

تمرین هفتم درس شناسایی الگو

سوال 1 (20 امتیاز)

به سوالات زیر پاسخ کوتاه دهید.

- 1) در مسائل طبقه بندی (classification) چند نوع رویکرد داریم. هر یک را توضیح دهید.
- 2) در صفحه hyperplane که مرز تصمیم LDF می باشد، جهتگیری و مکان (Location, orientation) صفحه چطور تعیین می شود؟
- 3) چرا در مسائل یادگیری به Cost function نیاز داریم؟
- 4) رویکردهای مختلف را در مواجهه با مسائلی که بیش از دو کلاس دارند را نام ببرید و مشکلات هر روش را توضیح دهید.
- 5) راه حلی را برای حل مشکل فضای مبهم در طبقه بندی را ارائه داده و توضیح دهید.
- 6) مفهوم overshoot را توضیح دهید.
- 7) مشکل الگوریتم نیوتون را توضیح دهید.

سوال 2 (30 امتیاز)

داده های هزار نفر از مشتری های یک خرده فروشی در شهر نیویورک در یک فایل CSV برای شما ضمیمه شده است. اطلاعات دیتا به شرح زیر می باشد:

Amount Spent: amount of money that the customer spent in this store.

Catalogs: the category of the item in which the customer has been buy.

Children: number of children that this customer have had.

Gender: gender of the customer.

History: this field regards to the activity of the customer.

Location: distance of the accommodation of the customer to this store.

Married: Marriage of the customer

Own Home: Customer's home is for himself or is rented?

Salary: annual salary of the customer

Age: Middle or young or old?

1. یک طبقه بند SVM طراحی کنید که سن مشتری را با توجه به اطلاعات مختلف وی طبقه بندی کند. در انتهای کد خود، confusion matrix را نیز محاسبه و تحلیل کنید.
 2. حال طبقه بند RBF-kernal SVM را نیز پیاده سازی کرده و نتیجه طبقه بندی را با قسمت 1 مقایسه کنید. (از پیاده سازی های آماده مانند کتابخانه sklearn می توانید استفاده کنید).
 3. تاثیر نرمالایز کردن داده ها را برای دو قبل تست کنید و مقایسه کنید.
- (برای دوستانی که با متلب کار میکنند، از لینک زیر می توانند در پیاده سازی کمک بگیرند.)

(<https://www.mathworks.com/help/stats/support-vector-machine-classification.html>)

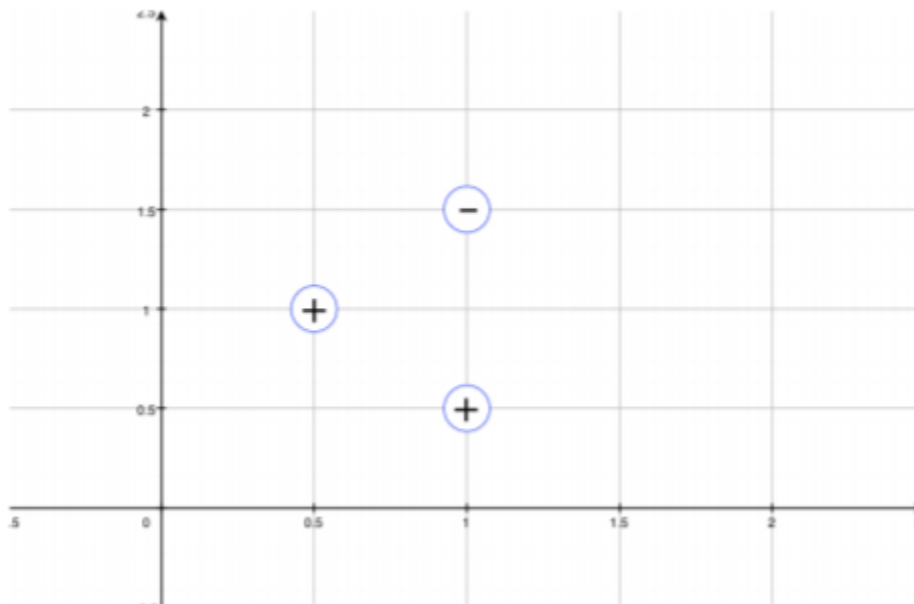
سوال 3 (20 امتیاز)

نشان دهید چنانچه تابع cost function برای soft margin SVM به صورت زیر تعریف شود، آنگاه اینکار می تواند تبدیل به یک مسئله class-separable تبدیل شود.

$$\frac{1}{2}||\mathbf{w}||^2 + \frac{c}{2} \sum_{i=1}^N \varepsilon^2$$

سوال 4 (10 امتیاز)

فرض کنید سه نقطه در یک فضای دو بعدی به صورت شکل زیر وجود دارد.



در این شکل یک linear SVM پیدا کنید که به صورت optimal دو کلاس مثبت ها و منفی ها را جدا کند.
روابط خود را نیز بنویسید.

سوال 5 (20 امتیاز)

در این تمرین می خواهیم با استفاده از linear perceptron، دو نوع مختلف گل را از هم جدا کنیم.

برای دیتاست این سوال، از دیتاست Iris دانشگاه UC-Irvine استفاده شده است که می توانید با استفاده از لینک زیر در مورد آن بیشتر مطالعه کنید. (<https://archive.ics.uci.edu/ml/datasets/iris>)

بخشی از این دیتاست برای شما در فایل iris_dataset.csv انتخاب شده است که شامل petal_length و sepal_length می باشد. با استفاده از اطلاعات طول گلبرگ و کاسبرگ می بایست شبکه عصبی را طراحی و کدنویسی کنید که دو نوع گل را از هم جدا کند. (مسئله دو کلاسه می باشد).

حتما در انتهای کد خود confusion matrix را محاسبه کرده و نتایج خود و خطاهای موجود را تحلیل کنید. چون در غیر این صورت نمره ای به شما تعلق نخواهد گرفت.

سوال 6 (30 امتیاز)

در این سوال می خواهیم با استفاده از دیتاست اصلی Iris مسئله چند کلاسه (Multi class) را با SVM حل کنیم. داده ها را با پسوند data. از لینک سوال قبل (<https://archive.ics.uci.edu/ml/datasets/iris>) دریافت کنید و مسئله طبقه بندی خود را انجام دهید. برای این کار، از دو رویکرد one-vs-rest و one-vs-one استفاده کنید.

حتما در انتهای کد خود confusion matrix را محاسبه کرده و نتایج خود و خطاهای موجود را تحلیل کنید.
چون در غیر این صورت نمره ای به شما تعلق نخواهد گرفت.

سوال 7 (20 امتیاز)

در این سوال میخواهیم یک مسئله Support Vector Regression حل کنیم. دیتاست Boston housing که برای دانشگاه UC Irvine است را برای اینکار در نظر گرفته ایم. داده ها را از لینک زیر می توانید دریافت کنید:

<http://archive.ics.uci.edu/ml/machine-learning-databases/housing/>

در این دیتاست شما با استفاده از 13 ویژگی (feature) مختلف، می بایست قیمت خانه را تخمین بزنید. توضیحات بیشتر با خود دیتاست در لینک فوق با پسوند data و names وجود دارد.

برای اجرای کدها با متلب و پایتون به لینک های زیر مراجعه کنید.

<https://www.mathworks.com/help/stats/support-vector-machine-regression.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

خطای مسئله regression را در انتها گزارش کرده و تحلیل کنید.

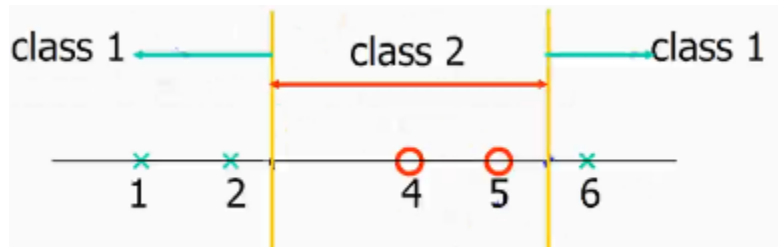
سوال 8 (30 امتیاز)

در اینجا می‌خواهیم با استفاده از الگوریتم LMS که مشابهت زیادی با پرسپترون دارد طبقه بندی را طراحی کنیم. در این سوال استفاده از کتابخانه مجاز نیست و شما باید تمامی مراحل را خودتان پیاده سازی کنید. برای آشنایی شما با این الگوریتم فایل pdf ضمیمه شده است. همچنین منابع بسیاری در اینترنت برای این الگوریتم وجود دارد که می‌توانید استفاده کنید. برای طبقه بندی از دیتاست Iris که در سوال 5 معرفی شد استفاده نمایید.

حتما در انتهای کد خود confusion matrix را محاسبه کرده و نتایج خود و خطاهای موجود را تحلیل کنید.
چون در غیر این صورت نمره ای به شما تعلق نخواهد گرفت.

سوال 9 (20 امتیاز)

فرض کنید در فضای یک بعدی زیر، دو کلاس رنگ سبز و قرمز را دارید. با استفاده از kernel SVM دو کلاس را از هم جدا کرده و support vector ها را به دست آورید. (Support vector ها را با روش SMO به دست آورید.) (در این سوال مجاز به استفاده از هیچ ابزار کتابخانه نیستید و باید خودتان مساله را به صورت کامل پیاده سازی کنید. در غیر این صورت نمره ای به شما تعلق نخواهد گرفت)



نکات تمرین:

- لطفا تمرین کامپیوتری خود را اگر چنانچه با زبان برنامه نویسی پایتون انجام می دهید (با پسوند .py) انجام دهید. (نه jupyter) (همچنین می توانید با Matlab هم پیاده سازی خود را انجام دهید).
- برای تمرین کامپیوتری، کد هر بخش را جدا و به طور مثال برای بخش a به صورت **q4_a.py** نام گذاری کنید.
- از 200 امتیاز کل سوالات، شما چنانچه به 100 امتیاز پاسخ دهید نمره کامل را خواهید گرفت و حداکثر نمره قابل کسب در این تمرین برای شما 150 امتیاز می باشد.
- لطفا نوشته های خود را کاملا خوانا و واضح بنویسید و در نوشتن گزارش برای تصحیح بهتر نوشته ها و روابط خود را در صورت امکان تایپ کنید.
- لطفا برای هر سوال از break page استفاده کنید تا هر سوال از ابتدای صفحه نوشته شود نه از انتهای سوال قبل.
- کدهای خود را در پوشه **codes** ذخیره کنید و گزارش نهایی خود را با فرمت **pdf** قرار دهید. (نه عکس)
- نام فایل نهایی خود را به صورت گفته شده نامگذاری کنید:
(LastName_Name_StudentNumber.zip)
- برای افرادی که کدهای مشابه دارند، باتوجه به policy استاد با افراد متقلب برخورد خواهد شد.

امیدوارم موفق و موید باشید،

باتشکر،

دهقان روزی

Dehghanr.mohammad@gmail.com

Dehghanr.mohammad@ut.ac.ir