

● مهلت تحویل تمرین: ۱۳۹۷/۰۷/۲۴

● تمرین کامپیوتری

۱. در کلاس درس دیدیم اگر رابطه‌ی میان دو متغیر تصادفی خطی و نویز برون^۱ از جنس غیر گاوسی باشد می‌توان جهت صحیح رابطه‌ی علی میان متغیرها را کشف کرد. همچنین دیدیم که غیرخطی بودن رابطه‌ی میان متغیرهای تصادفی نیز کمک می‌کند تا جهت صحیح رابطه‌ی علی را بیابیم.

(آ) یک مدل SCM دو متغیری را در نظر بگیرید که در آن $Y = X + bX^3 + \text{sign}(N)|N|^q$ که X و N متغیرهای تصادفی مستقل از توزیع نرمال باشند و پارامترهای b و q به ترتیب کنترل کننده‌ی خطی بودن و گاوسی بودن مدل هستند. با استفاده از مدل ارائه شده در بالا، نمودار توزیع های $p(x, y)$, $p(x|y)$, $p(y|x)$ برای دو مدل خطی و غیر خطی با نویز گاوسی را رسم کنید.

(ب) در این بخش می‌خواهیم تاثیر غیرخطی بودن و غیر گاوسی بودن در تشخیص جهت علی را بررسی کنیم. با استفاده از مدل بالا ۳۰۰ نمونه تولید کنید. سپس با استفاده از یک رگرسیون غیر خطی مانند SVR مدل را برازش کنید. اگر تابع برازش شده را \hat{f} بنامیم و

$$\hat{n} = y - \hat{f}(x)$$

آنگاه اگر باقی مانده \hat{n} بدست آمده مستقل از x باشد می‌توان گفت جهت مدل درست تشخیص داده شده است. برای بررسی استقلال این نمونه‌ها از تست آماری HSIC استفاده کنید و با سطح اطمینان ۲٪^۲ فرضیه صفر را رد کنید. این آزمایش را ۱۰۰ بار تکرار کنید و نمودار مربوط به درصد پذیرش مدل در دو جهت را رسم کنید (کد مربوط به تست HSIC در اختیار شما قرار گرفته است).

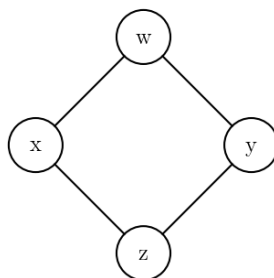
(ج) در این بخش قصد این را داریم که آزمون بالا را روی دو دسته داده دنیای واقعی بررسی کنیم:

○ مجموعه اول^۳ حاوی تعدادی نمونه از مدت زمان فوران و فاصله از فوران قبلی آیفشان است.

○ مجموعه دوم^۴ حاوی تعدادی نمونه از ویژگی‌های فیزیکی یک نوع صدف است. اطلاعات مربوط به این داده در این لینک^۵ قرار دارد.

داده‌ها را از لینکهای داده شده دانلود کنید. سپس در مجموعه داده اول سعی کنید جهت صحیح رابطه‌ی علی را میان مدت زمان فوران و فاصله فوران پیدا کنید. در مجموعه داده دوم سعی کنید جهت صحیح رابطه‌ی علی میان تعداد حلقه‌ها و طول صدف را بدست آورید.

۲. در این بخش قصد آن را داریم که روابط علی را در یک گراف جهت‌دار بدون دور بررسی کنیم. مجموعه داده‌ای که در اختیارتان قرار گرفته توسط یک گراف با ساختار زیر تولید شده است. جهت درست روابط در این گراف جهت‌دار را حدس بزنید.



فایل های تحویلی:

گزارش مختصری به همراه نتایج هر بخش در قالب pdf. آماده کنید و همراه با کد تمرین در قالب یک فایل فشرده در سایت کوئرا آپلود نمایید.

^۱Exogenous noise

^۲Confidence level

^۳<https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat>

^۴<https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data>

^۵<https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names>