گزارش پروژه

دانشکده مهندسی کامپیوتر

نام و نام خانوادگی دانشجو: سید احسان حسن بیگی

استاد: دکتر نجفی

پاییز ۱۴۰۳

فهرست مطالب

۱. مقدمه

در این گزارش به بررسی مقاله‌ی زیر می‌پردازیم:

Nearly Tight Sample Complexity Bounds for Learning Mixtures of Gaussians via Sample Compression Schemes [1]

به طور کلی یادگیری توزیع (distribution learning) یک مسئله‌ی معروف در حوزه‌ی یادگیری است و هدف آن تخمین توزیع از روی داده ای مشاهده شده است. یک جنبه‌ی مهم یادگیری توزیع، ارائه‌ی sample complexity مورد نیاز برای یادگیری آن خانواده از توزیع است. پیش از این کار، کران های ارائه شده برای خانواده هایی از توزیع ها که به صورت mixture هستند، به اندازه‌ی کافی tight نبودند. این مقاله کران بالا و پایین tight تری برای خانواده های توزیع که از جنس mixture هستند، به ویژه mixtures of gaussians و mixtures of axis-aligned gaussians، ارائه می‌دهد.

برای به دست آوردن کران بالا، مفهوم جدیدی به نام sample compressibility ارائه می‌شود که بیان می‌کند اگر بتوان پارامترهای توزیع را توسط تعداد اندکی از نمونه های دست‌چین شده encode کرد، یادگیری آن توزیع نیز با تعداد کمی از نمونه ها امکان پذیر است. همچنین نشان داده می‌شود که اگر توزیع های پایه، ویژگی sample compressibility را داشته باشند، خانواده‌ی mixture و product آن ها نیز این خاصیت را دارند و می‌توان برای یادگیری آن ها نیز کران tight ارائه کرد.

به طور خاص برای یادگیری توزیع mixture of k gaussians در فضای R^d، کران های بالا و پایین به یک order می‌رسند و می‌توان گفت که تعداد $\widetilde{\Theta}(kd^2/\varepsilon^2)$ سمپل برای یادگیری این خانواده از توزیع ها لازم و کافی است. همچنین شایان ذکر است که این کار بر روی یافتن الگوریتم بهینه تمرکز ندارد. با اینکه sample complexity ارائه شده بهینه است، اما time complexity الگوریتم، نسبت به بُعد فضا و تعداد mixture ها نمایی است.

۲. پیش‌نیازها و تعاریف

$\boldsymbol{\varepsilon - approximation}$: توزیع $\hat{g}$ یک $\varepsilon - approximation$ از توزیع g است، اگر $\|\hat{g} - g\|_1 \leq \varepsilon$.

در حقیقت در این مقاله هر وقت از فاصله بین دو توزیع صحبت می‌شود، منظور فاصله‌ی total variation (TV) می‌باشد که به صورت زیر تعریف می‌شود:

$$\mathrm{TV}(\mathrm{f}_1, \mathrm{f}_2) \coloneqq \sup_{B \in L} \int_B \big(f_1(x) - f_2(x)\big) dx = \frac{1}{2} \|f_1 - f_2\|_1$$

دلیل استفاده نکردن از معیاری مانند KL divergence آن است که این معیار نسبت به ساپورت توزیع ها از TV distance حساس تر است و برای توزیع هایی که مشابه اند اما ساپورت آن ها در نقاطی اشتراک ندارد، بی‌نهایت می‌شود. همچنین معیار TV از جنس فاصله بوده و نسبت به ترتیب توزیع ها متقارن است.

PAC-learning distributions

Realizable case:

$X_1, \ldots, X_m \sim g \quad (i.i.d\ samplse)$

$A(X_1, \ldots, X_m) \to \hat{g} \in \mathcal{G}$

$P(\|g - \hat{g}\|_1 \leq \varepsilon) \geq 1 - \delta$

Non-realizable case:

$X_1, \ldots, X_m \sim f \notin \mathcal{G} \quad (i.i.d\ samples)$

$A(X_1, \ldots, X_m) \to \hat{g} \in \mathcal{G}$

$P\left(\|f - \hat{g}\|_1 \leq \mathrm{c} \inf_{\mathrm{g}^* \in \mathcal{G}} \|g^* - f\|_1 + \varepsilon\right) \geq 1 - \delta$

$\boldsymbol{k - mix(F)}$

اگر F خانواده‌ی توزیع پایه باشد، آنگاه $k - mix(F)$ خانواده‌ی توزیع های mixture آن است که به صورت زیر تعریف می‌شود:

$$k - mix(F) \coloneqq \left\{ \sum_{i=1}^{k} w_i f_i \ : \quad w_i \geq 0, \qquad \sum_i w_i = 1, \qquad f_1, \ldots, f_k \in F \right\}$$

۳. روش های پیشین

روش های متنوعی برای یادگیری توزیع وجود دارد که در ادامه به شرح مختصر برخی از آن ها می‌پردازیم.

- روش maximum likelihood estimation یک روش پارامتریک برای تخمین پارامتر های توزیع از روی داده های مشاهده شده است. مشکل این روش آن است که برای خانواده‌ای مانند mixture of gaussians، به دست آوردن میانگین ها، ماتریس های کوواریانس و وزن های مربوط به هر توزیع پایه، در شرایطی که میانگین ها و کوواریانس ها بسیار به هم نزدیک هستند کار دشواری است. بنابراین در این روش به قیدی از جنس separability پارامترها نیاز است که لزوما برقرار نیست.
همچنین به طور کلی می‌توان نشان داد که entry-wise approximation نیز روش مناسبی نیست و به condition number وابسته است. به عنوان مثال می‌توان دو ماتریس کوواریانس $\Sigma, \hat{\Sigma}$ را در نظر گرفت که فاصله‌ی پارامترهای آن ها کم است:
$$\left|\Sigma - \hat{\Sigma}\right| < \varepsilon, \qquad \varepsilon = \frac{2}{\kappa(\Sigma) + 1}$$
اما فاصله‌ی توزیع های $\mathcal{N}(0, \Sigma),\ \mathcal{N}\left(0, \hat{\Sigma}\right)$ از حیث KL و یا TV بیشینه است.

- روش دیگری که در عمل زیاد استفاده می‌شود kernel density estimation است که نیازمند قیدهایی از جنس smoothness و boundedness می‌باشد. این در حالی است که خانواده‌ی گاوسی ها به طور کلی Lipschitz و یا bounded نیست.

- روش دیگری که برای یادگیری توزیع استفاده می‌شود histogram estimation است اما استفاده از این روش به sample complexity ای منجر می‌شود که نسبت به بُعد فضا نمایی است.

- روش دیگری که می‌توان استفاده کرد minimum distance estimation است که بر اساس uniform convergence و به دست آوردن VC dimension برای Yatracos class می‌باشد. این روش برای توزیع یک گاوسی d بُعدی به کران بهینه‌ی $O(d^2/\varepsilon^2)$ می‌رسد. اما برای خانواده k-mix(gaussian) در فضای d بُعدی به کران loose برابر با $O(k^4 d^4/\varepsilon^2)$ می‌رسد.

- همچنین تخمینگرهای بر پایه‌ی piecewise polynomials نیز نمی‌توانند به خوبی گاوسی های d بُعدی را تخمین بزنند و کران هایی که به دست می‌دهند به صورت نمایی به d وابسته است.

بهترین کرانی که قبل از این کار برای خانواده‌ی mixture of gaussians و mixture of axis-aligned gaussians وجود داشت به ترتیب $\tilde{O}(kd^2/\varepsilon^4)$ و $\tilde{O}(kd/\varepsilon^4)$ بودند که به واسه‌ی وجود ε^4 در مخرج، بهینه نبودند. بنابراین هیچ کدام از این روش ها برای یادگیری mixture ها بهینه نیستند.

۴. روش ارائه شده

۴.۱. مفهوم sample compressibility

ابتدا به معرفی مفهوم sample compressibility می‌پردازیم.

دیکودر J برای خانواده‌ی توزیع F تابعی است deterministic که تعداد متناهی سمپل (Z) و تعداد متناهی بیت دریافت کرده و عضوی از خانواده توزیع F را به دست می‌دهد:

$$\mathrm{J}: \bigcup_{\mathrm{n}=0}^{\infty} \mathrm{Z}^{\mathrm{n}} \times \bigcup_{\mathrm{n}=0}^{\infty} \{0,1\}^{\mathrm{n}} \rightarrow F$$

حال می‌گوییم F دارای (τ, t, m) $sample\ compressibility$ است اگر برای F دیکودر Jای وجود داشته باشد که به ازای هر $g \in F$ داشته باشیم:

به ازای هر $\varepsilon \in (0,1)$ اگر توالی S شامل $m(\varepsilon)$ تا سمپل $i.i.d$ از g داشته باشیم، وجود داشته باشد توالی L به طول حداکثر $\tau(\varepsilon)$ که از المان های S ایجاد شده است و توالی B که شامل حداکثر $t(\varepsilon)$ بیت است، به طوری که:

$$P(\| J(L, B) - g\|_1 \leq \varepsilon) \geq 1-\delta$$

بنابراین طبق این مفهوم، توزیع g توسط L, B انکود می‌شود و هرچه m کوچکتر باشد به این معناست که توزیع توسط تعداد نمونه های کمتری قابل انکود شدن است و بنابراین توزیع compressibility بالاتری دارد. همچنین دقت شود که S, L توالی هستند نه مجموعه و بنابراین می‌توانند شامل عضو تکراری باشند.

۴.۲. محاسبه‌ی sample compressibility برای چند توزیع نمونه

- $F = \{Unif(a,b)\}$
  $\mathrm{m} \in \mathrm{O}\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right), \quad t = 0, \quad \tau = 2$

- $F = \{\mathcal{N}(\mu, \sigma)\}$
  $\mathrm{m} \in \mathrm{O}\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right) \log \frac{1}{\varepsilon}, \quad t = 0, \quad \tau = 2$

۴.۳. قابل یادگیری بودن توزیع های دارای sample compressibility

حال می‌خواهیم نشان دهیم که sample compressibility برای توزیع نتیجه می‌دهد که آن توزیع learnable است. طبق قضیه‌ی ۳.۴ مقاله داریم:

الگوریتم deterministic ای وجود دارد که اگر M تا توزیع کاندید مانند $\{f_1, \ldots, f_M\}$ داشته باشیم و $\varepsilon > 0$ و تعداد $log(3M^2/\delta)/2\varepsilon^2$ تا سمپل $i.i.d$ از توزیع ناشناخته g داشته باشیم، این الگوریتم $j \in [M]$ ای خروجی می‌دهد که به احتمال حداقل $1 - \delta/3$ داشته باشیم:

$$\| f_j - g \|_1 \leq 3 \min_{i \in [M]} \| f_i - g \|_1 + 4\varepsilon$$

این نامساوی با پهن کردن grid در فضای پارامترها به دست می‌آید. به این صورت که ابتدا توسط confidence interval فضای پارامترهای توزیع را محدود می‌کنیم و سپس در فضای باقی مانده grid پهن کرده و عبارت اول سمت راست تساوی را $O(\varepsilon)$ می‌کنیم.

حال می‌توان به جای این کار از مفهوم sample compressibility ارائه شده استفاده کرد. می‌دانیم که اگر این ویژگی را داشته باشیم و m سمپل از توزیع g داشته باشیم، دیکودر یک توزیع در ε همسایگی g به ما می‌دهد. همچنین M نیز در این حوزه معادل حداکثر حالات خروجی دیکودر یعنی $m(\varepsilon)^{\tau(\varepsilon)} \times 2^{t(\varepsilon)}$ می‌شود.

حال طبق قضیه‌ی ۳.۵ مقاله می‌توان گفت:

اگر F دارای $sample\ compressibility$ (τ, t, m) باشد و داشته باشیم:

$$\tau'(\varepsilon) := \tau(\varepsilon/6) + t(\varepsilon/6)$$

آنگاه F توسط $\tilde{O}(m(\varepsilon/6) + \tau'(\varepsilon)/\varepsilon^2)$ سمپل قابل یادگیری است.

۴.۴. بسته بودن sample compressibility نسبت به mixture, product

حال می‌دانیم که اگر خانواده‌ی توزیع پایه و یا خانواده‌ی توزیع mixture دارای sample compressibility باشند، آنگاه مطابق sample complexity ذکر شده، قابل یادگیری اند. این در حالی است که بررسی sample compressible بودن توزیع های mixture نسبت به توزیع های پایه، پیچیده تر است. به این منظور لم های زیر در مقاله آورده شده اند که نشان می‌دهد sample compressibility نسبت به mixture و product توزیع ها بسته است:

لم ۳.۶ مقاله، درباره‌ی product توزیع پایه:

اگر F دارای $sample\ compressibility$ $(\tau(\varepsilon), t(\varepsilon), m(\varepsilon))$ باشد، آنگاه F^d دارای

$(d\tau(\varepsilon/d), dt(\varepsilon/d), m(\varepsilon/d)log(3d))$ $sample\ compressibility$

می‌باشد.

لم ۳.۷ مقاله، درباره‌ی mixture توزیع پایه:

اگر F دارای

$(\tau(\varepsilon), t(\varepsilon), m(\varepsilon))$ $sample\ compressibility$

باشد، آنگاه $k - mix(F)$ دارای

$$
\Big(k\tau(\varepsilon/3),\quad kt(\varepsilon/3) + k\ log_2(4k/\varepsilon),\quad 48m(\varepsilon/3)\ k\ log(6\mathrm{k})/\varepsilon \Big)\ sample\ compressibility
$$

می‌باشد.

در حالت کلی با فرض قابل یادگیری بودن توزیع پایه نمی‌توانستیم نتیجه بگیریم که توزیع mixture نیز قابل یادگیری است. حال اما با فرض sample compressible بودن توزیع پایه می‌توان sample compressible بودن توزیع mixture را نتیجه گرفت و با توجه به قضایای پیشین، به قابل یادگیری بودن این توزیع ها نیز رسید.

۴.۵. کران یادگیری برای توزیع های با پایه‌ی گاوسی

تا این جا خواصی که ذکر شد، به طور کلی و برای هر توزیع پایه‌ای صادق بود. حال اما می‌خواهیم به طور دقیق تر، خواص ذکر شده را برای توزیع پایه‌ی گاوسی بررسی کنیم.

قضیه‌ی ۱.۱ مقاله: خانواده‌ی k-mixture از گاوسی های d بُعدی، توسط $\tilde{O}(kd^2/\varepsilon^2)$ سمپل قابل یادگیری اند.

اثبات: طبق لم ۴.۱ مقاله می‌دانیم که خانواده‌ی گاوسی های d بُعدی دارای

$$
\Big(O(d\ log(2d)),\quad O\big(d^2 log(2d)\ log(d/\varepsilon)\big),\quad O\big(d\ log(2d)\big) \Big)\ sample\ compressibility
$$

می‌باشد. حال اگر این لم را با لم ۳.۷ مقاله که پیش تر ذکر شد ترکیب کنیم خواهیم داشت که خانواده‌ی k-mixture از گاوسی های d بُعدی دارای

$$
\Big(\; O(kd\,log(2d)),\quad O\big(kd^2\,log(2d)\,log(d/\varepsilon)+k\,log(k/\varepsilon)\big),\quad O(kd\,logk\,log(2d)/\varepsilon) \;\Big)\ sample\ compressibility
$$

می‌باشد.

حال اگر مقادیر τ, t, m به دست آمده را در قضیه‌ی ۳.۵ مقاله جایگذاری کنیم، طبق

$$m(\varepsilon)=\tilde{O}(dk/\varepsilon)\,,\qquad \tau'(\varepsilon)=\tilde{O}(d^2k)$$

خواهیم داشت که sample complexity برای یادگیری k-mixture گاوسی های d بُعدی برابر $\tilde{O}(kd^2/\varepsilon^2)$ می‌باشد.

قضیه‌ی ۱.۲ مقاله: هر روشی که برای یادگیری خانواده‌ی k-mixture از گاوسی های d بُعدی استفاده شود، حداقل به $\widetilde{\Omega}(kd^2/\varepsilon^2)$ سمپل نیاز دارد.

ترکیب قضیه‌ی ۱.۱ و ۱.۲ مقاله نشان دهنده‌ی بهینه بودن کران بالای ارائه شده است و می‌توان گفت که تعداد $\widetilde{\Theta}(kd^2/\varepsilon^2)$ سمپل برای یادگیری توزیع mixture of k gaussians در فضای R^d لازم و کافی است.

۴.۶. کران یادگیری برای توزیع های با پایه‌ی گاوسی axis-aligned

برای حالت mixture of gaussians نیز کران جدیدی به دست می‌آید.

قضیه‌ی ۱.۳ مقاله: خانواده‌ی k-mixture از گاوسی های axis-aligned که d بُعدی هستند، توسط $\tilde{O}(kd/\varepsilon^2)$ سمپل قابل یادگیری اند.

اثبات: به طور مشابه با ترکیب کردن لم های ۴.۱ و ۳.۶ و ۳.۷ مقاله می‌توان نشان داد که خانواده‌ی k-mixture از گاوسی های axis-aligned که d بُعدی هستند، دارای

$$
\Big(\; O(kd),\quad O(kd\,log(d/\varepsilon)\;+\;k\,log(k/\varepsilon)),
$$

$O(k\ logk\ log(3d)/\varepsilon)$

) $sample\ compressibility$ (

می‌باشد.

حال اگر مقادیر τ, t, m به دست آمده را در قضیه‌ی ۳.۵ مقاله جاگذاری کنیم، خواهیم داشت که sample complexity برای یادگیری k-mixture گاوسی های axis-aligned که d بُعدی هستند برابر $\tilde{O}(kd/\varepsilon^2)$ می‌باشد.

۵. نتیجه گیری

یکی از مسائل حل نشده در یادگیری توزیع، به دست آوردن sample complexity مورد نیاز برای یادگیری توزیع است. یک نگاه شهودی از حوزه یادگیری نظارتی بیان می‌کند که sample complexity مورد نیاز برای یادگیری هر concept ای، اعم از توابع یا توزیع ها، با بُعد فضای فرضیه تقسیم بر ε^2 نسبت دارد. در حالتی که درباره‌ی دسته بندی دوتایی صحبت کنیم، بُعد فضای فرضیه معادل VC dimension خواهد بود. برای یادگیری توزیع نیز، این بُعد معادل تعداد پارامتر های توزیع می‌باشد. قبل از این کار، برای توزیع هایی مانند گاوسی و گاوسی های axis-aligned که به ترتیب $O(d^2)$ و $O(d)$ پارامتر دارند، این شهود برقرار بود. اما کران هایی که برای mixture این توزیع ها وجود داشت، loose تر از این مقدار بودند. حال در این کار اثبات شد که این شهود برای mixture های این توزیع ها که به ترتیب $O(kd^2)$ و $O(kd)$ پارامتر دارند نیز برقرار است.

مفهوم جدید sample compressibility که در این کار ارائه شد، در مورد sample complexity مورد نیاز برای یادگیری توزیع، یک شرط کافی ارائه می‌دهد. با این که اثبات شد برای mixture of gaussians این شرط لازم نیز می‌باشد (یعنی در این حالت کران ارائه شده بهینه است)، اما در حالات دیگر لزوم این شرط اثبات نشده و به صورت یک مسئله‌ی حل نشده باقی می‌ماند. یک خاصیت جالب مفهوم sample compressibility آن است که نسبت به product و همچنین mixture توزیع های پایه بسته است. به طور کلی اگر بخواهیم از قابل یادگیری بودن توزیع پایه به قابل یادگیری بودن توزیع mixture برسیم، مسیر سختی را پیش رو خواهیم داشت. حال اما با معرفی مفهوم sample compressibility تنها کافی است وجود این ویژگی را برای خانواده‌ی توزیع پایه نشان دهیم تا وجود آن برای خانواده‌ی توزیع mixture نیز به دست آید و قابل یادگیری بودن آن اثبات شود.

۶. مراجع

[1] Ashtiani, H., Ben-David, S., Harvey, N., Liaw, C., Mehrabian, A., & Plan, Y. (2018). Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. Advances in Neural Information Processing Systems, 31.