



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 حتماً 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** حتماً 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؟

$$\begin{aligned}\varphi_1 &= [x, x^2]^T \\ \varphi_2 &= [2x, 2x^2]^T\end{aligned}$$

سوال دوم (۱۵ نمره)

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.

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

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)

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؟

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

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 می‌توانید به سایت زیر^۱ مراجعه کنید.

۱.

در ابتدا با استفاده از کرنل خطی داده‌ها را جدا کرده و ماتریس همبستگی و مقادیر $f1$ -precision, recall, score را بیان کنید.

در حالت کلی کرنل‌های خطی زمانی استفاده می‌شوند که داده‌ها به طور خطی جدایی پذیر باشند. معمولاً زمانی که تعداد زیادی از ویژگی در دیتاست موجود باشد، از این روش استفاده میشود زیرا سرعت ما را بسیار افزایش می‌دهد. از طرفی پارامترهای کمتری برای آموزش وجود دارد. برای انجام این کار روی دیتاست iris ابتدا دو ویژگی اول همه‌ی داده‌ها را برداشته و سپس با استفاده از کرنل خطی دسته‌بندی را انجام دهید. در انتها روی نموداری داده‌ها با لیبل کلاس‌های متناظر و همچنین خطوط جداکننده‌ی هر کلاس را نمایش دهید.

۲.

روش SVM را برای طبقه‌بندی دو ویژگی Petal Length و Petal Width بررسی کنید.

۳.

در مورد کرنل‌های rbf, linear, polynomial تحقیق کنید و بیان کنید هر کدام برای طبقه‌بندی کدام مجموعه داده مناسب می‌باشد. طبقه‌بندی را برای موارد قسمت قبل اعمال کنید و نتایج را تحلیل کنید.

۴.

در مورد هایپر پارامترهای Regularization, Gamma تحقیق کنید. هر کدام از هایپر پارامترها را ۳ مرتبه تغییر دهید و طبقه‌بندی را برای هر کدام تکرار کنید. تاثیر هر کدام از هایپر پارامترها را بر روی طبقه‌بندی تحلیل کنید.

۵.

بهترین پارامترهای قسمت قبل را با کمک grid search محاسبه کنید و برای کرنل‌هایی که در سوال سوم بررسی نموده‌اید طبقه‌بندی را مجدداً تکرار کرده و نتایج را تحلیل کنید.

۶.

در این بخش می‌خواهیم مسئله چند کلاسه را طبقه‌بندی 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 طبقه‌بندی 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 مفصل داریم: $K(x,y) = \theta(x) \times \theta(y)$. در حل مسئله از کرنل RBF به صورت $K(x,y) = \exp\left(-\frac{1}{2}\|x-y\|^2\right)$ 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 برپسب 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 طبقه‌بندها 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 $N \rightarrow \infty$ 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 موقعیت شغلی آنها شرح داده شده، در ستون دوم رتبه کاری و در ستون سوم میزان درآمد 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 داده ها 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 هتل 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 numerical و categorical را از هم تشخیص داده و برای هر کدام نرمال سازی های لازم را انجام دهید. توجه داشته باشید در این سوال هدف پیش بینی ویژگی Average Daily Rates هست.