به نام خدا

امنیت و حریم خصوصی در یادگیری ماشین (۴۰۸۱۶)

نیم‌سال اول سال تحصیلی ۱۴۰۵-۱۴۰۴

استاد درس: دکتر امیرمهدی صادق‌زاده

طراح: عرفان صبحائی، علیرضا فرج تبریزی

دانشکده مهندسی کامپیوتر

تمرین شماره ۲ | مهلت تحویل: ساعت ۲۳:۵۹:۵۹ ۲۰ آذر ۱۴۰۴

نکات و قواعد

۱. سوالات خود را زیر پیام مربوطه در Quera با گروه درس مطرح نمایید.

۲. لطفا مطابق تاکید پیشین، حتما آداب‌نامه‌ی انجام تمرین‌های درسی را رعایت نمایید. در صورت تخطی از آیین‌نامه، در بهترین حالت مجبور به حذف درس خواهید شد.

۳. از تحویل تمارین نظری به شکل تایپ شده (فایل Word-Latex یا هر نوع تایپ شده دیگری) بپرهیزید و حتما پاسخ دست نویس تحویل دهید.

۴. در صورتی که پاسخ‌های سوالات نظری را به صورت دست‌نویس آماده کرده‌اید، لطفا تصاویر واضحی از پاسخ‌های خود ارسال کنید. در صورت ناخوانا بودن پاسخ ارسالی، نمره‌ای به پاسخ ارسال شده تعلق نمی‌گیرد.

۵. همه‌ی فایل‌های مربوط به پاسخ خود را در یک فایل فشرده و با نام SPML_HW۲_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ اختلال کلی[1] در شبکه های عصبی عمیق (۲۰ نمره)

فرض کنید $F : \mathbb{R}^n \to \mathbb{R}^K$ یک شبکه عصبی کانولوشنی باشد که خروجی آن لاجیت‌ها است. برای ورودی تصادفی $x \sim \mathcal{D}$ با برچسب صحیح y، تابع هزینه را به‌صورت $L(x) = \ell(F(x), y)$ در نظر می‌گیریم که در آن ℓ یک تابع هزینه مشتق‌پذیر (مثلا آنتروپی متقاطع[2]) است. گرادیان تابع هزینه نسبت به ورودی برابر است با:

$$g(x) = \nabla_x L(x) = \nabla_x \ell(F(x), y).$$

فرض کنید برای یک اختلال کوچک $v \in \mathbb{R}^n$ با قید $\|v\|_2 \le \varepsilon$، تقریب مرتبه اول تیلور برقرار باشد:

$$L(x+v) - L(x) \approx \langle g(x), v \rangle.$$

در نتیجه، مسئله زیر را در نظر بگیرید:

$$\max_{\|v\|_2 \le \varepsilon} \mathbb{E}_{x\sim\mathcal{D}}\big[\, \langle g(x), v \rangle \,\big].$$

(الف) اختلال بهینه تحت قید ℓ_2

نشان دهید که بردار اختلال بهینه و مقدار بیشینه متناظر به صورت زیر هستند:

$$v^\star = \varepsilon \frac{\mu}{\|\mu\|_2}, \qquad \mu = \mathbb{E}_{x\sim\mathcal{D}}[\, g(x) \,], \qquad \mathbb{E}_{x\sim\mathcal{D}}\big[\langle g(x), v^\star \rangle\big] = \varepsilon \|\mu\|_2.$$

[1] Universal Perturbation

[2] Cross Entropy

(ب) میانگین و واریانس

بردار بهینه $v^\star$ را مطابق بخش (الف) در نظر بگیرید و متغیر تصادفی زیر را تعریف کنید:

$$Z = \langle g(x), v^\star \rangle.$$

میانگین و واریانس Z را به دست آورید و نشان دهید که:

$$\mathbb{E}[Z] = \varepsilon \|\mu\|_2, \qquad \mathrm{Var}(Z) = (v^\star)^\top \mathrm{Cov}(g(x))\, v^\star.$$

سوال ۲ یادگیری تخاصمی برای دسته‌بند خطی (۵۰ نمره)

مسئلهٔ یادگیری تخصامی یک دسته‌بند خطی با تابع ضرر زیر را در نظر بگیرید

$$\min_{w} \mathbb{E}_{x,y} \max_{\|\delta\|_p \le \epsilon} [\max(0, 1 - y w^\top (x+\delta)]. \tag{۱}$$

به طوری که x بردار ورودی، $y \in \{-1, +1\}$ دستهٔ هدف، δ تغییری جزئی[۳] هم‌بُعد ورودی و w وزن‌های قابل‌یادگیری مدل هستند.

بخش الف)

فرض کنید $p = \infty$. برای نمونهٔ ثابت x و وزن ثابت w نشان دهید که تغییر جزئی بهینه $\delta^*(x)$ که تابع ضرر را بیشینه می‌کند به شکل زیر است

$$\delta^* = -y\epsilon\, \mathrm{sign}(w).$$

بخش ب)

فرض کنید $p = \infty$. با استفاده از پاسخ بخش (الف)، مسئلهٔ بهینه‌سازی (۱) را به شکلی ساده‌تر بازنویسی کنید.

بخش ج)

مسئلهٔ بهینه‌سازی (۱) را برای حالت کلی $p \ge 1$ به شکلی ساده‌تر بازنویسی کنید. (راهنمایی: ابتدا به کمک بازنویسی بیشینه‌سازی داخلی به شکل دوگان مقدار بهینه را δ^* را یافته و سپس مسئله را بازنویسی کنید.)

سوال ۳ تحلیل همگرایی در روش افزایش گرادیان تصویرشده (PGA) (۱۰۰ نمره)

فرض کنید یک دسته‌بند $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ با پارامترهای θ داریم. همچنین تابع ضرر مشتق‌پذیری به شکل

$$L(x, y; \theta) := \ell(f_\theta(x), y)$$

مفروض است. به ازای یک نمونهٔ ثابت (x, y) مسئلهٔ بیشینه‌سازی زیر را در نظر بگیرید

$$\max_{\|\delta\|_\infty \le \epsilon} L(x+\delta, y; \theta),$$

در مورد تابع L فرض می‌کنیم که نسبت به تغییرات δ مقعر و $L-$ smooth است؛ یعنی داریم

$$\|\nabla_\delta L(x+\delta_1, y; \theta) - L(x+\delta_2, y; \theta)\|_2 \le L\, \|\delta_1 - \delta_2\|_2 .$$

از طرف دیگر، می‌دانیم که هر گام از الگوریتم افزایش گرادیان تصویرشده (PGA)[۴] به شکل زیر است

$$\delta_{t+1} = \Pi_{\|\cdot\|_\infty \le \epsilon} \left(\delta_t + \alpha \nabla_\delta L(x+\delta_t, y; \theta)\right), \qquad \delta_0 = 0,$$

که α نرخ یادگیری و نماد $\Pi_{\|\cdot\|_\infty \le \epsilon}$ نشان‌دهنده عمل‌گر تصویر بر $\ell_\infty -$ ball است.

[۳] perturbation
[۴] Projected Gradient Ascent

بخش الف)

فرض کنید

$$\mathcal{C} = \{\delta : \|\delta\|_\infty \le \epsilon\}.$$

به ازای هر بردار $z \in \mathcal{C}$ و هر بردار $u \in \mathbb{R}^d$ که تصویر آن بر مجموعهٔ $\mathcal{C}$ که به شکل $p = \Pi_{\mathcal{C}}(u)$ نمایش داده می‌شود، نامساوی زیر را ثابت کنید

$$\langle u - p, z - p\rangle \le 0 \tag{۲}$$

بخش ب)

فرض کنید

$$F(\delta) := L(x + \delta, y; \theta)$$

که ورودی بهینهٔ آن را با نماد $\delta^* = \arg\max_{\delta\in\mathcal{C}} F(\delta)$ نشان می‌دهیم.

۱. نشان دهید برای گام مشخص t از اجرای الگوریتم، نامساوی زیر برقرار است:

$$2\langle \delta_{t+1} - \delta_t,\, \delta^* - \delta_{t+1}\rangle = \|\delta_t - \delta^*\|_2^2 - \|\delta_t - \delta_{t+1}\|_2^2 - \|\delta_{t+1} - \delta^*\|_2^2 \tag{۳}$$

۲. به کمک روابط (۲) و (۳) نشان دهید نامساوی زیر برقرار است:

$$\langle \nabla F(\delta_t), \delta^\star - \delta_{t+1}\rangle \le \frac{1}{2\alpha}\Big(\|\delta_t - \delta^\star\|^2 - \|\delta_t - \delta_{t+1}\|^2 - \|\delta_{t+1} - \delta^\star\|^2\Big) \tag{۴}$$

بخش ج)

۱. به کمک تعریف تابع مقعر نشان دهید که برای هر دو ورودی دلخواه x و y داریم

$$F(x) - F(y) \le \langle \nabla F(y), x - y\rangle. \tag{۵}$$

۲. بدون اثبات می‌پذیریم ک برای تابع F که مقعر و $L-$ smooth است رابطهٔ زیر برقرار است

$$F(u) \ge F(v) + \langle \nabla F(v), u - v\rangle - \tfrac{L}{2}\|u - v\|_2^2. \tag{۶}$$

به کمک روابط (۵) و (۶) کرانی بالایی به شکل زیر برای $F(\delta^*) - F(\delta_{t+1})$ بیابید:

$$\cancel{F(\delta^*) - F(\delta_{t+1}) \le \frac{1}{2\alpha}\Big(\|\delta_t - \delta^*\|^2 - \|\delta_{t+1} - \delta^*\|^2\Big)}. \tag{۴}$$

$$F(\delta^\star) - F(\delta_{t+1}) \le \langle \nabla F(\delta_t), \delta^\star - \delta_{t+1}\rangle + \frac{L}{2}\|\delta_{t+1} - \delta_t\|^2$$

بخش د)

فرض کنید $\alpha \le \frac{1}{L}$. نشان دهید

$$F(\delta^*) - F(\delta_{t+1}) \le \frac{1}{2\alpha}\Big(\|\delta_t - \delta^*\|_2^2 - \|\delta_{t+1} - \delta^*\|_2^2\Big). \tag{۷}$$

بخش هـ)

۱. برای مقدار $\sum_{t=0}^{T-1}\big(F(\delta^\star) - F(\delta_{t+1})\big)$ اثبات کنید که کران بالای زیر برقرار است:

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(F(\delta^\star) - F(\delta_{t+1})\big) \le \frac{2\epsilon^2 d}{\alpha T}.$$

(راهنمایی: توجه کنید که $\delta_0 = 0$ است و از آنجایی که $\|\delta^*\|_\infty \le \epsilon$ می‌تواند به سادگی نشان داد که $\|\delta^*\|_2^2 \le 4\epsilon^2 d$ برقرار است.)

۲. ثابت کنید که تابع $F(\cdot)$ در الگوریتم مذکور ناکاهشی است یعنی برای هر گام t داریم $F(\delta_{t+1}) \ge F(\delta_t)$.

بخش و)

۱. نامساوی زیر را ثابت کنید

$$F(\delta^*) - F(\delta_T) \leq \mathcal{O}(1/T)$$

۲. این کران بالا را به زبان ساده تفسیر کنید و توضیح دهید که چرا در شبکه‌های عصبی عمیق این کران لزوماً برقرار نیست؟

سوال ۴ پایداری دسته‌بند خطی و هموارسازی تصادفی (۳۰ نمره)

فرض کنید برای مسئلهٔ دسته‌بندی با برچسب‌های $y \in \{-1, 1\}$ از یک مدل دسته‌بند خطی استفاده می‌کنیم. خروجی این مدل به صورت

$$f(x) = \text{sign}(w^\top x + b)$$

است که در آن w شامل وزن‌های مدل است.

Theorem 0: If $f : \mathbb{R}^d \to [0, 1]^K$ is L-lipschitz, then f is ϵ-robust at x with $\epsilon = \frac{1}{2L}(P_A - P_B)$, where $P_A = \max_i f_i(x)$, $P_B = \max_{j \neq i} f_j(x)$, and $f_k(x)$ is the k-th element of the probability vector $f(x)$.

بخش الف)

شعاع تضمین مقاومت این مدل در برابر اغتشاش جمع‌شونده با ورودی را بیابید.

بخش ب)

فرض کنید مشابه با روش هموارسازی تصادفی، ورودی‌های x را ابتدا با یک نویز گاوسی به صورت

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

جمع می‌کنیم و سپس به عنوان ورودی به دسته‌بند خطی می‌دهیم. حال تابع زیر را تعریف می‌کنیم:

$$g(x) = \arg\max_{c \in \{-1,1\}} \mathbb{P}\big[f(x + \epsilon) = c\big].$$

که در آن تابع g محتمل‌ترین کلاس برای حالات مختلف خروجی دسته‌بند هنگام جمع با نویز را نشان می‌دهد. نشان دهید همواره داریم

$$g(x) = f(x).$$

به عبارت دیگر، نشان دهید هرگاه یک نویز گاوسی با میانگین صفر با ورودی یک دسته‌بند خطی جمع شود، فارغ از پارامتر واریانس آن، همواره محتمل‌ترین کلاس برای خروجی دادهٔ جدید، همان خروجی دسته‌بند به ازای ورودی اصلی است.

موفق باشید.