سعید غفوری **۹۵۰۷۸۹۴**

تعریف پروژه درس مدل‌های گرافیکی احتمالاتی

موضوع پروژه: پیاده سازی یک مدل سازی از گراف‌های نمایی اتفاقی (ERGM)[1]

مقدمه

یکی از معضلات موجود در کلیه مسائل یادگیری ماشین عدم دسترسی به مجموعه داده[2] کافی است و در بسیاری از موارد نیاز به مطالعه داده‌هایی که توسط یک الگوریتم تولید داده اتفاقی ایجاد شده‌اند می‌تواند مفید واقع شود اما تولید این داده‌ها به صورت کاملا اتفاقی نمی‌تواند به ما شهود دقیقی در آزمایش‌های انجام شده بر روی مساله ما بدهد بنابراین نیاز به تولید داده شبه اتفاقی[3]باشد منظور از داده شبه‌اتفاقی شبه‌اتفاقی داده‌ای است که با وجود این‌که به صورت اتفاقی تولید شده است اما با این حال در بعضی خواص دارای تنظیم می‌باشند مثلا می‌توانیم توزیع آماری داده‌ها یا مثلا تعداد آن‌ها را مشخص کنیم.

بسیاری از کارهای انجام شده در یادگیری ماشین به طور مستقیم و یا غیرمستقیم به بررسی و تحلیل گراف‌ها می‌پردازند و ماهیت داده‌های آن‌ها دارای توصیفی گراف مانند می‌باشد. بنابراین برای تولید مجموعه داده آن‌ها همانطور که گفته شد به دلیل این‌که گرافی که کاملا اتفاقی تولید شده باشد نمی‌تواند اطلاعات مناسبی به ما بدهد نیاز به تولید گراف‌های شبه اتفاقی داریم.

گراف‌های نمایی اتفاقی (ERGM)

یکی از دسته گراف‌های معروف که برای ایجاد گراف‌های اتفاقی با خواصی مدنظر الگوریتم ما استفاده می‌شود گراف‌های نمایی اتفاقی می‌باشد که در ادامه به معرفی آن‌ها میپردازیم.

برای شهود بهتر بدون از دست دادن کلیت مساله فرض میکنیم گرافی که در ادامه تولید می‌کنیم گراف رابطه دوستی در یک شبکه اجتماعی باشد.

اگر در گراف مورد بررسی ما دو گره[4] i و j دارای رابطه باشند این مورد ممکن است ناشی از وجود یک اتصال[5] مستقیم و یا یک ارتباط غیر مستقیم که ناشی از اتصال هردوی این گره‌ها به گره سوم k باشد. اگر بخواهیم در گرافی که قصد ایجاد آن را داریم این ارتباطات را در نظر بگیریم ممکن است با شرایطی رو به رو شویم که در آن تعداد زیادی از گره‌ها با هم در ارتباط باشند زیرا که این رابطه تعدی‌وار دوستی‌ها ممکن است همینطور تا ارتباط بین تمام گره‌ها ادامه پیدا کند. هدف در اینجا این است که علاوه بر در نظر گرفتن صرف تعداد گره‌ها در گراف اتفاقی روابطی از قبیل ارتباطات بین گره‌ها را در نظر بگیریم.

ما در اینجا به عنوان مثال برای گرافی که قصد تولیدش را داریم دو نوع رابطه زیر را بین نودها در نظر می‌گیریم:

- تعداد اتصالات در گراف
- تعداد مثلث‌ها در گراف

---

[1] Exponential Random Graphs
[2] Dataset
[3] Pseudo random
[4] node
[5] link

منظور از مثلث سه گره‌ای است که از طریق یک رابطه تعدی به هم متصل‌اند همان‌طور که در شکل ۱ مشاهده می‌کنید.

این دو معیار را طبق رابطه زیر فرموله‌سازی می‌کنیم:

$$\beta_L \#links(g) + \beta_T \#triangles(g)$$

ما می‌خواهیم که احتمال تشکیل هر گراف مرتبط با این رابطه باشد:

$$\beta_L L(g) + \beta_T T(g)$$

پس قرار می‌دهیم:

$$\Pr(g) \sim \beta_L L(g) + \beta_T T(g)$$

از آن‌جایی که رابطه بالا یک تناسب است و تساوی مستقیم نیست می‌توانیم در سمت راست رابطه $exponential$ مقدار موجود را قرار دهیم. یعنی خواهیم داشت:

$$\Pr(g) \sim \exp[\beta_L L(g) + \beta_T T(g)]$$

طبق قضیه همرسلی-کلیفورد[6] داریم:

هر مدل از گراف‌ها را می‌توان با استفاده از ترکیبی از رابطه‌های آماری بین نودهای آن‌ها بیان کرد.

به طور مثال خانواده گراف‌های اردوش-رینی[7] که رابطه زیر بین گره‌های آن‌ها برقرار است را در نظر بگیرید:

$$P = probabilithy\ of\ a\ link \quad L(g) = number\ of\ links\ in\ g$$

$$\Pr[(g)] = p^{L(g)}(1-p)^{\frac{n(n-1)}{2} - L(g)}$$

می‌توانیم اعمال ریاضی زیر را روی آن انجام دهیم:

$$\Pr[(g)] = p^{L(g)}(1-p)^{\frac{n(n-1)}{2} - L(g)}$$

$$= [\frac{p}{1-p}]^{L(g)}(1-p)^{\frac{n(n-1)}{2}}$$

$$= \exp[\log\left(\frac{p}{1-p}\right)L(g) - \log\left(\frac{1}{1-p}\right)n(n-1)/2]$$

$$= \exp[\beta_1 s_1(g) - c]$$

همان‌طور که مشاهده می‌شود پس از ساده‌سازی موفق شدیم که این دسته از گراف‌ها را به صورت رابطه گفته شده دربیاوریم.

به جهت آن‌که بتوانیم رابطه گفته شده را به صورت احتمالی بنویسیم آن را بر فاکتور نرمال‌سازی زیر که مجوع همان رابطه برای سایر گره‌های گراف است تقسیم می‌کنیم:

---

[6] Hammersly-Cliford

[7] Erdos-Reyni

$$\Pr(g) = \frac{\exp[\beta_L L(g) + \beta_T T(g)]}{\sum_{g'} \exp[\beta_L L(g') + \beta_T T(g')]}$$

که پس از محاسبه $exponential$ داریم:

$$\Pr(g) = \exp[\beta_L L(g) + \beta_T T(g) - c]$$

برای هر نوع از گراف که داشته باشیم با استفاده از تخمین پارامتر از طریق محاسبه $max\ likelihood$ از رابطه بالا و تخمین پارامترهای مربوطه که بنابه رابطه آماری انتخاب شده برای تخمین گراف تعیین می‌شوند توزیع مطلوب که در بهترین حالت بتواند شرایط اولیه روابط آماری مشخص شده توسط ما را ارضا کند را مشخص کنیم. مثلا در رابطه بالا پارامترهایی که نیاز به تخین آن‌ها داریم β_L و β_T می‌باشد. در راه این تخمین می‌توانیم از روش‌های مختلف نمونه برداری مانند نمونه برداری گیبز[8] یا متروپولیس هیستینگ[9] استفاده کنیم.

تعریف پروژه

مراحلی که در پروژه انجام خواهیم داد به این ترتیب خواهد بود:

۱. ابتدا دو تا واحد آماری زیر را برای گراف تعریف خواهیم کرد.
   a. تعداد اتصالات یا همان تعداد یال‌های گراف
   b. تعداد مثلث‌ها
۲. سپس رابطه نمایی که در بالا گفته شد را با توجه به این دو فاکتور حساب میکنیم.
۳. از آن‌جایی که برای محاسبه ضرایب رابطه نیاز به تولید تمام حالات داریم به جای آن از یکی از روش‌های نمونه گیری استفاده خواهیم کرد.
۴. روش نمونه گیری مورد استفاده ما برای نمونه گیری MCMC خواهد بود.
۵. در نهایت با استفاده از نمونه‌های به دست آمده مقدار بهینه پارامترهای گفته شده را به روشی تکراری[10] حساب خواهیم کرد.
۶. در خروجی مقدار بهینه $max\ likelihood$ را به ازای نمونه‌های تولید شده در خروجی چاپ خواهیم کرد.

ابزار مورد استفاده

زبان برنامه نویسی ما متلب خواهد بود و به جز ابزار متلب برای نمایش گراف‌ها (صرفا نمایش و بیان گراف) از ابزار دیگری به جز ابزار معمول محاسباتی متلب استفاده نخواهد شد.

[8] Gibbs sampling
[9] Metropolis Hasting
[10] iterative