دانشگاه صنعتی شریف

دانشکده‌ی مهندسی کامپیوتر

گزارش پیاده‌سازی ماشین بردار پشتیبان

عنوان:

SVM

نویسنده :

علی قربان‌پور

استاد :

دکتر آرش عبدی هجران‌دوست

نیم‌سال دوم سال تحصیلی ۱۳۹۸ ـ ۱۳۹۹

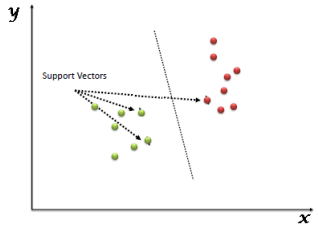
فهرست مطالب

مقدمه

کسب مهارت در زمینه الگوریتم‌های یادگیری ماشین، کاری دشوار و غیر قابل انجام محسوب نمی‌شود. بسیاری از افراد مبتدی، یادگیری ماشین را با یادگیری رگرسیون آغاز می‌کنند. رگرسیون مبحثی ساده و یادگیری آن آسان است. اما همه مسائل را نمی‌توان با رگرسیون حل کرد. بسیاری از مسائل با روشی غیر از رگرسیون حل می‌شوند.

الگوریتم‌های یادگیری ماشین مانند انواع سلاح‌های جنگی از قبیل تبرزین، شمشیر، خنجر و چاقو هستند. همه آن‌ها وسایل تیز و برنده‌ای هستند؛ اما هر یک کاربرد خاص خود را دارند و باید در جایگاه خاصشان مورد استفاده قرار گیرند. الگوریتم‌های یادگیری ماشین نیز بسیار متنوع هستند و هر یک برای حل نوع خاصی از مسائل قابل استفاده به حساب می‌آیند. می‌توان تصور کرد «رگرسیون» مانند شمشیری است که قادر به قطعه‌قطعه کردن و برش دادن داده‌ها به صورت کارا و مؤثر است، ولی توانایی کار کردن با داده‌های دارای پیچیدگی بالا را ندارد.

«ماشین بردار پشتیبان» (SVM) یک الگوریتم نظارت‌شده یادگیری ماشین است که هم برای مسائل طبقه‌بندی و هم مسائل رگرسیون قابل استفاده است؛ با این حال از آن بیشتر در مسائل طبقه‌بندی استفاده می‌شود. در الگوریتم SVM، هر نمونه داده را به عنوان یک نقطه در فضای بعدی-n روی نمودار پراکندگی داده‌ها ترسیم کرده n) تعداد ویژگی‌هایی است که یک نمونه داده دارد) و مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص می‌کند. سپس، با ترسیم یک خط راست، داده‌های مختلف و متمایز از یکدیگر را دسته‌بندی می‌کند.



شکل ۱: ماشین بردار پشتیبان

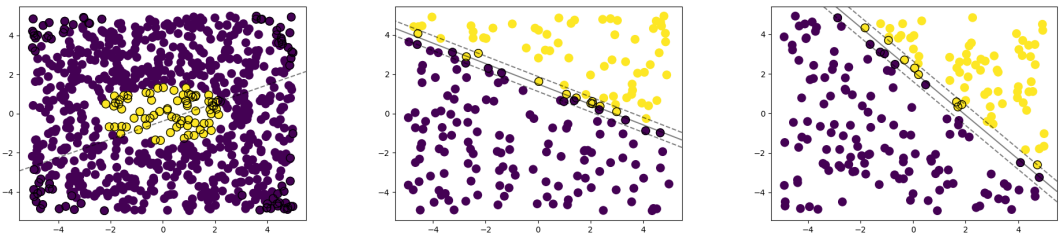
۱ بخش اول

در بخش اول هدف آن است که تعدادی نقطه در صفحه‌ی دوبعدی تولید کنیم. سپس به شکلی دلخواه آن‌ها را به دو دسته‌ی مختلف تقسیم کرده و برچسب‌گذاری بر روی آن‌ها انجام دهیم. سپس با استفاده از ماشین بردار پشتیبان این فضا ره به دو دسته تقسیم کنیم و با خطی این دو گروه را از یکدیگر جدا کنیم. در این بخش به بررسی انواع شکل داده‌های ورودی و همچنین تغییر پیچیدگی توابع و تعداد نقاط می‌پردازیم. همچنین با تغییر هسته انواع دسته‌بندی را مشاهده می‌کنیم.

۱ـ۱ انواع هسته

با تغییر انواع هسته [۱] ، کارکرد ماشین بردار به عنوان جداکننده تغییر می‌کند. برای داده‌های غیر خطی برای مثال نمی‌توان از جداکننده خطی استفاده کرد. در تصویر آخر این بخش ناتوانی این هسته را به خوبی مشاهده می‌کنید. با انتخاب مناسب هسته به دسته‌بندی صحیح داده‌ها می‌توان پرداخت. مثال‌های زیر از انواع هسته با نقاط مختلف را ببینید.

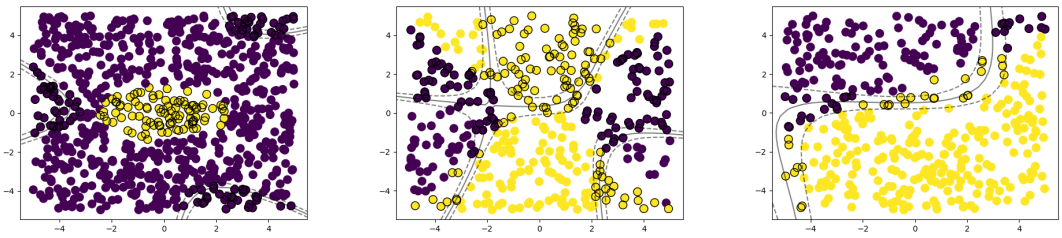
linear ۱ـ۱ـ۱



شکل ۱ـ۱: تقسیم نقاط دوکلاسه با هسته‌ی خطی

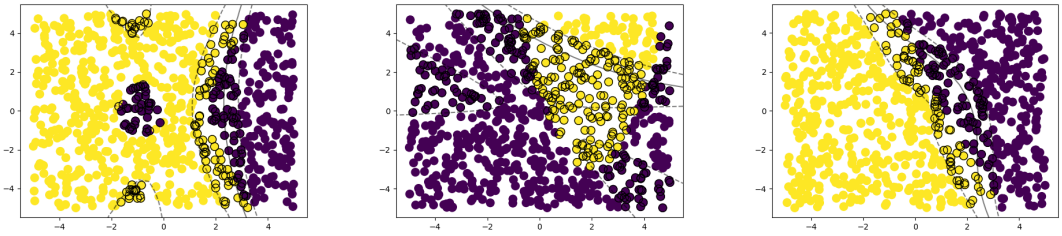
[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

sigmoid ۱ـ۱ـ۲



شکل ۱ـ۲: تقسیم نقاط دوکلاسه با هسته‌ی سیگموئید

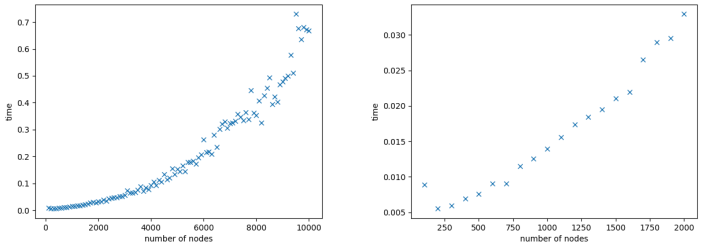
ploy ۱ـ۱ـ۳



شکل ۱ـ۳: تقسیم نقاط دوکلاسه با هسته‌ی چندجمله‌ای

هر چه تابع مولد نقاط پیچیده‌تر می‌شود، به هسته‌های پیچیده‌تری نیاز خواهیم داشت برای تقسیم‌بندی. به عبارتی کارایی هسته‌ها در محدوده‌ای مشخص است. با هسته‌ی خطی نمی‌توان داده‌های چندجمله‌ای یا پیچیده‌تر را کلاس‌بندی کرد.

حال تعداد نقاط را به مرور بر روی یک تابع مشخص زیاد می‌کنیم تا کارکرد ماشین بردار پشتیبان را بسنجیم.

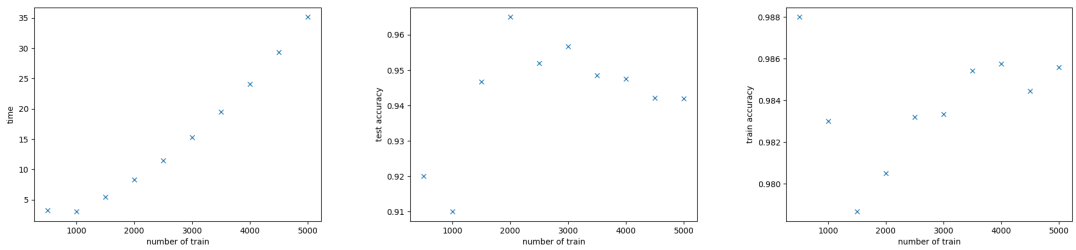


شکل ۱ـ۴: رشد زمان اجرا با زیاد شدن تعداد نقاط ورودی

همانطور که از نمودارها برمی‌آید این رشد برای داده‌های با ابعاد کوچک خطی و برای ابعاد بالاتر به صورت نمایی با ضریب پایین رشد می‌کند که میتوان تقریب خطی مناسبی در هر بازه با خطای اندک بر آن زد.

۲ بخش دوم

در بخش دوم به پیاده‌سازی کلاس‌بندی همان پایگاه داده‌ای که در شبکه عصبی استفاده کردیم می‌پردازیم. پایگاه داده MNIST که یک پایگاه داده تصویری متشکل از ۷۰،۰۰۰ تصویر دست‌نوشته‌ی اعداد لاتین است. می‌خواهیم روند تغییرات دقت داده‌های آموزش، دقت داده‌های آزمایش و زمان اجرا را متناسب با تغییرات تعداد داده‌های آموزش و آزمایش بسنجیم. برای این کار با ۵۰۰ داده آموزش و ۵۰ داده آزمایش شروع می‌کنیم. در هر مرحله به هر کدام به ترتیب ۵۰۰ و ۵۰ داده اضافه می‌کنیم و از ابتدا اجرا را پی می‌گیریم. روند تغییرات این پارامترهای ارزیابی در قبال تغییر تعداد داده‌های آموزش در نمودارهای زیر آمده است.

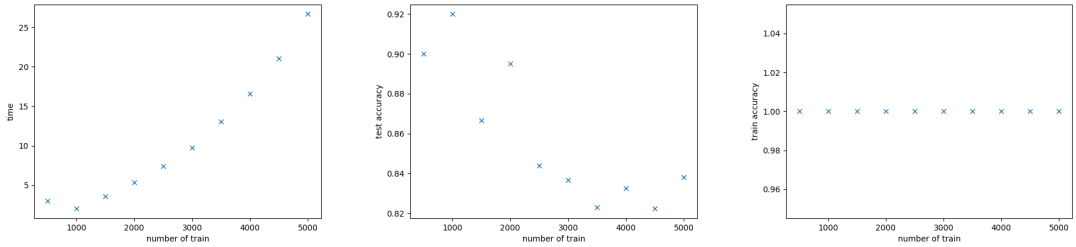


شکل ۲ـ۱: روند تغییرات پارامترهای ارزیابی شبکه با افزایش تعداد ورودی با هسته‌ی rbf

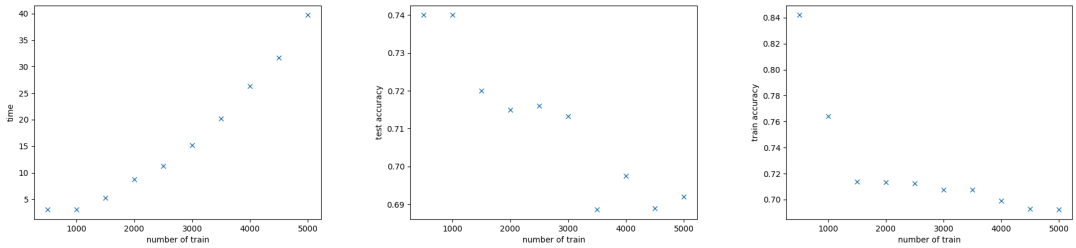
با توجه به نمودارها به طور عمومی دقت داده‌های آزمایش با افزایش تعداد داده‌های آموزش افزایش می‌یابد که قابل انتظار نیز است. همچنین مانند بخش قبل زمان اجرا به صورت نسبتا خطی افزایش می‌یابد که قابل توجه است. همچنین با نگاهی به میزان دقت ماشین بردار پشتیبان در داده‌های تست می‌توان دید که دقت قابل توجهی دارند.

حال همان کاری که در بخش قبل انجام دادیم را با استفاده از سایر هسته‌ها نیز تکرار می‌کنیم تا دقت سایر هسته‌ها را بر روی

این پایگاه داده بسنجیم و پارامترهای ارزیابی ماشین بردار را مشاهده کنیم.



شکل ۲ـ۲: روند تغییرات پارامترهای ارزیابی شبکه با افزایش تعداد ورودی با هسته‌ی linear

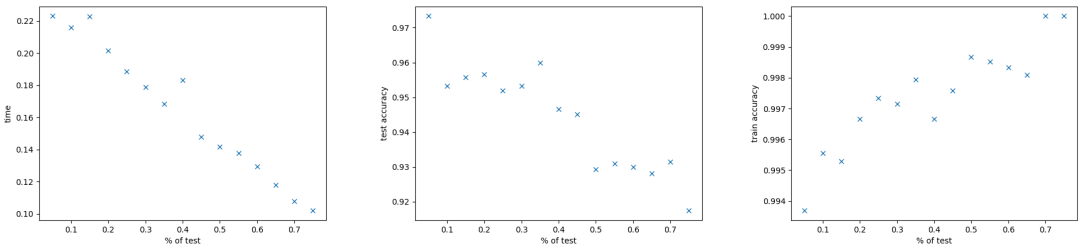


شکل ۲ـ۳: روند تغییرات پارامترهای ارزیابی شبکه با افزایش تعداد ورودی با هسته‌ی sigmoid

ارزیابی هسته‌ها نشان می‌دهد که هسته‌ی rbf بالاترین دقت را در داده‌های آموزش و آزمایش دارد. بعد از آن دقت هسته‌ی خطی چیزی در حدو ۸۵ درصد است، اما این دقت در داده‌های آموزش ۱۰۰ درصد است و این عدد فارغ از تعداد داده‌های آموزش می‌باشد. به عبارتی در این نوع هسته اوورفیتینگ رخ می‌دهد. در نهایت نیز هسته‌ی sigmoid با کمترین دقت در هر دو دسته‌ی داده کمترین ارزش را برای انتخاب هسته در این پایگاه داده دارد. توجه کنید کارایی هسته‌های ابداً مطلق نیست و متناسب با نوع داده‌های مورد بحث بسیار متغییر است. باید داده‌ها را شناخت و بهترین هسته را برای آن‌ها انتخاب کرد.

۳ بخش سوم

در این بخش به دسته‌بندی داده‌های تصویری می‌پردازیم. در ابتدا تصاویر موجود در فایل ارسالی را به فرمت اعداد csv تبدیل می‌کنیم. بدین ترتیب می‌توانیم به خوبی با این داده‌ها کار کنیم و به دسته‌بندی و آموزش و تست بپردازیم. در گام نخست این داده‌ها را به دو دسته‌ی آموزش و آزمایش تقسیم می‌کنیم. نسبت این تقسیم از پارامترهای قابل تغییر است. روند تغییرات پارامترهای ارزیابی به نسبت تغییر این نسبت را به خوبی مشاهده می‌کنید.



شکل ۳ـ۱: تقسیم داده‌ها به دو گروه و تاثیر نسبت این تقسیم بر پارامترهای ارزیابی

در این بخش نیز به شکل قسمت قبل، دقت ماشین بردار هنگامی که از هسته‌ی sigmoid استفاده می‌کند نسبت به دو حالت دیگر به شکل چشم‌گیری پایین‌تر است. در دو هسته‌ی خطی و rbf دقت حدود ۹۵ درصد است که قابل قبول است. اما دقت سیگموئید در حدود ۴۰ درصد است. شاید اصلا این هسته برای داده‌های تصویری مناسب نباشد. :))

جمع‌بندی

چندین نکته مهم را می‌توان از این بخش برداشت کرد. نخست آنکه نسبت داده‌های ورودی به زمان اجرا چیزی از مرتبه‌ی خطی در داده‌های کوچک و نمایی در داده های بزرگتر است. مورد بعدی انتخاب درست هسته است. هسته‌های مختلف بر روی یک پایگاه داده اختلاف دقتی در حدود ۶۰ درصدی دارند. باید داده‌ها را شناخت و هسته‌ی مناسب با آن را انتخاب کرد. مورد آخر هم آنکه در قیاس با شبکه‌های عصبی مزایا و معایبی دارد. مزایای آن زمان بنسبت خوب اجرا و همچنین دقت بالای آن است. اما مشکل آن شاید قابل تطبیق نبودن با داده‌های جدید است به نحوی که نیاز است دوباره از نو محاسبات انجام شود. در حالی که برای آپدیت شبکه‌ی عصبی به چنین افزایش بار محاسباتی نیاز نداریم.

امیدوارم از مطالعه‌ی این پژوهش بهره‌ی کافی برده باشید.