درس مدل‌های احتمالاتی گرافی

پروژه دوم

یادگیری و استنتاج در مدل تخصیص پنهان دیریکله

۱- شرح پروژه

با توجه به گزارش مربوط به مدل‌های عنوان و گزارش مربوط به یادگیری مبتنی بر نمونه برداری گیبز برای مدل تخصیص پنهان دیریکله که در اختیار شما قرار گرفته است، الگوریتم یادگیری مربوطه را پیاده سازی نموده و آزمایش‌های زیر را بر روی دو مجموعه داده‌ای که در ادامه مشخص شده است اجرا و نتایج به دست آمده را تحلیل نمایید. در بررسی های خود از معیارهای مختلفی نظیر زمان آموزش مدل، زمان استنتاج برای نمونه جدید، درستنمائی مدل برای اسناد آموزشی (Perplexity) ، عناوین استخراج شده از اسناد و ... استفاده کنید.

۱-۱- اندازه پارامتر توزیع‌های دیریکله متقارن

در [1] از توزیع متقارن دیریکله استفاده شده است که از یک پارامتر تنهای α در توزیع اولیه θ و از یک پارامتر تنهای β در توزیع اولیه ϕ استفاده می کند. با تغییر مقدار این دو پارامتر تاثیر آن را بر روی مدل یاد گرفته شده بررسی کنید.

۱-۲- تعداد عناوین

یکی از پارامترهای مدل LDA تعداد عناوین (T) است که می‌بایست از قبل مشخص شود. با تغییر تعداد عناوین در نظر گرفته شده در مدل تاثیر این پارامتر بر روی مدل را بررسی نمایید. (نمودار تعداد T به $Perplexity$ را رسم کنید)

۱-۳- نمونه برداری

پارامترهای مدل را می‌توان با یک نمونه $\boldsymbol{z}$ و یا با نمونه‌های بیشتر از $\boldsymbol{z}$ انجام داد. همچنین بعد از رسیدن فرآیند نمونه‌برداری به حالت mixing می‌توان از کلیه نمونه‌های تولید شده استفاده کرد یا این که از هر n نمونه متوالی تولید شده یکی را انتخاب کرد. حالت‌های مختلف نمونه‌برداری را بررسی و نتایج را با یکدیگر مقایسه کنید.

۱-۴- خوشه بندی اسناد

از بردار $\theta^{(d)}$ محاسبه شده برای اسناد در مجموعه داده ۲ برای خوشه‌بندی این اسناد استفاده کرده و نتایج به دست آمده را بررسی کنید.

۱-۵ استنتاج تغییراتی (نمره اضافه)

پیاده‌سازی روش استنتاج تغییراتی (Variational) که در [2] به آن اشاره شده است و مقایسه نتایج با روش استنتاج مبتنی بر نمونه برداری نمره اضافه دارد.

۲- مجموعه داده‌ها

مجموعه داده ۱: اسناد با نمایش گرافیکی

به عنوان مجموعه داده اول از اسناد و عناوین معرفی شده در بخش "A graphical example" مرجع [1] استفاده کنید.

مجموعه داده ۲:

مجموعه داده ضمیمه شده است.

۳- موارد تحویلی

مواردی که می‌بایست تحویل داده شوند عبارتند از:

- پیاده سازی کامل آزمایش‌ها
- گزارش

۴- نکات مهم

کمک گرفتن از کدهای آماده در پیاده‌سازی بلامانع است اما هر دانشجو می‌بایست کد مربوطه را به طور کامل پیاده سازی نماید و تسلط کامل بر همه بخش‌های کد داشته باشد.

منابع

[1] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National Academy of Sciences 101, no. suppl 1 (2004): 5228-5235.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.