به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده‌ی مهندسی کامپیوتر

تمرین سری چهارم یادگیری ماشین

دکتر احسان ناظرفرد

طراح سوال:

محمدرضا امامی ناصری

سید اردلان قریشی

دی ۱۳۹۹

توضیحات مهم:

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان #StudentId_HW4.zip بارگذاری نمایید (به عنوان مثال 99131000_HW4.zip).
- مهلت انجام تمرین تا **ساعت ۲۳:۵۵** روز **سه‌شنبه** مورخ **۳۰ دی** می‌باشد و به هیچ وجه تمدید نمی‌شود.
- تمرین بدون گزارش فاقد ارزش می‌باشد و **نمره‌ای به آن تعلق نمی‌یابد**.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید (به بهترین گزارش نمره تشویقی تعلق می‌گیرد).
- مطابق قوانین دانشگاه هرگونه کپی برداری **ممنوع** می‌باشد و در صورت مشاهده نمره‌ی **هر دو طرف** صفر در نظر گرفته 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 هرگونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

MLAUTFALL99@gmail.com

سوالات تشریحی

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

الف) ماشین‌های بردار پشتیبان[۱] پارامتریک[۲]اند.

ب) مقدار حاشیه[۳]ی به دست آمده برای دو ماشین بردار پشتیبان با هسته[۴]های متفاوت که برای داده‌های یکسان آموزش دیده‌اند، می‌تواند معیاری برای میزان کارایی مدل باشد.

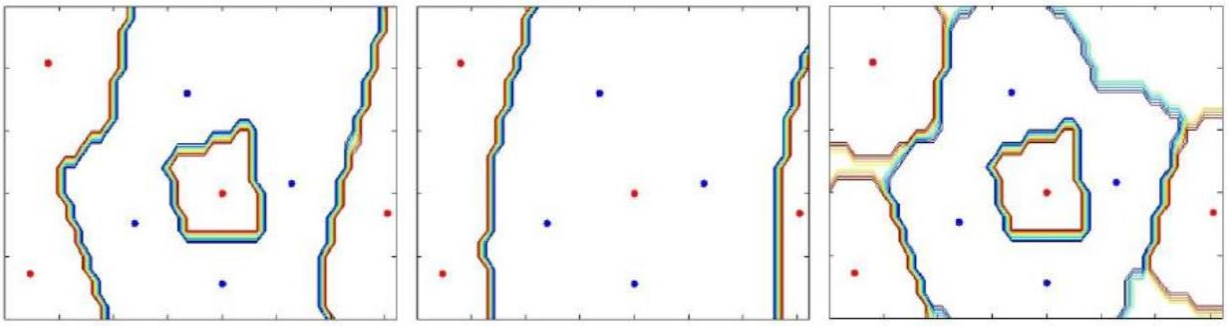
ج) ماشین‌های بردار پشتیبان همواره در برابر بیش برازش مقاوم[۵]اند.

د) وجود داده‌های پرت[۶] و نویز بر روی ماشین‌های بردار پشتیبان بی‌تاثیر است.

ه) الگوریتم آدابوست[۷] 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- برای حل مسئله‌ی دسته‌بندی دو کلاسه از روش بردار پشتیبان با هسته‌ی RBF با σ ={0.2, 1, 10} استفاده کرده‌ایم. مشخص کنید که هر کدام از شکل‌های زیر حاصل دسته‌بندی با کدام مقدار سیگما است؟



۳- تفاوت دو روش hard voting و soft voting در الگوریتم‌های مبتنی بر رای‌گیری[۸] چیست؟

[1] Support Vector Machines (SVMs)
[2] parametric
[3] margin
[4] kernel
[5] robust
[6] outlier
[7] adaboost
[8] voting

۴- فرض کنید جدول زیر مربوط به یکی از داده‌های آزمون است. در هر یک از دو روش hard voting و soft voting، کلاس پیش‌بینی شده توسط الگوریتم کدام است؟ ($w_1=2$، $w_2=1$، $w_3=2$)

	Probabilities		
Classifier	**Class 1**	**Class 2**	**Class 3**
Classifier 1	0.1	0.5	0.4
Classifier 2	0.6	0.3	0.1
Classifier 3	0.4	0.3	0.3

سوالات پیاده‌سازی

توضیحات مهم:

- در روند اجرا انتخاب مقادیر برای تقسیم داده‌ها به مجموعه آموزش، ارزیابی و... به عهده دانشجو می‌باشد.
- حتما پارامترهای انتخاب شده برای برنامه خود و هرگونه شرایطی که درنظر گرفته‌اید را در گزارش خود بیاورید.
- برای بهبود سرعت برنامه توصیه می‌شود از عملیات ماتریسی استفاده کنید.
- در هر مرحله، نتایج خود را تحلیل کنید.
- کدهای خود را برای خوانایی بیشتر **کامنت گذاری** کنید.
- گذاشتن عنوان برای نمودارها و برچسب گذاری محورهای نمودار الزامی می‌باشد.
- **توجه: در این تمرین برای تمامی بخش‌های پیاده‌سازی، مجاز به استفاده از کتابخانه‌های آماده هستید.**

۱- مجموعه داده‌ی Parkinson.data که در فایل تمرین وجود دارد را بارگذاری کرده و داده‌ها را با استفاده از مدل SVM و کرنل[۹]های زیر دسته‌بندی کرده و به سوالات پاسخ دهید.[۱۰]

الف) کرنل خطی[۱۱]

ب) کرنل چند جمله‌ای[۱۲] (پارامترهای r و d)

ج) RBF[۱۳] (پارامتر گاما)

د) سیگموید[۱۴] (پارامتر r)

۱,۱) 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، حداقل ۴ مقدار متفاوت در نظر بگیرید.)

۱,۲) تاثیر پارامترهای هر کرنل بر کارایی مدل‌ها را تحلیل کنید.

۱,۳) آیا روشی هوشمند برای تنظیم پارامترها وجود دارد؟ به طور خلاصه توضیح دهید.

[9] Kernel
[10] More info: SVM Kernels (Link)
[11] Linear
[12] Polynomial
[13] Radial Basis Function
[14] Sigmoid

۲- مجموعه داده‌ی pima_indians_diabetes.csv که در فایل تمرین وجود دارد را بارگذاری کرده و به کمک مدل‌های زیر داده‌ها را دسته‌بندی کنید.

۲,۱) به ازای حداقل ۳ مقدار برای هر یک از پارامترهای زیر، دسته‌بند جنگل تصادفی[15] را بر روی این مجموعه داده آموزش دهید و دقت مدل بر روی مجموعه آموزش و آزمون را گزارش و بهترین مدل را مشخص کنید.

(n_estimators, max_features, max_depth)

۲,۲) نقش و تاثیر پارامترهای یاد شده بر عملکرد مدل‌ها را تحلیل کنید.

۲,۳) از هر یک از روش‌های ترکیبی و یا پایه‌ی دلخواه استفاده کرده و سعی کنید دقت بر روی مجموعه داده‌ی آزمون را افزایش دهید. (برای این بخش تنها ۳ مدل نهایی کفایت می‌کند. اما بهتر است دقت به دست آمده، بیشتر یا مساوی دقت بهترین مدل بخش پیشین باشد. بهترین دقت مشمول نمره امتیازی خواهد بود.)

با آرزوی موفقیت!

[15] random forest classifier