تمرین سری سوم

درس: یادگیری ماشین- پاییز ۱۴۰۳

استاد درس: دکتر فاطمه زمانی

دستیاران آموزشی: حسین آقاگل زاده، ابوالفضل حسینی فر

دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی نوشیروانی بابل

- این تمرین دارای سه بخش **تئوری**، **پیاده سازی** و **مطالعاتی** است.
- پاسخنامه بخش تئوری می تواند به صورت دست نویس یا تایپی باشد. در هر صورت اما می بایست در قالب **یک فایل pdf** تحویل گردد.
- پاسخ بخش پیاده سازی می بایست حاوی **کد پیاده سازی** شده **به همراه گزارش آن در قالب فایل pdf** باشد. از آوردن کد در گزارش اجتناب کنید (مگر آوردن بخشی از آن از نظر شما ضرورت داشته باشد).
- پاسخ بخش مطالعاتی می بایست به صورت تایپ شده در قالب یک فایل pdf باشد.
- گزارش بخش پیاده سازی می بایست میزان تلاش شما رو با ٤ اولویت مهم **۱- شفافیت ۲- درستی ۳-زیبایی ٤- کوتاهی**، نشان بده
- بخش تئوری بدون استفاده از کد انجام بشه مگر دلیل موجهی وجود داشته باشه.
- کل پاسخ (کد و فایل های pdf) در **یک فایل zip** با فرمت زیر ارسال شود.

HW3_شماره دانشجویی.zip

تئوری

مسئله ۱

یک مسئله دسته بندی (طبقه بندی) دو کلاسه را در نظر بگیرید که نمونه های آموزشی آن به صورت زیر است.

$$\text{Class} - 1: \begin{bmatrix} 1 & 9 \\ 5 & 5 \\ 1 & 1 \end{bmatrix} \quad \text{Class} + 1: \begin{bmatrix} 8 & 5 \\ 13 & 1 \\ 13 & 9 \end{bmatrix}$$

(هر سطر یک نمونه و هر ستون یک ویژگی است)

نقاط را رو نمودار دو بعدی نشان دهید و بدون انجام محاسبات راه حل مناسبی که روش SVM پیدا می کند را به نمایش در بیاورید. بردار های پشتیبان را مشخص کنید.

مسئله ۲

در رابطه با مقادیر مختلفی که variable slack می تواند در روش SVM داشته باشد، با توجه به رابطه زیر بحث کنید. با توجه به مقدار variable slack تخصیص داده شده در فاز آموزش، کدام نقاط به نادرستی طبقه بندی می شوند؟

$$\mathbf{w}^T \mathbf{x}_i + w_0 \geq 1 - \xi_i \quad \text{for} \quad y_i = +1$$
$$\mathbf{w}^T \mathbf{x}_i + w_0 \leq -1 + \xi_i \quad \text{for} \quad y_i = -1$$

- برای پاسخ به این سؤال به کتاب Bishop مراجعه کنید و گزارشی از مبحث مربوطه بنویسید.

مسئله ۳

الف) مفهوم کلی کرنل و دلایل استفاده از روش های مبتی بر کرنل را بیان کنید.

ب) یک کرنل گوسی را با رابطه ی $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2}\right)$ را در نظر بگیرید. اگر با استفاده از تابع نگاشت مربوط با این کرنل، نقاط $x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ و $x_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ (هر سطر مقدار یک ویژگی است) را به فضای جدید انتقال دهیم فاصله این دو نقطه در فضای جدید چقدر خواهد بود؟

مسئله ۴

برای یک کرنل چند جمله ای درجه ۲ با مقدار ثابت ۱ ($K(x,y) = (x^T y + 1)^2$) نشان دهید:

$$K(x,y) = \varphi(x)^T \varphi(y)$$

به صورتی که:

$$\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2).$$

مسئله ۵

الف) برای یک کرنل معتبر، اثبات کنید:

$$K(x,y)^2 \leq K(x,x)K(y,y)$$

ب) مجموعه داده ی $S = \{x_q\}_{q=1}^{Q}$ مفروض است. تحت تبدیل ϕ این نقاط از فضای d_1 بعدی به فضای d_2 بعدی انتقال پیدا کرده اند. اگر میانگین نقاط در این فضای جدید و کرنل مربوط با تبدیل ϕ را به ترتیب با μ_ϕ و K_ϕ نشان دهیم، اثبات کنید:

$$||\mu_\phi|| = \frac{1}{Q}\sqrt{\sum_{m=1}^{Q}\sum_{n=1}^{Q} K_\phi(x_m, x_n)}$$

مسئله ۶

کرنل $K(x,y) = (x^T y + 1)^2$ را در نظر بگیرید. اثبات کنید که فضای جدید ایجاد شده توسط تابع نگاشت این کرنل، یک فضای $\frac{(d+1)(d+2)}{2}$ بعدی است (d تعداد ابعاد فضای اولیه است).

پیاده سازی

مسئله ۱

در این مسئله محدودیت در استفاده از پکیج های آماده ندارید.

مجموعه داده Spam Turkish را در نظر بگیرید. پس از دریافت این مجموعه داده در ابتدا آن را به برنامه اضافه کنید. این دیتاست به عنوان یک مجموعه داده منظم به در قالب 2 ستون است که ستون اول آن یک متن مربوط به ایمیل و ستون دوم نشان دهنده اسپم بودن یا نبودن آن ایمیل است.

https://archive.ics.uci.edu/dataset/530/turkish+spam+v01

الف) برخی از درایه های این مجموعه خالی است (missing value) و این موضوع در مراحل بعدی مشکلاتی را به دنبال خواهد داشت. از این رو سطر های شامل درایه های خالی را پاک کنید. ٪۷۰ از داده ها را به عنوان داده های آموزشی و مابقی را به عنوان داده های آزمایشی در نظر بگیرید.

ب) از انجا که نوع ویژگی ورودی این مجموعه، متنی است و ماشین با عدد کار میکند با کمک کتابخانه nltk هر درایه متنی از این مجموعه را به بردار عددی متناظر تبدیل کنید.

پ) با کمک طبقه بند SVM دقت داده های آموزشی و آزمایشی را با استفاده از هسته ی خطی، با مقادیر پارامتر C زیر بدست آورید و در هر مورد تعداد بردار های پشتیبان مربوط به هر کلاس را نیز ذکر کنید.

C=0.001, 0.01, 0.1, 1, 10, 100, 1000

تحلیلی از نتایج با تمرکز بر ارتباط بین مقادر C، تعداد بردار های پشتیبان، دقت حاصل شده و overfitting ارائه دهید.

ت) با کمک طبقه بند SVM دقت داده های آموزشی و آزمایشی را با استفاده از هسته های چندجمله ای و RBF با پارامترهای مختلف با مقدار C=1 محاسبه کنید.

- پارامتر های هسته چند جمله ای:

درجه: ۱، ۲، ۳

عدد ثابت: ۱+، ۰، ۱-

$$K(x,y) = (a + x^T y)^d$$

- پارامتر های هسته RBF (γ):

1، 1/k، $1/k^2$

K تعداد ویژگی است.

$$K(x,y) = \exp(-\gamma||x-y||^2)$$

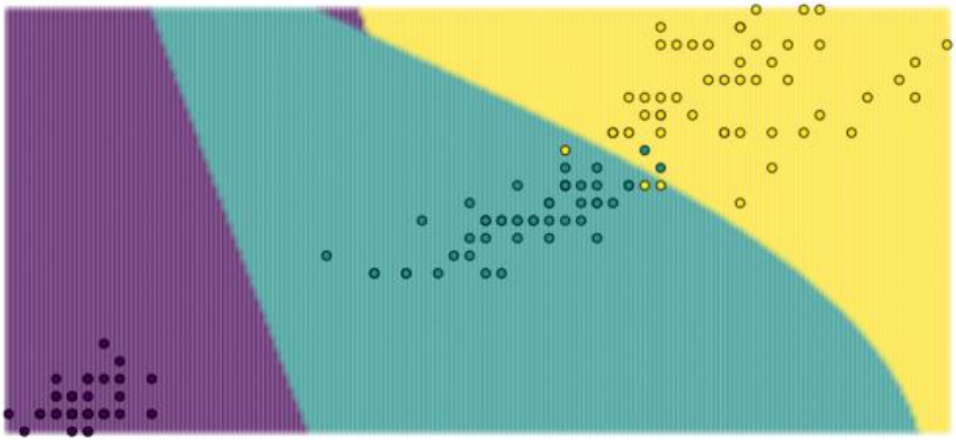
مسئله ۲

در این مسئله محدودیت در استفاده از پکیج های آماده ندارید.

در این مسئله بر روی مجموعه داده iris کار خواهید کرد. بدون تقسیم بندی دادگان به آموزش و تست، با استفاده از SVM در حالت های زیر، دقت را گزارش و ناحیه ی کلاس های مختلف را رسم کنید. برای رسم ناحیه ها فقط از دو ویژگی Petal Length و Petal Width استفاده کنید.

- SVM with linear kernel, one-vs-rest
- SVM with linear kernel, one-vs-one
- SVM with rbf kernel, one-vs-rest
- SVM with polynomial kernel (d=3), one-vs-rest

- نمونه ای از نمایش ناحیه کلاس های مختلف بر اساس دو ویژگی مربوطه:



مطالعاتی

مقاله ی پیوست شده را مطالعه کنید و برداشت خود از آن را شرح دهید (حداقل ۲ صفحه).

- بخشی از مقاله که از قبل در کلاس فرا گرفته اید را کوتاه تر بیاورید.

موفق باشید 😊