810100088

تمرین 4 محمد

آرین باستانی

سوال اول

1) اگر تعداد این داده‌های ناموجود کم باشد می‌توان آن‌ها را با چیزهایی مانند میانگین و میانه پر کرد یا اگر در سطری، تعداد این فیچرهای ناموجود زیاد باشد، آن سطر را حذف کرد.
ولی اگر تعداد داده‌های خذفی زیاد باشد، این فیچر را حذف می‌کنیم و یا آن را با میانگین می‌کنیم.

2) این مشکل زمانی پیش میاید که تعداد نمونه‌ها در هر کلاس نامتعادل باشد. برای حل این مشکل می‌توان از دو راه زیر استفاده کرد:

i) Oversampling : تعداد نمونه‌های مربوط به کلاس کم نمونه را افزایش دهیم (با روشی مانند تولید رندوم داده)

ii) Undersampling : نمونه‌های مربوط به کلاس با نمونه زیاد را حذف کنیم (برای این روش هم می‌توان از حذف رندوم استفاده کرد)

3) می‌توان نمونه‌های نویزی را حذف کرد و یا داده‌ها با استفاده از فیلترهای میانگین‌گیری تصفیه شوند.
همچنین با استاندارد کردن و یا نرمالایز کردن می‌توان تاثیر نویز را کاهش داد.
همچنین تصفیه‌ی داده‌ها با بردن به فضای فوریه (تبدیل فوریه) می‌توان انجام داد.

4) با توجه به کوریلیشن فیچرها می‌توان از بین این فیچرها، آنهایی را نگه داشت که تاثیر بهتری در مدل ما دارند و بقیه را حذف کنیم.

سوال دوم :

برای درک ارتباط دو فیچر و یا تاثیر هر فیچر بر روی target ، کورولیشن گیری همیشه کمک می کند .

برای (least Square method) ، بدین صورت ضرایب بدست میاید که باید طوری ضرایب انتخاب شوند که ~~اختلاف~~ مربع های اختلاف مقادیر اصلی و مقادیر پیش بینی شده کمینه شود.

برای Gradient desent ، اینکار را اینقدر تکرار می کنیم که به کمترین گرادیان ممکن برسیم :

- شیب را (مقدار گرادیان) را ~~به~~ برای هر پارامتر محاسبه می کنیم .
- مقادیر ~~جدید~~ پارامترها را در جهت کمترین مقدار گرادیان تغییری دهیم .

از روش های دیگری توان Least Absoulute Deviation و روش نیوتون را نام برد که اولی مجموع اختلاف مقادیر اصلی و پیش بینی شده را کم می کند و با این روش ضرایب را بدست میاورد و روش دوم نیز گفته می شود که بهترین ضرایب را پیدا خواهد کرد .

سوال سوم :

مقادیر پیش بینی شده		مقادیر اصلی
$T_p = 300$	$F_p = 30$	
$F_N = 20$	$T_N = 200$	

$$Accuracy = \frac{300+200}{300+200+20+30} = \frac{500}{550} = \frac{10}{11} \simeq 0,91$$

$$Precision = \frac{300}{300+30} = \frac{300}{330} = 10/11 \simeq 0,91$$

$$Recall = \frac{300}{300+20} = \frac{300}{320} = 15/16 \simeq 0,94$$

$$f_1\text{-score} = 2 \times \frac{\frac{10}{11} \times \frac{15}{16}}{\frac{10}{11} + \frac{15}{16}} = \frac{2 \times 176 \times 150}{88 \times 325} = \frac{12}{13} \simeq 0,92$$

KNN : فاصله‌های اقلیدسی:

$d_1(x,y) \Rightarrow$ فاصله اقلیدسی (x,y) تا $(1/2,0)$

$d_1(0,0) = 1/2$ و $d_1(0,1) = \frac{\sqrt{5}}{2}$ و $d_1(1,0) = \frac{1}{2}$ و $d_1(1,1) = \frac{\sqrt{5}}{2}$ و

$d_1(0,-1) = \frac{\sqrt{5}}{2}$ و $d_1(1,-1) = \frac{\sqrt{5}}{2}$ و $d_1(2,0) = 3/2$ و $d_1(3,0) = \frac{5}{2}$

بنابراین به علت وجود $(0,1)$ و $(0,0)$ و $(1,0)$ و $(0,-1)$ که این نمونه به کلاس دوم تعلق دارد.

فاصله‌های منهتن:

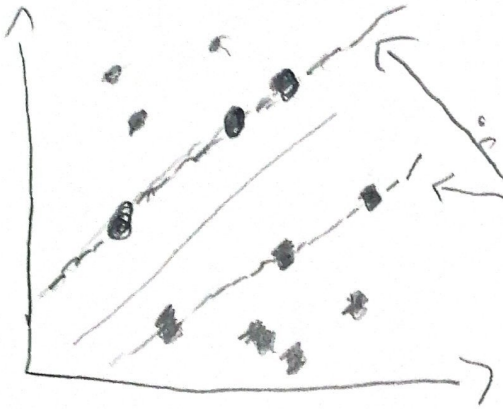
$d_2(x,z) \Rightarrow$ فاصله‌ی منهتن (x,y) تا $(1/2,0)$

$d_2(0,0) = 1/2$ و $d_2(0,1) = 3/2$ و $d_2(1,0) = 1/2$ و $d_2(1,1) = 3/2$ و

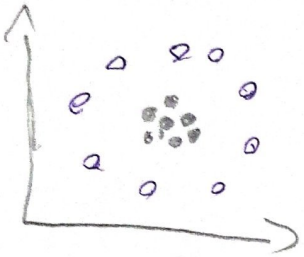
$d_2(0,-1) = 3/2$ و $d_2(1,-1) = 3/2$ و $d_2(2,0) = 3/2$ و $d_2(3,0) = 5/2$

بنابراین به علت وجود نقاط $(0,1)$ و $(0,0)$ و یکی از نقاطی که فاصله‌ی $3/2$ دارند باز هم در کلاس دوم قرار می‌گیرد.

SVM :



- به نقاطی گفته می‌شود که با خط جدا کننده نزدیکترند. مثال:



- برای داده‌های که بصورت خطی تفکیک پذیر نیستند مانند

و همچنین برای داده‌های خیلی بزرگ یا نویزی تنها مناسب نیست

- هسته یک تابع ریاضی است که داده ها را به بعد بالاتر تبدیل می‌کند.
این تبدیل این امکان را می دهد که در فضای جدید، الگوریتم خطی را روی داده ها اجرا کنیم. به عبارت دیگر کرنل ها به ما اجازه می دهند که با استفاده از الگوریتم های خطی، مسائل غیر خطی را حل کنیم.

- هدف اصلی Hard SVM، حاشیه‌ای بین دو کلاس را جدا کند کردن است و هیچ اشتباه طبقه بندی در آن مجاز نیست و تمام داده ها باید بدرستی دسته بندی شوند. اما در Soft SVM، به داده ها اجازه می دهیم در حاشیه‌ی SVM قرار بگیرند یا حتی در داخل حاشیه باشند.
Soft SVM به ما امکان می دهد که با اجازه‌ی اشتباه طبقه بندی، مدلی با عمومیت بیشتر داشته باشیم و برای داده‌های noisy یا دارای همپوشانی مناسب‌تر است.

• خطی را پیدا می‌کند که دو کلاس را از هم جدا می‌کند و از Support Vector های دو طرف بیشترین فاصله را داشته باشد (مثلا اگر Support Vector ها و خط پیدا شده هر سه موازی باشند، خط پیدا شده در وسط Support Vector ها قرار خواهد گرفت تا بیشترین فاصله را داشته باشد)