تمرین سری ۳ واحد درسی داده کاوی مجازی

جناب آقای دکترفراهانی و جناب آقای دکتر خردپیشه

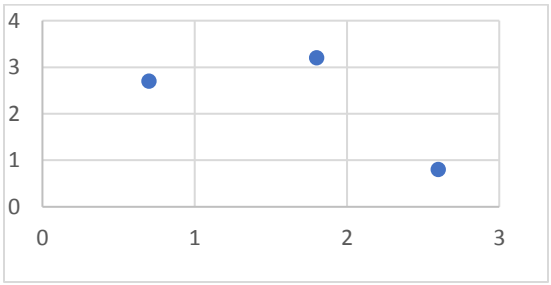
TA: علی شریفی

پاسخ و گردآوری: سلاله شیخیان

1. در خصوص کرنل های پرکاربرد روش SVM تحقیق کنید. به صورت کلی چرا ما از ایده کرنل در بحث SVM بهره می بریم . آیا می توان در خصوص کرنل ها و استفاده ی آنها حکم کلی داد . به طور مثال بگوییم از کرنل RBF در این مواقع خاص استفاده میکنیم .

باید توجه نمود که نمیتوان به طور قطعی معین کرد که در هر مسئله کدام کرنل بهتر است اما بر اساس تجربه و بررسی کارایی کرنل های مختلف، برای برخی مسائل کرنل های مشخصی را پیشنهاد میدهند. اما این پیشنهاد بهترین بودن کرنل را تضمین نمی کند.

در واقع هدف این است که کرنل ایده ساخت ویژگی های جدیدی ارائه دهد که برای ساخت توابع غیر خطی از آن استفاده نماییم. با داشتن x، مجموعه جدید ویژگی ها را بر اساس شباهت آن با نقاط راهنمای l^1 و l^2 و l^3 انتخاب می شود.



الف) کرنل گوسی:

$$f_1 = sim(x, l^1) = exp\left(-\frac{||x - l^1||^2}{2\sigma^2}\right)$$

تابع بالا میزان شباهت نقاط به l^1 را نشان می دهد. اگر x و l^1 کاملا منطبق باشند خروجی تابع یک خواهد بود. (بیشترین شباهت) و با افزایش اختلاف دو عدد خروجی به سمت صفر میل میکند. برای l^2 و l^3 هم میتوانیم توابع زیر را تعریف کنیم:

$$f_2 = sim(x, l^2) = exp\left(-\frac{||x - l^2||^2}{2\sigma^2}\right)$$

$$f_3 = sim(x, l^3) = exp\left(-\frac{||x - l^3||^2}{2\sigma^2}\right)$$

پس برای هر نقطه شباهت آن با هر سه نقطه راهنما را محاسبه می کنیم. یعنی 3 ویژگی اضافه میکنیم.

$$b + w_1 f_1 + w_2 f_2 + w_3 f_3 \geq 0$$

الگوریتم به ازای هر داده یک نقطه راهنما در نظر می گیرد و اگر m داده داشته باشیم، m نقطه راهنما خواهیم داشت. و مسئله از n بعدی به m بعدی افزایش می یابد.

این مسئله خطی جدایی پذیر است اگر داده ها مستقل خطی باشند.

$$x = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Longrightarrow f = \begin{bmatrix} f_0 = 1 \\ f_1 = k(x, l^1) \\ f_2 = k(x, l^2) \\ \vdots \\ f_m = k(x, l^m) \end{bmatrix}$$

پیش پردازش داده x با استفاده از توابع کرنل:

$$z = \varphi(x) = (\varphi_1(x), \varphi_2(x), \dots, \varphi_k(x))$$

$$g(z) = w^T z + b$$

$$g(x) = w^T \varphi(x) + b$$

مرز تصمیم گیری:

$$w = \sum_{t=1}^{m} a^t y^t z^t = \sum_{t=1}^{m} a^t y^t \varphi(x^t)$$

دسته بندی داده جدید:

$$g(x) = w^T \varphi(x) + b = \left(\sum_{t=1}^{m} a^t y^t \varphi(x^t)^T \varphi(x) \right) + b = \left(\sum_{t=1}^{m} a^t y^t k(x^t, x) \right) + b$$

در فرمول بالا می بینیم که اگر کرنلی مثل k وجود داشته باشد، نیازی به انجام محاسبه ضرب داخلی نیست.

برای این کار ماتریس گرم را برای kمی سازیم. این ماتریس متقارن است.

ب) کرنل چند جمله ای

این کرنل در پردازش تصویر پرکاربرد است. معادله آن به صورت زیر است:

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$$

که در آن d درجه چند جمله ای است.

پ) تابع پایه شعاعی گاوسی (RBF)

این کرنلی برای اهداف عمومی کلربرد دارد. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود نداشته باشد، مورد استفاده قرار می گیرد. معادله آن به صورت زیر است

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2)$$

ت) کرنل RBF لاپلاس

این هم یک کرنل برای اهداف عمومی است. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود ندارد استفاده می شود .معادله آن به صورت زیر است:

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$$

ث) کرنل تانژانت هیپربولیک(tanh)

می توانیم از آن در شبکه های عصبی استفاده کنیم. معادله مربوط به آن عبارت است از:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$$

در برخی موارد (نه همیشه $k > 0$) و $c < 0$.

ج) کرنل سیگموئید

می توان این کرنل را در شبکه های عصبی مورد استفاده قرار داد. معادله مربوط به آن عبارت است از:

$$k(x,y) = \tanh(\alpha x^T y + c)$$

چ) کرنل تابع بِسِل (Bessel)از نوع اول

ما می توانیم از آن برای حذف مقطع عرضی در توابع ریاضی استفاده کنیم. معادله آن عبارت است از:

$$k(x,y) = \frac{J_{v+1}(\sigma\|x-y\|)}{\|x-y\|^{-n(v+1)}}$$

که J تابع بسل از نوع اول است.

ح) کرنل پایه شعاعی ANOVA

ما می توانیم از آن در مسائل رگرسیون استفاده کنیم. معادله مربوط به آن عبارت است از:

$$k(x,y) = \sum_{k=1}^{n} \exp(-\sigma(x^k - y^k)^2)^d$$

خ) کرنل splineخطی بصورت یک بعدی

این کرنل، هنگام کار با بردارهای بزرگ داده پراکنده ، کاربرد زیادی دارد. این کرنل اغلب در دسته بندی متن مورد استفاده قرار می گیرد. کرنل spline همچنین در مسائل رگرسیون عملکرد خوبی دارد. معادله آن عبارت است از:

$$k(x,y) = 1 + xy + xy\ min(x,y) - \frac{x+y}{2}\ min(x,y)^2 + \frac{1}{3}\min(x,y)^3$$

2. قبلا با دیتاست کلاس بندی قیمت موبایل در کگل کار کرده ایم . بر روی دیتاست ، روش SVM را اجرا کنید . (استفاده از پکیج ها همانند sklearn مجاز است.)

برای اجرای SVM ابتدا داده تست و آموزش را جدا میکنیم و متغیر ها آموزش و برچسب Y را میسازیم. با استفاده از پکیج sklearn روش SVM را اجرا میکنیم. امتیاز 0.9533333333333334 بدست می آید به این معنی که در داده های تست با دقت 95 درصد برچسب ها به درستی تشخیص داده شده است. با استفاده از ماتریس زیر میتوان مشاهده کرد که داده های کدام دسته به درستی تشخیص داده شده و یا خطا داشته است:

```
[141   3   0   0]
[  7 138   0   0]
[  0   7 143   8]
[  0   0   3 150]]
```

در روش بالا از کرنل BRF استفاده شده از در بخش های بعدی نتیجه حاصل از کرنل های دیگر را بررسی خواهیم کرد.

3. برای سوال ۲ حداقل ۵ حالت مختلف از قبیل کرنل ها و پارامترها را بررسی کنید و نتایج آن را گزارش دهید.

برای این کار کرنل های RBF، linear، poly، sigmoid را با مقادیر مختلف برای پارامتر C و GAMMA بررسی کرده ایم.

در هر ستون از جدول زیر میتوانید با توجه به پارامتر ها و نوع کرنل امتیاز Svm روی داده های تست را مشاهده نمایید.

ضریب کرنل برای rbf ، polyو sigmoid است.

هرچه مقدار گاما بیشتر باشد، الگوریتم تلاش می‌کند برازش را دقیقاً بر اساس مجموعه داده‌های تمرینی انجام دهد و این امر موجب تعمیم یافتن خطا و وقوع مشکل بیش برازش (Over-Fitting)می‌شود.

در بخش 4 تمرین ها به برسی پارامتر c میپدازیم.

Gamma=scale	C=1	C=0.01	C=100
RBF	0.953	0.581	0.968
linear	0.966	0.968	0.971
poly	0.955	0.763	0.978
sigmoid	0.185	0.24	0.17

Gamma=0.000005	C=1
RBF	0.946
linear	0.966
poly	0.966
sigmoid	0.24

4. برای سوال ۲ سعی کنید مبحث margin soft و margin hard را بررسی کنید و نتایج آن را گزارش دهید

در واقع هایپرپارامتر C مقدار مجازات مدل برای هر نمونه ای را که اشتباه طبقه بندی میکند را تعیین میکند. اگر حاشیه خیلی نرم باشد یعنی سختگیری ما نسبت به میزان خطا کمتر اشد، احتمال Over-Fitting بیشتر میشود اما اگر حاشیه کاملا سخت باشد ممکن است مدل نتواند مرزی تشخیص دهد. میتوانید نتیجه حاصل از تغییر پارامتر C را ببینید.

5. مهندسی ویژگی یکی از بخش های مهم در فرایندهای علم داده میباشد . بر روی دیتاست موارد زیر را اجرا کنید .

(آ) بر روی فیچر power battery از روش binning استفاده کنید . (حداقل سه اندازه مختلف برای بین ها در نظر بگیرید و حتی سایز بین ها را نامساوی در نظربگیرید .)

یه حالت دسته بندی مختلف با انداره های 3 و 9 و29 ساخته شد.

(ب) بر فیچرهای کتگوریکال در دیتاست encoding hot one را اعمال کنید . چرا ما باید به صورت کلی از این کدگذاری بهره ببریم .

encoding hot on را برای فیچر های blue ، dual_sim ،touch_screen،wifi و n_cores اعمال کرده ایم در بخش بعدی نتیجه svm با این تغییرات را بررسی میکنیم.
(ج) بررسی کنید آیا استفاده از تبدیل هایی از قبیل transform log و یا تبدیل نمایی در اینجا کاربرد دارد . به صورت کلی چرا از این دست تبدیلات بهره میبریم . (در این بخش شما مجاز هستید اگر تبدیل دیگری را مناسب میدانید اعمال کنید این بخش
نمره امتیازی برای شما خواهد داشت . حتما دلیل استفاده از تبدیل استفاده شده را بیان کنید .)

(د) یک فیچر جدید به نام مساحت یا حجم گوشی بسازید .

فیچر مساحت ساخته شد اما باعث کاهش قدرت تشخیص اگوریتم شد و به جای آن از روی متغیر طول و عرض تاچ ، قطر آن محاسبه گردید.

6. برای هریک از حالت های سوال ۵ یک مدل SVM بسازید و بررسی کنید یکبار هم هر ۵ حالت را باهم اعمال کنید و مدل SVM روی آنها اجرا کنید . حاصل این مدل ها را گزارش کنید .

(الف) در حالت با سه دسته تغییری در خروجی حاصل نشد و امتیاز الگوریتم 0.953 شد. با حذف فیچر اصلی power battery امتیاز به 0.798 کاهش پیدا کرد. برای دو حالت دسته های 9 و 29 تایی هم تغییری در خروجی حاصل نشد.

(ب) با اعمال encoding hot on روی ویژگی های ذکر شده، امتیاز الگورتم به 0.958 افزایش یافت.

(د) ساخت فچر مساحت باعث کاهش شدید امتیاز الگویتم شد

(ه) در این حالت تمام موارد را با هم اعمال میکنیم و با بهترین کرنل ، امتیاز الگوریتم 0.9716666666666667 بدست آمد.