آنجایی که تابع $(.)_+$، تابعی محدب و غیرکاهشی است و $\mu + f_i(x)$ نسبت به μ و $f_i(x)$ محدب است، $\left(\mu + f_i(x)\right)_+$ نیز محدب خواهد بود و در نتیجه مسئله محدب است.

می‌توان شرط $\mu > 0$ را با $\mu \ge 0$ جایگزین کرد. زیرا اگر $\mu = 0$ باشد، شرط $\Sigma_{i=1}^{m}\left(\mu + f_i(x)\right)_+ \le \mu(m-k)$ به معنای آن است که تمام شروط $f_i(x) \le 0$ برقرار باشند.

c) جواب مسئله به صورت زیر بدست می‌آید:

$$\lambda = 282.98, f^* = -8.45, \text{number of satistified constraints} = 66$$

سوال ۴:

a) برای اعمال بایاس، از پارامتری جدا استفاده شده است.

i) در مرحله ابتدا گرادیان تابع را نسبت به پارامترهای β و b به طور جداگانه محاسبه و سپس مقادیر β_i و b_i را به‌روزرسانی می‌کنیم.

$$f(\beta, b) = g(\beta, b) + h(\beta, b) = \frac{1}{2N}\|X\beta + b - y\|^2 + \frac{\lambda}{2}\|\beta\|_2^2$$

$$\nabla f_\beta(\beta, b) = \frac{1}{B}X^T(X\beta + b - y) + \lambda\beta$$

$$\beta_{i+1} = \beta_i - t\nabla f_\beta(\beta, b) = \beta_i - t\left(\frac{1}{B}X^T(X\beta_i + b_i - y) + \lambda\beta_i\right)$$

$$\nabla f_b(\beta, b) = \frac{1}{B}1^T(X\beta + b - y)$$

$$b_{i+1} = b_i - t\nabla f_b(\beta, b) = b_i - t\left(\frac{1}{B}1^T(X\beta_i + b_i - y)\right)$$

ii) موارد خواسته شده در فایل کد انجام شده اند.

iii) نمودارهای مربوط به ۱۶ حالت مورد نظر در فایل‌های ارسالی آمده است.

با بررسی نمودارهای رسم شده، دو نکته مشخص می‌شود. اولا مقدار step-size، باید نه خیلی کوچک و نه خیلی بزرگ باشد؛ زیرا اگر خیلی کوچک باشد، سرعت همگرایی الگوریتم کم خواهد بود و اگر خیلی بزرگ باشد، الگوریتم حول نقطه بهینه نوسان می‌کند و نمی‌توان خیلی به مقدار بهینه نزدیک شد. بنابراین مقدار step-size، باید مقداری متوسط باشد. ثانیا به طور کلی مقدار batch-size، هر چه کوچکتر باشد، بهتر است و سرعت همگرایی الگوریتم بیشتر خواهد بود؛ البته در مواردی که مقدار step-size خیلی بزرگ است، با انتخاب مقادیر بزرگ برای batch-size، می‌توان تا حدودی از نوسان حول نقطه بهینه جلوگیری کرد که به ازای کاهش سرعت الگوریتم تمام می‌شود.

b) برای اعمال بایاس، از پارامتری جدا استفاده شده است.

i) از قضیه زیر برای حل مسئله استفاده می‌کنیم.

قضیه: فرض کنید $f: \mathbb{E} \to \mathbb{R}$ به صورت $f(x) = g(x)$ داده شده باشد که در آن $g: \mathbb{R} \to (-\infty, \infty]$ تابع خوش تعریف بسته و محدبی است که برای آن $dom(g) \subseteq [0, \infty)$ است. آنگاه داریم:

$$\mathrm{prox}_{f,t}(x) = \begin{cases} \{u \in \mathbb{E}: \|\mathbb{E}\| = \mathrm{prog}_{g,t}(0)\} & x = 0 \\ \mathrm{prog}_{g,t}(\|x\|)\dfrac{x}{\|x\|} & x \neq 0 \end{cases}$$

بر اساس این قضیه با در نظر گرفتن $f(x)=\lambda\|x\|$ داریم:

$$\operatorname{prox}_{f,t}(x)=\begin{cases}0 & x=0\\ (\|x\|-\lambda)_+\dfrac{x}{\|x\|} & x\neq 0\end{cases}$$

$$\operatorname{prox}_{f,t}(x)=\left(1-\frac{\lambda t}{\max\{\|x\|,\lambda t\}}\right)x$$

حال از آنجایی که در مسئله مورد نظر $h(\beta)=\Sigma_{j=1}^{J}\lambda w_j\|\beta_j\|_2=\Sigma_{j=1}^{J}h_j(\beta_j)$ مجموعی از چند تابع مستقل نسبت به متغیرهای مستقل است، داریم:

$$\min_{\beta,b}\left(\frac{1}{2N}\|X\beta+b-y\|^2+\lambda\Sigma_{j=1}^{J}w_j\|\beta_j\|_2\right)$$

$$\min_{b}\left(\Sigma_{j=1}^{J}\min_{\beta_j}\left(\frac{1}{2N}\|X_j\beta_j+b-y_j\|^2+\lambda w_j\|\beta_j\|_2\right)\right)$$

در نتیجه می‌توان تابع را به طور جداگانه کمینه کرد. در نتیجه داریم:

$$\operatorname{prox}_{h,t}(x)=\left(\operatorname{prox}_{h_1,t}(x_1),\dots,\operatorname{prox}_{h_J,t}(x_J)\right)$$

ii) با توجه به الگوریتم داریم:

$$\beta^{i+1}=\operatorname{prox}_{h,t}\left(\beta^i-t\nabla g(\beta^i)\right)$$

$$\beta_j^{i+1}=\left(1-\frac{\lambda w_j t}{\max\left\{\left\|\beta_j^i-t\nabla g(\beta^i)_j\right\|_2,\lambda w_j t\right\}}\right)\left(\beta_j^i-t\nabla g(\beta^i)_j\right)$$

$$b^{i+1}=b^i-t\left(\frac{1}{B}1^T\left(X\beta^i+b^i-y\right)\right)$$

iii) موارد خواسته شده در فایل کد انجام شده اند.

iv) با توجه به خروجی، گروه‌های Age، RPDE و PPE انتخاب می‌شوند.

v) موارد خواسته شده در فایل کد انجام شده اند. به وضوح سرعت همگرایی و میزان تنک بودن در Group LASSO بیشتر است.

vi) موارد خواسته شده در فایل کد انجام شده اند. به وضوح وضوح سرعت همگرایی و میزان تنک بودن در الگوریتم شتاب‌یافته بیشتر است. خروجی در این الگوریتم تقریبا ۱۰ برابر به مقدار بهینه نزدیک‌تر است.