دانشگاه تهران

دانشکده علوم مهندسی

الگوریتم‌ها و محاسبات

یادگیری ماشین – دکتر سایه میرزایی

تمرین پنجم

بهار ۰۰

یادگیری غیرنظارتی[1]

سوال اول: خوشه‌بندی K-میانگین[2]

۱- در مورد روش K-میانگین و روابط آن تحقیق کنید و یک مثال ساده برای آن طراحی کنید. آیا الگوریتم همگرا می‌شود؟ (در صورت مثبت بودن جواب با روابط ریاضی اثبات کنید)

۲- داده‌های Wine را دریافت کنید و الگوریتم K-میانگین را با ۱۰۰ بار تکرار و تعداد خوشه‌های ۳، ۵ و ۷ پیاده‌سازی کنید. خوشه‌های به‌دست آمده به همراه نقطه میانگین را رسم کنید.

۳- در مورد معیارهای شباهت درونی و بیرونی تحقیق کنید و حداقل دو مورد از هر کدام را بررسی کنید.

۴- مناسب‌ترین عدد برای تعداد خوشه‌ها را با استفاده از یک معیار شباهت درونی و یک معیار شباهت بیرونی به‌دست آورید.

۵- روشی برای پیداکردن تعداد بهینه خوشه‌بندی ارائه کنید.

سوال دوم: خوشه‌بندی سلسله مراتبی[3] (می‌توانید این سوال را بدون برنامه‌نویسی حل کنید)

جدول زیر نشان‌دهنده داده‌های کتگوریکال ۵-بعدی است:

Point	X_1	X_2	X_3	X_4	X_5
$\mathbf{x}_1^T$	1	0	1	1	0
$\mathbf{x}_2^T$	1	1	0	1	0
$\mathbf{x}_3^T$	0	0	1	1	0
$\mathbf{x}_4^T$	0	1	0	1	0
$\mathbf{x}_5^T$	1	0	1	0	1
$\mathbf{x}_6^T$	0	1	1	0	0

مشابهت بین داده‌های کتگوریکال را می‌توان با استفاده از تعداد ویژگی‌های مشابه و متفاوت محاسبه کرد. فرض کنید n_{11} نشان‌دهنده تعداد ویژگی‌هایی باشد که نقاط x_i و x_j هر دو ۱ هستند، همچنین فرض کنید n_{10} نشان‌دهنده تعداد ویژگی‌هایی باشد که x_i مقدار ۱ دارد و x_j مقدار ۰، به همین صورت n_{01} و n_{00} را نیز تعریف می‌کنیم.

برای هر جفت از داده‌ها معیارهای شباهت زیر را به‌دست آورید و جدول مناسب را رسم کنید.

- Simple matching coefficient: $SMC(x_i, x_j) = \frac{n_{11}+n_{00}}{n_{00}+n_{11}+n_{01}+n_{10}}$
- Jaccard coefficient: $JC(x_i, x_j) = \frac{n_{11}}{n_{11}+n_{01}+n_{10}}$

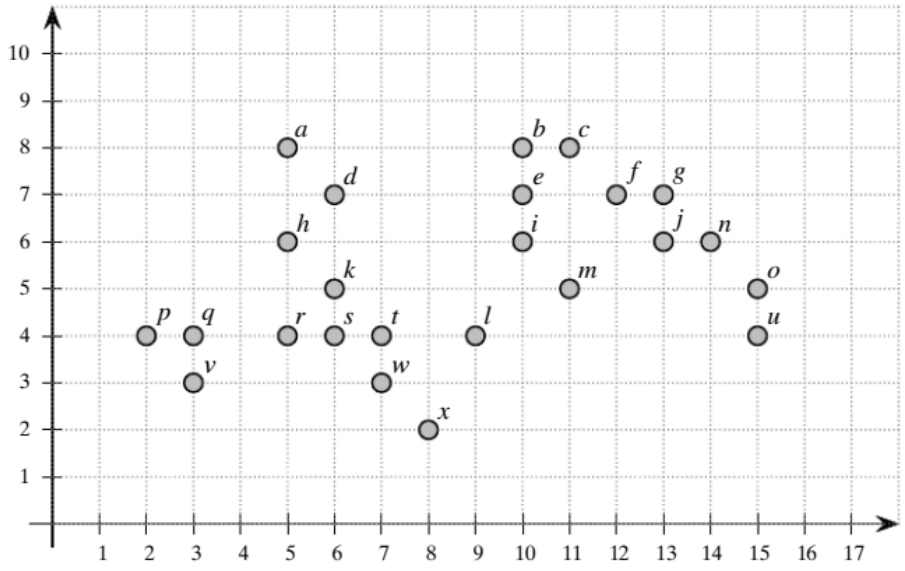
دندوگرام تولید شده توسط الگوریتم خوشه‌بندی سلسله‌مراتبی Single link را با یکی از معیارهای شباهت به‌دست آمده رسم کنید.

---

[1] Unsupervised learning
[2] k-means
[3] hierarchical clustering

سوال سوم: خوشه‌بندی با DBSCAN (می‌توانید این سوال را بدون برنامه‌نویسی حل کنید)

شکل زیر را در نظر بگیرید و به سوالات پاسخ دهید. (از فاصله اقلیدسی میان نقاط، $minpts = 3$ و $\epsilon = 2$ استفاده کنید)



الف) نقاط مرکزی را مشخص کنید. (محاسبات را حداقل برای دو نقطه بنویسید)

ب) نقاط را با استفاده از الگوریتم DBSCAN خوشه‌بندی کنید و نقاط نویزی را نیز مشخص کنید.

سوال چهارم: PCA (برای حل این سوال استفاده از np.linalg.svd مجاز است)

۱- با استفاده از روش PCA ویژگی‌های داده‌های Iris که با آن‌ها آشنا هستیم را به دو ویژگی کاهش دهید و پس از انتقال همه داده‌ها به فضای جدید آن‌ها را رسم کنید.

۲- برداروِیژه و مقداروِیژه اول و دوم داده‌های Iris را با Power Method به‌دست آورید و پس از انتقال همه داده‌ها به فضای جدید آن‌ها را رسم کنید.

۳- ویژگی‌های داده‌ها را با استفاده از یک کرنل غیرخطی با روش PCA به دو ویژگی کاهش دهید و پس از انتقال همه داده‌ها به فضای جدید آن‌ها را رسم کنید. (استفاده از توابع KPCA آماده مشکلی ندارد)

۴- در مورد AutoEncoder ها تحقیق کنید (حداقل ۱ صفحه). با استفاده از AutoEncoder ویژگی‌های داده‌های Iris را ابتدا با یک تابع فعال‌ساز خطی و سپس با یک تابع فعال‌ساز غیرخطی به دو ویژگی کاهش دهید و پس از انتقال همه داده‌ها به فضای جدید آن‌ها را رسم کنید. نتایج به‌دست آمده با تابع فعال‌ساز خطی را با قسمت ۱ و نتایج به‌دست آمده با تابع فعال‌ساز غیرخطی را با قسمت ۳ مقایسه کنید.

۵- الگوریتم LDA را با PCA به‌طور کامل مقایسه کنید. هر کدام در چه مواردی کاربرد بیشتری دارند؟

نکات

- **تمرین‌ها را در سامانه ایلرن تحویل بدهید.**
- لطفا گزارش خود را به زبان فارسی تهیه کنید و تمامی نکات، فرض‌ها و فرمول‌ها در آن ذکر شوند. گزارش در روند تصحیح تمرین‌ها از اهمیت ویژه‌ای برخوردار است.
- کپی‌کردن کدهای آماده موجود در اینترنت و یا استفاده از کدهای همکلاسی‌ها تقلب محسوب می‌شود.
- استفاده از کتابخانه‌های آماده پایتون به جز *Numpy*، *Pandas* و *Matplotlib* غیرمجاز است، تنها برای بارگذاری داده‌ها mat. می‌توانید از کتابخانه‌های دیگر استفاده کنید.
- در صورت مشاهده تقلب نمرات تمامی افراد شرکت‌کننده در آن صفر لحاظ می‌شود.
- پس از به اتمام رسیدن مهلت تحویل تمرین، تاخیر تا یک هفته با کسر ۳۰ درصد نمره لحاظ خواهد شد.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق گروه کلاسی یا ایمیل mo.bakhtyari@ut.ac.ir با دستیار آموزشی در تماس باشید.