محمدرضا امیرزاده
۴۰۱۲۰۶۹۹۹

تمارین سری نهم درس SPML

۶ شنبه Saturday 27 May 2023 ۷ ذی‌القعده ۱۴۴۴ خرداد

(۱) الف)

$$\max_{D_{poison}} l_{attack}(\theta, D_{test})$$

$$s.t \quad \theta = \arg\min_{\theta'} l_{train}(\theta', D_{train} + D_{poison})$$

هدف داوگان مسدم یعنی D_{poison} با عمل کردن مسند بیشینه سازی و سطحی بالا به دست می‌آید.

ب) مشکل روش ارائه شده این است که تابع هش داده‌شده داده‌ها را به صورت یکتا به یک دسته نگاشت می‌کند، و واضح همپوشانی داده‌ها پس از تقسیم‌بندی آن‌ها وجود دارد. به عبارت دیگر اگر $h(s)$ اعداد غیر از اعداد اول تولید کند، باقی مانده $h(s)$ بر بیش از یک دسته i صدق می‌کند و در نتیجه این داده به چندین دسته تعلق خواهد گرفت. با توجه به این نکته می‌توان تضمین با توجه به این روش ارائه داد، زیرا اگر یک نمونه متعلق به یکی از K دسته تعلق داشته باشد باعث می‌شود که نتیجه حداکثر یکی از دسته‌بندها عوض شود. پس در پایان این تقسیم‌بندی می‌توان به صورت زیر عمل کرد.

$$P_i^S = \{ s \in S \mid h(s) = K \bmod i \}$$

همچنین توجه کنید که باید راهکاری برای حالتی که بیش از یک دسته شرکا تعداد رأی یکسانی دارند ارائه شود (به عنوان مثال انتخاب دسته با کمترین اندیس) که در اینجا ارائه نشده است.

(۲) الف) از آنجایی که در این الگوریتم، کلاس هدف مشخص نشده است و به عنوان ورودی داده نشده است بنابراین الگوریتم untargeted است و برای هدف‌مند کردن آن کافی است دو خط زیر را تغییر دهیم:

4 while $P_y = \max_{y'} P_{y'}$ $\Rightarrow$ while $P_t \neq \max_{y'} P_{y'}$

8 if $P'_y < P_y$ $\Rightarrow$ if $P'_t > P_t$

ب) دلیل انتخاب بردارهای orthonormal آن است که تغییرات در هر یک از جهت‌ها توسط برخی از جهات دیگر خنثی نشود. همچنین از اینکه در بیش از ε حرکت کنیم جلوگیری می‌کند، زیرا انتخاب این بردارها بدون جایگذاری است، بنابراین L_∞ آشفتگی از ε فراتر نمی‌رود.

ج) در این الگوریتم به این گونه است که اگر مقدار P'_y از یک آستانه معین بیشتر به P_y از این رو باشد آنگاه می‌توان اینطور برداشت کرد که اگر در جهت مخالف حرکت کنیم، احتمال گفته شده کاهش خواهد یافت. بنابراین می‌توان از این تکنیک به عنوان یک heuristic برای کاهش تعداد کوئری‌ها استفاده کرد.

د) $q_t^T \cdot q_{t'} = 0 \Rightarrow \|\delta_T\|_2^2 = \|\sum_{t=1}^{T} \alpha_t q_t\|_2^2 = \sum_{t=1}^{T} \|\alpha_i q_t\|^2$

$= \sum_{t=1}^{T} \alpha_t^2 \|q_t\|^2 \leq T\varepsilon^2 \Rightarrow \|\delta\|_2 \leq \sqrt{T}\varepsilon$

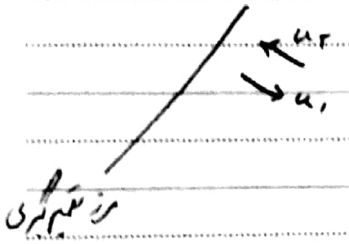
وقتی که روی تعداد کوئری‌ها محدودیت وجود دارد، برای ازای این L_2 distance نیاز است که ε را متناسب بزرگ‌تری انتخاب کنیم. اما اگر بخواهیم L_∞ کم باشد ولی L_2 حفظ شود، باید تعداد کوئری‌ها افزایش یابد که در نتیجه آن T افزایش یابد. بنابراین یک معامله بین تعداد کوئری و میزان آشفتگی وجود دارد.

سه‌شنبه
Tuesday
30 May 2023
۱۰ ذی‌القعده ۱۴۴۴
خرداد

(۳) الف) ما علامت گرایان به معتقدی هستیم که علامت گرادیان در جهتی که و از آن می‌یابد یک هست، در جهتی که و کاهش می‌یابد ۱- .

ما در این شکل دیدی می‌توان دریافت که با حرکت در جهت u_2، به مرز تقسیم‌گیری نزدیک شده و در نتیجه مقدار و کاهش می‌یابد. ولی با حرکت در جهت u_1 از مرز تقسیم‌گیری دور شده و مقدار و از این می‌یابد. بنابراین علامت گرادیان در جهت u_1 باید برابر 1 و در جهت u_2 باید برابر 1- باشد.

$$\operatorname{sgn} g(\theta+\varepsilon u)-g(\theta)=\begin{cases}+1 & : f\left(x+g(\theta)\frac{\theta+\varepsilon u}{\|\theta+\varepsilon u\|}\right)=1 \\ -1 & : \text{o.w}\end{cases}$$



ب) اگر نتند از علامت گرادیان استفاده کنیم، نیاز به تعداد کوئری‌های بیشتری برای تخمین گرادیان داریم زیرا علامت گرادیان نسبت به خود گرادیان اطلاعات کمتری دارد.
اما به‌هرحال برای تولید نمونه‌های خصمانه، استفاده از علامت گرادیان نیز کفایت می‌کند.

(۴) الف) Time dependant prior : می‌توان گفت گرادیان‌هایی که از طریق روش NES در هر step از حمله PGD محاسبه می‌شوند، همبستگی یا correlation بسیار زیادی با هم دارند. در واقع cosine similarity گرادیان‌های متوالی حدود ۰.۹ است که شامل این مشاهدات، برای استفاده از این مشاهده، در این مقاله از گرادیان تخمینی در مرحله $t-1$ به عنوان یک prior برای گرادیان مرحله t استفاده شده است.

Data dependant prior : از آنجایی که در روش‌های تصویری، مشابهت محلی نمایی یا Spatial local similarity (پیکسل‌هایی که از لحاظ مکانی نزدیک هم هستند، مقدار مشابهی دارند)، بنابراین این مشابهت در گرادیان ما نیز وجود دارد. به طوری که تغییر هرگاه در مختصه (i,j) و (k,l) در داده ورودی x نزدیک و مشابه باشند آنگاه $\nabla_x L(x,y)_{kl} \approx \nabla_x L(x,y)_{ij}$. برای استفاده از این مشاهده، در این مقاله تعدد $\nabla_x L(x,y)$ که هم‌شکل با x است را با یک کرنل $(K,K,1)$ و استراید $(K,K,1)$، Averge Pooling می‌گیرد تا به این ترتیب اندازه سطح با ضریب K^2 کوچک شود، نهایت به صورت معکوس سطح upscale می‌شود، و به سایز اصلی برمی‌گردد.

ب) اگر تقریب کنیم $l_t(v) = -\langle \nabla L(x,y), v \rangle$ آنگاه، ربطی که برای تخمین گرادیان این تابع که با Δ_t نمایش داده می‌شود، استفاده از spherical gradient estimator می‌شود. به این صورت که از دو کوئری روی جهت تصادفی u که به صورت نمونه‌برداری تقریبی با antithetic sampling انتخاب می‌شوند بهره می‌گیریم:

$$\Delta_t \approx \frac{l_t(v+\delta u) - l(v-\delta u)}{\delta} u \quad ; \quad u \sim N(0, \frac{1}{d} I)$$

برای محاسبه عبارت بالا تقریب می‌کنیم $\begin{cases} q_1 = v + \delta u \\ q_2 = v - \delta u \end{cases}$ پس تابع loss l_t را در نقاط q_1 و q_2 به صورت زیر تخمین می‌زنیم:

$$\begin{cases} l_t(q_1) = -\langle \nabla L(x,y), q_1 \rangle \approx \dfrac{L(x,y) - L(x+\varepsilon q_1, y)}{\varepsilon} \\ l_t(q_2) = -\langle \nabla L(x,y), q_2 \rangle \approx \dfrac{L(x,y) - L(x+\varepsilon q_2, y)}{\varepsilon} \end{cases}$$

در تخمین بالا از مفهوم مشتق جهتی (x,y) ما در جهت q استفاده کردیم. در نهایت Δ_t به صورت زیر محاسبه می‌شود:

$$\Delta_t = \frac{l_t(q_1) - l_t(q_2)}{\delta} u = \frac{L(x+\varepsilon q_2, y) - L(x+\varepsilon q_1, y)}{\delta}$$

بنابراین توانستیم Δ_t را تنها با دو کوئری به L محاسبه کنیم.