دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۰ـ۹۹

مدرس: حمید بیگی

تمرین سری اول | یادگیری عمیق | زمان تحویل: ۶ فروردین‌ماه

نکات زیر را رعایت کنید:
فایل گزارش را به همراه تمامی کدها در یک فایل فشرده و با عنوان #HW1_STD در سایت Quera.ir بارگذاری نمایید.
بخش‌های پیاده‌سازی مربوط به هر سوال را در نوت‌بوک ارائه‌شده و فایل‌های پایتون مربوط به آن تکمیل کنید و در یک پوشه قرار دهید. سوالات خود را از طریق Piazza مطرح کنید.

مسئله‌ی ۱. Linear regression (۱۲ نمره)

فرض کنید n داده آموزش به صورت $D = (x^1, y^1), ..., (x^n, y^n)$ در اختیار داریم که هر کدام از x ها، d بعدی می‌باشد. می‌خواهیم از رگرسیون خطی با تابع هزینه SSE استفاده کنیم که به فرم زیر است:

$$J(w) = \sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)})^2$$

(آ) (۲/۵ نمره) رابطه بهینه w را به دست آورید.

(ب) (۵.۱ نمره) مشکلات استفاده مستقیم از رابطه‌ی قسمت قبل را بیان کنید و برای آن‌ها راه حلی ارائه دهید.

(ج) (۱/۵ نمره) اگر به تابع هزینه جمله‌ی منظم ساز $\| w \|^2$ را بیافزاییم، فرم بسته پاسخ بهینه w را به دست آورید.

(د) (۲/۵ نمره) رگرسیون خطی وزن دار، تعمیمی روی رگرسیون خطی است که در آن به هر یک از داده‌ها وزنی اختصاص داده می‌شود:

$$J(w) = \sum_{i=1}^{n} F_i (y^{(i)} - w^T x^{(i)})^2$$

فرم بهینه‌ی w را برای این تابع هزینه به دست آورید.

(ه) (۴ نمره) اگر مسئله را به صورت احتمالاتی بنویسیم، خواهیم داشت:

$$\hat{w} = argmin_w E_{x,y}[(y - w^T x)^2]$$

مقدار بهینه‌ی w را بر حسب ماتریس خودهمبستگی $R = E_x[xx^T]$ و بردار همبستگی $c = E_{x,y}[xy]$ محاسبه کنید. سپس خطا را به صورت جمع دو خطای ساختاری و تقریب، تفکیک کنید و تعبیر هر یک را بیان نمایید.

مسئله‌ی ۲. perceptron (۱۱ نمره)

به سوالات زیر بر اساس دسته بند پرسپترون پاسخ دهید:

(آ) (۳ نمره) نشان دهید که ترتیب داده‌ها در بردار وزن حاصل از نسخه‌ی تک‌نمونه‌ی این روش می‌تواند اثرگذار باشد.
یعنی نشان دهید اگر روش پرسپترون را روی داده‌ها اجرا کنیم و در هر چرخه یک داده را بررسی کنیم و بردار وزن را بروزرسانی کنیم، ترتیب بررسی داده‌ها در زمان آموزش در بردار نهایی می‌تواند اثرگذار باشد.

(ب) (۲ نمره) به طور شهودی نمودار تابع هزینه ی مربوط به batch perceptron و پرسپترون تک نمونه را مقایسه و توصیف کنید.

(ج) (۶ نمره) فرض کنید مسأله‌ی دسته‌بندی را برای داده‌هایی از دو کلاس حل کرده‌ایم و بردار w^* نتیجه شده است به طوری که همه‌ی داده‌ها را به درست با حاشیه‌ی γ دسته‌بندی می‌کند یعنی داریم $w^{*T}x_iy_i > \gamma, \forall i$ با دانستن این حقیقت که تمام داده‌ها در ابرکره‌ای با شعاع R قرار دارند، ثابت کنید تعداد گام‌های لازم برای رسیدن به این بردار نهایی حداکثر $\frac{R^2\|w^*\|_2^2}{\gamma^2}$ گام بوده است. از استقرا روی بردار در هر گام استفاده کنید وفرض کنید بردار وزن اولیه بردار تماما صفر باشد.

مسئله‌ی ۳. backpropagation (۶ نمره)

به دو سوال زیر در رابطه با back propagation پاسخ دهید:

(آ) (۳ نمره) کنید تابعی داریم که یک ورودی دو بعدی $x = (x_1, x_2)$ را به عنوان ورودی می‌گیرد و دارای دو پارامتر $w = (w_1, w_2)$ است که $f(x, w) = \sigma(\sigma(x_1w_1)w_2 + x_2)$ و $\sigma(x) = \frac{1}{1+e^{-x}}$.
ما از backpropagation استفاده می‌کنیم تا مقدار درست پارامترها را تخمین بزنیم. در ابتدا مقدار هر دو پارامتر را صفر قرار دهید و فرض کنید $x_1 = 1, X_2 = 0, y = 5$. سپس شبکه عصبی متناظر با مسئله را کشیده و مقدار $\frac{\partial f}{\partial w_2}$ را بیابید. حال مقدار $\sigma(x_1w_1)w_2 + x_2$ را به عنوان o_2 و x_1w_1 را به عنوان o_1 بازنویسی کنید.(یعنی شبکه عصبی باید دو خروجی داشته باشد.)

(ب) (۳ نمره) اگر نرخ یادگیری برابر ۰/۵ باشد، مقدار w_2 را بعد از یک مرحله به روز رسانی توسط الگوریتم انتشار به عقب به دست آورید.

مسئله‌ی ۴. Activation Functions (۸ نمره)

۱. (۴ نمره) توابع فعال‌سازی sigmoid σ و tanh به شکل زیر تعریف می‌شوند:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

(آ) (۲ نمره) ماتریس ژاکوبین $\partial y/\partial z$ را برای تابع فعال‌سازی tanh که بر روی تمام عناصر یک لایه اجرا شده است، به دست آورید.

(ب) (۱ نمره) یکی از مشکلات تابع sigmoid اشباع[۱] است. توضیح دهید که آیا استفاده از tanh این مشکل را حل می‌کند؟

(ج) (۱ نمره) توضیح دهید که استفاده از tanh به جای sigmoid چرا و چگونه باعث بهبود بهینه‌سازی می‌شود؟

۲. (۱ نمره) مشکل vanishing gradient را توضیح دهید و راه حلی برای رفع این مشکل ارائه دهید.

۳. (۳ نمره) توضیح دهید که چگونه می‌توان از سرریز در محاسبات softmax جلوگیری کرد؟ ادعای خود را اثبات کنید.

مسئله‌ی ۵. Regularization (۱۲ نمره)

۱. (۲ نمره) توضیح دهید که چرا در شبکه‌هایی که batch normalization استفاده می‌شود، ضریب یادگیری را می‌توان افزایش داد.

۲. (۳ نمره) به سوالات زیر در مورد Drop-out پاسخ دهید:

(آ) توضیح دهید که چرا Drop-out مانند منظم‌ساز عمل می‌کند؟

(ب) توضیح دهید که چرا Drop-out عملکردی شبیه ensemble-learning دارد؟

(ج) استفاده از Drop-out در حین test و train چه تفاتی دارد و این تفاوت به چه علت است؟

۲. (۴ نمره) روش‌های multi-task learning و parameter sharing را به طور مختصر توضیح دهید و بگویید که چگونه باعث افزایش generalization می‌شوند.

۳. (۱ نمره) توضیح دهید که چرا منظم‌سازها بر روی بایاس‌های شبکه اعمال نمی‌شوند.

۴. (۲ نمره) دو نحوه‌ی منظم‌سازی زیر را با هم مقایسه کنید.

- اضافه کردن جمله‌ی منظم‌ساز به تابع هزینه

$$L(w) + \alpha \|w\|_2^2$$

- منظم‌ساز بیشینه‌ـ‌نرم (استفاده از جمله‌ی منظم‌ساز به عنوان قید مسئله)

$$\min_w L(w) \quad \texttt{s.t.} \|w^{[l]}\|_2^2 \leqslant c, l = 1, ..., L$$

مسئله‌ی ۶. Optimization (۱۵ نمره)

۱. (۵ نمره) فرض کنید که g گرادیان تابع f و H ماتریس هسین[۲] آن باشد. نشان دهید که در الگوریتم گرادیان کاهشی ضریب یادگیری بهینه‌ی η^* از رابطه‌ی $\frac{\vec{g}^T\vec{g}}{\vec{g}^T H\vec{g}}$ به دست می‌آید.

---

[۱] saturation

[۲] Hessian

۲. (۲ نمره) نقاط زینی[۳] در بهینه‌سازی چه مشکلی ایجاد می‌کنند؟ توضیح دهید که چرا در ابعاد بالا تعداد این نقاط از نقاط بهینه‌ی محلی بیشتر است؟

۳. (۸ نمره) روش‌های GD، momentum، Nestrov-momentum، RMS-Prob و ADAM را با هم مقایسه کنید و بگویید هر کدام چه مشکلی در روش‌های قبلی را حل می‌کنند و چه مزایا و معایبی دارند.

مسئله‌ی ۷. MLP (عملی ـ ۴۰ نمره)

۱. در این سوال هدف پیاده‌سازی شبکه‌ی عصبی چندلایه با استفاده از numpy است. فایل نوتبوک Q7a را مطالعه کنید و طبق دستورالعمل‌های آن، قسمت‌های مشخص شده در فایل‌های پایتون و نوتبوک را تکمیل کنید. در پایان سوال با استفاده از ماژول‌هایی که پیاده‌سازی کرده‌اید مسئله‌ی طبقه‌بندی تصاویر مجموعه داده‌ی CIFAR10 را حل خواهید کرد.

۲. هدف این قسمت سوال آشنایی و کار با فریمورک pytorch است. برای پاسخ به این سوال نوتبوک Q7b را مطالعه کرده و قسمت‌های خواسته شده را پیاده‌سازی کنید.

موفق باشید.

[۳] saddle point