پروژه SVM

درس شناسایی الگو

استاد درس : دکتر مرتضی آنالویی

دستیار آموزشی : سامان گودرزی

مهلت ارسال پروژه : جمعه 11 آبان‌ماه 1403

مقدمه

در این پروژه به تحلیل و پیاده‌سازی ماشین بردار پشتیبانی (SVM) در دو بخش پرداخته می‌شود.

در بخش اول، مدل SVM در یک سناریوی ساده و دوکلاسه با داده‌های دوبعدی بررسی می‌شود. هدف از این بخش، مشاهده‌ی عملکرد مدل به‌صورت بصری و بررسی تأثیر هایپرپارامترها و انواع کرنل‌ها بر روی مرزهای تصمیم‌گیری و تفکیک‌پذیری در فضای ویژگی‌ها است.

در بخش دوم، مدل SVM بر روی دیتاست واقعی MNIST پیاده‌سازی می‌شود و از آن برای دسته‌بندی تصاویر دست‌نوشته‌ی ارقام استفاده می‌گردد. این بخش امکان بررسی اثربخشی SVM در تحلیل داده‌های پیچیده‌تر و با ابعاد بالا را فراهم می‌آورد.

نکات

- گزارش کاملی از پروژه خود در هر قسمت ارائه دهید.
- پاسخ شما شامل یک فایل PDF برای گزارش پروژه و یک فایل ipynb است.
- سلول‌های نتایج را در فایل ipynb ذخیره کنید.
- استفاده از لایبرری های SVM در هر بخش مجاز است و نیاز به پیاده‌سازی SVM نیست. می‌توانید از لایبرری SVC استفاده کنید.

بخش اول

در این بخش 3 دیتاست مختلف به شما داده شده است. موارد خواسته شده در هر

قسمت را انجام دهید. در هر بخش مرز تصمیم را با خط چین نمایش دهید (پلات کنید).

الف) دیتاست تفکیک پذیر خطی :

- از SVM با کرنل خطی استفاده کنید. مقادیر مختلف C را تست کنید و تغییر در مرزهای تصمیم SVM را بررسی کنید. بعد از نمایش بردارهای پشتیبان، بررسی کنید تغییر در پارامتر C، چه تغییری در بردارهای پشتیبان دارد.(تغییرات C را مضرب 10 در نظر بگیرید. مثلا 0.1، 1، 10، ...)
- با استفاده از Logistic Regression دسته بندی را انجام دهید و با SVM مقایسه کنید.

ب) دیتاست تفکیک ناپذیر خطی :

- از SVM با کرنل خطی استفاده کنید. سپس با استفاده از کرنل چند جمله‌ای مرز تصمیم را بهبود دهید. در نهایت با استفاده از کرنل گاوسی مرزبندی را به طور کامل انجام دهید به گونه‌ای که کلاس ها به درستی از یکدیگر جدا شوند.

ج) پیدا کردن بهترین مدل :

- این دیتاست تفکیک ناپذیر خطی است که دارای یک مجموعه ولیدیشن نیز می‌باشد. مدلی بیابید که دقتش روی مجموعه ولیدیشن بالای 92% باشد.

بخش دوم

در این بخش، هدف استفاده از مدل ماشین بردار پشتیبانی (SVM) برای دسته‌بندی داده‌های واقعی موجود در دیتاست MNIST است. این دیتاست شامل تصاویر دست‌نوشته‌ی ارقام ۰ تا ۹ است.

الف) مقدمه

ابتدا دیتاست MNIST را دریافت و آن را به دو قسمت آموزشی و تستی تقسیم کنید. سپس، با بهره‌گیری از تکنیک‌های کاهش ابعاد مانند **PCA**، داده‌ها را به فضای دوبعدی منتقل کرده و به‌صورت بصری نمایش دهید. می‌توانید برای راحتی قسمتی از داده‌ها را نمایش دهید اما توزیع هر کلاس باید یکسان باشد (از هر رقم به اندازه یکسان نمایش دهید). این مرحله به شما امکان می‌دهد که توزیع و ساختار داده‌های تصویری را در فضای دوبعدی مشاهده کرده و درک بهتری از نحوه‌ی پراکندگی ارقام مختلف در این دیتاست واقعی کسب کنید. (می‌توانید پیکسل ها را به صورت سطری یا ستونی کنار یکدیگر قرار دهید تا بردار ویژگی تصویر موردنظر را بسازید).

ب) SVM چند کلاسه

1. مدل SVM به‌طور کلی برای دسته‌بندی‌های دوکلاسه طراحی شده است؛ بنابراین برای دسته‌بندی چندکلاسه (مانند دیتاست MNIST که شامل ۱۰ کلاس مختلف است)، نیاز به رویکردهای خاصی داریم. در این بخش، دو روش متداول برای گسترش SVM به مسائل چندکلاسه، یعنی **one-vs-one** و **one-vs-all**، مورد بررسی قرار می‌گیرد. در مورد این دو روش تحقیق کنید و روش کار هریک را توضیح دهید.
2. سپس **با استفاده از SVM دوکلاسه**، دسته‌بندی چندکلاسه با روش های بالا انجام دهید. استفاده از لایبرری‌های SVM چند کلاسه در این بخش امکان پذیر **نیست**.
3. در هنگام دسته‌بندی با روش one-vs-all، کلاس مورد بررسی (one)، تعداد کمتری داده نسبت به بقیه کلاس ها (rest) یا (all) دارد. با توجه به مسئله بهینه‌سازی SVM توضیح دهید این عدم توازن چه تاثیری بر عملکرد مدل می‌گذارد و چگونه

می‌توان این مورد را برطرف کرد.

ج) روش Crammer-Singer

یکی دیگر از روش‌های معروف در حل مسئله‌ی چندکلاسه با استفاده از SVM، **روش Crammer-Singer** است. این روش یک رویکرد چندکلاسه مستقیم برای SVM است که به جای تبدیل مسئله به چندین مسئله دوکلاسه (مانند روش‌های one-vs-one و one-vs-all)، به‌طور مستقیم به حل مسئله چندکلاسه می‌پردازد. در این روش، همه کلاس‌ها همزمان در یک مدل حضور دارند و هدف یافتن یک مجموعه مرزهای تصمیم‌گیری است که هر نمونه را به درستی به کلاس مربوطه تخصیص دهد. با استفاده از لایبرری این بخش را پیاده‌سازی کنید.

نکته:

برای قسمت‌های ب و ج، هایپرپارامترهای مختلف را امتحان کنید و نتیجه را گزارش کنید.

د) نمایش و معیار های ارزیابی

از مجموعه دادگان تستی، ۱۰ نمونه را به صورت تصادفی انتخاب و نمایش دهید. برای هر نمونه، لیبل واقعی و همچنین لیبل‌های پیش‌بینی شده توسط سه مدل SVM مختلف خود را نشان دهید. دقت کنید که لیبل‌های پیش‌بینی شده باید مربوط به مدل‌هایی باشند که در قسمت‌های ب و ج، بهترین نتیجه را داشته‌اند (بهترین کرنل‌ها و هایپرپارامترها را برای نمایش لیبل‌ها استفاده کنید.)

برای ارزیابی آزمایش‌هایی که در قسمت‌های ب و ج انجام داده‌اید، از معیارهایی که در ادامه اشاره شده‌است، استفاده کنید. این معیارها به شما کمک می‌کنند تا دقت و کارایی مدل SVM خود را بهتر بسنجید. ابتدا توضیح مختصری در مورد هر یک از معیارها بیان

کنید.

- دقت (Accuracy)
- ماتریس درهم‌ریختگی
- F1-Score
- Precision
- Recall