رضا عبدلی 810100251

1

برای حل مشکلات مطرح شده می‌توان از روش‌های مختلفی استفاده کرد:

برای حل مشکل وجود نداشتن یک یا چند ویژگی در داده‌های آموزش، می‌توان از روش‌هایی مانند تکمیل داده‌ها با میانگین مقادیر موجود، استفاده از مدل‌های پیش‌پردازش مانند Imputation یا کار با روش‌هایی مانند KNN Imputation استفاده کرد

برای حل مشکل نامتوازن بودن توزیع داده‌ها در کلاس‌ها، می‌توان از روش‌هایی مانند Undersampling و Oversampling استفاده کرد. در Oversampling، تعداد نمونه‌های کلاس‌های کمتر افزایش می‌یابد و در Undersampling تعداد نمونه‌های کلاس‌های بیشتر کاهش می‌یابد. همچنین، می‌توان از روش‌های مبتنی بر وزن برای کاهش تاثیر کمترین کلاس‌ها و افزایش تاثیر بیشترین کلاس‌ها استفاده کرد.

برای حل مشکل وجود نویز در داده‌ها، می‌توان از روش‌هایی مانند فیلترینگ (Filtering) یا روش‌های پیش‌پردازش مانند Smoothing استفاده کرد. همچنین، می‌توان از مدل‌های مقاوم به نویز مانند Random Forest یا SVM استفاده کرد.

برای حل مشکل وجود ویژگی‌های correlated، می‌توان از روش‌های مختلفی مانند Principal Component Analysis (PCA) یا روش‌های انتخاب ویژگی‌های مهم استفاده کرد. در PCA، ویژگی‌های مرتبط با یکدیگر ترکیب می‌شوند و ویژگی‌های جدیدی به دست می‌آیند که باعث کاهش ابعاد داده و کمک به جلوگیری از overfitting می‌شوند.

2

رای درک ارتباط بین ساعت مطالعاتی دانشجویان و نمره آزمون، می‌توان از یک مدل رگرسیون خطی با متغیرهای دوگانه استفاده کرد، که علاوه بر ساعت مطالعاتی، متغیر دیگری برای تاثیر انکار ناپذیر آزمون دادن نیز در نظر بگیرد. این مدل به صورت زیر می‌تواند باشد:

نمره آزمون= B 0 + تاثیر آزمون دادن * β 2 + ساعت مطالعاتی * B 1

حال باید ضرایب B0وB1 وB2 رو حساب کنیم

روش :Least Squares

Least Squares یکی از روش‌های متداول برای تخمین ضرایب مدل است. در این روش، ما سعی می‌کنیم مقدار مجموع مربعات اختلاف بین مقادیر پیش‌بینی شده توسط مدل و مقادیر واقعی مشاهده شده را کمینه کنیم.

برای مدل خطی، مجموع مربعات اختلاف بین مقادیر پیش‌بینی شده yj و مقادیر مشاهده شده yi به این شکل هست:

$$\sum_{i=1}^{n} (yj - yi)$$

که اینجا n تعداد نمونه‌هاست.

ضرایب به صورتی که این مجموع را کمینه کنند، به صورت تئوری می‌توانند با استفاده از روش‌های ریاضیاتی به‌دست آیند، مانند روش معادلات ناپایدار (به روش نرمال معادلات)، اما در مواردی که تعداد متغیرها زیاد است، این روش ممکن است مشکلاتی داشته باشد.

روش Gradient Descent:

در روش Gradient Descent، ما تلاش می‌کنیم تا با به‌روزرسانی تدریجی پارامترها، به مقدار کمینه برسیم. در اینجا، به ازای هر پارامتر، مشتق جزئی تابع هزینه نسبت به آن پارامتر را محاسبه می‌کنیم و سپس مقدار پارامتر را به سمت معکوس مشتق جزئی (با ضریب کوچکی که به آن learning rate می‌گویند) به‌روزرسانی می‌کنیم.

روش‌های دیگر:

علاوه بر این دو روش، روش‌هایی مانند روش نرم‌افزارهای بهینه‌سازی مانند Adam و RMSprop نیز وجود دارند که عموماً بهتر از Gradient Descent عمل می‌کنند، زیرا با استفاده از ترکیبی از momentum و adaptive learning rate، سریعتر به نقطه‌ی کمینه همگرا می‌شوند.

به‌طور خلاصه، مدل ریاضیاتی مناسب می‌تواند مدل رگرسیون خطی با متغیرهای دوگانه باشد، که ضرایب آن با استفاده از روش Least Squares، Gradient Descent یا روش‌های بهینه‌سازی دیگر به‌دست می‌آید.

3

confusion matrix:

	Predicted ham	Predicted spam
Actual ham	TN 200	FP 30
Actual spam	FN 20	TP 30

حالامعیارهای ارزیابی را محاسبه کنیم:

Accuracy:

$$\frac{TP + TN}{TP + TN + FP + FN} = ACCURACY = \frac{500}{550} \approx 0.909$$

precision:

$$\frac{TP}{TP + FP} = precision = \frac{300}{330} \approx 0.909$$

Recall:

$$\frac{TP}{TP + FN} = RECALL = \frac{300}{320} \approx 0.9375$$

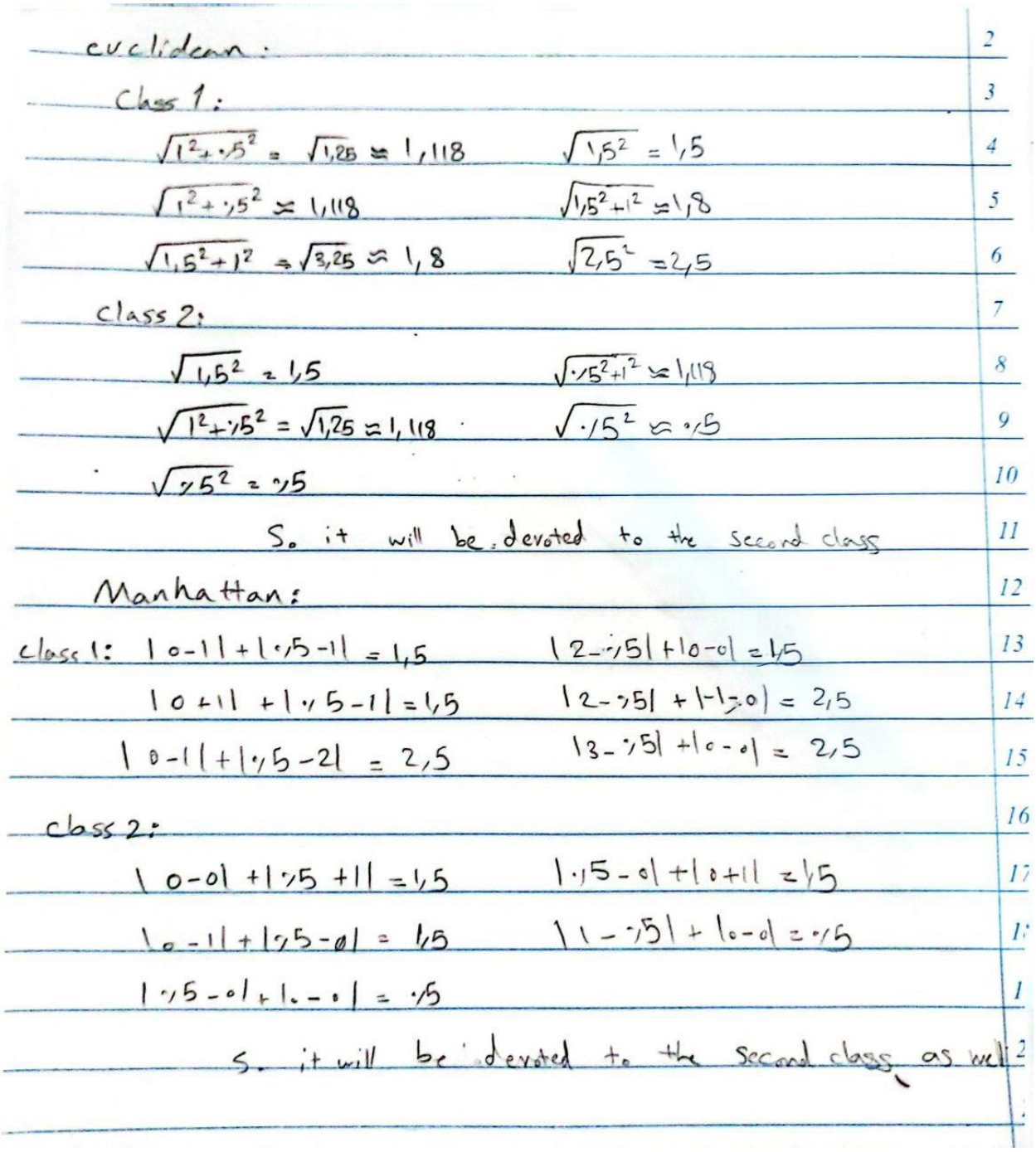
F1-score:

$$\frac{precision * recall * 2}{precision + recall} = F1 = \frac{1.7015}{1.84} \approx 0.92$$

KNN

1

The image below:

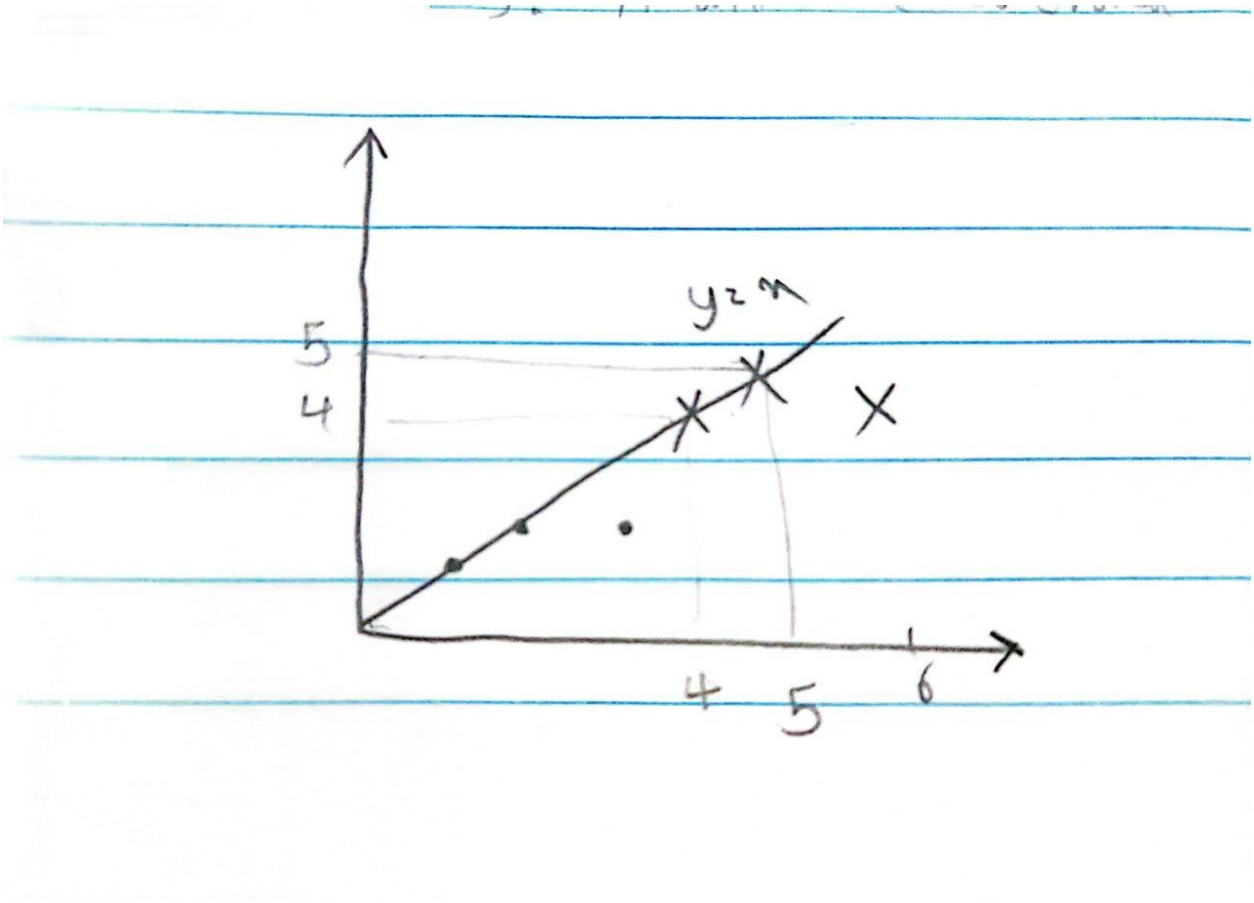


:Support Vector Machine

1

نقاط پشتیبان یا ""support vectors، نقاطی هستند که بهترین انتخاب برای مرز تصمیم‌گیری در ماشین بردار پشتیبان ((SVMهستند. آنها نقاطی هستند که به مرز تصمیم‌گیری در SVM نزدیک‌ترینند. این نقاط تعیین‌کننده‌ی ماکزیمم حاشیه بین دو کلاس هستند

ادامه صفحه بعد



طبقه‌بندی SVM برای داده‌هایی که ویژگی‌هاشان با شلوغی یا نویز مواجه هستند، مناسب نیست. اگر کلاس‌ها به شدت با هم اشتراک داشته باشند و قابل جدا کردن به صورت خطی نباشند، SVM به خوبی عمل نمی‌کند.

Kernelها در SVM، توابعی هستند که داده‌های ورودی را به یک فضای بعد بالاتر تبدیل می‌کنند تا جداپذیری کلاس‌ها را بهبود ببخشند. این توابع به SVM کمک می‌کنند تا مرز تصمیم‌گیری پیچیده‌تری را در فضای ویژگی اصلی پیدا کند. Kernel ها وظیفه دارند تا محاسبه محصول داخلی نقاط داده‌ها را در فضای بعد بالاتر انجام دهند بدون اینکه به صورت صریح تبدیل را محاسبه کنند. انواع معمول kernel ها شامل linear، polynomial، radial basis function (RBF) و sigmoid هستند.

تفاوت بین یک طبقه‌بند سافت مارژین (soft margin classifier) و یک طبقه‌بند هارد مارژین (hard margin classifier) در نحوه مدیریت اشتباه‌ها و عرض مارژین است:

طبقه‌بند هارد مارژین: این طبقه‌بند سعی می‌کند مارژین بیشینه بین دو کلاس را پیدا کند بدون اینکه اشتباهی در دسته‌بندی اتفاق بیفتد. این مدل فقط زمانی عملکرد خوبی دارد که داده‌ها به طور کامل قابل جداسازی باشند. این مدل به شدت به اورفیتینگ حساس است.

طبقه‌بند سافت مارژین: این طبقه‌بند اجازه می‌دهد تا چند اشتباه در دسته‌بندی داشته باشد تا یک مارژین گسترده‌تر پیدا کند. هدف این طبقه‌بندی، بالانس بین بیشینه کردن مارژین و کاهش خطاهای دسته‌بندی است. این مدل به نویزها و داده‌های پرت حساسیت کمتری دارد.

استفاده از SVM در مسائل رگرسیون به عنوان ماشین بردار پشتیبان برای رگرسیون (Support Vector Regression)نامیده می‌شود. در SVR، هدف یافتن یک تابع است که تقریباً مطابق با نگاشت از ویژگی‌های ورودی به خروجی پیوسته باشد. به جای بیشینه کردن مارژین مانند طبقه‌بندی، هدف SVR این است که تعداد بیشتری نمونه را داخل یک مارژین مشخص نگه دارد و در عین حال خطای کمینه را حفظ کند