به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری چهارم یادگیری ماشین

دانشگاه تهران

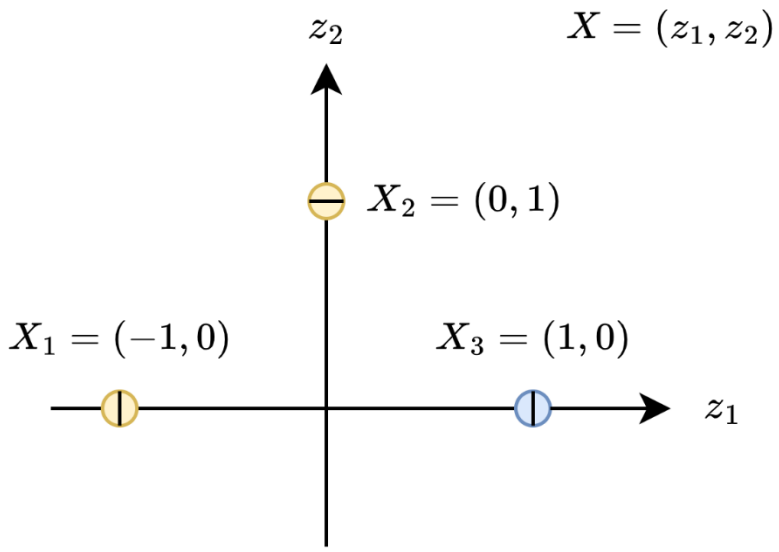
پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، 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 می‌باشند.
4. برای سوالات شبیه سازی، 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 **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**10 نمره امتیازی**) می توانید کسب کنید.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل yara.mohamadi@gmail.com، سوال خود را مطرح کنید.

سوال 1: (15 نمره)

داده‌های شکل زیر را که به دو کلاس تقسیم شده‌اند در نظر بگیرید. با نوشتن روابط موجود در مساله dual problem، به صورت دستی مقادیر آلفاها را بدست آورده، SV ها را مشخص کرده، و معادله جداساز خطی را نیز بدست آورید.



شکل 1. نمودار داده‌های مساله

سوال 2: (15 نمره)

کرنل: توابع کرنل به صورت غیر مستقیم یک تابع مپینگ $\phi(.)$ بدست میآورند که نمونه ورودی $x \in \mathbb{R}^d$ را به یک فضای ابعاد بالای Q میبرد. این کار با استفاده از یک ضرب داخلی انجام میشود به صورت:

$$Q: K(x_i, x_j) \equiv < \phi(x_i), \phi(x_j) >$$

الف) ثابت کنید که کرنل یک تابع متقارن است. یعنی $K(x_i, x_j) = K(x_j, x_i)$. (در دو الی سه خط)

ب) فرض کنید که از کرنل RBF استفاده میکنیم که به فرم $K(x_i, x_j) = \exp\left(-\frac{1}{2}\|x_i - x_j\|^2\right)$ است. میدانیم که یک مپینگ ناشناخته داریم به فرم $\phi(.)$. اثبات کنید که برای هر دو نمونه ورودی x_j و x_i، مجذور فاصله اقلیدوسی مربوط به این دو نقطه در فضای Q کوچکتر از 2 است. به عبارتی ثابت کنید $\|\phi(x_i) - \phi(x_j)\|^2 \leq 2$.

SVM: با کمک یک تابع کرنل، SVM یک ابرصفحه در فضای Q میسازد که حاشیه (Margin) بین دو کلاس را بیشینه کند. میتوان در نهایت نتیجه طبقه بندی نمونه x را با توجه به مثبت و یا منفی بودن تابع زیر بدست آورد:

$$< \widehat{W}, \phi(x) > + b = \sum_{i \in SV} y_i \, \alpha_i \, K(x_i, x) + b = f(x; \alpha, b)$$

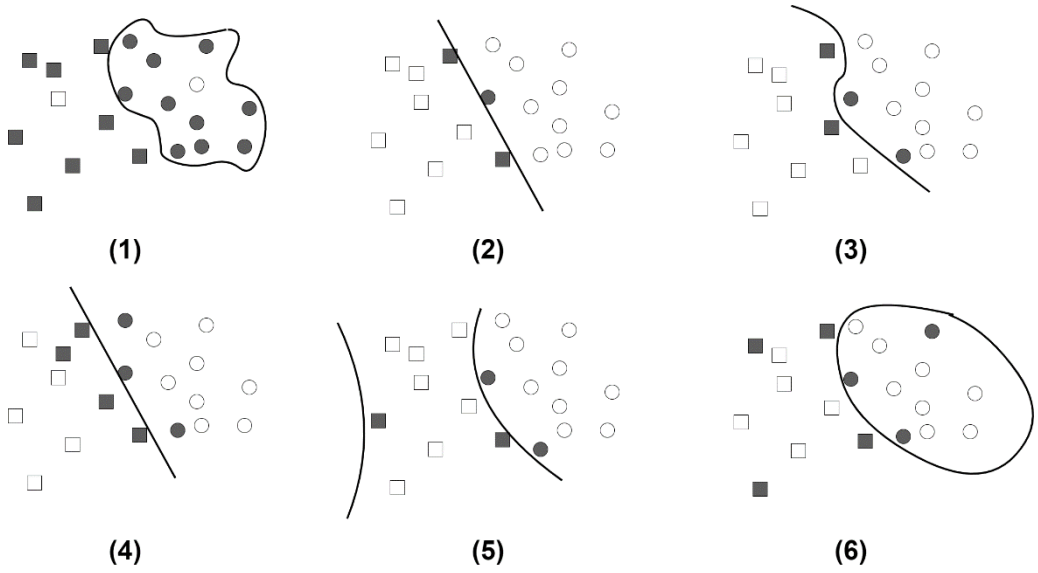
در تابع بالا b و $\widehat{W}$ پارامترهای طبقهبند در فضای Q هستند، SV مجموعه support vector ها را تشکیل میدهد، و α_i نشان‌دهنده ضرایب هرکدام از support vector هاست.

پ) فرض کنید که دوباره از کرنل RBF استفاده میکنیم. همچنین فرض کنید که داده‌های train در فضای Q به صورت خطی جداپذیر هستند و SVM یک خط با بیشترین حاشیه پیدا کرده که کاملا نقاط دو کلاس را از هم جدا کرده است. ثابت کنید که اگر یک نقطه جدید انتخاب کنیم که فاصله بسیار زیادی از تمام نقاط train یعنی x_i ها داشته باشد (اینجا منظور فاصله در همان فضای اولیه داده ها یعنی $\mathbb{R}^d$ است)، آنگاه پیشبینی مدل فقط تابعی از مقدار b خواهد بود، یعنی: $f(x_{far}; \alpha, b) \approx b$.

سوال 3: (15 نمره)

در تصویر زیر، چند نمونه SVM با مرزهای تصمیمگیری متفاوت میبینید. در این شکلها دو کلاس مربع و دایره داریم و SV ها به رنگ مشکی هستند. تعیین کنید که هر کدام از سناریوهای زیر مربوط به کدام تصویر است؟ برای هر انتخاب خود دلیل کوتاه بیاورید.

- Soft margin خطی با c=1
- Soft margin خطی با c=10
- Hard margin با کرنل $k(x_i, x_j) = x_i.x_j + (x_i.x_j)^2$
- Hard margin با کرنل $k(x_i, x_j) = exp(-10\|x_i - x_j\|^2)$
- Hard margin با کرنل $k(x_i, x_j) = exp(-\frac{1}{10}\|x_i - x_j\|^2)$



شکل 2. هر مورد، مربوط به یکی از مدلهای بالاست. یکی از موارد اضافی ست.

سوال 4: (**15 نمره**)

فرض کنید یک دیتاست یک بعدی حاوی 3 نقطه داریم به صورت زیر (y نشان‌دهنده لیبل هاست):

$$x_1 = 0, \quad y_1 = -1$$
$$x_2 = -1, \quad y_2 = +1$$
$$x_3 = +1, \quad y_3 = +1$$

همانطور که مشاهده میشود کلاسهای این نقاط به صورت خطی جداپذیر نیستند.

الف) فرض کنید که از کرنلی به فرم $\phi(x) = [1, x, x^2]$ استفاده کنیم. آیا میتوانید به صورت شهودی برداری پیدا کنید که این نقاط را در فضای جدید به صورت خطی جدا کند؟

ب) اگر داشته باشیم $W = (w_1, w_2, w_3)^T$ ، آنگاه طبقه‌بند max-margin SVM معادله مقابل را حل میکند:

$$min_{w,b} \frac{1}{2} \|W\|_2^2$$
$$s.t. \quad y_i(W^T \phi(x_i) + b) = 1 \qquad i = 1,2,3$$

با استفاده از روش ضرایب لاگرانژ نشان دهید که جواب معادله بالا برای دیتاست این سوال برابر $\widehat{W} = (0,0,2)^T$ و $b = -1$ است. اندازه حاشیه (Margin) را نیز بدست آورید.

سوال 5: (15 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)

الف) $N = 5$

ب) $N = 9$

پ) هنگامی که $N \to \infty$ میشود دقت چقدر میشود؟ (فقط جواب نهایی را بگویید، نیاز به محاسبات نیست) آیا در واقعیت با زیاد کردن تعداد طبقه‌بندها میتوانیم به این دقت برسیم؟ چرا؟

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

سوال 6: پیاده‌سازی (**20 نمره**)

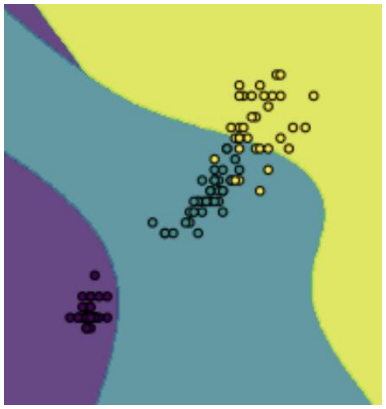
در این سوال محدودیتی برای استفاده از پکیجهای مورد نیازتان ندارید.

الف) ابتدا مختصرا توضیحی در مورد هر کدام از کرنلهای موجود در کلاس svm.SVC کتابخانه SKlearn دهید و بگویید هر کدام برای چه نوع داده یا مساله‌ای بهتر عمل میکند.

ب) روی دیتاست iris، با استفاده از الگوریتم SVM برای هر کدام از روشهای زیر، ناحیه کلاسهای مختلف نسبت به ویژگی‌های Petal Length و Petal Width را رسم کنید.

- SVM with Linear Kernel, one-vs-rest
- SVM with Linear Kernel, one-vs-one
- SVM with RBF Kernel, one-vs-rest
- SVM with Polynomial Kernel (d=5), one-vs-rest

پ) برای هر کدام، دقت روی دیتاست ترین و تست و ماتریس آشفتگی روی دیتاست تست را نیز نشان دهید و مدلها را با هم مقایسه کنید.



شکل 3. نمونه‌ای از تصویری که باید گزارش بدهید

سوال 7: پیاده‌سازی (15 نمره)

در این سوال محدودیتی برای استفاده از پکیجهای مورد نیازتان ندارید.

در این سوال میخواهیم با SVM رگرسیون انجام دهیم. ابتدا با استفاده از قطعه کد زیر داده های مساله را بسازید:

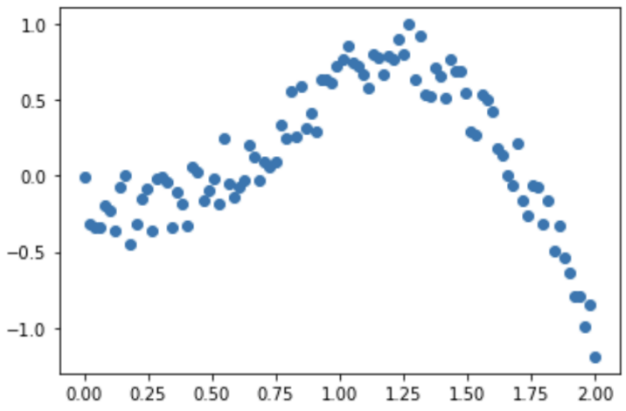
```
import random
import numpy as np

x = np.linspace(0, 2,100)
er = np.random.random_sample(size = 100)/2 - 0.5

y= np.sin(x**2) + er
```

شکل 4. داده‌ها را بر این اساس بسازید.

داده‌های ساخته شده باید به شکل زیر باشند:



شکل 5. داده‌ها باید این شکلی باشند.

الف) با استفاده از cross-validation، تابع بالا را توسط کرنل‌های Linear، Polynomial (d=3)، و RBF تخمین بزنید. بهترین مقادیر gamma و C را برای هر کرنل پیدا کنید

ب) خطای تخمین بهترین مدل از هر کدام از این کرنل‌ها را با هم مقایسه کنید. نمودار تخمین‌زده شده توسط هر کرنل را به همراه داده‌های ورودی رسم کنید.

پیروز باشید.