سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها هست، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. کدهای ارسال شده **بدون گزارش** فاقد نمره هستند.
4. برای سوالات شبیه‌سازی، فقط از دیتاست داده شده استفاده کنید.
5. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW4_StudentNumber داشته باشد.
6. از بین سوالات **شبیه‌سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره کل سوالات تمرین ۱۰۰ نمره هست.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه‌سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل elahe.bvakili97@gmail.com، سوال خود را مطرح کنید.

سوال اول (۱۰ نمره)

برای سوالات زیر پاسخ مناسب ارائه دهید.

۱. در صفحه Hyperplane که مرز تصمیم LDF می باشد، جهت گیری و مکان صفحه چطور تعیین میشود.؟

۲. در صورت وجود نویز در داده‌ی آموزش، آیا روش SVM توانای تفکیک آن‌ها را دارد؟ (بررسی کنید که چطور تفکیک انجام می شود)

۳. مفهوم کلی کرنل را بیان کنید. فرض کنید از دو کرنل چند جمله‌ای درجه دو زیر برای تعریف طبقه‌بندی SVM استفاده شده است. کدام یک از طبقه‌بندها حاشیه بزرگ‌تری را نتیجه می‌دهد؟ چرا؟

$$\varphi_1 = [x. x^2]^T$$
$$\varphi_2 = [2x. 2x^2]^T$$

سوال دوم (۱۵ نمره) Maede

اگر کرنل های معتبر $K_1(x.y)$ و $K_2(x.y)$ را داشته باشیم، اعتبار کرنل های زیر را بررسی کنید.

۱.

$$K(x.y) = \exp\left(-\frac{||x-y||^2}{\sigma^2}\right)$$

۲.

$$K(x.y) = b \times K_1(x.y) + a \times K_2(x.y)$$

برای موارد $(a>0, b>0), (a<0, b>0), (a>0, b<0), (a<0, b<0)$ بررسی کنید.

۳.

$$\text{f}\, K(x.y) = f(x)f(y)$$ تابع حقیقی

۴.

$$K(x.y) = K_1(g(x).g(y)$$

G(x) تابع دلخواه

۵.

$$K(x.y) = K_1(x.y) \times K_2(x.y)$$

سوال سوم (۵ نمره)

Maede

چنانچه تابع Cost function برای Soft margin SVM به صورت زیر تعریف شود، نشان دهید که آیا اینگار می تواند تبدیل به یک مسئله class-separable شود؟

$$\frac{1}{2}\left|\left|w\right|\right|^2 + \frac{c}{2}\sum_{i=1}^{N}\varepsilon^2$$

سوال چهارم (۲۰ نمره)

در این سوال قرار است با مجموعه داده iris کار کنید.
این دیتاست شامل ۳ کلاس است که در هر کدام ۵۰ عضو وجود دارد که هر کلاس به یک نوع از گل زنبق اشاره دارد. هر عنصر شامل اطلاعات طول و عرض کاسبرگ و طول و عرض گلبرگ است. هم چنین داده لیبل خورده و کلاس آن نیز مشخص است. برای توضیحات بیشتر می‌تونید به سایت زیر[1] مراجعه کنید.

۱.
در ابتدا با استفاده از کرنل خطی داده ها را جدا کرده و ماتریس همبستگی و مقادیر precision, recall, f1-score را بیان کنید.
در حالت کلی کرنل های خطی زمانی استفاده می شوند که داده ها به طور خطی جدایی پذیر باشند. معمولا زمانی که تعداد زیادی از ویژگی در دیتاست موجود باشد، از این روش استفاده میشود زیرا سرعت مارا بسیار افزایش می دهد. از طرفی پارامتر های کمتری برای آموزش وجود دارد.
برای انجام این کار روی دیتاست iris ابتدا دو ویژگی اول همه ی داده هارا برداشته و سپس با استفاده از کرنل خطی دسته بندی را انجام دهید. در انتها روی نموداری داده ها با لیبل کلاس های متناظر و همچنین خطوط جدا کننده ی هر کلاس را نمایش دهید.

۲.
روش SVM را برای طبقه بندی دو ویژگی Petal Length و Petal Width بررسی کنید.

۳.
در مورد کرنل های rbf, linear, polynomial تحقیق کنید و بیان کتید هر کدام برای طبقه بندی کدام مجموعه داده مناسب می باشد. طبقه بندی را برای موارد قسمت قبل اعمال کنید و نتایج را تحلیل کنید.

۴.
در مورد هایپر پارامتر های Regularization, Gamma تحقیق کنید. هرکدام از هایپر پارامتر هارا 3 مرتبه تغییر دهید و طبقه بندی را برای هرکدام تکرار کنید. تاثیر هرکدام از هایپر پارامتر ها را بر روی طبقه بندی تحلیل کنید.

۵.
بهترین پارامترهای قسمت قبل را با کمک grid search محاسبه کنید و برای کرنل هایی که در سوال سوم بررسی نموده اید طبقه بندی را مجددا تکرار کرده و نتایج را تحلیل کنید.

۶.
در این بخش می خواهیم مسئله چند کلاسه را طبقه بندیم کنیم. برای این کار از رویکرد های one-vs-rest و one- vs-one و ۳ کرنل rbf, linear, polynomial(d=3) استفاده کنید و طبقه بندی را انجام بدهید.

[1] https://archive.ics.uci.edu/ml/datasets/iris

سوال پنجم (۱۵ نمره)

۱.
کرنل $k(0,0)$ بیانگر نگاشت $\theta(0)$ از مجموعه داده $x \in R^d$ به فضای θ است، به صورتی که در فضای مفصد داریم: $K(x.y) = \theta(x) \times \theta(y)$. در حل مسئله از کرنلRBF به صورت $K(x.y) = \exp\left(-\frac{1}{2}\left|\left|x-y\right|\right|^2\right)$ استفاده شود رابطه‌ی زیر را اثبات کنید.

۲.
ابعاد مجموعه داده $x, y \in R^n$ را در نظر بگیرید. اثبات کنید که فضای جدید ایجاد شده توسط تابع نگاشت زیر دارای بعد $\frac{1}{2} * (n+1)(n+2)$ است.

$$K(x,y) = (x^T y + 1)^2$$

۳.
یک Hard linear SVM برای یک مسئله طبقه بندی دو دسته ای و در فضای دو بعدی با n داده، آموزش داده شده است. نتیجه این طبقه بندی به چند SV نیاز دارد؟ در صورتی که یک داده با برچسب دلخواه به مجموعه داده های قبلی اضافه کنیم و مجددا طبقه بندی را انجام بدهیم، حداکثر به چند SV نیاز می باشد؟ (با دلیل)

سوال ششم (۱۵ نمره)

فرض کنید در یک مساله طبقه‌بندی دو کلاسه یک مدل ensemble دارید که از N طبقه‌بند ضعیف ساخته شده و Majority Vote انجام میدهد. به این صورت که کلاسی انتخاب میشود که حداقل $\frac{N+1}{2}$ طبقه‌بندها به آن رای دهند. با فرض این که دقت هر کدام از طبقه‌بند ها 51٪ باشد و خطای آنها از هم مستقل باشد، برای هر کدام از حالات زیر، دقت مدل ensemble را بدست آورید. (راهنمایی: میتوانید مساله را به فرم پرتاب N سکه ناصاف با احتمال شیر و خط 51٪ و 49٪ مدل کنید) (15)

۱.

$$N = 5$$

۲.

$$N = 9$$

۳.

هنگامی که $N \to \infty$ میشود دقت چقدر میشود؟ آیا در واقعیت با زیاد کردن تعداد طبقه‌بندها میتوانیم به این دقت برسیم؟ چرا؟

۴.

حالت $N = 5$ را دوباره برای زمانی که دقت طبقه‌بندها 50٪ باشد تکرار کنید. چه نتیجه‌ای میگیرید؟

سوال هفتم (۲۰ نمره)

۱.
در این سوال میخواهیم به کمک مجموعه داده میزان درآمد افراد بر اساس موقعیت شغلی آنها، یک مسئله Support Vector Regression را حل کنیم. این مجموعه داده شامل سه ستون است که در ستون اول موقعیا شغلی آنها شرح داده شده، در ستون دوم رتبه کاری و در ستون سوم میزان در آمد آورده شده است. در این سوال به کمک سه کرنل rbf, linear, polynomial میزان درآمد را پیش بینی کنید و در خروجی مقادیر تخمین زده شده و واقعی را در یک نمودار نمایش دهید.

۲.
در این سوال میخواهیم مسئله Support Vector Regression را حل کنیم که هدف آن پیش بینی هزینه اقامت در هتل به ازای ویژگی های مختلف می باشد که مجموعه دادگان در فایل Hotel.zip قرار داده شده است. از داده های فایل H1.csv به عنوان آموزش و داده های فایل H2.csv به عنوان داده های تست استفاده کنید. در این سوال از تمامی ویژگی های numerical و categorical استفاده کرده و هزینه اقامت در ختل را به ازای تمامی این ویژگی ها پیش بینی کنید. به ازای داده های تست مقدار هزینه اقامت در هتل را پیش بینی نمایید و در یک فایل CSV ذخیره نمایید که شامل سه ستون مقدار واقعی داده، مقدار پیش بینی شده و اختلاف مقدار واقعی و تخمین زده شده است. این فایل را به همراه دیگر فایل ها در پوشه تمرین قرار دهید.
در ابتدا ویژگی های categoricalو numerical را از هم تشخیص داده و برای هرکدام نرمال سازی های لازم را انجام دهید. توجه داشته باشید در این سوال هدغ پیش بینی ویژگی Average Daily Rates هست.