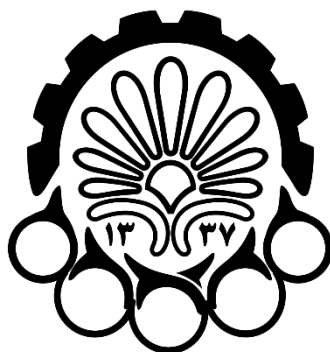


به نام خدا



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

تمرین درس یادگیری ماشین-سری چهارم

فردین آیار

شماره دانشجویی: ۹۹۱۳۱۰۴۰

استاد: دکتر ناظر فرد

دانشکده کامپیوتر- زمستان ۹۹

سوالات تشریحی

(۱)

الف) درست/نادرست؛ ماشین‌های بردار پشتیبان پارامتریک هستند. زیرا برای ساخت مدل، تعداد محدودی پارامتر در نظر می‌گیرد که با افزایش تعداد داده‌های آموزش، افزایش نمی‌یابند. به عنوان مثال در کرنل خطی، بردار وزن w به صورت مجموعه‌ای از ضرایب است که صرف‌نظر از تعداد داده‌ها ثابت است. (هرچند ممکن است تعداد بردارهای پشتیبان افزایش یابد، اما تعداد پارامترها در مدل نهایی ثابت است) در این بین SVM با هسته RBF یک استثنا است؛ RBF داده‌ها را به فضای بی‌نهایت بعدی می‌برد و در آن فضا، همه‌ی داده‌ها می‌توانند در شکل نهایی مدل تاثیرگذار باشند. به بیان بهتر با افزایش تعداد داده‌های آموزش، شکل مدل پیچیده‌تر می‌شود که یعنی تعداد پارامترهای آن افزایش می‌یابد. در این حالت SVM غیرپارامتریک خواهد بود.

ب) نادرست؛ مدلی که دارای هسته‌های متفاوت هستند، در فضاهای متفاوتی قرار دارند و لزوماً حاشیه بیشتر نشان دهنده کارایی بیشتر نیست. به عنوان مثال در صورت وجود داده‌های نویزی، هسته RBF نسبت به هسته خطی احتمالاً حاشیه بیشتری ایجاد کند. اما ممکن است به علت بیش‌برازش کارایی آن کمتر باشد.

ج) نادرست؛ در مدل خطی در صورت تنظیم نبودن پارامتر C (در Soft-Margin SVM) و وجود داده‌های نویزی، SVM دچار بیش‌برازش خواهد شد. همچنین در صورت استفاده از هسته‌های غیرخطی امکان وجود بیش‌برازش وجود خواهد داشت. به عنوان مثال در هسته RBF با کاهش پارامتر سیگما، پتانسیل بیش‌برازش در SVM افزایش خواهد یافت.

د) نادرست؛ همانطور که در مورد قسمت ج گفته شد، در صورت تنظیم نبودن پارامتر C و یا استفاده از هسته‌های غیرخطی، داده‌های نویزی می‌توانند منجر به مدل‌های ناکارآمد شوند.

ه) درست؛ می‌توان ثابت کرد خطای این الگوریتم، به صورت نمایی (نسبت به تعداد گام‌ها t) به صفر میل خواهد کرد.

و) وزن دسته‌بندها در مرحله t از رابطه $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$ بدست می‌آید. این رابطه زمانی منفی می‌شود که $\epsilon_t < 1 - \epsilon_t$ که یعنی خطای مرحله t از ۰.۵ بیشتر شود. می‌دانیم خطای دسته‌بندی‌کننده از ۰.۵ بیشتر نخواهد بود؛ بنابراین وزن‌های اختصاص داده‌شده همواره نامنفی است.

(۲)

در هسته‌ی RBF، پارامتر سیگما نشان‌دهنده‌ی محدوده هر یک از داده‌های آموزش می‌باشد. اگر مقدار سیگما کوچک باشد، مرز تصمیم تنها به نقاطی بستگی دارد در نزدیکی مرز تصمیم قرار دارند. از سوی دیگر بزرگ بودن پارامتر سیگما باعث می‌شود مرز تصمیم به محدوده بزرگتری از نقاط وابسته باشد. به طور خلاصه سیگمای کوچکتر باعث افزایش واریانس و سیگمای بزرگتر باعث افزایش بایاس می‌شود.

با توجه به توضیحات فوق، شکل سمت راست مربوط به سیگما ۰.۲، شکل وسط مربوط به سیگما ۱۰ و شکل سمت چپ مربوط به سیگما ۱ است.

(۳)

در روش hard voting نظر هر دسته‌بندی‌کننده درباره برچسب داده پرسیده می‌شود و سپس بدون در نظر گرفتن احتمالات خروجی، برچسب داده با توجه به نظر اکثریت، تعیین می‌شود.

در روش **soft voting** برخلاف روش قبل، احتمال تعلق داده به هر کلاس از هر دسته‌بندی کننده دریافت می‌شود. برچسب داده برابر است با کلاسی که بیشترین میانگین وزن دار احتمالات را دارد. منظور از ((وزن)) اهمیت هر دسته‌بند است.

(۴)

ابتدا به بررسی روش **hard voting** می‌پردازیم. در این روش احتمالات کلاس‌ها را در نظر نمی‌گیریم و خروجی هر الگوریتم، کلاسی است که بیشترین احتمال را دارد. بنابراین خروجی دسته‌بندهای یک تا ۳ به ترتیب عبارتند از **class1**، **class2** و **class1**. خروجی روش **hard voting** کلاسی است که بیشترین فراوانی را دارد؛ یعنی **class1**.

برای روش **soft voting** میانگین وزن دار احتمالات را برای هر کلاس محاسبه می‌کنیم.

$$\begin{cases} p(x \in \text{class1}) = \frac{2 \times 0.1 + 1 \times 0.6 + 2 \times 0.4}{2 + 1 + 2} = 0.32 \\ p(x \in \text{class2}) = \frac{2 \times 0.5 + 1 \times 0.3 + 2 \times 0.3}{2 + 1 + 2} = 0.38 \\ p(x \in \text{class3}) = \frac{2 \times 0.4 + 1 \times 0.1 + 2 \times 0.3}{2 + 1 + 2} = 0.3 \end{cases}$$

بنابراین داده مربوط به **class2** است.

سوالات پیاده‌سازی

(۱)

کد مربوط به این سوال در فایل **1.py** قرار دارد. داده‌ها را بعد از وارد کردن شافل می‌کنیم و ستون مربوط به ویژگی **name** را، به علت بی ارتباط بودن، حذف می‌کنیم.

۱-۱) برای هر پارامتر ۵ مقدار متفاوت در نظر گرفته شده است. لازم به ذکر است از آنجا که دیتاست در ابتدای کد شافل می‌شود و تعداد نمونه‌های موجود در دیتاست نسبتاً کم است؛ در اجراهای متفاوت ممکن است نتایج متفاوتی بدست آید.

کرل خطی:

accuracy	f1
0.847458	0.901099

کرل چندجمله‌ای:

d	r	accuracy	f1
2	0	0.813559	0.884211
2	1	0.813559	0.884211
2	3	0.813559	0.884211

2	5	0.813559	0.884211
2	10	0.813559	0.884211
4	0	0.813559	0.886598
4	1	0.830508	0.895833
4	3	0.830508	0.895833
4	5	0.830508	0.895833
4	10	0.830508	0.895833
8	0	0.79661	0.875
8	1	0.79661	0.866667
8	3	0.847458	0.894118
8	5	0.847458	0.891566
8	10	0.830508	0.875
10	0	0.79661	0.87234
10	1	0.779661	0.857143
10	3	0.830508	0.878049
10	5	0.830508	0.878049
10	10	0.813559	0.864198
12	0	0.79661	0.87234
12	1	0.79661	0.869565
12	3	0.864407	0.904762
12	5	0.79661	0.85
12	10	0.762712	0.825

مطابق جدول فوق مقادیر بهینه برای کرنل چندجمله‌ای $d=12$ و $r=3$ می‌باشد. در این حالت مقدار Accuracy برابر با 0.8644 و مقدار F1-score برابر با 0.9047 می‌باشد.

کرنل RBF:

gamma	accuracy	f1
0.001	0.762712	0.847826
0.01	0.779661	0.865979
0.1	0.745763	0.851485
1	0.728814	0.843137
10	0.728814	0.843137

بهترین مقدار برای پارامتر گاما 0.01 می‌باشد که مقادیر Accuracy و F1 متناظر با آن 0.7796 و 0.8659 می‌باشد.

کرنل سیگموئید:

r	accuracy	f1
0	0.728814	0.843137

1	0.728814	0.843137
3	0.728814	0.843137
5	0.728814	0.843137
10	0.728814	0.843137

در این حالت همانطور که مشاهده می‌شود مقدار ضریب γ تاثیری روی نتیجه نداشته است و مقدار بهینه وجود ندارد.

(۲-۱)

کرنل چندجمله‌ای:

به طور کلی به ازای مقدار ثابت γ ، افزایش پارامتر d باعث افزایش دقت و سپس کاهش آن می‌شود (به علت بیش‌برازش). از طرف مقابل نیز، افزایش پارامتر γ با ثابت نگه داشتن d ، موجب افزایش و سپس کاهش دقت می‌شود. البته این افزایش برای درجات بالاتر بیشتر مشهود است؛ به طوری که در درجه ۲، تغییر پارامتر γ تاثیر خاصی روی دقت ندارد اما در درجه ۱۲، افزایش γ سبب تغییرات زیادی در دقت می‌شود.

کرنل RBF:

در اینجا منظور از پارامتر گاما، معکوس سیگما است. در سوال ۲ تشریحی توضیح داده شد که با افزایش سیگما، از واریانس بالا به سمت بایاس بالا حرکت می‌کنیم. بنابراین با افزایش گاما، از بایاس بالا به سمت واریانس بالا حرکت می‌کنیم. در این سوال نیز مطابق انتظار، با شروع از مقدار یک هزارم برای گاما، دقت ابتدا افزایش یافته و سپس به علت کم‌برازش، کاهش می‌یابد.

کرنل سیگموئید:

در کرنل سیگموئید پارامتر γ تاثیری روی نتایج این مسئله نداشته است.

(۳-۱) برای این کار می‌توان از روش‌های ارزیابی (مانند cross-validation) و الگوریتم‌های بهینه‌سازی برای جستجوی پارامترهای بهینه استفاده کرد.

(۲)

(۱-۲) کد مربوط به این قسمت در فایل 2.py قرار دارد. دو مدل که بهترین نتیجه را داشته‌اند در جدول زیر به صورت **bold** مشخص شده‌اند. منظور از بهترین مدل، بیشترین دقت روی مجموعه آزمون است.

n_estimators	max_features	max_depth	train accuracy	test_accuracy
10	2	2	0.756052	0.748918
10	2	5	0.86406	0.796537
10	2	8	0.94041	0.796537
10	5	2	0.780261	0.792208
10	5	5	0.860335	0.805195
10	5	8	0.960894	0.787879

10	8	2	0.772812	0.78355
10	8	5	0.873371	0.792208
10	8	8	0.951583	0.800866
100	2	2	0.765363	0.766234
100	2	5	0.865922	0.792208
100	2	8	0.962756	0.809524
100	5	2	0.769088	0.787879
100	5	5	0.865922	0.792208
100	5	8	0.979516	0.800866
100	8	2	0.776536	0.770563
100	8	5	0.867784	0.792208
100	8	8	0.975791	0.805195
1000	2	2	0.761639	0.770563
1000	2	5	0.862197	0.792208
1000	2	8	0.96648	0.809524
1000	5	2	0.765363	0.774892
1000	5	5	0.867784	0.800866
1000	5	8	0.977654	0.805195
1000	8	2	0.769088	0.78355
1000	8	5	0.875233	0.792208
1000	8	8	0.979516	0.796537

۲-۲) با توجه به جدول فوق، پارامتر **max_depth** بیشترین تاثیر را در افزایش دقت دارد و دو پارامتر دیگر تاثیر زیادی در افزایش دقت نداشته اند. به طور کلی همه پارامترهای فوق در حد پایین خود می توانند موجب کم برآزش و در حد بالای خود موجب بیش برآزش شوند.

۲-۳) کد مربوط به این بخش در فایل 2-3.py قرار دارد. برای این بخش از سه الگوریتم متفاوت استفاده می کنیم:

الف) SVM با هسته RBF و گاما ۰.۰۰۰۰۰۹ - دقت بدست آمده روی مجموعه آزمون: ۰.۷۷۹۲

ب) الگوریتم Bagging با ۴۰ دسته بند رگرسیون لاجستیک - دقت بدست آمده روی مجموعه آزمون: ۰.۷۹۲۲

ج) الگوریتم Adaboost با ۵۰ مرحله درخت تصمیم بدون محدودیت عمق - دقت بدست آمده روی مجموعه آزمون: ۰.۷۹۶۵

بنابراین بهترین عملکرد مربوط به مدل ج است.