وزارت علوم، تحقیقات و فناوری
دانشگاه تحصیلات تکمیلی علوم پایه
گاوازنگ، زنجان

دانشکده علوم رایانه و فناوری اطلاعات

درس: مدل‌های گرافی احتمالاتی
پروژه پایانی

استاد: دکتر نریمانی

دانشجو: امیرحسین صفری

شماره دانشجویی: ۱۴۰۱۴۱۲۱

بهار ۱۴۰۲

فرض کنیم شبکه‌ای از تعاملات ژن‌ها در سلول داریم. این تعاملات به‌صورت روابط علت-معلولی هستند و در نتیجه می‌توانیم این سیستم را به‌صورت یک گراف جهت‌دار که در آن گره‌ها همان ژن‌ها و یال‌های جهت‌دار نشاندهنده‌ی روابط بین ژن‌ها هستند نشان دهیم. ما در سلول قادر هستیم غلظت هر ژن را اندازه‌گیری کنیم، و اگر برای مثال در گراف تعامل ژن‌ها از ژن A به ژن B یالی وارد شده باشد، در داده‌ها نیز میزان غلظت ژن B وابسته به میزان غلظت ژن A است. این وابستگی را می‌توان به‌صورت یک توزیع شرطی در نظر گرفت.

اگر ما داده را داشته باشیم ولی ساختار شبکه (یعنی همان مدلی که بر روی داده‌ها صادق است) را نداشته باشیم، می‌توانیم همانطور که در این درس دیدیم، با استفاده از بیشینه درست‌نمایی[1]، مدل بهینه که بر روی داده‌ها صادق است را به‌دست بیاوریم.

با توجه به آنچه گفته شد، مراحل زیر را انجام دهید:

الف) به‌طور دقیق و با استفاده از آنچه در این کلاس یاد گرفتید بنویسید چگونه می‌توان درست‌نمایی هر مدل با داشتن یک داده را محاسبه نمود؟ **(۲ نمره)**

$$P(\text{model} \mid \text{data}) = ??$$

الف) همانطور که می دانیم p(model | data) از رابطه‌ی زیر بدست می‌آید:

Given a prior belief that a probability distribution function is $p(\theta)$ and that the observations x have a likelihood $p(x|\theta)$, then the posterior probability is defined as

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)}p(\theta),$$[6]

where $p(x)$ is the normalizing constant and is calculated as

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

[منبع wikipedia posterior probability]

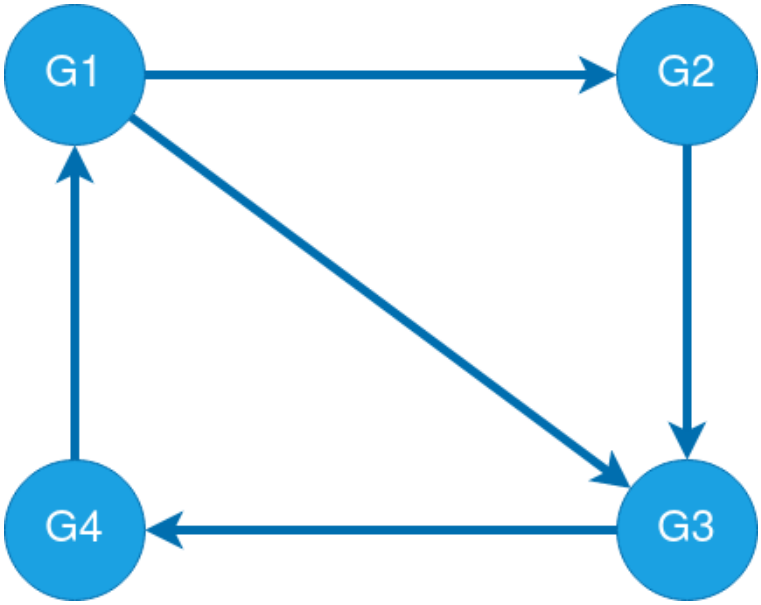
لذا احتمال posterior برابر با حاصل ضرب likelihood در احتمال پیشین تقسیم احتمال داده ها می‌باشد. همانطور که میدانیم محاسبه‌ی احتمال داده ها $p(x) = \int p(x|\theta)p(\theta)d\theta$ (چون یک مقدار ثابت است و تقسیم بر کل فرمول می‌شود و برای همه‌ی حالات یکسان است) انجام نمی‌شود.

همینطور در مسئله‌ی ما مقدار احتمال پیشین (Prior probability) به دلیل نبود اطلاعات قبلی وجود ندارد؛ لذا برای محاسبه‌ی احتمال posterior تنها لازم است تا احتمال Likelihood را محاسبه کنیم.

برای محاسبه‌ی Likelihood یک مدل، یک مدل فرضی به صورت رندوم ایجاد می‌کنیم و با توجه به داشتن دیتاست (داده‌های مسئله) مقدار Likelihood را محاسبه می‌کنیم؛ بدین صورت که با توجه به مدل، روابط بین نودها(ژن‌ها) را در

نظر گرفته و احتمال آن روابط را محاسبه می‌کنیم. برای محاسبه‌ی احتمال روابط از شمارش در دیتاست داده شده استفاده می‌کنیم؛ بدین صورت که حالات مختلف روابط را شمارش می‌کنیم و تقسیم بر احتمال والدها (در روابط) می‌کنیم. (برای بدست آوردن احتمال والدها نیز از شمارش حالات مختلف آن در دیتاست استفاده می‌کنیم.)

برای مثال مدل گرافی زیر را در نظر بگیرید:



و دیتاست فرضی برای مثال ما برابر است با:

	G1	G2	G3	G4
T1	0	1	1	0
T2	1	1	0	0
T3	0	0	1	0
T4	1	1	1	0

حال برای محاسبه‌ی احتمال مدل مثال ما، همانطور که گفته شد، ابتدا روابط بین نودها (ژن‌ها) را ایجاد می‌کنیم.

روابط مدل گرافی بالا به صورت زیر است:

G4 => G1
G1 => G2
G1, G2 => G3
G3 => G4

حال به محاسبه‌ی احتمال روابط با شمارش در دیتاست می‌پردازیم.
لذا باید احتمالات زیر را محاسبه نمود:

P(G1 | G4)
P(G2 | G1)
P(G3 | G2, G1)
P(G4 | G3)

برای مثال برای محاسبه‌ی P(G1 | G4) باید احتمال P(G1, G4) را تقسیم بر احتمال P(G4) کنیم. (چون P(G1,) = P(G1 | G4) / (G4))
برای محاسبه‌ی P(G1, G4) باید تمام حالاتی که P(G1, G4) دارد را در دیتاست بشماریم؛ یعنی تعداد حالات [0, 0] و [0, 1] را باید بشماریم.

	G1	G4
T1	0	0
T2	1	0
T3	0	0
T4	1	0

بعد از شمارش داریم:

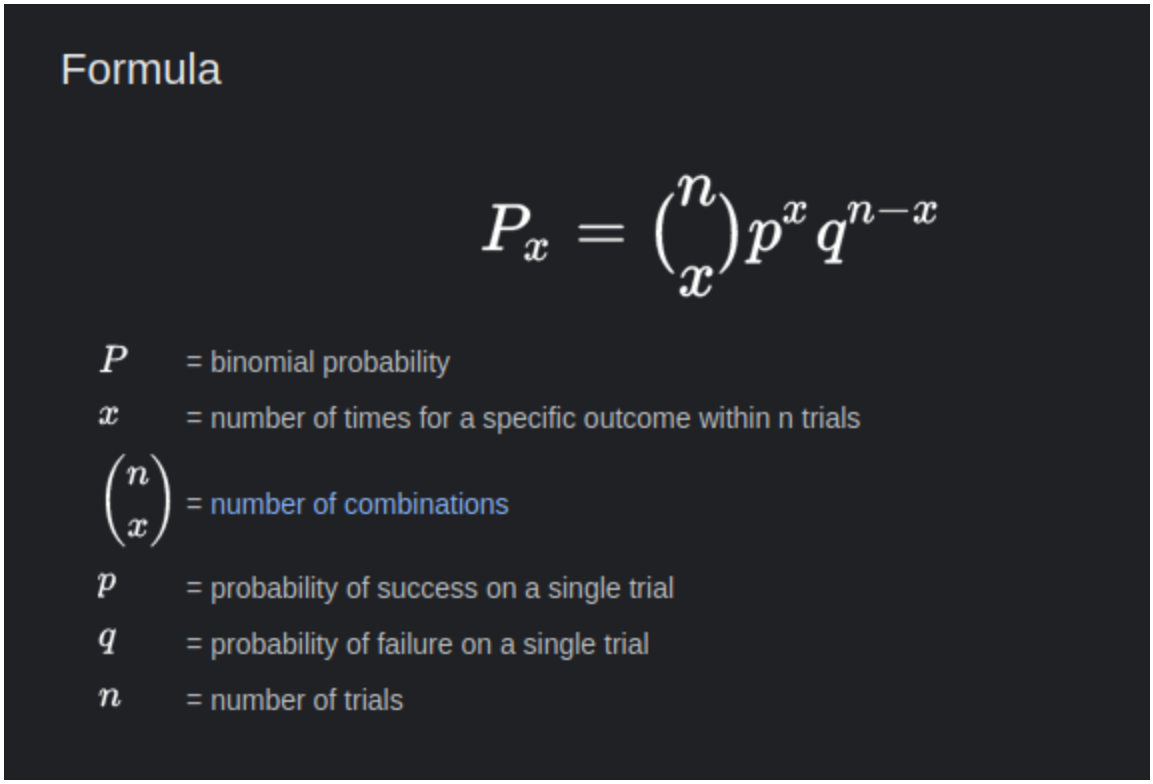
تعداد [0, 0] = ۲
تعداد [0, 1] = ۲

حال برای محاسبه‌ی احتمال مخرج به همین صورت داریم:
حالات P(G4) برابر است با [0]

تعداد [0] : ۴

حال برای محاسبه‌ی احتمال، مقادیر را بر یکدیگر تقسیم می‌کنیم؛ بدین صورت که باید تعداد حالات را بر تعداد حالات والد مربوطه تقسیم کنیم.

در نهایت با توجه به این نکته که توزیع دیتای ما به صورت صفر و یک است، لذا از توزیع Binomial تبعیت کرده و باید برای محاسبه‌ی احتمال نهایی از فرمول این توزیع استفاده کنیم.



[منبع Binomial distribution]

لذا مقدار احتمال را به توان تعداد بار تکرار آن حالت می‌کنیم.

لذا برای P(G1 | G4) داریم:

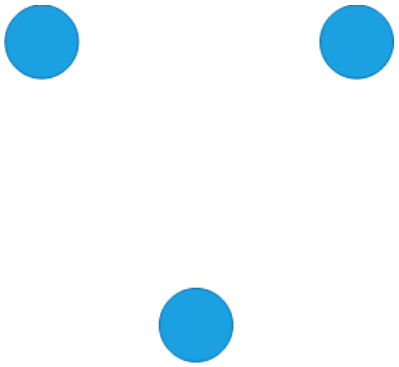
(۱/۲)^۸۲ × (۱/۲)^۸۲

همینطور احتمال بقیه‌ی نودها را محاسبه می‌کنیم.
در نهایت برای بدست آوردن Likelihood مدل فرضی، احتمالات نودها (ژن‌ها) را در یکدیگر ضرب می‌کنیم. لذا داریم:

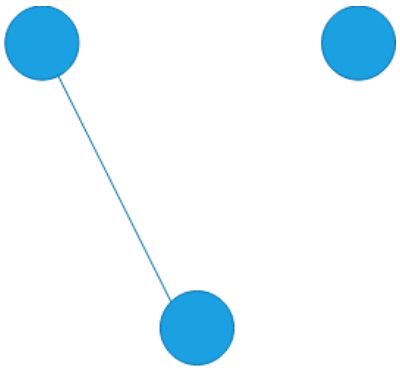
Likelihood = P(G1 | G4) × P(G2 | G1) × P(G3 | G2, G1) × P(G4 | G3)

ج) فرض کنید که ما به دنبال یافتن گراف تعاملات ژن‌ها هستیم به‌طوریکه در یک مجموعه داده بیشترین درست‌نمایی (امتیاز) را داشته باشد. برای یافتن این گراف در حالت ساده، باید از بین تمام گراف‌های موجود جستجو کنیم و گراف با بیشترین درست‌نمایی (امتیاز) را بیابیم. تعداد ساختارهای گراف ممکن وقتی N گره داشته باشیم را حساب کنید. **(۱ نمره)**

ج) برای محاسبه‌ی تعداد ساختارهای گراف ممکن ابتدا برای ۳ = N مثال را حل می‌کنیم.

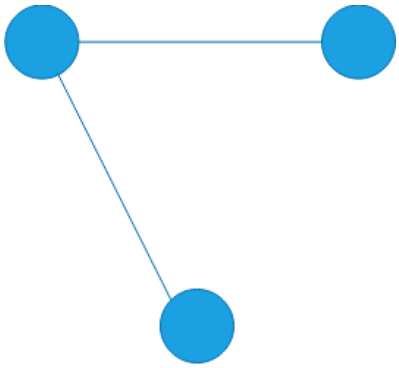


حالت بدون یال:
تعداد حالات: ۱



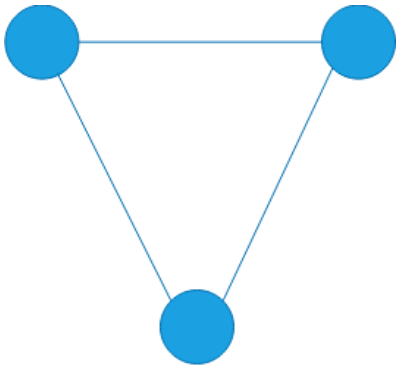
حالت یک یال:
تعداد حالات:

$$\binom{3}{1} \times 2$$



حالت دو یال:
تعداد حالات:

$$\binom{3}{2} \times 2^2$$



حالت سه یال:
تعداد حالات:

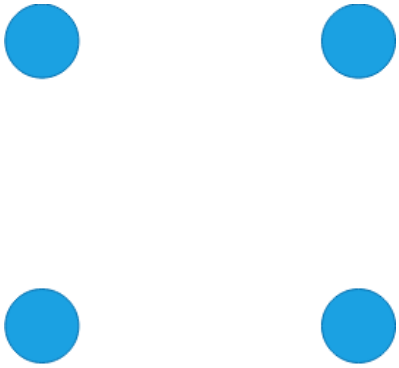
$$\binom{3}{3} \times 2^3$$

==========

در این حالت ۳ برابر با تعداد یال‌ها در ۳ = N است.

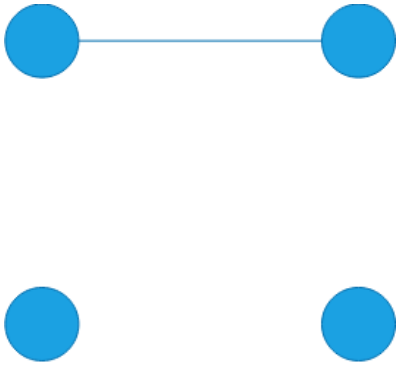
برای حالت ۴ = N داریم:

$$\binom{4}{2} = 6$$



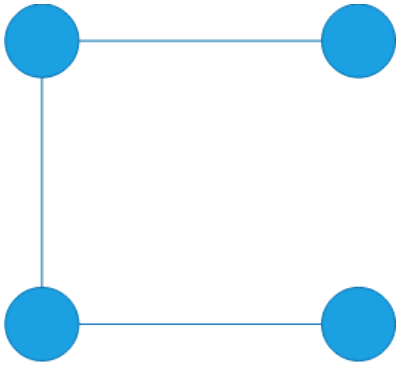
حالت بدون یال:
تعداد حالات:

$$\binom{6}{0} \times 2^0$$



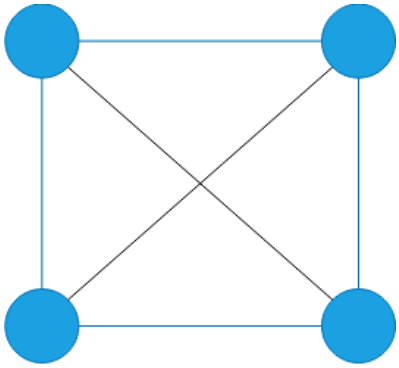
حالت یک یال:
تعداد حالات:

$$\binom{6}{1} \times 2^1$$



حالت سه یال:
تعداد حالات:

$$\binom{6}{3} \times 2^3$$



حالت شش یال:
تعداد حالات:

$$\binom{6}{6} \times 2^6$$

لذا در واقع ۲ حالت برای جهت‌های یال‌ها می‌باشد و با انتخاب تعداد یال و ضرب تعداد حالات هر یال می‌توانیم تعداد کل ساختارهای گراف را محاسبه کرد.

در نهایت باید جمع تمامی حالات ساختار گراف‌ها را محاسبه کرد. برای حالت کلی داریم:

$$\binom{\binom{N}{2}}{0} \times 2^0 + \binom{\binom{N}{2}}{1} \times 2^1 + \ldots + \binom{\binom{N}{2}}{\binom{N}{2}} \times 2^{\binom{N}{2}} = (2+1)^{\binom{N}{2}}$$

که نتیجه‌ی مساوی از عبارت بسط چندجمله‌ای نتیجه شده است:

$$\binom{n}{0} \times a^0 + \binom{n}{1} \times a^{n-1}b + \ldots + \binom{n}{n} \times a\, b^n = (a+b)^{(n)}$$

که در آن داریم:

$$\begin{aligned} a &= 2 \\ b &= 1 \\ n &= \binom{n}{2} \end{aligned}$$

با اجازه استاد گرامی، توضیحات کد (قسمت های ب، د، ه) حضوری تحویل داده شده و نیازی به داکیومنت کد نیست.