«به نام خدا»

امیررضا آذری — نظریه یادگیری ماشین

۹۹۱۰۱۰۸۷ — تمرین سوم

1

الف) در تابع هزینه داده شده ϵ ناحیه‌ای است که خطا در آن صرف نظر می‌شود. همچنین تابع ضرر داده شده به علت قدر مطلق قابل مشتق گیری نیست. حال متغیرهای slack را ξ و ξ^* تعریف می‌کنیم که بتوانند خطا نقاط خارج از ناحیه ϵ را مشخص کنند.

$$y_i - \langle w, x_i \rangle - \epsilon \le \xi_i$$

$$\langle w, x_i \rangle - y_i - \epsilon \le \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0 \quad i = 1, 2, \dots, n$$

حال فرم اصلی را می‌توانیم به صورت زیر بنویسیم:

$$\min_{w \in R^m, \xi \in R^n, \xi^* \in R^n} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

که ۳ رابطه بالاتر را برآورده می‌کنند. پس نشان دادیم به یک مسئله QP رسیدیم.

ب) لاگرانژین به شکل زیر خواهد بود:

$$L = L(w, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\beta_i \xi_i + \beta_i^* \xi_i^*)$$

$$- \sum_{i=1}^{n} \alpha_i(\epsilon + \xi_i - y_i + \langle w, x_i \rangle) - \sum_{i=1}^{n} \alpha_i^*(\epsilon + \xi_i^* + y_i - \langle w, x_i \rangle)$$

$$* \rightarrow \alpha_i, \alpha_i^*, \beta_i, \beta_i^* \ge 0, \quad (i = 1, \dots, n)$$

داریم :

$$\min_{w,\xi,\xi^*} \max_{a,a^*,\beta,\beta^*} L(w,\xi,\xi^*,a,a^*,\beta,\beta^*) = \max_{a,a^*,\beta,\beta^*} \min_{w,\xi,\xi^*} L(w,\xi,\xi^*,a,a^*,\beta,\beta^*)$$

چون مسئله QP و strong duality برقرار است.

حال مشتق جزئی نسبت به w، ξ_i و ξ_i^* داریم:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} (a_i - a_i^*) x_i = 0$$

$$\left.\begin{aligned} \frac{\partial L}{\partial \xi_i} &= C - a_i - \beta_i = 0 \\ \frac{\partial L}{\partial \xi_i^*} &= C - a_i^* - \beta_i^* = 0 \end{aligned}\right\} \Rightarrow \begin{aligned} \beta_i &= C - a_i \geq 0 \\ \beta_i^* &= C - a_i^* \geq 0 \end{aligned}$$

حال به کمک عبارات بالا و عبارت صفحه قبل داریم:

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\beta_i\xi_i + \beta_i^*\xi_i^*) - \sum_{i=1}^{n} a_i(\epsilon + \xi_i - y_i + \langle w, x_i\rangle) - \sum_{i=1}^{n} a_i^*(\epsilon + \xi_i^* + y_i - \langle w, x_i\rangle)$$

$$= \frac{1}{2}\Big\|\sum_{i=1}^{n}(a_i - a_i^*)x_i\Big\|^2 + \sum_{i=1}^{n}\xi_i\underbrace{(C - \beta_i - a_i)}_{0} + \sum_{i=1}^{n}\xi_i^*\underbrace{(C - \beta_i^* - a_i^*)}_{0} - \epsilon\sum_{i=1}^{n}(a_i + a_i^*) + \sum_{i=1}^{n} y_i(a_i - a_i^*)$$

$$+ \sum_{i=1}^{n}(a_i^* - a_i)\langle w, x_i\rangle \quad \left(\Big\langle\sum_{j=1}^{n}(a_j - a_j^*)x_j ,\ x_i\Big\rangle\right)$$

$$= -\frac{1}{2}\sum_{i,j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)\langle x_i, x_j\rangle - \epsilon\sum_{i=1}^{n}(a_i + a_i^*) + \sum_{i=1}^{n} y_i(a_i - a_i^*)$$

بنابراین، dual به صورت زیر خواهد شد:

$$\max_{a,a^*} -\frac{1}{2}\sum_{i,j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)\langle x_i, x_j\rangle - \epsilon\sum_{i=1}^{n}(a_i + a_i^*) + \sum_{i=1}^{n} y_i(a_i - a_i^*)$$

$$\text{s.t.} \quad a_i, a_i^* \in [0, C]$$

DATE / / SUBJECT:

ب)

بله؛ زیرا تابع هدف درجه ۲ و قیود خطی هستند و تغییر هدف quadratic دارد.

ت)

شروط KKT را می‌نویسیم:

$$(I) \quad \alpha_i(\epsilon+\xi_i-y_i+\langle w,x_i\rangle)=0$$

$$\alpha_i^*(\epsilon+\xi_i^*+y_i-\langle w,x_i\rangle)=0$$

$$\beta_i\xi_i=0$$
$$\beta_i^*\xi_i^*=0 \qquad \forall i=1,\dots,n$$

برای (I) می‌دانیم $\alpha_i \neq 0$، بنابراین $(\epsilon+\xi_i-y_i+\langle w,x_i\rangle)=0$.

اگر $\xi_i=0$ باشد یعنی x_i در مرز قرار دارد و x_i یک margin support vector می‌شود. اگر $\xi_i>0$،

یعنی خارج مرز هستیم و x_iها MSV نخواهند بود و non-margin support vector است.

به صورت مشابه می‌توان برای α_i^* و ξ_i^* استدلال کرد. برای $\alpha_i^*>0$ و $\xi_i^*=0$، $\xi_i^*>0$ هم به طور مشابه دو وضعیت

margin support vector و non-margin support vector برای x_iها برقرار است.

ث) پیش‌بینی و $f(x)=\langle w,x\rangle$ و $w=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)x_i \rightarrow f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\langle x_i,x\rangle$

آیا می‌شود از کرنل بهره برد؟ بله:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)k(x_i,x)$$

ج)

تغییر ϵ:

ϵ نقش برعکس C را دارد. مقدار کوچک ϵ باعث می‌شود یک SVM به صورت hard داشته باشیم که سعی دارد خطای کمتری به تابع SVM یادگرفته شده fit بکند. که پیچیدگی مدل را بالاتر می‌برد

CLASS memo

پس ε های کوچک باعث می‌شود support vector های بیشتری داشته باشیم. کمتر بودن ε باعث می‌شود تا به جواب sparse برسیم.

ε	complexity	Bias	Var
↑	↓	↑	↓
↓	↑	↓	↑

تغییر C

این تغییر نشان دهنده این است که ما چقدر خطاها را Penalize می‌کنیم. پس مقدار بالای C می‌تواند موجب دادن پنالتی بالا به خطاهای کم و اورفیت شدن بشود و مقدار کم آن به معنی نادیده گرفتن ارور و آندرفیت شدن است.

C	complexity	Bias	Var
↑	↑	↓	↑
↓	↓	↑	↓

۲

$|H| = 1000$

$\varepsilon = 0.05$

$\delta = 0.05$

طبق قضیه Haussler PAC Bound داریم: m تا داده آموزشی کافی وجود دارد که تضمین کند هر فرضیه consistent که ارور حداکثر ε با احتمال $1-\delta$ دارد 8

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln(1/\delta)\right)$$

$$\Rightarrow m \geq \frac{1}{0.05}\left(3\ln(10) + \ln\left(\frac{1}{0.05}\right)\right) \approx 198.069$$

پس حداقل به $\boxed{199}$ نمونه نیاز داریم.

DATE / / SUBJECT:

اثبات این قضیه در اسلاید‌ها دکتر سلیمی‌ای موجود است. به اثبات سریع آن اشاره می‌کنیم.

* فرض کنید در فضای H ، k تا H داریم، true error $> \epsilon$.

$$\Rightarrow P(h_i(x_j) = c(x_j)) \le 1-\epsilon \Rightarrow P(h_i \text{ is consistent on m data}) \le (1-\epsilon)^m$$

$$\Rightarrow P(\text{حداقل یک h این ویژگی را داشته باشد}) \le k(1-\epsilon)^m \le |H|(1-\epsilon)^m \le |H| e^{-\epsilon m} \le \delta$$

$$\Rightarrow \ln|H| - \epsilon m \le \ln\delta \rightarrow m \ge \frac{1}{\epsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right) \quad \square$$

۳

الف)

گراف محاسباتی:

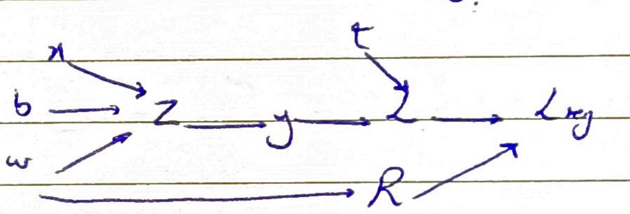
$Z = wx + b$

$y = \sigma(z)$

$L = \frac{1}{2}(y-t)^2$

$R = \frac{1}{2}w^2$

$L_{reg} = L + \lambda R$



$$\frac{\partial L_{reg}}{\partial L} = 1 , \quad \frac{\partial L_{reg}}{\partial R} = \lambda$$

$$\frac{\partial L}{\partial y} = y - t , \quad \frac{\partial L}{\partial t} = t - y$$

$$\frac{\partial y}{\partial z} = \sigma'(z) = \sigma(z)(1-\sigma(z))$$

$$\frac{\partial R}{\partial w} = w , \quad \frac{\partial z}{\partial w} = x , \quad \frac{\partial z}{\partial x} = w , \quad \frac{\partial z}{\partial b} = 1$$

$$\frac{\partial L_{reg}}{\partial b} = 1 \times (y-t) \times \sigma(z)(1-\sigma(z)) \times 1 = (y-t)\sigma(z)(1-\sigma(z)) = (\sigma(wx+b)-t)\,\sigma(wx+b)(1-\sigma(wx+b))$$

CLASS memo

$$\frac{\partial L_{reg}}{\partial w} = \lambda w + 1 \times (y-t) \times \sigma(z)(1-\sigma(z)) \times x$$

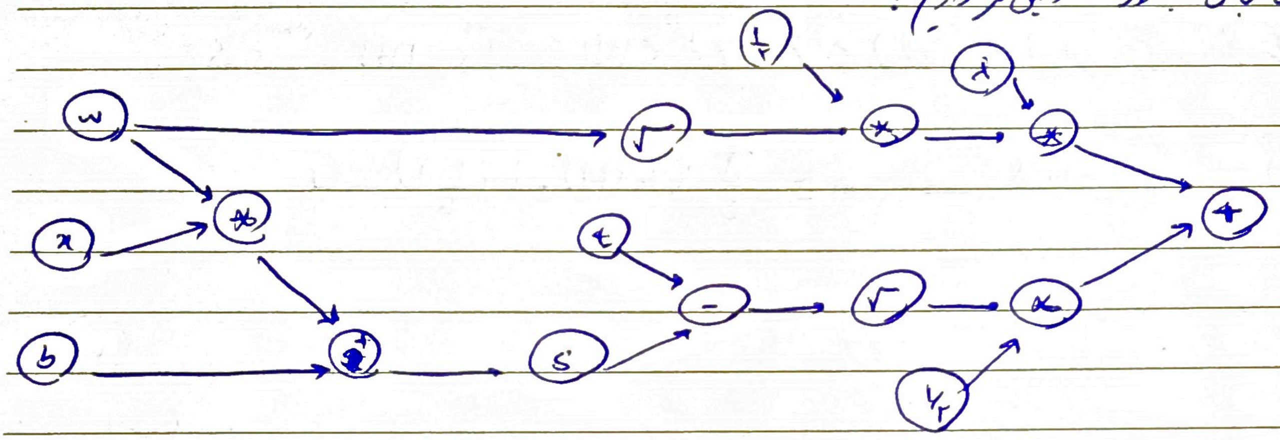
$$= \lambda w + (\sigma(wx+b)-t)\,\sigma(wx+b)(1-\sigma(wx+b))\,x$$

حال پارامترهای شبکه (b,w) قابل آپدیت هستند. که در واقع می‌توان به شکل زیر آپدیت کرد:

$$w = w - \alpha \frac{\partial L(y,\hat{y})}{\partial w}$$

$$b = b - \alpha \frac{\partial L(y,\hat{y})}{\partial b}$$

برای گراف محاسباتی به صورت دقیق‌تر داریم:



ب)

اگر مقداردهی رندوم نباشد، تقارن بین نورون‌ها به وجود می‌آید و نورون‌ها از همان ابتدا با هم تفاوتی نخواهند داشت. این عدم تفاوت باعث می‌شود به سمت استخراج فیچرهای مختلف نروند. پس مهم است که مقداردهی اولیه به شکل تصادفی باشد. (در این حالت در واقع تعداد زیاد نورون بی‌خاصیت شده و شبکه زیاد یاد نمی‌گیرد)

اگر مقادیر پارامترها از ابتدا بزرگ باشند، ممکن است گرادیان‌های به دست آمده در back prop خیلی زیاد بشوند و شاهد تغییرات زیاد و ناگهانی خواهیم بود و امکان دارد converge نکنیم. در واقع شبکه نمی‌تواند train بشود و تابع هزینه را به خوبی optimize نمود.

ج)

با ۲ مقداردهی متفاوت جلو می‌رویم:

DATE / / SUBJECT:

$w = ۰٫۰۵$

$b = ۰٫۱$

$x = ۵$

$y = ۲$

$$\rightarrow \frac{\partial L_{xy}}{\partial w} = \big(\underbrace{\sigma(۰٫۳۵)}_{۰٫۵۸۹} - ۲\big)\big(\sigma(۰٫۳۵)\big)\big(۱ - \sigma(۰٫۳۵)\big) \times ۵ + ۰٫۰۵۸$$

$$= ۰٫۰۵۸ - ۱٫۷۱$$

$$\rightarrow \frac{\partial L_{xy}}{\partial b} = \big(\sigma(۰٫۳۵) - ۲\big)\big(\sigma(۰٫۳۵)\big)\big(۱ - \sigma(۰٫۳۵)\big) = -۰٫۳۴۲$$

$$\Rightarrow w = ۰٫۰۵ - ۰٫۱\,(۰٫۰۵۸ - ۱٫۷۱)$$

$$b = ۰٫۱ - ۰٫۱\,(-۰٫۳۴۲) = \underline{۰٫۱۳۴}$$

$\Longrightarrow$

$w = ۰٫۱$

$b = ۰٫۲$

$x = ۱۰$

$y = ۱$

$$\rightarrow \frac{\partial L_{xy}}{\partial w} = ۷۱ \text{ــ شبیه بالا} = ۰٫۱۸ - ۰٫۴۱۲$$

$$\rightarrow \frac{\partial L_{xy}}{\partial b} = -۰٫۰۴۱۲$$

$$\Rightarrow w = ۰٫۱ - ۰٫۱\,(۰٫۱۸ - ۰٫۴۱۲)$$

$$b = ۰٫۲ - ۰٫۱\,(-۰٫۰۴۱۲) = \underline{۰٫۲۰۴۱}$$