پروژه نهایی درس مدل‌های گرافی احتمالاتی
بهار ۱۴۰۲

- این پروژه باید به‌صورت انفرادی انجام شود.
- هر روز تأخیر در ارسال هر بخش، منجر به کاهش ۰٫۱ از نمره آن بخش می‌گردد.
- مستند کردن مراحل پروژه و مرتب بودن کد و مستندات لازم است.
- این پروژه ۷ نمره از کل نمره شما را دارد (بارم بندی از ۱۰ است یعنی ۳ نمره هم بعنوان نمره اضافه برای شما در نظر گرفته شده است).

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)**

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

+ برای پاسخ به این سوال، هم می‌توانید از ابتدا خودتان فکر کنید (قطعاً با داده‌های کلاس قابل حل است)، و یا می‌توانید از مقالات استفاده کنید. اگر از مقالات استفاه می‌کنید، پس از ذکر مرجع، خلاصه‌ی مفید و خوانایی از مقاله را ذکر کنید. ترجمه یا کپی/پیست مجاز نیست.

ب) تابعی که برای محاسبه‌ی درست نمایی نوشته‌اید را در زبان پایتون پیاده سازی کنید. **(۳ نمره)**

+ اگر نتوانستید تابع را خودتان پیاده سازی کنید، از یکی از کتابخانه‌های موجود برای ادامه‌ی کار استفاده نمایید (نمره قسمت ب را نخواهید گرفت). هم می‌توانید از توابع امتیازدهی برای شبکه‌های بیزی استفاده کنید و هم می‌توانید از توابع امتیازی که برای (شبکه‌های تنظیم ژنی[2]) تعریف شده‌اند استفاده نمایید.

[1] Maximum likelihood
[2] Gene regulatory network

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)**

د) از آنجا که با توجه به محاسبه‌ی پاسخ ج خواهید دید که جستجو در چنین فضایی عملاً امکان پذیر نیست، نیاز است که از یک روش دیگر برای یافتن بهترین ساختار استفاده شود. کار شما این است که بر اساس الگوریتم متروپلیس-هستینگز[3] جستجو را انجام دهید. **(۳ نمره)**

+ راهنمایی: از یک گراف تصادفی شروع کنید، سپس با استفاده از یک تابع پروپوزال (که در اینجا باید یک ساختار مجاور را پیشنهاد دهد که از ساختار فعلی دور نیست)، یک گراف جدید پیشنهاد دهید. سپس درست‌نمایی را برای گراف جدید محاسبه کنید و با استفاده از آن احتمال پذیرش این ساختار را محاسبه کنید. در نهایت یا این ساختار پذیرفته می‌شود یا خیر. به همین صورت تا همگرا شدن ادامه دهید. برای بررسی همگرا شدن طبق درس از روش‌های گفته شده استفاده کنید.

ه) در این قسمت برای داشتن شبکه‌ی صحیح از فایل هم‌نام داده‌ها که به انتهای آن کلمه‌ی gold اضافه شده است استفاده کنید. این فایل ساختار واقعی شبکه را نشان می‌دهد (ژن‌ها به صورت $G1$، $G2$ و ... نامگذاری شده‌اند). شبکه‌ای که خودتان در نهایت پس از اجرای الگوریتم متروپلیس-هستینگز به دست آورده‌اید را با این شبکه مقایسه کنید و میزان مثبت کاذب، مثبت درست، منفی کاذب و منفی درست را گزارش کنید. **(۱ نمره)**

فایل‌های داده برای دو شبکه‌ی ده (شروع نام با 1_10) و صد ژنی (شروع نام با 1_100) داده شده است. برای هر شبکه یک فایل داده و یک فایل که ساختار واقعی شبکه را نشان می‌دهد (در نام کلمه‌ی gold وجود دارد) داده شده است.

موفق باشید
زهرا نریمانی

[3] Metropolis-Hastings