به نام خدا

تحلیل داده‌های حجیم زیستی

نیم‌سال دوم ۹۹-۰۰

مدرس: دکتر مطهری

تمرین سری پنجم — موعد تحویل: ۰۲ تیر

مسئله‌ی ۱. نرمال‌سازی (۵۰ نمره)

در این بخش، می‌خواهیم به مقایسه روش‌های مختلف نرمال سازی داده‌های RNA-seq بپردازیم. در ضمیمه صورت سوالات دو فایل GSE60450_Lactation-GenewiseCounts.txt (که شامل اطلاعات تعداد readها و طول ژن برای هر نمونه است) و SampleInfo.txt (که شامل اطلاعات مربوط به گروه هر نمونه است) ارائه گردیده. این مجموعه داده شامل ۶ گروه ۲تایی از نمونه‌هاست.

الف) شما باید داده‌های ارائه شده را به کمک سه روش CPM، TPM و DeSeq2، نرمال‌سازی کرده و سپس به کمک روش‌های PCA و خوشه‌بندی سلسله مراتبی، اقدام به نمایش و خوشه‌بندی داده‌های نرمال شده نمایید. در نتیجه در این بخش باید ۶ نمودار (خروجی PCA و خوشه‌بندی سلسله‌مراتبی برای هر کدام از روش‌های نرمال‌سازی عنوان شده) ارائه گردد.

ب) در این بخش نیاز است تا به مقایسه روش‌های نرمال سازی بپردازید و نقاط ضعف و قوت هر کدام را به صورت جداگانه بیان فرمایید. لازم به ذکر است که ارجاع‌دهی به مراجع مورد استفاده ضروری است.

نکته ۱: شما می‌توانید جهت پیاده‌سازی از زبان‌های R و Python استفاده نمایید.

نکته ۲: جهت دریافت اطلاعات بیشتر در ارتباط با مجموعه داده ارائه شده می‌توانید به Galaxy Training مراجعه فرمایید.

مسئله‌ی ۲. یادگیری ساختار گرافی (۵۰ نمره)

در این بخش می‌خواهیم به بررسی روش Glasso بپردازیم. ما یک مدل گرافی گوسی ایجاد کرده‌ایم و از آن نمونه گرفته‌ایم. ماتریس precision مدل مربوطه در فایل PrecisionMatrix.csv و نمونه‌های گرفته شده در فایل Samples.csv قرار دارد.

الف) ابتدا با روش نمونه‌برداری بدون جایگزین از نمونه‌های موجود در فایل Samples.csv، دسته‌های ۱۰تایی، ۱۰۰تایی و ۱۰۰۰تایی از نمونه‌ها را ایجاد کنید. سپس به کمک روش Glasso و بر اساس مقادیر مختلف پارامتر تنظیم کننده، اقدام به شناسایی ساختار گراف نمایید. یال‌های گراف به دست آمده در حالات مختلف (با تعداد نمونه‌های مختلف) را با یال‌های گراف اصلی مقایسه نمایید. در ادامه منحنی‌های مربوط به نتایج به دست آمده برای حالات مختلف را در نموداری که محور عمودی آن TPR[1]

[1] True Positive Rate

و محور افقی آن FPR[۲] می‌با شد، ر سم نمایید. نهایتا سطح زیر نمودار را برای هر کدام از حالت‌ها به صورت جداگانه محا سبه و گزارش فرمایید.

خروجی: نموداری که محور عمودی آن TPR و محور افقی آن FPR اســت و در آن ســه منحنی مربوط به حالات اســتفاده از ۱۰، ۱۰۰ و ۱۰۰۰ نمونه، وجود دارد + اندازه ناحیه زیر منحنی برای هر سه منحنی

ب) تابع $\boldsymbol{f(x) = e^x}$ را بر روی داده‌های مورد بررســی در بخش قبل اعمال کرده و بررســی‌های عنوان شــده در بخش "الف" را مجددا بر روی داده‌های جدید انجام دهید.

خروجی: نموداری که محور عمودی آن TPR و محور افقی آن FPR اســت و در آن ســه منحنی مربوط به حالات اســتفاده از ۱۰، ۱۰۰ و ۱۰۰۰ نمونه، وجود دارد + اندازه ناحیه زیر منحنی برای هر سه منحنی

ج) با اســتفاده از داده‌های مربوط به گره‌های ۱ تا ۱۰ موجود در فایل Samples.csv و به کمک روش‌های انتخاب مدل AIC[۳] و BIC[۴] به منظور شنا سایی مقدار منا سب پارامتر تنظیم‌کننده مدل Glasso، ساختار گرافی را در این حالات شنا سایی کرده و نمایش دهید.

خروجی: دو ساختار گرافی، یکی برای حالت استفاده از روش AIC برای تنظیم پارامتر و دیگری برای حالت استفاده از روش BIC

---

[2] False Positive Rate

[3] Akaike Information Criterion

[4] Bayesian Information Criterion