گزارش ابتدایی پروژه‌ی درس تئوری یادگیری ماشین

بهراد منیری
bemoniri@ee.sharif.edu
شماره‌ی دانشجویی: ۹۵۱۰۹۵۶۴

۱ مقدمه

در این پروژه به بررسی مقاله‌ی [۲] می‌پردازیم. در این مقاله، مسئله کاهش بعد غیرخطی تفسیرپذیر Interpretable Kernel Dimension Reduction (IKDR) بررسی شده است. در این مسئله تلاش بر این است که بعد از کاهش بعد، هر بعد ارتباط کاملاً مشخص و قابل تفسیری با ابعاد مسئله‌ی اصلی داشته باشند. بعد از معرفی فرمول‌بندی دقیق این مسئله، این مقاله تلاش می‌کند تا الگوریتم Iterative Spectral Method (ISM) که در مقاله‌ی [۱] برای حل مسئله‌ی IKDR مطرح شده است را بهبود ببخشد و همچنین نتایج و گارانتی‌های تئوری موجود در آن را به خانواده‌های بسیار بزرگی از کرنل‌ها تعمیم دهد. مقاله‌ی [۱] نتایج و گارانتی‌های همگرایی الگوریتم خود را تنها در صورتی که از کرنل گوسی استفاده شود ارائه کرده بود. در این گزارش ابتدایی، خلاصه‌ای از نتایج و ایده‌های اساسی بازگو خواهند شد.

۲ معرفی اولیه و ایده‌ی اصلی

برای کاهش بعد داده به صورت غیرخطی، اولین ایده‌ای که به ذهن می‌رسد این است که ابتدا داده‌ها را به یک فضای هیلبرت نگاشت دهیم (مثلاً به کمک یک کرنل) و سپس در آن فضای بعد بالا، با استفاده از روشی مثل PCA کاهش بعد را انجام دهیم. این روش دو ایراد عمده دارد. ایراد اول این است که برای کاهش بعد supervised قابل استفاده نیست. دومین مشکل، تفسیرناپذیر بودن آن است. تفسیر ناپذیر بودن به آن معناست که بعد از اعمال PCA در فضای بعد بالا و نگه داشتن جهت‌های با واریانس بالا در آن، واضح نیست هر بعد باقی مانده دقیقاً به چه معناست و ارتباط هر کدام از آن ابعاد با ویژگی‌های اولیه بسیار پیچیده است. در روش مطرح شده، داده‌ها را به صورت $\Phi(X)W$ کاهش بعد می‌دهیم. یک راه دیگر این است که تلاش کنیم ماتریسی (در همان فضای بعد پایین) پیدا کنیم که بعد از اعمال آن بر دیتا، ماتریس $\Phi(XW)$ اطلاعات زیادی از خود $\Phi(X)$ را در بر داشته باشد. در این حالت هر دو مشکل فوق بر طرف می‌شوند. هر ویژگی، بعد از کاهش بعد را می‌توان به صورت ترکیبی خطی از ویژگی‌های اولیه دید. همچنین می‌توان مسئله‌ی supervised را به صورت زیر فرمول‌بندی کرد:

$$\max_{W} \ \mathrm{DM}(XW, Y) \quad \text{s.t.} \quad W^T W = I$$

در این فرمول بندی، DM یک متر وابستگی (مثلاً اطلاعات متقابل) است.

۳ معرفی معیار HSIC

HSIC معیاری برای سنجیدن استقلال دو متغیر تصادفی است. این معیار مشابه اطلاعات متقابل فاصله‌ای بین توزیع‌های P_{XY} و $P_X P_Y$ است. اطلاعات متقابل از فاصله‌ی KL Divergence استفاده می‌کند، در مقابل، HSIC از متر Maximum Mean Discrepancy (MMD) بهره می‌برد. ایده‌ی اصلی این متر آن است که در ابتدا، نگاشتی بر مبنای یک کرنل از فضای توزیع‌های احتمال به یک RKHS به شکل زیر ساخته می‌شود:

$$\begin{cases} \mathcal{F}: \mathcal{P} \to \mathcal{H} \\ \mathcal{F}(P) = \mathbb{E}_P\Big[k(.,X)\Big] \end{cases}$$

در ادامه نیز از متر فضای هیلبرت ساخته شده توسط کرنل، به عنوان متر بین دو توزیع استفاده می‌کنند.

$$\mathrm{MMD}(P,Q) = \Big|\Big|\mathbb{E}_P\Big[k(.,X)\Big] - \mathbb{E}_Q\Big[k(.,X)\Big]\Big|\Big|_{\mathcal{H}}$$

اگر نگاشت بین فضای توزیع‌ها و فضای هیلبرت، نگاشتی یک به یک باشد، در این نگاشت هیچ اطلاعاتی از توزیع از بین نمی‌رود و به راحتی می‌توان نشان داد که عبارت فوق، تمام خواص یک متر را دارا می‌باشد. به کرنل‌هایی که در آن‌ها نگاشت مذکور یک به یک است، کرنل مشخصه می‌گویند. کرنل گوسی مشهورترین کرنل مشخصه است. تعریف می‌کنیم:

$$\mathrm{HSIC}(X,Y) = \mathrm{MMD}(P_X P_Y, P_{XY})$$

۱.۳ تخمین HSIC از روی داده

حال فرض کنید مجموعه‌ی $\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ با توزیع مشترک P_{XY} داده شده است و قصد داریم بررسی کنیم که آیا X و Y مستقل هستند یا خیر. فرض کنید $X \in \mathbb{R}^{n\times d}$ و $Y \in \mathbb{R}^{n\times c}$ بردارهای مشاهدات باشند. ماتریس‌های گرام مربوط به X و Y را $K_X \in \mathbb{R}^{N\times N}$ و $K_Y \in \mathbb{R}^{N\times N}$ می‌نامیم. ماتریس H را نیز ماتریس centering بگیرید، $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. در این صورت تخمین‌گیر زیر، یک تخمین‌گر نا اریب از HSIC است:

$$\mathrm{HSIC}(X,Y) = \frac{1}{(n-1)^2}\mathrm{Tr}(K_X H K_Y H)$$

۴ تعریف مسئله‌ی IKDR

فرض کنید دیتاست $X \in \mathbb{R}^{n\times d}$ با k فیچر و n نمونه باشد و $X \in \mathbb{R}^{n\times k}$ لیبل‌های این نمونه‌ها باشد (k تعداد کلاس‌ها است). با استفاده از دو کرنل k_x و k_y داده‌شده، دو ماتریس گرام K_X و K_Y ساخته شده است ($\mathbb{R}^{n\times n}$). در مسئله‌ی IKDR قصد داریم ماتریس $W \in \mathbb{R}^{d\times q}$ را به نحوی بسازیم که $\mathrm{HSIC}(XW,Y)$ کمینه شود (توجه داریم که $q < d$ و فرض کنید $W^TW = I$). مسئله‌ی فوق را می‌توان به فرم مسئله‌ی بهینه‌سازی زیر نوشت:

$$\max_W \mathrm{Tr}(\Gamma K_{XW}) \quad \text{s.t.} \quad W^TW = I$$

به عنوان مثال، با کرنل گوسی مسئله به فرم زیر در می‌آید:

$$\max_W \quad \Gamma_{ij}\exp\Big(-\frac{(W^Tx_i - W^Tx_j)^2}{2\sigma^2}\Big) \quad \text{s.t.} \quad W^TW = I$$

این مسئله در حالت کلی، مسئله‌ای شدیداً غیر محدب است. این مقاله الگوریتمی برای حل این مسئله ارائه کرده و گارانتی‌هایی برای همگرایی الگوریتم ارائه می‌دهد.

۵ الگوریتم ISM

ایده‌ی اصلی الگوریتم این است که مشابه PCA یک ماتریس «کوواریانس» معادل برای هر کرنل معرفی کرده و ماتریس W را برابر بردارهای ویژه‌ی غالب آن قرار دهد.
مشاهده می‌شود که در جدول فوق، برخی از ماتریس‌های کوواریانس تابع W هستند. برای این دسته از ماتریس‌ها، ایده‌ی مقاله این است که در ابتدا تقریب تیلور درجه‌ی دوم را در نظر می‌گیرد و W بهینه را می‌یابد. به کمک آن W، ماتریس Φ را می‌سازد و این روند را تا همگرایی ادامه می‌دهد.

کرنل	ماتریس کواریانس
Linear	$\Phi = X^T \Gamma X$
Squared	$\Phi = X^T \mathcal{L}_\Gamma X$
Polynomial	$\Psi = \Gamma \odot K_{XW,p-1}$ ، $\Phi = X^T \Psi X$
Gaussian	$\Psi = \Gamma \odot K_{XW}$ ، $\Phi = -X^T \mathcal{L}_\Psi X$
Multiquadratic	$\Psi = \Gamma \odot K_{XW}^{(-1)}$ ، $\Phi = X^T \mathcal{L}_\Psi X$

جدول ۱: ماتریس کواریانس معادل برای کرنل‌های معروف

۶ صورت قضایای اصلی تئوریک

در ابتدا، این مقاله ثابت می‌کند که بردارهای ویژه‌ی غالب Φ شرایط لازم مرتبه اول (KKT) و شرایط لازم مرتبه دوم (شرط لاگرانژین) را ارضا می‌کنند. سپس اثبات می‌کند که الگوریتم ISM نیز به همان بردارهای ویژه Φ همگراست. این اثبات‌ها برای دسته‌ای بزرگ از کرنل‌ها به نام کرنل‌های ISM انجام شده‌اند، این پیشرفتی بزرگ است که این مقاله به نسبت کارهای قبل داشته است. این مقاله نشان می‌دهد که هر Conic Combination از کرنل‌های ISM یک کرنل ISM است و ماتریس کوواریانس مربوط به این Conic Combination نیز Conic Combination ماتریس‌های کوواریانس کرنل‌های اصلی است.

مراجع

[1] Wu, C., Ioannidis, S., Sznaier, M., Li, X., Kaeli, D., and Dy, J. Iterative spectral method for alternative clustering. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (Playa Blanca, Lanzarote, Canary Islands, 2018), vol. 84 of *Proceedings of Machine Learning Research*, PMLR.

[2] Wu, C., Miller, J., Chang, Y., Sznaier, M., and Dy, J. Solving interpretable kernel dimensionality reduction. In *Advances in Neural Information Processing Systems 32*. 2019.