به نام خدا

# امنیت و حریم خصوصی در یادگیری ماشین (۴۰۸۱۶)

نیم‌سال اول سال تحصیلی ۱۴۰۵-۱۴۰۴

استاد درس: دکتر امیرمهدی صادق‌زاده

طراحان: فیروزه ابریشمی، علیرضا فرج تبریزی

دانشکده مهندسی کامپیوتر

تمرین شماره ۱ | مهلت تحویل: ساعت ۲۳:۵۹:۵۹ ۵ 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_HW۱_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

## سوال ۱ لیپشیتس (۹۰+۷۰ نمره)

### بخش اول) خواص پایه‌ای توابع لیپشیتس (۳۰+۳۰ نمره)

در این بخش، چند خاصیت بنیادی توابع لیپشیتس را که به عنوان ابزارهای ریاضی برای تحلیل شبکه‌های عصبی به کار می‌روند، اثبات و بررسی می‌کنیم.

۱. جمع توابع لیپشیتس (۱۰ نمره): ثابت کنید اگر $f : \mathbb{R}^n \to \mathbb{R}^m$ یک تابع $L_f$ـلیپشیتس و $g : \mathbb{R}^n \to \mathbb{R}^m$ یک تابع $L_g$ـلیپشیتس باشد، آنگاه تابع مجموع $h(x) = f(x) + g(x)$ نیز لیپشیتس است. کران بالای ثابت لیپشیتس آن $(L_h)$ چیست؟

۲. ترکیب توابع لیپشیتس (۱۰ نمره):

(الف) اثبات خاصیت ترکیب: ثابت کنید اگر $g : \mathbb{R}^n \to \mathbb{R}^m$ یک تابع $L_g$ـلیپشیتس و $f : \mathbb{R}^m \to \mathbb{R}^k$ یک تابع $L_f$ـلیپشیتس باشد، آنگاه تابع مرکب $h(x) = f(g(x))$ نیز یک تابع لیپشیتس است. کران بالای ثابت لیپشیتس آن $(L_h)$ چیست؟

(ب) کران ثابت لیپشیتس یک لایه: یک لایه شبکه عصبی به صورت $f_k(x) = \sigma(W_k x)$ را در نظر بگیرید که در آن $\sigma$ یک تابع فعال‌سازی[1] $L_\sigma$ـلیپشیتس و $W_k$ ماتریس وزن است. با استفاده از خاصیت ترکیب توابع که در بند (الف) اثبات کردید، نشان دهید که ثابت لیپشیتس این لایه $(L_k)$ توسط $L_k \leq L_\sigma \cdot \|W_k\|_2$ کران‌دار می‌شود. (یادآوری: ثابت لیپشیتس تبدیل خطی $x \mapsto Wx$ برابر با نورم طیفی ماتریس، یعنی $\|W\|_2$ است).

[1] Activation Function

۳. ارتباط با مشتق (۱۰+۳۰ نمره):

(الف) حالت یک‌بعدی: برای یک تابع مشتق‌پذیر $f : \mathbb{R} \to \mathbb{R}$، ثابت کنید که $f$ یک تابع $L$ـلیپشیتس است اگر و تنها اگر برای تمام مقادیر $x$ داشته باشیم $|f'(x)| \leq L$.

(ب) شرط کافی برای لیپشیتس بودن در حالت کلی: (امتیازی، اما در صورت استفاده در بخش‌های بعدی باید آن را ثابت کنید.)
ثابت کنید اگر تابع $f : \mathbb{R}^n \to \mathbb{R}^m$ در یک مجموعه محدب[۲] $\mathcal{D}$ مشتق‌پذیر باشد و نرم طیفی[۳] ماتریس ژاکوبین[۴] آن برای تمام $x \in \mathcal{D}$ کران‌دار باشد (یعنی $\|J_f(x)\|_2 \leq L$)، آنگاه $f$ در آن مجموعه یک تابع $L$ـلیپشیتس است.
راهنمایی اثبات:

- (گام ۱) تابع کمکی $g(t) = f(y + t(x - y))$ را برای $t \in [0, 1]$ تعریف کنید.
- (گام ۲) با استفاده از قضیه اساسی حسابان نشان دهید $f(x) - f(y) = \int_0^1 g'(t)\, dt$.
- (گام ۳) مشتق $g'(t)$ را با قاعده زنجیره‌ای بر حسب ژاکوبین $J_f$ بیابید. سپس از طرفین انتگرال نورم بگیرید و با استفاده از نامساوی مثلثی برای انتگرال‌ها ($\|\int h(t)dt\|_2 \leq \int \|h(t)\|_2 dt$) و فرض کران‌دار بودن $\|J_f\|_2 \leq L$ اثبات را کامل کنید.

## بخش دوم) تحلیل ثابت لیپشیتس برای اجزای شبکه عصبی (۶۰+۴۰ نمره)

با استفاده از خواص بخش قبل، ثابت لیپشیتس را برای چند جزء متداول در شبکه‌های عصبی عمیق تحلیل می‌کنیم.

۱. تحلیل لایه‌های نرمال‌سازی (۲۰+۲۰ نمره): در این سوال، خواص لیپشیتس دو مکانیزم نرمال‌سازی رایج را بررسی می‌کنیم.

(الف) نرمال‌سازی دسته‌ای[۵]: در زمان استنتاج[۶]، آماره‌های میانگین ($\mu \in \mathbb{R}^d$) و واریانس ($\sigma^2 \in \mathbb{R}^d$) ثابت هستند. فرض کنید ورودی لایه یک ماتریس دسته $X \in \mathbb{R}^{B \times d}$ باشد (که $B$ اندازه دسته و $d$ تعداد ویژگی‌ها است). تبدیل نرمال‌سازی برای این ماتریس به صورت درایه به درایه روی ویژگی‌ها اعمال می‌شود:

$$\mathrm{BN}(X) = \mathrm{Scale}(X, \gamma) + \beta$$

که در آن $\gamma, \beta \in \mathbb{R}^d$ پارامترهای برداری هستند.

i. نشان دهید که این تبدیل را می‌توان به فرم ماتریسی فشرده‌ی زیر (که یک تبدیل آفین روی فضای ماتریس‌هاست) نوشت:

$$\mathrm{BN}(X) = X\Lambda + \mathbf{1}_B \mathbf{b}^T$$

در این رابطه $\Lambda \in \mathbb{R}^{d \times d}$ و $\mathbf{b} \in \mathbb{R}^d$ را بر حسب پارامترهای $(\gamma, \beta, \mu, \sigma^2)$ تعیین کنید و ساختار ماتریس $\Lambda$ را مشخص نمایید. ($\mathbf{1}_B$ بردار ستونی شامل تمام یک‌ها به طول $B$ است).

ii. ثابت لیپشیتس دقیق این تبدیل را نسبت به نرم فروبنیوس ($\|\cdot\|_F$) بیابید. به عبارتی کوچکترین مقدار $K$ را پیدا کنید که برای هر دو ماتریس ورودی $X$ و $Y$ داشته باشیم:

$$\|\mathrm{BN}(X) - \mathrm{BN}(Y)\|_F \leq K\|X - Y\|_F$$

iii. با توجه به پاسخ قسمت قبل، اگر واریانس تخمین‌زده شده برای یکی از ویژگی‌ها (مثلاً ویژگی $j$ـام) بسیار کوچک باشد ($\sigma_j^2 \to 0$)، این موضوع چه تأثیری بر پایداری عددی و حساسیت خروجی نسبت به تغییرات ورودی دارد؟

(ب) نرمال‌سازی لایه[۷]: (امتیازی) تابع نرمال‌سازی لایه با ترم $\epsilon > 0$ در مخرج به صورت $\mathrm{LN}_\epsilon(x) = \frac{x - \mu_x}{\sqrt{\sigma_x^2 + \epsilon}}$ تعریف می‌شود. ثابت کنید که این تابع $1/\sqrt{\epsilon}$ـلیپشیتس است. یعنی:

$$\|\mathrm{LN}_\epsilon(x) - \mathrm{LN}_\epsilon(y)\|_2 \leq \frac{1}{\sqrt{\epsilon}}\|x - y\|_2$$

۲. لایه کانولوشنی[۸] (۲۰ نمره):

[2] Convex Set
[3] Spectral Norm
[4] Jacobian Matrix
[5] Batch Normalization
[6] inference time
[7] Layer Normalization
[8] Convolutional Layer (CNN)

(الف) کانولوشن یک‌بعدی: نشان دهید که یک لایه کانولوشن یک‌بعدی با کرنل به طول $k$ لیپشیتس است (محاسبه ضریب لازم نیست).

(ب) کانولوشن دوبعدی: نشان دهید که یک لایه کانولوشن دوبعدی با کرنل $k$ نیز لیپشیتس است (محاسبه ضریب لازم نیست).

۳. مکانیزم توجه در ترنسفورمر [9] (۲۰+۲۰ نمره): در این سوال، مکانیزم توجه را به صورت تابعی از یک ورودی $x$ در نظر می‌گیریم. فرض کنید بردارهای query و key از ورودی $x$ توسط ماتریس‌های وزن $W_q$ و $W_k$ به صورت $q = W_q x$ و $k = W_k x$ تولید می‌شوند. تابع امتیازدهی $f(x) = q^T k$ را در نظر بگیرید.

(الف) نمایش به صورت فرم درجه دو: ابتدا نشان دهید که $f(x)$ یک فرم درجه دو[10] بر حسب $x$ است. یعنی آن را به صورت $f(x) = x^T M x$ بنویسید و ماتریس $M$ را بر حسب $W_q$ و $W_k$ مشخص کنید.

(ب) اثبات غیر لیپشیتس بودن (کلی): با استفاده از فرم درجه دو که به دست آوردید، ثابت کنید که تابع $f(x)$ به صورت کلی[11] لیپشیتس نیست.

(ج) اثبات لیپشیتس بودن (محلی): (امتیازی) ثابت کنید که $f(x)$ به صورت محلی[12] لیپشیتس است. یعنی نشان دهید که برای هر ناحیه کران‌دار (مانند یک گوی $\|x\|_2 \leq R$)، یک ثابت $L_R$ (وابسته به $R$) وجود دارد که تابع در آن ناحیه $L_R$ـلیپشیتس است.

## سوال ۲ بررسی حمله DeepFool (۹۰ نمره)

هدف DeepFool یافتن کوچک‌ترین اغتشاش در نرم ۲ است که classifier را وادار می‌کند کلاس خروجی را تغییر دهد:

$$\Delta(x; \hat{k}) = \min_{r \in \mathbb{R}^d} \frac{1}{2}\|r\|_2^2, \quad \text{s.t.} \quad \hat{k}(x+r) \neq \hat{k}(x).$$

### بخش اول) داده دو کلاسه (۴۰ نمره)

در این بخش، دسته‌بند باینری را بررسی می‌کنیم: ابتدا برای مدل خطی جواب فرم بسته را بدست می‌آوریم، سپس با تقریب تیلور برای هر تابع دلخواه این محاسبه را تکرار می‌کنیم.

۱. مدل خطی دو کلاسه (۲۰ نمره): با فرض $\hat{k}(x) = \text{sign}(f(x))$، $f(x) = w^\top x + b$ و $F = \{x \in \mathbb{R}^d : \ w^\top x + b = 0\}$ مسئله بهینه‌سازی زیر را حل کنید. (راهنمایی: برای حل جبری از روش لاگرانژ و برای حل هندسی از تصویر کردن نقطه بر ابر صفحه استفاده کنید.)

$$r^*(x_0) = \arg\min_{r \in \mathbb{R}^d} \frac{1}{2}\|r\|_2^2 \quad \text{s.t.} \quad \text{sign}(f(x_0 + r)) \neq \text{sign}(f(x_0)).$$

۲. مدل دلخواه دو کلاسه (۲۰ نمره): فرض کنید $f : \mathbb{R}^d \to \mathbb{R}$ در $x_i$مشتق پذیر است. با فرض برقرار بودن تقریب مرتبه اول تیلور در هر گام برای تغییرات کوچک، مسئله بهینه‌سازی در گام $i$ام را حل کنید:

$$f(x_i + r_i) \approx f(x_i) + \nabla f(x_i)^\top r_i, \quad \arg\min_{r_i \in \mathbb{R}^d} \frac{1}{2}\|r_i\|_2^2 \quad \text{s.t.} \quad f(x_i) + \nabla f(x_i)^\top r_i = 0$$

### بخش دوم) دسته‌بند خطی و داده چند کلاسه (۵۰ نمره)

در اینجا فقط به بررسی حالت خطی می‌پردازیم، مشابه بخش قبل می‌توان با فرض برقراری تقریب مرتبه اول تیلور برای هر مدل دسته‌بند دلخواه نیز جواب فرم بسته بدست آورد که خروجی محاسبات در الگوریتم ۲ مقاله مربوطه موجود است.

۱. مدل خطی (ماتریسی) چند کلاسه (۴۰ نمره): بگذارید $f(x) = W^\top x + b \in \mathbb{R}^C$ با $W = [w_1, \ldots, w_C] \in \mathbb{R}^{d \times C}$ و $b \in \mathbb{R}^C$ و $\hat{k}(x) = \arg\max_k f_k(x)$. همچنین $c_0 := \hat{k}(x_0)$.

(الف) صورت‌بندی (۱۰ نمره): ابتدا صورت مسئله بهینه‌سازی مقید مناسب برای این حالت را بنویسید.

[9] Transformer
[10] Quadratic Form
[11] globally
[12] locally

(ب) نزدیک‌ترین ابر صفحه تصمیم‌گیری (۲۰ نمره): اگر $\hat{\ell}(x_0)$ را نزدیک ترین مرز تصمیم‌گیری به نقطه $x_0$ در نظر بگیریم نشان دهید

$$\hat{\ell}(x_0) = \arg\min_{k \neq c_0} \frac{|\, f_k(x_0) - f_{c_0}(x_0)\,|}{\|\, w_k - w_{c_0}\,\|_2}.$$

(ج) پیداکرن اغتشاش بهینه (۲۰ نمره): اگر $H_{\hat{\ell}(x_0)} = \{x : \; f_{\hat{\ell}(x_0)}(x) = f_{c_0}(x)\}$، نتیجه بگیرید

$$r^*(x_0) = \frac{|\, f_{\hat{\ell}(x_0)}(x_0) - f_{c_0}(x_0)\,|}{\|\, w_{\hat{\ell}(x_0)} - w_{c_0}\,\|_2^2}\,(w_{\hat{\ell}(x_0)} - w_{c_0}),$$

موفق باشید.