بسمه تعالی

تمرین سوم

ماشین بردار پشتیبان[1]

استاد درس: دکتر آرش عبدی هجراندوست

تدریسیار آموزشی: فربد داودی

✓ نکات تمرین :

- ❖ مهلت تحویل 1401/01/18 ساعت 23:55
- ❖ مهلت ارسال به هیچ وجه قابل تغییر نیست .
- ❖ مواردی که بعد از تاریخ فوق ارسال شوند قابل قبول نبوده و نمره ای نخواهند داشت .

[1] Support Vector Machine (SVM)

- انجام تمرین تک نفره است. لطفا به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر گرفته خواهد شد .
- کل محتوای ارسالی زیپ شود و نام فایل زیپ ارسالی HW3_StudentNumber_YourName باشد .
- محتوای ارسالی دارای راهنما (read me) جهت تسهیل اجرا باشد .
- زبان برنامه نویسی دلخواه است. (پیشنهاد : پایتون)
- در صورت استفاده از زبان پایتون فایل کد ترجیحا به فرمت ipynb. بوده و **فایل کد حتما به صورت اجرا شده آپلود گردد و از وجود خروجی سلول ها اطمینان حاصل نمایید** .
- موارد ارسال شده در تاریخی که بعدا مشخص خواهد شد و متعاقبا اعلام می‌گردد به صورت آنلاین نیز تحویل گرفته خواهند شد (صرفا آنچه در LMS طبق تاریخ فوق تحویل داده شده است بعدا به صورت حضوری تست شده و توضیح داده می شود).
- تنها تکالیفی که به LMS و قبل از مهلت ارسال، فرستاده می شوند بررسی خواهند شد .
- در صورت داشتن هر گونه سوال می‌توانید سوال خود را در گروه تلگرامی درس مطرح کنید .
- حداقل **یک ساعت قبل از مهلت ارسال** را احتیاطا هدف قرار دهید، تا مشکلات غیرقابل پیش بینی مانند موارد زیر باعث عدم آپلود پاسخ ها در LMS و ارسال آنها از طریق ایمیل نشوند :

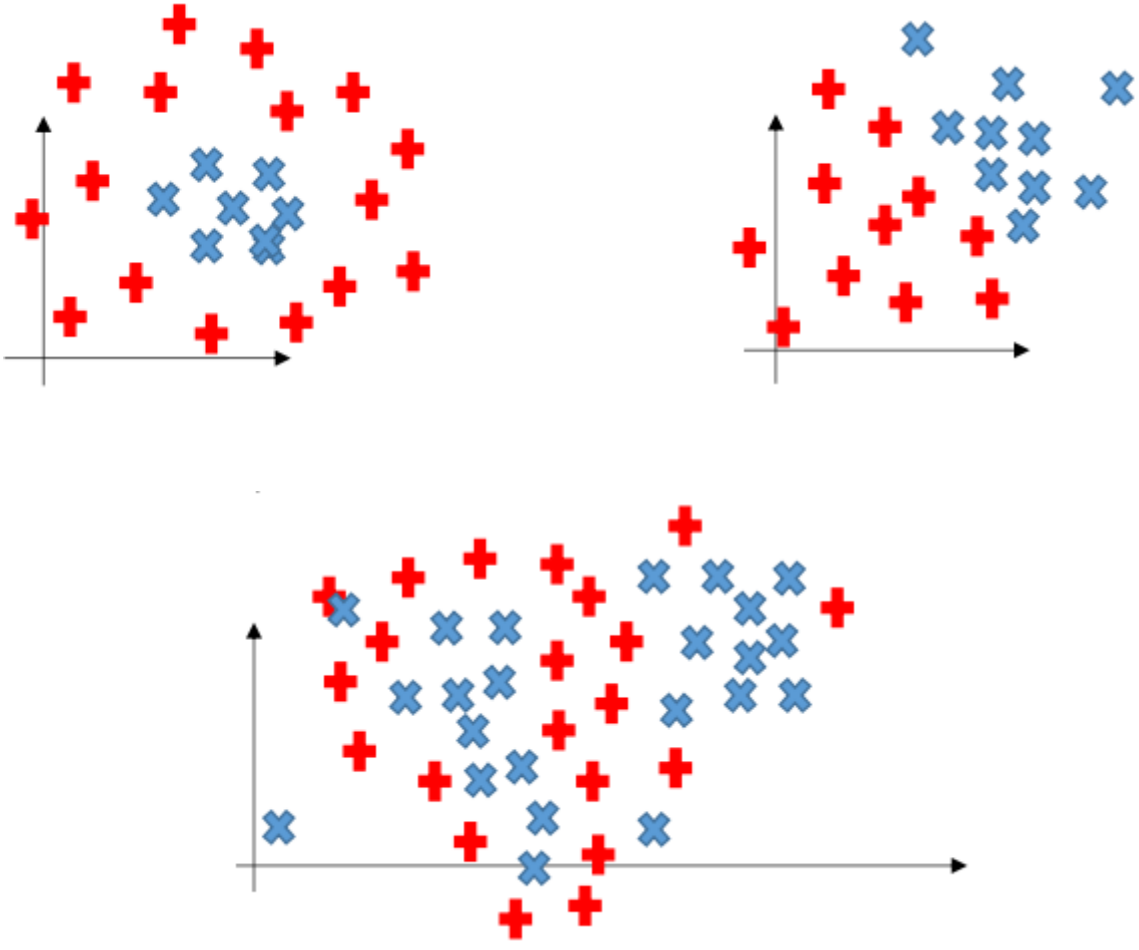
(قطعی اینترنت، تنظیم نبودن دقیق ساعت سایت با ساعت گرینویچ، کرش سیستم عامل و نیاز به فرمت، بارش زیبای شهاب سنگ از آسمان و ...)

شرح تمرین:

بخش اول:

ابتدا چند مساله دو کلاسه از خیلی ساده تا خیلی پیچیده طراحی کنید. به عنوان مثال، تعدادی نقطه (500 تا 5000 نقطه) در فضای 2 بعدی برای دو کلاس فرض کرده و نقاط را در نمودار دو بعدی رسم کنید.

مثال شکل های زیر می توانند چند مثال پیشنهادی باشند:



یکبار با استفاده و کمک از توابعی نقاطی را تولید کنید که هر بار بتوانید نقاطی و دسته هایی با ویژگی ها و شکلهای متمایز ایجاد و آزمایش کنید و یکبار نقاط را به صورت دستی تولید و آزمایش کنید.

سپس با ماشین بردار پشتیبان، اقدام به دسته بندی داده های ساختگی خود کنید. بدون هسته، با انواع هسته ها، پارامترهای مختلف برای هر هسته و ... را تست کنید و **برای هر کدام از مجموعه داده‌ها بهترین پارامترها را پیدا کنید.**

خط جداکننده‌ای که ماشین بردار پشتیبان یافته است را در کنار نقاط آموزشی به همراه خط **Margin** در یک نمودار رسم کنید برای این کار میتوانید از کتابخانه‌های آماده بهره بگیرید.

سعی کنید یک تابع هسته را با ترکیب توابع هسته ساده تر مشابه آنچه در کلاس درس آموختید، خودتان پیاده سازی کنید و یا همچنین با ایده ای خلاقانه تابع هسته‌ای را خودتان طراحی و پیاده کنید و رده بند را با آن تابع هسته بر داده ها اعمال کنید و نتایج را با هم مقایسه کنید. سعی کنید تابع هسته ای که طراحی کردید نتیجه بهتری از بقیه توابع هسته داشته باشد.

-رفته‌رفته پیچیدگی داده‌ها را بیشتر کرده و عملیات بالا را روی داده‌ها تکرار کنید. پیچیدگی داده‌ها چه تاثیری در انتخاب هسته و پارامترهای مربوطه خواهند داشت؟ به طور کامل شرح دهید.

-آیا می‌توان به طور قطع در مورد نوع هسته خاصی اظهار نظر کرد و مطمئن بود همواره بهترین خواهد بود؟ به طور کامل شرح دهید.

- در مورد پارامتر تنظیم[2] روی داده‌ها توضیح دهید و نتایج مختلف را به سبب تغییر میزان آن به طور منحصر به فرد به غیر از مابقی پارامتر ها به طور کامل شرح دهید.

بخش دوم:

یک پایگاه داده استاندارد برای رده بندی انتخاب کنید. می توانید از سایت Mendeley یا این لینک مجموعه داده ای را به دلخواه انتخاب کنید و آنرا نرمال سازی کنید تا در بخش های بعدی از آن استفاده کنید. توصیه می شود مجموعه داده کوچکی با کمترین سایز و حجم انتخاب شود تا فرایند آموزش زمان‌بر نباشد. توضیح مختصری درباره پایگاه داده انتخاب شده (تعداد داده و نمونه، تعداد ویژگی، چالشهای آن، دقت تقریبی روشهای Baseline روی آن و ...) در گزارش خود بیاورید.

لازم به ذکر است که مثل روال سابق می‌بایست مجموعه‌ای از داده‌ها را برای آموزش و مجموعه‌ای را برای آزمون در نظر بگیرید.

در این بخش نیز با تمامی هسته‌ها و تغییر انواع پارامترها (C و gamma) عملیات رده‌بندی را انجام داده **و سپس نیز بهترین پارامترها و هسته را پیدا کنید**.

در مورد متغیر های slack به سوالات زیر پاسخ دهید:

اگر مقدار این متغیر بین 0 و 1 باشد به چه معناست؟ اگر بزرگ تر از 1 باشد چطور؟

کم یا زیاد بودن تعداد متغیر های slack چه ارتباطی با میزان بایاس و واریانس و حاشیه خط جدا کننده و تعداد بردار های پشتیبان دارد؟ در مدل خود مقادیر مختلفی برای پارامتر C در نظر بگیرید و مقادیر بایاس و واریانس را در هر حالت چاپ کنید و نتایج را در گزارش خود تحلیل و برسی کنید.

[2] Regularization Term

پس از یافتن بهترین هسته و بهترین پارامتر ها، آموزش را با این هسته و پارامتر ها در حالت های زیر انجام و **نتایج را با هم مقایسه و تحلیل کنید:**

کاهش ویژگی با استفاده از روش PCA به 30 ویژگی

کاهش ویژگی با استفاده از روش PCA به 100 ویژگی

روش ها و تکنیک های کاهش ابعاد را در اینترنت جست و جو کرده و مطالعه کنید. سپس یک روش را به دلخواه انتخاب کرده (بجز PCA و T-SNE) و با استفاده از آن کاهش ویژگی را انجام داده و سپس نتایج را با هم مقایسه و تحلیل کنید.

پس از انجام آزمایش های فوق، به سوالات زیر پاسخ دهید:

کاهش ویژگی با استفاده از روش PCA و کم یا زیاد بودن تعداد ویژگی ها در این روش چه تاثیری در عملکرد مدل می گذارد؟

روش PCA و روش انتخاب شده را با هم مقایسه کنید و مزایا و معایب هر یک را نسبت به دیگری در گزارش خود شرح دهید.

نمودار ROC و مقادیر Test accuracy و Precision و Recall و AUC و معیار f1 و ماتریس درهم ریختگی (Confusion Matrix) و میزان خطا و بایاس و واریانس را برای مدلی که بهترین هسته و بهترین پارامتر ها را دارد در گزارش خود نشان دهید.

با استفاده از تکنیک cross validation با مقادیر k=3,5,7,10 برای مدل با هسته و پارامترهایی که در مرحله قبل بهترین عملکرد را داشتند رده بندی را انجام داده و نتایج را تحلیل کنید.

پس از انجام آزمایش های فوق، در گزارش خود به سوالات زیر به طور کامل پاسخ دهید:

-آیا اگر مجموعه دادگان دارای عدم تعادل باشد(یعنی از هر کلاس به تعداد تقریبا یکسان داده موجود نباشد، برای یک کلاس تعدادی خیلی کم و برای دیگری تعدادی بسیار و یا تا حدی بیشتر)، آیا رده‌بندی ماشین بردار پشتیبان دچار مشکل می‌شود و رده‌بندی توسط آن تحت الشعاع قرار می‌گیرد؟چرا؟ به طور کامل توضیح دهید.

- راه حل مشکل بالا یعنی عدم تعادل مجموعه دادگان چیست و تاثیرات کلی آن را روی عملیات رده‌بندی توسط هر رده‌بندی را توضیح دهید.

بخش سوم:

در این بخش از SVM برای تقریب تابع استفاده می کنیم. ابتدا مانند بخش اول تعدادی نقطه (1000 تا 5000 نقطه) در فضای چند بعدی تشکیل دهید (تعداد ابعاد می تواند بین 2 تا 10 بعد باشد). سپس کار های زیر را انجام دهید:

ابتدا یک تابعی دلخواه در نظر بگیرید که ورودی حداقل 2 ویژگی داشته باشد و داده های آموزشی(نقاط تشکیل شده) را به آن تابع بدهید و سپس خروجی های تابع را که برچسب های داده های ورودی هستند به همراه ورودی ها در یک مجموعه داده ذخیره کنید.

در مرحله بعد سعی کنید SVM را با هسته ها و پارامتر های مختلف روی این مجموعه داده فیت کنید و نتایج کامل خروجی را چاپ کنید.

در مرحله بعد فرض کنید که تابع اولیه که در نظر گرفتید را ندارید و فقط یک مجموعه داده در اختیار دارید. تقریب تابع را انجام دهید. این تابعی که تقریب زدید چقدر به تابعی که در ابتدا فرض کردید شباهت دارد؟

تابعی که با SVM به دست آوردید و تابعی که تقریب زدید را رسم کنید. چقدر به یکدیگر شبیه اند؟ این کار را با هسته ها و تغییر پارامتر های مختلف SVM انجام دهید. SVM با کدام هسته و چه پارامتر هایی به تابعی که تقریب زدید نزدیک تر است؟ چرا؟

توابعی که در ابتدا فرض می کنید را از ساده به پیچیده تغییر دهید و نتایج را با هم مقایسه و تحلیل کنید.

بخش چهارم:

مجموعه داده ای که در بخش دوم برای رده بندی انتخاب کردید را در نظر بگیرید و ابتدا آن را نرمال سازی کنید و سپس داده ها را به دو بخش آموزش و آزمون تقسیم کنید و با توابع هسته ذکر شده در بخش دو تقریب تابع را انجام دهید و معیارهای دقت را به عنوان خروجی چاپ کرده و با یکدیگر مقایسه کنید. کدام تابع هسته با انواع پارامترهای داده شده بهترین نتایج خروجی را دارد (دقت،.....)؟ چرا؟

آنچه تحویل داده می‌شود:

1) کداجرایی برنامه با توضیحات لازم برای اجرا
2) **گزارش کاملی از مسیر انجام کار، چالش‌هایی که مواجه شده‌اید، اجراهایی که گرفتید و نتایجی که حاصل شده است. گزارش کار از اهمیت بالایی برخوردار است، حجم آن و فرمت استاندارد آن اهمیت ندارد، اما باید نشان دهنده مسیر انجام پروژه، چالشها، راه حلها و نتایج کار شما باشد.**