بسمه تعالی

دانشکده مهندسی برق و کامپیوتر

تمرین سری چهارم درس یادگیری ماشین

دانشگاه تهران

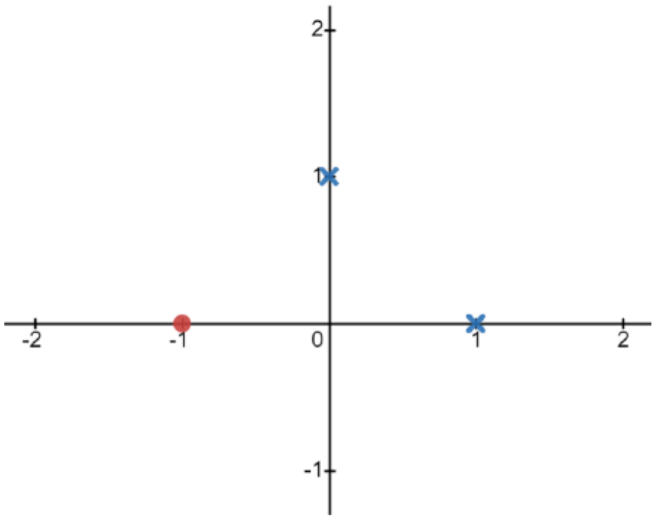
پردیس دانشکده‌های فنی

سلام بر تمام دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده از کنید. شکل ها به طور واضح و در فرمت درست گزارش شوند.
4. از بین سوالات شبیه سازی حتما به هر دو مورد پاسخ داده شود. حداکثر تا نمره ۱۱۰ (۱۰ نمره امتیازی) لحاظ خواهد شد.
5. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر می‌شود.
6. در صورت داشتن سوال، از طریق ایمیل farbodmoosavi@ut.ac.ir، سوال خود را مطرح کنید.

---

1. داده‌های شکل زیر را که به دو کلاس تقسیم شده اند در نظر بگیرید. با نوشتن روابط موجود، معادله‌ی جداساز خطی را براساس SVM برای آن‌ها پیدا کنید. SVها را نیز مشخص کنید. (15 نمره)



---

2. فرض کنید کرنلی به صورت $K(x_i, x_j) = \exp(-\frac{1}{a}\|x_i - x_j\|^c)$ تعریف شده است. نشان دهید به ازای هر دو ورودی دلخواه x_i و x_j در فضای feature space رابطه زیر برقرار است. همچنین در این رابطه، k که یک عدد ثابت است را نیز بیابید. (15 نمره):

$$\|\varphi(x_i) - \varphi(x_j)\|^2 \leq k$$

حال اگر کرنل را به صورت $K(x_i, x_j) = \tanh(ax_i^T x_j + b)$ تعریف کنیم رابطه بالا به چه صورت می شود؟

3. دسته‌ای داده به صورت جدول زیر در اختیار داریم: (20 نمره)

Class	x
+	0
-	-1
-	1

الف) آیا دو کلاس مشخص شده به صورت خطی جداپذیرند؟

ب) فرض کنید هر نقطه از این فضا را به یک نقطه در فضای سه بعدی با تابع φ، که در زیر تعریف شده است، نگاشت کنیم. آیا کلاس‌ها در این حالت جداپذیر خطی هستند؟ در صورت جداپذیر بودن یک صفحه جداکننده را پیدا کنید.

$$\varphi(x) = [1, \sqrt{2}x, x^2]^T$$

ج) یک متغیر برای کلاس ها به صورت $y_i \in \{-1, 1\}$ در نظر بگیرید، که کلاس هرکدام از x_i ها را نشان می‌دهد و داریم $W = (w_1, w_2, w_3)^T$. با یک طبقه بند max-margin SVM مساله زیر را حل کنید:

$$\min_{w,b} \frac{1}{2} \|W\|^2_2 \; s.t.$$

$$y_i(W^T \varphi(x_i) + b) \geq 1, i = 1, 2, 3$$

با استفاده از روش ضرایب لاگرانژ W، b و اندازه حاشیه (margin) را بیابید.

---

4. الف) در مساله Soft Margin SVM شرایط ناتساوی به صورت زیر می‌باشد (20 نمره):

$$x_i.w + b \geq +1 - \xi_i \quad for \quad y_i = +1$$
$$x_i.w + b \leq -1 + \xi_i \quad for \quad y_i = -1$$
$$\xi_i \geq 0 \quad \forall i$$

که در آن $x_i, i = 1, \dots, M$ کل داده‌ها هستند، که در دو کلاس 1+ و 1- قرار دارند. نشان دهید $\sum_{i=1}^{M} \xi_i$ حد بالای خطای طبقه‌بند می‌باشد.

ب) ثابت کنید که برای یک کرنل معتبر رابطه زیر برقرار است:

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

---

5. (شبیه سازی) در این سوال به اعمال طبقه‌بند به کمک Support Vector Machines بر روی مجموعه داده‌ای که با اسم q5.csv است می‌پردازیم. (25 نمره) (در این سوال مجاز به استفاده از کتابخانه ها هستید)

الف) در مورد کرنل‌های rbf و linear و polynomial تحقیق کنید و بیان کنید هر کدام برای طبقه‌بندی چه مجموعه داده‌ای مناسب هستند.

ب) در این مرحله با هر کدام از روش‌های زیر و ویژگی‌های داده شده طبقه‌بندی را انجام دهید و برای هرکدام دقت طبقه‌بندی داده‌های آموزش و تست را در یک جدول گزارش کنید (تا 4 رقم اعشار) و با هم مقایسه کنید.

- SVM with RBF Kernel, C = 1, 100, 1000
- SVM with Linear Kernel, C = 1, 100, 1000
- SVM with Polynomial Kernel, C = 1, 100
- SVM with Sigmoid Kernel, C = 1, 100

پ) در این مرحله قصد داریم که پارامترهای بهینه (Hyper parameter) طبقه‌بند را بیابیم. برای این کار از GridSearch استفاده می‌کنیم. سعی کنید که از بین پارامترهای طبقه‌بند که زیر مشخص شده است با روش گفته شده بهترین پارامترها را بیابید و گزارش کنید. (گزارش پارامترها شامل نوع Kernel، C، gamma می‌باشد.)

- Kernel: RBF, C = [1, 10, 100, 500], gamma = [0.1, 0.3, 0.5, 0.7, 0.9]
- Kernel: Linear, C = [1, 10, 100, 1000]
- Kernel: Polynomial, degree = [2, 3, 4],C = [1, 10, 100, 500], gamma = [0.01, 0.03, 0.05]

ت) بهترین طبقه بند بخش ب و بخش پ را بر روی داده‌ها اعمال کنید و دقت هرکدام و ماتریس کانفیوژن را گزارش کنید و با هم مقایسه کنید.

---

6. (شبیه سازی) در این سوال قصد داریم با استفاده از روش Ensemble – Learning یک دسته داده دو کلاسه را طبقه‌بندی کنیم. (در این سوال مجاز به استفاده از کتابخانه‌ها هستید) (15 نمره)

الف) ابتدا در مورد روش Ensemble-Learning Voting Based کمی توضیح دهید.

ب) دیتاست cancer.csv مربوط به شناسایی سرطان سینه در بیماران است. در ستون اول صرفا یک ID از هر بیمار موجود است. در 9 ستون بعدی هرکدام یک ویژگی عددی برای هر بیمار مشخص شده است و در ستون آخر نیز کلاس ها مشخص شده اند که با دو عدد 2 و 4 که نشان دهنده خوشخیم یا بدخیم بودن غده است مشخص شده‌اند. با استفاده از سه روش Logistic regression، Decision Tree و SVM و استفاده از Voting Based Ensemble سعی کنید این داده‌ها را طبقه‌بندی کنید و دقت طبقه‌بند را بیابید. برای این کار نیاز به جدا کردن داده‌ها به داده‌های آموزش و تست نیست و از روش 10 fold استفاده کنید و میانگین دقت را گزارش کنید.