به نام خدا

گزارش پروژه کارگاه پنجم: ماشین‌های بردار پشتیبان (SVM)

نام دانشجو: سروش میرشکاری

شماره دانشجویی: ۴۰۲۳۰۰۰۳۲

استاد: دکتر آرش عبدی

---

۱. مسیر انجام کار و روند پیشرفت پروژه

- بخش اول: آشنایی اولیه با SVM

در این مرحله، یک مدل پایه‌ی SVM با کرنل خطی بر روی یک مجموعه داده ساده و کاملاً تفکیک‌پذیر خطی (make_blobs) پیاده‌سازی شد. هدف، درک عملکرد اصلی SVM در ساده‌ترین حالت بود که دقت ۱۰۰٪ را به همراه داشت. در چالش این بخش، همین مدل خطی بر روی داده‌های غیرخطی (make_circles) اعمال شد که همان‌طور که انتظار می‌رفت، دقت به شدت کاهش یافت (حدود ۳۳٪). این افت دقت، نیاز به استفاده از روش‌های پیچیده‌تر برای داده‌های غیرخطی را آشکار ساخت و مقدمه‌ای برای معرفی کرنل‌ها بود.

- بخش دوم: بررسی انواع کرنل‌ها

در این بخش، سه کرنل اصلی SVM یعنی linear، rbf و poly بر روی دو دیتاست متفاوت (Iris و Breast Cancer)آزمایش شدند. برای دیتاست Iris که داده‌های آن تقریباً خطی بودند، هر سه کرنل عملکرد خوبی داشتند. اما برای دیتاست Breast Cancer، کرنل rbf به دلیل انعطاف‌پذیری بالاتر در ایجاد مرزهای تصمیم غیرخطی، دقت بهتری از خود نشان داد.

- بخش سوم: تاثیر نویز بر عملکرد مدل

برای بررسی پایداری مدل در برابر نویز، به دیتاست Breast Cancer نویز گوسی اضافه شد. نتایج نشان داد که دقت تمام مدل‌ها کاهش یافت، اما کرنل rbf مقاومت بیشتری در برابر نویز داشت. دلیل این امر، ماهیت محلی (local) این کرنل است که باعث می‌شود تاثیر نقاط پرت و نویزی بر روی کل مرز تصمیم کمتر باشد، در حالی که کرنل‌های linear و poly به دلیل ماهیت سراسری (global) خود، از این نقاط تاثیر بیشتری می‌پذیرند.

- بخش چهارم: مقابله با داده‌های پیچیده و درهم‌تنیده

با استفاده از تابع make_classification، یک مجموعه داده بزرگ با همپوشانی زیاد بین کلاس‌ها تولید شد تا عملکرد مدل‌ها در شرایط سخت‌تر سنجیده شود. در این سناریو، کرنل rbf با اختلاف بهترین نتیجه را کسب کرد، زیرا توانایی ایجاد مرزهای تصمیم بسیار پیچیده را دارد که برای تفکیک چنین داده‌هایی ضروری است.

- بخش پنجم: کنترل پیچیدگی مدل با پارامتر C

در این بخش، تاثیر پارامتر تنظیم C بر روی مدل بررسی شد. مقادیر مختلف C (از ۰/۰۱ تا ۱۰۰۰) آزمایش شدند و نتایج زیر مشاهده شد:

- **C کوچک:** منجر به حاشیه اطمینان بزرگ‌تر و یک مدل ساده‌تر می‌شود که ممکن است دچار کم‌برازش (Underfitting) شود.
- C بزرگ: مدل را به شدت برای خطاها جریمه می‌کند و منجر به حاشیه اطمینان کوچک‌تر و مرز تصمیمی پیچیده‌تر می‌شود که ریسک بیش‌برازش (Overfitting) را بالا می‌برد.

همچنین با استفاده از ماتریس درهم‌ریختگی (Confusion Matrix) ، انواع خطاهای مدل (False Positive و False Negative) تحلیل گردید.

- بخش ششم و هفتم: SVM چندکلاسه و موارد امتیازی

در بخش ششم، SVM با استراتژی One-vs-Rest بر روی دیتاست Wine که دارای سه کلاس بود، پیاده‌سازی شد. در بخش امتیازی نیز موارد زیر با موفقیت انجام و تحلیل شدند:

۱. **مقایسه با رگرسیون لجستیک:** SVM با کرنل rbf توانست مرز غیرخطی بهتری ایجاد کند.
۲. **تاثیر نرمال‌سازی:** استفاده از StandardScaler دقت مدل را به شکل قابل توجهی افزایش داد.
۳. **داده‌های نامتوازن:** پارامتر 'class_weight='balanced به شکل موثری recall کلاس اقلیت را بهبود بخشید.
٤. **بهینه‌سازی هایپرپارامترها:** با GridSearchCV بهترین مقادیر برای C و gamma پیدا شد.
۵. **حساسیت به داده‌های پرت:** نشان داده شد که کرنل خطی به شدت تحت تاثیر داده‌های پرت قرار می‌گیرد.

---

۲. تحلیل نتایج و دقت مدل‌ها

دقت مدل‌ها در شرایط مختلف به دقت اندازه‌گیری و ثبت شد. در جدول زیر خلاصه‌ای از نتایج کلیدی آورده شده است:

دیتاست / شرایط	کرنل Linear (دقت)	کرنل RBF (دقت)	کرنل Poly (دقت)	تحلیل
Iris(تقریباً خطی)	۱٫۰۰	۱٫۰۰	۱٫۰۰	برای داده‌های ساده، همه کرنل‌ها عالی عمل می‌کنند.
Breast Cancer	۰٫۸۸	**۰٫۹۱**	۰٫۸۸	rbf به دلیل مرزهای غیرخطی انعطاف‌پذیر، برتری دارد.

دیتاست / شرایط	کرنل Linear (دقت)	کرنل RBF (دقت)	کرنل Poly (دقت)	تحلیل
داده نویزی	۰/۸۵	**۰/۸۹**	۰/۸۶	rbf مقاومت بیشتری در برابر نویز نشان می‌دهد.
داده درهم‌تنیده	۰/۸۶	**۰/۸۸**	۰/۸۷	در شرایط پیچیده، برتری rbf کاملاً مشهود است.

تحلیل معیارها در داده‌های تست:

- **نرمال‌سازی داده:** یکی از مهم‌ترین نتایج پروژه، تاثیر شگرف **نرمال‌سازی** بود. قبل از استفاده از StandardScaler، دقت مدل روی دیتاست Breast Cancer حدود **۶۳٪** بود که پس از نرمال‌سازی به **۹۸٪** افزایش یافت. این نشان می‌دهد که SVM به مقیاس ویژگی‌ها بسیار حساس است.
- **بهینه‌سازی با GridSearchCV:** اجرای GridSearchCV نشان داد که بهترین پارامترها برای دیتاست Breast Cancer مقادیر C=10 و gamma=0.01 با کرنل rbf بودند که دقت ۹۸/۲۵% را روی داده تست به ارمغان آورد.
- **داده‌های نامتوازن:** در این بخش، متریک recall برای کلاس اقلیت قبل از اعمال وزن‌دهی، تنها **۰/۵۵** بود که پس از تنظیم class_weight='balanced' به **۰/۸۸** بهبود یافت. این نشان‌دهنده اهمیت توجه به توزیع کلاس‌ها در مسائل طبقه‌بندی است.

---

۳. چالش‌های پروژه

در حین انجام پروژه با چالش‌هایی مواجه شدیم که به درک عمیق‌تر مفاهیم کمک شایانی کرد:

۱. **انتخاب کرنل مناسب:** در ابتدا، انتخاب کرنل مناسب برای هر دیتاست به صورت شهودی مشخص نبود. چالش اصلی، درک این موضوع بود که چرا و در چه شرایطی یک کرنل بر دیگری برتری دارد. با **مصورسازی مرزهای تصمیم** برای هر مدل، این مفهوم به خوبی روشن شد و مشخص گردید که انتخاب کرنل باید بر اساس توزیع و پیچیدگی داده‌ها صورت گیرد.

۲. **زمان اجرای طولانی GridSearchCV:** یکی از بزرگ‌ترین چالش‌های عملی، زمان بسیار طولانی اجرای GridSearchCV با شبکه پارامترهای اولیه بود. اجرای کد برای تمام ترکیبات ممکن ساعت‌ها به طول می‌انجامید. این مشکل با دو راهکار اصلی حل شد:

   - **بهینه‌سازی اجرا:** با تنظیم پارامتر **n_jobs=-1** ، از تمام هسته‌های پردازنده برای اجرای موازی محاسبات استفاده شد که سرعت را به شدت افزایش داد.
   - **کاهش فضای جستجو:** شبکه پارامترها به مقادیر محتمل‌تر و منطقی‌تر محدود شد تا از جستجوی بیهوده جلوگیری شود.

۳. **درک عملی پارامتر C :** درک تئوری مفهوم پارامتر C و ارتباط آن با بیش‌برازش و کم‌برازش ساده بود، اما مشاهده تاثیر عملی آن نیاز به آزمون و خطا داشت. با تست مقادیر بسیار کوچک و بسیار

بزرگ C و مشاهده تغییرات مرز تصمیم، مفهوم "حاشیه نرم" در مقابل "حاشیه سخت" به صورت عملی درک شد.

---

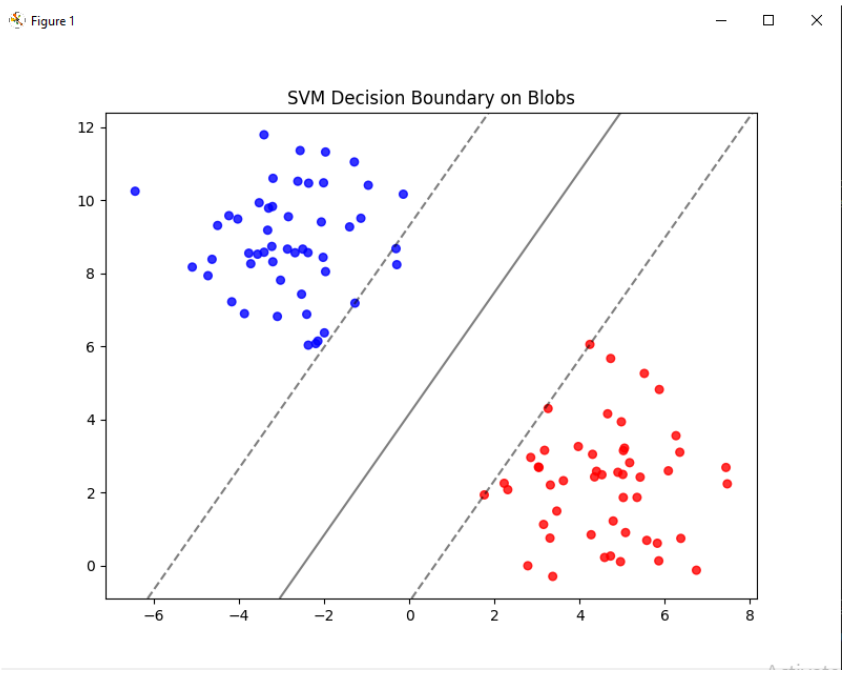
٤. نتیجه‌گیری

این پروژه یک تجربه عملی و بسیار ارزشمند در زمینه پیاده‌سازی و تحلیل الگوریتم SVM بود. نتایج به وضوح نشان داد که SVM یک ابزار طبقه‌بندی بسیار قدرتمند و انعطاف‌پذیر است، اما عملکرد آن به شدت به **آماده‌سازی داده‌ها**، **انتخاب صحیح کرنل** و **تنظیم دقیق هایپرپارامترها** بستگی دارد.

مهم‌ترین دستاوردهای این پروژه عبارتند از:

- درک عمیق تفاوت عملکرد کرنل‌های مختلف در مواجهه با داده‌های خطی، غیرخطی، نویزی و پیچیده.
- مشاهده تاثیر حیاتی نرمال‌سازی داده‌ها بر روی دقت مدل‌های حساس به مقیاس مانند SVM.
- کسب مهارت عملی در بهینه‌سازی هایپرپارامترها با استفاده از ابزارهایی مانند GridSearchCV.
- یادگیری نحوه مدیریت چالش‌های عملی مانند زمان اجرای طولانی و انتخاب پارامترهای بهینه.

در نهایت، این پروژه دیدگاه جامعی نسبت به نقاط قوت و ضعف SVM ارائه داد و نشان داد که چگونه با استفاده صحیح از ابزارهای موجود، می‌توان مدلی بسیار دقیق و کارآمد برای مسائل طبقه‌بندی ساخت.



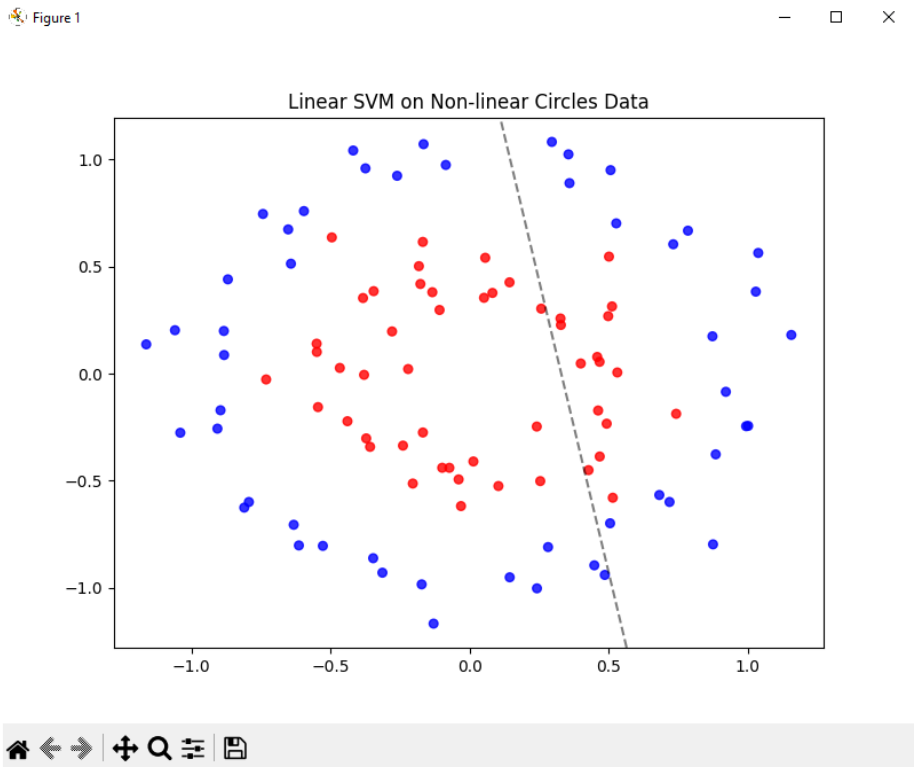


Figure 1
Linear SVM on Non-linear Circles Data
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

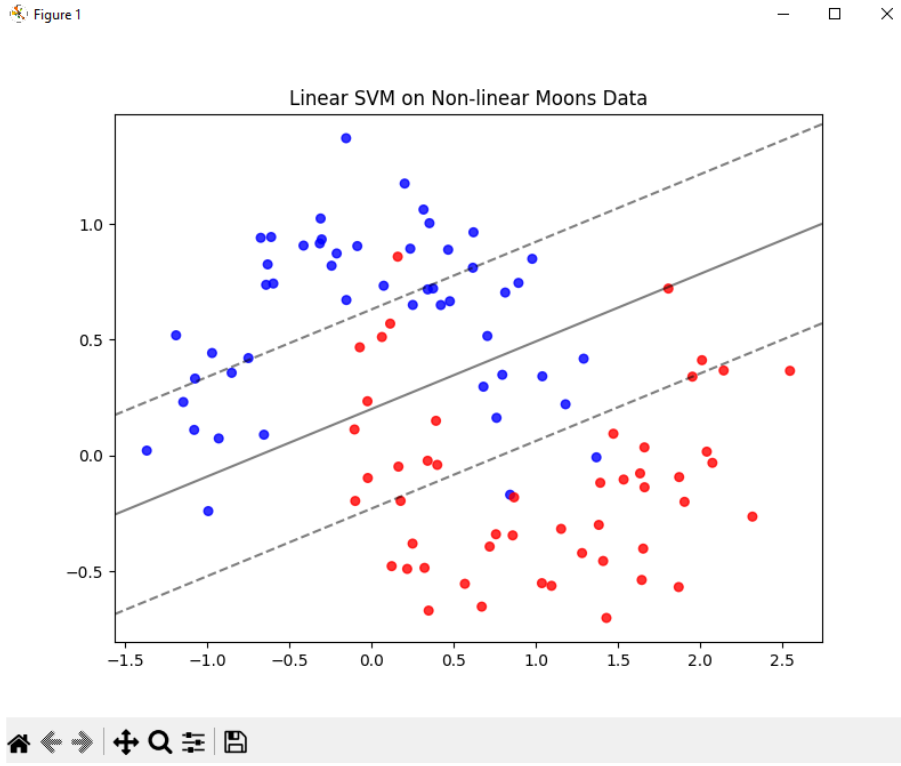


Figure 1
Linear SVM on Non-linear Moons Data
1.0
0.5
0.0
−0.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

o