به نام خدا

یادگیری ماشین برای بیوانفورماتیک — نیمسال دوم ۰۰ـ۹۹

تمرین سری سوم

موعد تحویل: ۲۳ اردیبهشت — نمره: ۲۰+۱۰۶

سوال ۱: Clustering (26 points)

۱.۱. (۹ نمره) به سوالات زیر با ذکر دلیل پاسخ دهید:

(آ) (۳ نمره) آیا الگوریتم K-means با معیار فاصله اقلیدسی،حالت خاصی ازالگوریتم EM است که در آن از k تابع گوسی با واریانس یکسان برای هربعد استفاده می‌کنیم؟

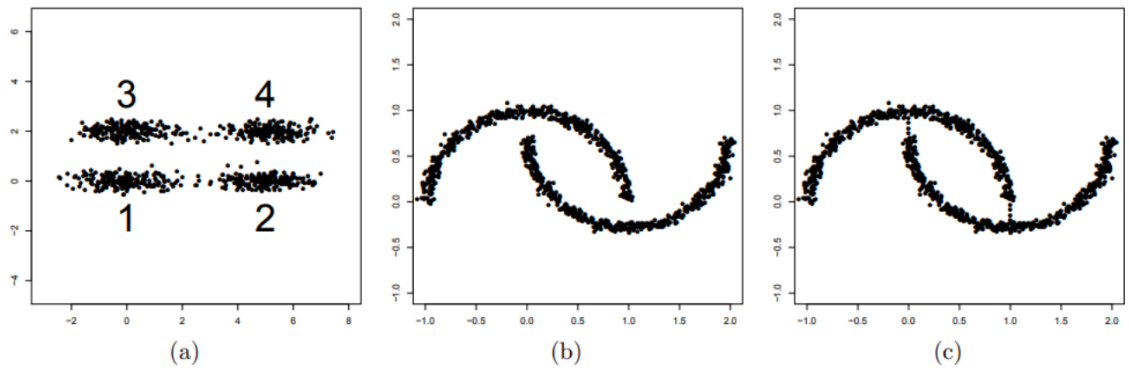
(ب) (۴ نمره) درخوشه‌بندی سلسله مراتبی با کدامیک از معیارهای شباهت خوشه‌ای امکان دارد که داده‌ای در یک خوشه به داده‌ای از خوشه دیگر،نزدیکتر از داده‌ای درخوشه خودش باشد.

(ج) (۲ نمره) به یک مجموعه از راس‌ها p-cluster گفته می‌شود اگر حداقل p درصد از یال‌های این راس‌ها به راس‌های داخل این مجموعه متصل باشند.اگر ما خوشه‌ها را p-cluster های گراف درنظر بگیریم،آیا نتیجه خوشه‌بندی با این تعریف مطلوب است؟

۱.۲. (۶ نمره) با توجه به مجموعه داده‌های شکل۱ به سوالات زیر پاسخ دهید:

(آ) (۳ نمره) در مجموعه داده a اگر با استفاده از خوشه‌بندی سلسله مراتبی با $K = 2$ خوشه‌بندی صورت گیرد،با استفاده از هرکدام از معیارهای شباهت خوشه‌ای single_link و complete_link و average_link ، ۴دسته مشخص شده به چه خوشه‌ای تعلق می‌گیرند. ؟

(ب) (۳ نمره) کدامیک از سه معیار فاصله درصورت وجود میتوانند داده‌ها در دوشکل c و b را با موفقیت جدا کند؟پاسخ خود را به صورت خلاصه توضیح دهید.



شکل ۱ : سوال۱.۲

۱.۳. (۵ نمره) برای ترکیب دو خوشه در خوشه‌بندی سلسله مراتبی روشی به نام Ward هست که با محاسبه تابع هزینه مورد استفاده در خوشه‌بندی k-means یعنی یافتن کمینه مربع مجذور فاصله ،دوخوشه را ترکیب می‌کند.تابع هزینه مورد استفاده، تابع زیر است:

$$\text{cost}(T) = \sum_{x \in S} min_{t \in T} ||x - t||^2$$

ثابت کنید برای هر دوخوشه C و C’ تابع هزینه ترکیب این دو خوشه ،معادله زیر حاکم است:

$$\text{cost}(C \cup C') = \text{cost(C)} + \text{cost(C')} + \frac{|c||c'|}{|c|+|c'|} ||mean(C) - mean(C')||^2$$

۱.۴. (۶ نمره) جدول فاصله برای ۶ شی زیر را در نظر بگیرید:

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

جدول ۱ : داده‌های مسئله ۵

(آ) نمودار درختی برای خوشه‌بندی[1] به روش Single-linkage را رسم کنید.

(ب) نمودار درختی برای خوشه‌بندی به روش Complete-linkage را رسم کنید.

(ج) دو مقدار از جدول بالا را چنان تغییر دهید که نمودارهای دو سوال قبل مشابه شوند.

سوال ۲: Feature Selection(+10 points)

۲.۱. با مراجعه به بخش ۲ مقاله Ben-Dor et al. معیار TNoM در انتخاب ویژگی تک‌متغیره را توضیح دهید.

۲.۲. روش‌های انتخاب ویژگی چه مزیتی نسبت به روش‌های استخراج ویژگی (Feature extraction e.g., PCA) دارند.

۲.۳. مثالی از عدم کارایی هر یک از دو معیار همبستگی پیرسون و Mutual information در انتخاب ویژگی تک‌متغیره بزنید.

سوال ۳: Dimension Reduction(15 points)

۳.۱. ماتریس X که نمونه‌ها در سطرهای آن قرار گرفته اند در نظر بگیرید. نشان دهید یافتن راستاهایی که داده‌های بازسازی شده روی آن‌ها فاصله‌ی کمی با داده‌های اصلی دارند معادل یافتن راستاهای با واریانس زیاد است.به عبارت دیگر نشان دهید رابطه

$$argmin_D||X - XDD^T||_F$$
$$w.r.t. D^T D = I$$

نتیجه می‌دهد

$$(X^T X)D = \lambda D$$

۳.۲. ماتریس X که نمونه‌ها در ستون‌های آن قرار گرفته‌اند با یک کرنل به یک فضای غیرخطی برده‌ایم و اکنون ماتریس $K = \phi^T(X)\phi(X)$ را در اختیار داریم.از PCA می‌دانیم بردار v که داده‌ها در جهت آن بیشترین واریانس را دارند در رابطه زیر صدق می کند

$$\phi(X)\phi^T(X)v = \lambda v$$

با استفاده از K بازتاب داده‌ها بر بردار v را بیابید.

۳.۳. (NMF)Non-negative Matrix Factorization یکی از روش‌های کاهش ابعاد است که همزمان قابلیت خوشه‌بندی داده‌ها را نیز دارد. NMF با داشتن ماتریس X که داده‌های آن در ستون‌ها قرار گرفته و ویژگی‌هایشان مقادیر مثبت دارند مساله زیر را حل می‌کند

$$argmin_{W\geqslant 0, H\geqslant 0}||X - WH||_F$$

اگر ابعاد ماتریس X $m \times n$ بوده و ابعاد ماتریس W $m \times p$ باشد که p بعد نهایی است، کدام ماتریس داده‌های تبدیل‌یافته را در خود جای می‌دهد.

سوال ۴: Clustering&Feature selection(Practical)(65+10)

[1] Clustering

در این بخش یک دیتاست مرتبط با تشخیص بیماری COVID-19 و همچنین یک نوتبوک شامل دو بخش کاهش ابعاد و خوشه‌بندی در اختیار شما قرار گرفته است. در این نوتبوک از شما خواسته شده الگوریتم PCA ، K-Means و EM (Expectation–maximization) را پیاده سازی کنید و الگوریتم‌های دیگر در زمینه خوشه‌بندی همچون DBSCAN یا کاهش ابعاد همچون KPCA را از کتابخانه scikit-learn استفاده کنید.
در این نوتبوک شما ابتدا ابعاد داده‌ای که در اختیارتان قرار گرفته است را با استفاده از روش مناسب به دو کاهش می‌دهید و سپس خوشه‌بندی روی آن با روش مناسب را انجام می‌دهید و سپس میزان تفکیک پذیری خوشه‌ها از یکدیگر را رسم می‌کنید تا از میزان تفکیک پذیری آن‌ها شهود پیدا کنید. در نهایت میزان تفکیک پذیری روش ها را با هم مقایسه می‌نمایید.

۵.۱. (۱۰+۲۰ نمره) کاهش ابعاد و انتخاب ویژگی

(آ) (۲۰ نمره) الگوریتم PCA را با استفاده از numpy پیاده‌سازی کنید و الگوریتم های ذکر شده در نوتبوک را از کتابخانه scikit-learn گرفته و بر روی داده اعمال کنید تا در نهایت بعد داده به ۲ کاهش یابد.

(ب) (۱۰ نمره) (امتیازی) با استفاده از شبکه‌های عصبی و ساختار خودکدگذار (Autoencoder) ابعاد داده را به دو کاهش دهید. برای جزئیات بیشتر به این لینک مراجعه کنید.

۵.۲. (۴۵ نمره) خوشه‌بندی

(آ) (۳۰ نمره) ابتدا الگوریتم‌های K-Means و EM را طبق ساختار موجود در نوتبوک پیاده سازی کنید و بر روی داده‌های کاهش بعد یافته اعمال کنید.سپس الگوریتم‌های دیگر مانند DBSCAN را طبق ترتیب بیان شده از کتابخانه scikit-learn گرفته و بر روی داده‌های کاهش بعد یافته اعمال کنید. (ارزیابی روش خوشه‌بندی را با استفاده از برچسب‌های داده‌های تست انجام دهید)

(ب) (۱۵ نمره) نتایج حاصل از بخش قبل را تصویرسازی کرده و میزان تفکیک پذیری خوشه‌ها را با روش‌های مختلف موجود در این نوتبوک مقایسه نمایید.

(ج) در این بخش خوشه‌بندی را در تحلیل داده‌های با ابعاد بالا به کار خواهید برد.