توضیحات مهم:

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان studentID_HW1.zip بارگذاری نمایید(بعنوان مثال 99131000_HW1.zip).
- مهلت انجام تمرین تا ساعت ۲۳:۵۵ تاریخ ۱۴۰۱/۰۳/۱۵ می‌باشد و **به هیچ وجه تمدید نمی‌شود.**
- استفاده از کتابخانه‌های رایج در یادگیری ماشین بلا مانع است.
- برای سهولت در انجام تمرین می‌توانید از پلتفرم کولب گوگل استفاده نمایید.
- ملاک اصلی انجام تمرین گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. یک فایل pdf تهیه کرده و برای هر سوال، ورودی، خروجی و توضیحات مربوطه را بصورت جامع گزارش کنید.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید. (به بهترین گزارش نمره‌ی تشویقی تعلق می‌گیرد.)
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره‌ی طرفین **صفر** می‌شود.
- شما مجاز هستید برای تمامی تمرین‌ها ۷ روز در کل و با سقف حداکثر ۳ روز برای هر تمرین، تاخیر بدون کسر نمره داشته باشید. به ازای هر روز تاخیر بیشتر، ۱۰٪ از نمره‌ی تمرین مربوطه کسر می‌شود.
- در صورت داشتن هر گونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

Z.Khalvandi77@gmail.com
Najmeh.Mohammadbagheri77@gmail.com

بخش اول: سوالات تشریحی

۱- صحت هر یک از موارد زیر را بررسی کرده و دلایل خود را توضیح دهید.

الف) ماشین‌های بردار پشتیبان[1] پارامتریک[2] اند.

ب) مقدار حاشیه[3]ی به دست آمده برای دو ماشین بردار پشتیبان با هسته[4] های متفاوت که برای داده‌های یکسان آموزش دیده‌اند، میتواند معیاری برای میزان کارایی مدل باشد.

ج) ماشین‌های بردار پشتیبان همواره در برابر بیش برازش مقاوم‌اند.

د) وجود داده‌های پرت و نویز بر روی ماشین‌های بردار پشتیبان بی‌تاثیر است.

ه) الگوریتم آدابوست با استفاده از هر نوع دسته‌بند ضعیف و یا ترکیب چند دسته‌بند ضعیف، در نهایت به خطای آموزش صفر می‌رسد.

و) وزن‌های اختصاص داده شده به دسته‌بندها در الگوریتم آدابوست همواره نامنفی هستند.

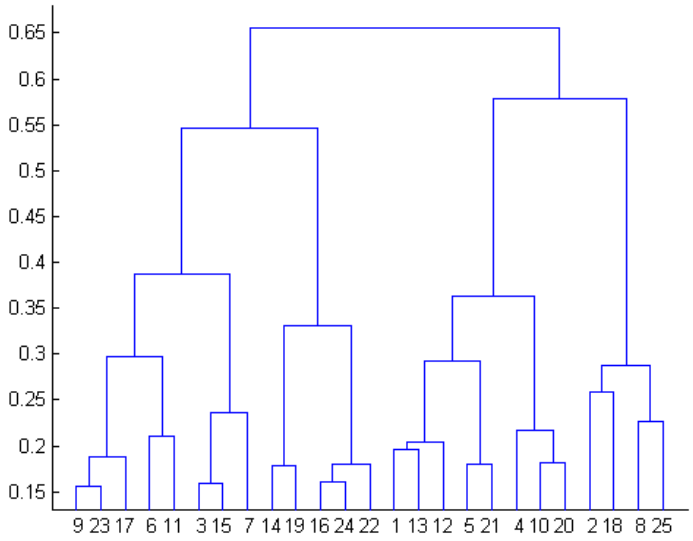
۲- آیا در ساخت یک درخت تصادفی (Random Forest) لازم است که از هرس کردن استفاده کنیم؟ دلیل پاسخ خود را شرح دهید.

۳- چرا درخت‌های ساخته‌شده در جنگل تصادفی را تصادفی می‌نامیم؟

۴- چگونه می‌توان از ماشین بردار پشتیبان برای رگرسیون استفاده کرد؟

۵- نمودار زیر برای مقادیر شباهت در خوشه‌بندی سلسله مراتبی بدست آمده است. با توجه به این نمودار بهتر است داده‌ها را به چند خوشه تقسیم کنیم؟

[1] SVM
[2] Parametric
[3] Margin
[4] Kernel



۶ – در شکل زیر جدول یک مختصات شش نقطه را در فضای دو بعدی و جدول دو فاصله بین هر دو نقطه را نمایش می‌دهد. این داده‌ها را در حالتی که معیار ادغام single link باشد، به روش سلسله مراتبی خوشه‌بندی کنید. نمودار ون و دندروگرام[5]را رسم کنید.

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

[5] Dendrogram

بخش دوم: پیاده‌سازی

ماشین بردار پشتیبان

۱ -مجموعه داده‌ی Parkinson.data که در فایل تمرین وجود دارد را بارگذاری کرده و داده‌ها را با استفاده از مدل SVM و کرنل‌های زیر دسته‌بندی کرده و به سوالات پاسخ دهید.

الف) کرنل خطی

ب) کرنل چند جمله‌ایی (پارامترهای d,r)

ج) کرنل rbf (پارامتر گاما)

د) سیگموید(پارامتر r)

یک) معیار Accuracy و F1-Measure را برای هر یک از دسته‌بندهای بالا به دست آورده و مقادیر بهینه را مشخص کنید.(برای هر یک از پارامترهای یاد شده، حداقل۴ مقدار متفاوت در نظر بگیرید)

دو) تاثیر پارامترهای هر کرنل بر کارایی مدل‌ها را تحلیل کنید.

سه) آیا روشی هوشمند برای تنظیم پارامترها وجود دارد؟ به طور خلاصه توضیح دهید.

Ensemble Method

۲- هدف از این بخش انجام عمل رگرسیون با استفاده از الگوریتم جنگل تصادفی است. در این سوال مجموعه داده‌ایی برای پیش‌بینی قیمت خانه در نظر گرفته‌شده است. باید بتوانید با استفاده از ویژگی‌های خواسته‌شده قیمت خانه را پیش‌بینی کنید.

الف) پیش‌پردازش و پاکسازی‌های لازم را بر روی داده‌ها انجام دهید و روش خود را گزارش کنید.

ب) داده‌ها را با مقادیر سه ویژگی 'Latitude', 'Longitude', 'MedHouseVal' رسم کنید.

پ) به کمک مدل جنگل تصادفی متغیر 'MedHouseVal' را پیش‌بینی کنید و مقدار خطای مدل را پیش‌بینی کنید. برای هر پارامترهای تابع (n_estimators, max_depth) حداقل سه مقدار مختلف را آزمایش و تاثیر آن‌ها را بر نتیجه‌ی پیش‌بینی بررسی کنید.

خوشه‌بندی

۳- با استفاده از یک کتابخانه‌ی آماده که در آن الگوریتم خوشه‌بندی k_means وجود دارد، موارد زیر را پیاده سازی نمایید.

الف) ابتدا دو تصویر flower.jpg, butterfly.jpg را بارگذاری کرده و نمایش دهید. هر پیکسل از تصویر با سه ویژگی R,G,B مقداردهی می‌شود. در این مسئله هر پیکسل تصویر یک نمونه‌ی داده است که سه ویژگی دارد. در این بخش باید بتوانید با استفاده از مقادیر پیکسل‌ها خوشه‌بندی را انجام دهید. خوشه‌بندی را با تعداد ۲و۳و۵و۸و۱۰و۱۵و۳۰ انجام دهید و نتایج را گزارش کنید. به این ترتیب که برای هر پیکسل مقدار پیکسل مرکز خوشه‌ایی که در آن قرار دارد را جایگزین کنید.

ب) در این بخش مجموعه داده‌ی Shill Bidding Dataset.csv را بارگذاری کنید.

ب -۱) یکی از روش‌های تعیین تعداد خوشه‌های بهینه در الگوریتم k-means استفاده از روش elbow است؛ این روش را توضیح دهید.

ب -۲) تعداد خوشه‌ها را از ۱ تا ۱۰ تغییر دهید و الگوریتم را اجرا نمایید. با توجه به روش elbow بهترین تعداد خوشه، برای خوشه‌بندی مجموعه داده را مشخص نمایید و دلیل انتخاب خود را توضیح دهید.

۴ – در این بخش می‌خواهیم اثر شکل داده‌ها بر روی عملکرد الگوریتم‌های مختلف خوشه‌بندی را بررسی کنیم، به این ترتیب که سه مجموعه داده‌ی مختلف موجود است که باید بر روی هر مجموعه داده سه روش مختلف خوشه‌بندی اعمال شود.

الف) توزیع سه مجموعه داده‌ی داده شده را در فضای دو بعدی نمایش دهید.

ب) بر روی هر مجموعه داده الگوریتم‌های k_means، DBSCAN و guassian mixtures را اجرا و معیار purity را در هر حالت محاسبه کنید و نتایج خوشه‌بندی هر الگوریتم را بر روی هر مجموعه داده نمایش دهید.

راهنمایی :

تعداد خوشه‌ها برای الگوریتم k_means را در هر حالت با توجه به توزیع داده‌ها انتخاب کنید. همچنین مقادیر مناسب پارامترهای دو روش دیگر به کمک توزیع داده‌ها و با آزمودن مقادیر مختلف تنظیم کنید.