## بخش اول: پرسش‌های تشریحی

سوال ۱) صحت هر یک از موارد زیر را بررسی کرده و دلیل خود را توضیح دهید.

الف) ماشین‌های بردار پشتیبان پارامتریک هستند.
ب) ماشین‌های بردار پشتیبان در برابر بیش برازش مقاوم هستند.
ج) وجود داده‌های پرت و نویزی بر روی ماشین‌های بردار پشتیبان بی‌تاثیر است.
د) الگوریتم آدابوست با استفاده از هر نوع دسته‌بند ضعیف و یا ترکیب چند دسته‌بند ضعیف در نهایت به خطای آموزش صفر می‌رسد.
ه) وزن‌های اختصاص داده‌شده به دسته‌بندها در الگوریتم آدابوست همواره نامنفی هستند.
و) الگوریتم آدابوست در برابر داده نویز حساس است.

سوال ۲) مجموعه داده آموزشی زیر را در نظر بگیرید:

| X1 | X2 | class |
|---|---|---|
| 1 | 1 | + |
| 2 | 2 | + |
| 2 | 0 | + |
| 0 | 0 | - |
| 0 | 1 | - |
| 1 | 0 | - |

الف) داده‌ها را در فضای دوبعدی رسم کنید. آیا کلاس‌های {+و-} جداپذیر خطی هستند؟

ب) معادله خط جداکننده را به دست آورید.

ج) اگر یکی از بردارهای پشتیبان را حذف کنیم، آیا اندازه حاشیه تغییر می‌کند؟

* د) آیا گزاره زیر صحیح است؟ در صورت غلط بودن یک مثال نقض بیاورید و در صورتی که صحیح است یک اثبات کوتاه ارائه دهید.

" در یک مدل SVM با حذف یکی از بردارهای پشتیبان، اندازه حاشیه تغییری نمی‌کند یا افزایش می‌یابد"

سوال ۳) مجموعه داده زیر در فضای یک بعدی را در نظر بگیرید:

| x | class |
|---|---|
| 0 | + |
| -1 | - |
| +1 | - |

الف) آیا کلاس‌های {+و-} جداپذیر خطی هستند؟

ب) فرض کنید هر نقطه را با کرنل $\phi(x) = [1. \sqrt{2}\, x. x^2]^T$ به فضای ۳ بعدی انتقال بدهیم. آیا کلاس‌های مثبت و منفی حالا جداپذیر خطی هستند؟ اگر پاسخ شما مثبت است، یک ابرصفحه که کلاس‌ها را از هم جدا می‌کند بیابید.

سوال ۴) شکل ۱ مرزهای تصمیم حاصل از SVM با استفاده از کرنل‌ها و خطاهای slack متفاوت را نشان می‌دهد. دو کلاس از داده‌ها داریم که با مربع و دایره مشخص شده‌اند و مربع‌ها و دایره‌های خاکستری، بردارهای پشتیبان را نشان می‌دهند. با ذکر دلیل مشخص کنید که هر یک از مواردی که در ادامه آورده شده است، مربوط به کدام شکل است. ( دقت کنید که یکی از شکل ها اضافی است)

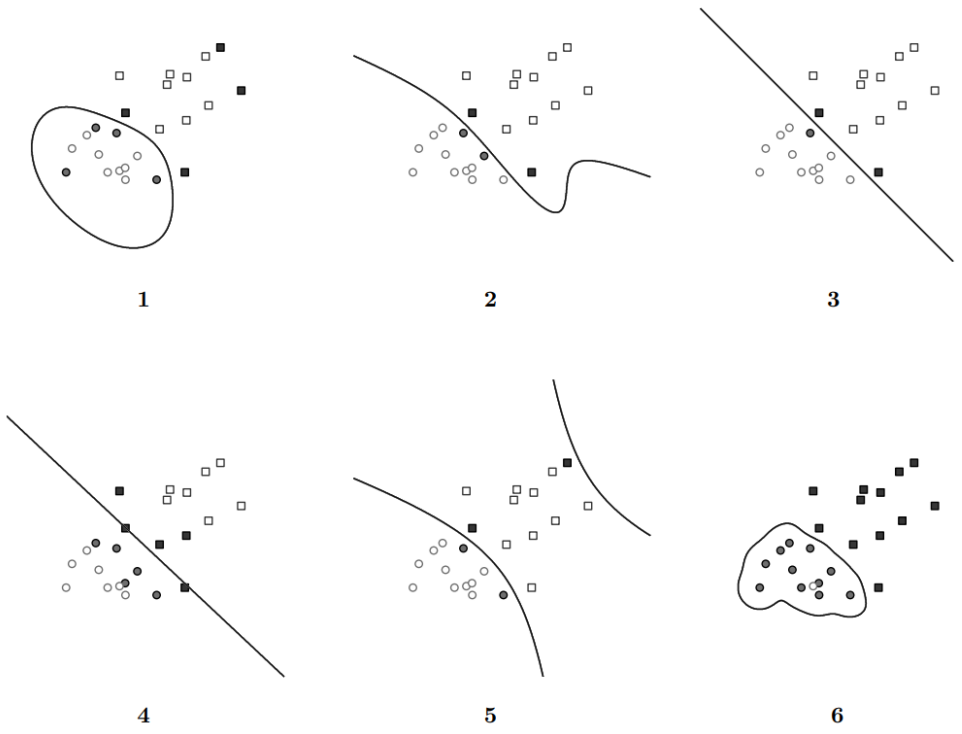
الف) SVM خطی ، soft-margin ، C=0.1

ب) SVM خطی ، soft-margin ، C=10

ج) hard-margin SVM با کرنل: $K(u.v) = u \cdot v + (u \cdot v)^2$

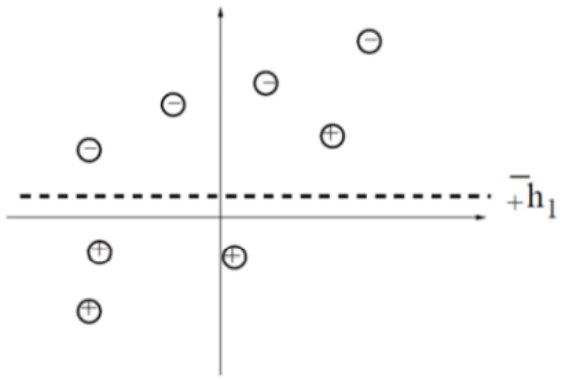
د) hard-margin SVM با کرنل: $K(u.v) = \exp(-\frac{1}{4}\left||u - v|\right|^2)$

ه) hard-margin SVM با کرنل: $K(u.v) = \exp(-4\left||u - v|\right|^2)$




شکل ۱

سوال ۵) در شکل۲ الگوریتم Adaboost را اجرا کرده‌ایم و در اولین تکرار الگوریتم، مرز تصمیم $h_1$ حاصل شده است. وزن $\alpha_1$ که توسط الگوریتم Adaboost به $h_1$ اختصاص داده می‌شود و هم چنین مقدار $e_1$ را بدست آورید. ( وزن اولیه هر داده برابر $\frac{1}{8}$ می‌باشد)




شکل ۲

## بخش دوم: پیاده‌سازی

**سوال اول:** مصورسازی کرنل‌های SVM (استفاده از کتابخانه آماده مجاز است)

در این سوال قصد داریم که مرز تصمیم SVM با کرنل‌های مختلف را ترسیم کنیم. مجموعه داده این سوال دیتاست iris می‌باشد که از لینک زیر قابل دریافت است:

https://archive.ics.uci.edu/ml/datasets/iris

برای اینکه مصورسازی قابل انجام باشد، مجموعه داده دوبعدی نیاز داریم به همین دلیل برای پیش‌بینی کلاس یک گل، به جای استفاده از ۴ ویژگی، اطلاعات کاسبرگ[1] و گلبرگ[2] را جداگانه استفاده می‌کنیم.

الف) در دو نمودار جداگانه که یکی برحسب طول و عرض کاسبرگ و یکی برحسب طول و عرض گلبرگ است کلاس‌های داده‌ها را (با رنگ‌های مجزا) نمایش دهید.

ب) برای هر یک از ویژگی‌های کاسبرگ و گلبرگ، داده‌ها را با استفاده از روش SVM و با کرنل‌های زیر دسته‌بندی کنید. سپس مرز تصمیم را در نمودار رسم کنید.

- کرنل خطی
  برای پارامتر C ، این ۵ مقدار را در نظر بگیرید و مرز تصمیم را رسم کنید: [0.01,0.1,1,10,100]

- کرنل چندجمله ای (پارامترC و درجه چندجمله‌ای)
  برای هر یک از پارامترها حداقل ۳ مقدار مختلف را بررسی کنید.

- کرنلRBF (پارامترهای گاما و C)
  برای هر یک از پارامترها حداقل ۳ مقدار مختلف را بررسی کنید.

- کرنل sigmoid (پارامترهای گاما و C)
  برای هر یک از پارامترها حداقل ۳ مقدار مختلف را بررسی کنید.

ج) معیار accuracy و F1-measure را برای هر یک از حالت‌های بالا به دست آورید و گزارش کنید.

د) تاثیر پارامتر(های) هر کرنل بر عملکرد مدل را تحلیل کنید.

**سوال دوم:** بررسی عملکرد مدل های ترکیبی

داده های مورد استفاده در این تمرین مربوط به تصاویر اعداد دست نوشته انگلیسی است که در فولدر dataset قرار دارد. از هر تصویر، ۱۶ ویژگی استخراج شده که این ویژگی ها به همراه برچسب تصویر متناظر، در اختیار شما قرار داده شده است. هدف، طراحی دسته بند مناسب برای این داده ها است. لازم به ذکر است که داده های یادگیری و تست به صورت مجزا و به ترتیب در فایل data_train.csv و data_test.csv قرار دارد.

[1] sepal
[2] petal

الف) با استفاده از روش Random Forest یک دسته بند مناسب طراحی کنید. دسته بندی را با استفاده از ترکیب نتیجه‌ی ۱۵ درخت تصمیم انجام دهید. هر درخت تصمیم را نیز با استفاده از ۳ ویژگی که به صورت تصادفی انتخاب می شود، آموزش دهید. برای آموزش درخت‌های تصمیم می توانید از کتابخانه scikit-learn استفاده کنید. بیشینه‌ی عمق درخت ها را نیز ۳ در نظر بگیرید. دقت دسته بندی و همچنین ماتریس درهم ریختگی را برای داده های تست در گزارش خود ارائه کنید.

ب) با استفاده از روش AdaBoost یک دسته بند مناسب طراحی کنید. دسته بندی را با استفاده از ترکیب نتیجه‌ی ۱۰ درخت تصمیم انجام دهید. برای آموزش درخت‌های تصمیم همچون قبل می توانید از کتابخانه‌یscikit-learn استفاده کنید. دقت دسته بندی را برای داده های تست در گزارش خود ارائه کنید.

ج) با استفاده از روشAdaBoost قسمت قبل را با ۵، ۲۰ و۵۰ دسته بند درخت تصمیم تکرار کرده و برای هر کدام دقت دسته بندی را برای داده های تست در گزارش خود ارائه کنید.

د) در چند سال اخیر روش های Gradient Boosting بسیار مورد توجه قرار گرفته اند. یکی از کتابخانه هایی که برای این روش ها منتشر شده است، کتابخانه‌ی XGBoost است. با بررسی این کتابخانه، یک دسته بند مناسب برای داده های این تمرین طراحی کنید. انتخاب پارامترهای بهینه در این قسمت به عهده شما است. پس از پایان یادگیری، دقت دسته بندی را بر روی داده های تست برای دسته بند پیشنهادی خود در گزارش ارائه کنید.

پایدار باشید

## نکات مربوط به تحویل تمرین

- کدهای خود را ترجیحا به زبان پایتون و در محیط jupyter پیاده‌سازی کنید.
- نظم و خوانایی در نوشتن گزارش و کدها از اهمیت بالایی برخوردار است. کدهای خود را تا حد امکان کامنت‌گذاری کنید.
- در پیاده‌سازی بخش‌های مختلف امکان استفاده از کتابخانه‌های آماده مربوط به الگوریتم‌های یادگیری ماشین رابه طور کلی ندارید. مگر در مواردی که در صورت سوال ذکر شده باشد.
- برای خواندن داده‌ها می‌توانید از کتابخانه pandas و برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. برای محاسبه معیارهای ارزیابی مانند دقت، ماتریس درهم‌ریختگی و تقسیم داده‌ها به مجموعه‌های آموزش و آزمون نیز استفاده از کتابخانه آماده مجاز است.
- در صورتی که داده‌ها را به دسته‌های آموزشی، تست (ویا validation) تقسیم می‌کنید، درصد هر کدام را در گزارش ذکر کنید.
- فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت HW01_StdNumber.zip که StdNumber شماره دانشجویی شماست، در سامانه بارگذاری کنید.
- سوالات ستاره دار(*) دارای نمره اضافی بر تمرین است.
- مطابق قوانین دانشگاه هر گونه کپی‌برداری ممنوع است و در صورت مشاهده، نمره **هر دو طرف صفر** داده می‌شود.
- در صورت وجود هر گونه سوال یا ابهامی با ایمیل درس در تماس باشید:

ml.ce.aut@gmail.com