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 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 Rezatalakoob@yahoo.com، سوال خود را مطرح کنید.

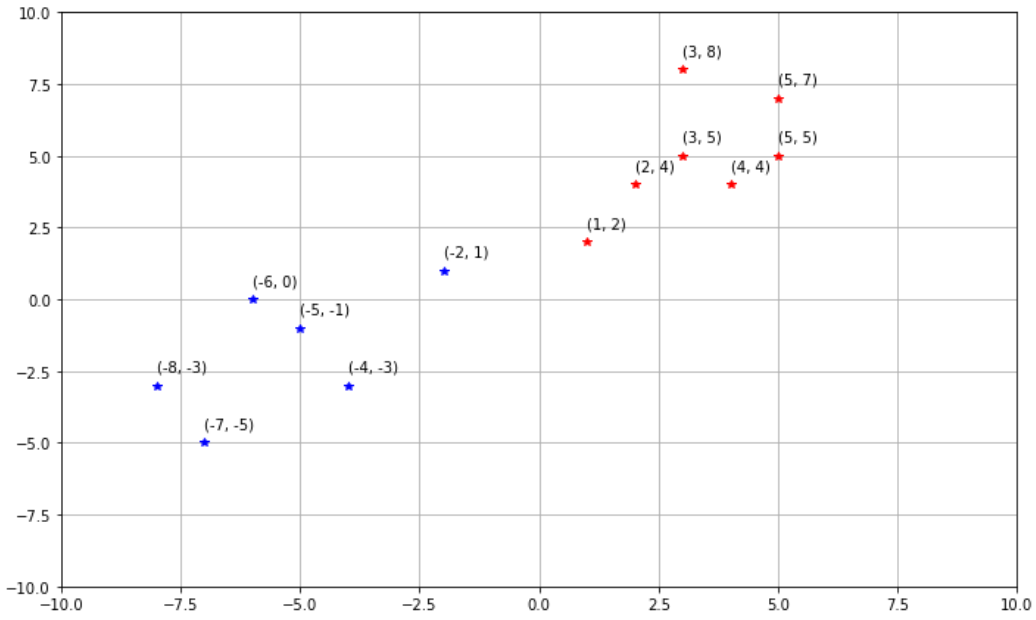
تمرین دوم درس یادگیری ماشین

پاییز ۱۴۰۰

1. سوال اول (۱۵ نمره)

داده های زیر به دو کلاس آبی و قرمز تقسیم شده اند. داده های آبی را کلاس یک و قرمز را کلاس دو در نظر بگیرید. با فرض توزیع گاوسی برای هر کلاس و با در نظر گرفتن ماتریس هزینه ی Λ ، معادله ی مرز تصمیم مابین این دو کلاس را محاسبه کنید. تمامی مقادیر لازم (میانگین ، احتمال پیشین و ...) را از مشاهدات زیر بدست آورید. (بعد از محاسبه ی مرز برای تحقیق درستی جواب خود ، نمودار را در سایت https://www.desmos.com/calculator رسم کنید.)

$$\Lambda = \begin{bmatrix} 1 & 5 \\ 2 & 1 \end{bmatrix} \quad , \quad \Lambda_{ij} = \lambda(\alpha_i \mid \omega_j) \ , \ \alpha_i : deciding \ \ \omega_i$$



شکل ۱) نمودار نقاط دو کلاس آبی و قرمز

2. سوال دوم (۱۵ نمره)

در مسئله طبقه بندی چند کلاسه Multiclass classification :

الف) نشان دهید که قانون تصمیم گیری Bayes احتمال خطا را کمینه می کند. (۸ نمره)

ب) نشان دهید که در حالت M کلاسه حد بالای احتمال خطا بصورت زیر می باشد: (۵ نمره)

$$P_e \leq \frac{M-1}{M}$$

ج) راه حلی برای رسم نمودار ROC در حالت چند کلاسه ارائه دهید (۲ نمره)

3. سوال سوم (۲۰ نمره)

الف) با استفاده از روش ضرایب لاگرانژ، فاصله نقطه ای مانند x_0 را از ابرصفحه ی $w^T x + b = 0$ به دست آورید.

ب) مجددا با استفاده از ضرایب لاگرانژ و ایده ای که از بخش قبل گرفتید، فاصله ی نقطه ی x_0 را از بیضی گون $x^T Ax = 1$ بیابید. اگر به نظرتان فرم بسته ای به عنوان راه حل موجود نیست، سعی کنید روشی را برای حل عددی این سوال ارائه بدهید. هرچند بایستی تا جای ممکن روابط موجود را ساده کرده و سپس این روش را ارائه دهید.

ج) برای شرایط زیر، مختصات نزدیکترین نقطه روی بیضی گون به نقطه ی داده شده را حساب کنید. در اینجا از روش تحلیلی (و نه روش عددی که در بخش قبل به دست آوردید) استفاده کنید. (برای محاسبات سنگین می توانید از ابزارهای حل معادله استفاده کنید و برای تحقیق پاسخ خود می توانید نمودارهای مربوطه را در سایت desmos رسم کنید.)

A) $x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

B) $x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, A = \begin{bmatrix} 2 & 3 \\ 1 & 5 \end{bmatrix}$

در این سوال تنها مجاز به استفاده از روش های بهینه سازی مبتنی بر ضرایب لاگرانژ هستید. استفاده از سایر روش ها نمره ای نخواهد داشت.

4. سوال چهارم (۲۰ نمره)

الف)مسئله دو کلاسه که $p(x|w_1)$ توزیع $N(\mu, \sigma^2)$ دارد و $p(x|w_2)$ توزیع یکنواخت بین a و b دارد را در نظر بگیرید. نشان دهید که احتمال خطای تصمیم گیری Bayesian حد بالایی بصورت $G(\frac{b-\mu}{\sigma}) - G(\frac{a-\mu}{\sigma})$ دارد که G(.) همان توزیع نرمال $N(0, 1)$ می باشد.

ب) فرض کنید که فضای ویژگی x دارای ستون های کاملا مستقل از هم باشد. با توجه به فرض استقلال ویژگی ها در طبقه بند naïve Bayes آیا این طبقه بند عملکرد بهینه بر روی این داده خواهد داشت یا روش های دیگر می توانند به جواب بهتری دست پیدا کنند.؟ علت جواب خود را توضیح دهید.

5. سوال پنجم (20 نمره)

الف) برای عبارات زیر در صورت درستی اثبات آورده و در صورت غلط بودن مثال نقضی بیاورید.(۴ نمره)

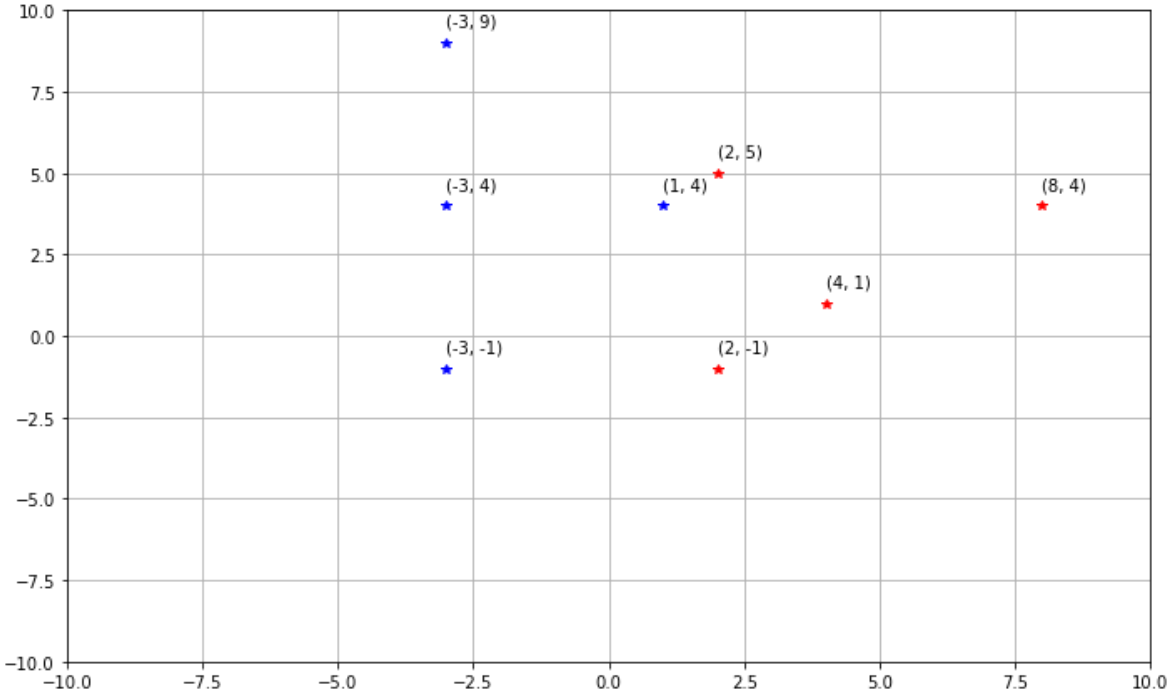
a) If P(a|b,c) = P(b|a,c) , then P(a|c) = p(b|c)

b) If P(a|b) = P(a) , then P(a|b,c) = p(a|c)

ب) بعد از چکاپ سالانه دکتر به شما می گوید که یک خبر بد و یک خبر خوب دارد. خبر بد این است که تست شما برای یک بیماری حاد ، مثبت بوده است در حالیکه این تست ۹۹ درصد دقیق می باشد. (بدین معنی که احتمال مثبت شدن تست هنگامی که شخصی بیماری را دارد ۰,۹۹ هست و هم چنین احتمال منفی شدن تست وقتی که شخص بیمار نباشد نیز ۰,۹۹ است.) خبر خوب اما این است که این یک بیماری نادر است که تنها یک نفر از هر ۱۰۰۰۰ نفر در سن شما آن را می گیرد. خبر نادر بودن این بیماری چگونه به شما کمک می کند. احتمال اینکه شما واقعا به این بیماری مبتلا شده باشید چند درصد است ؟ نتیجه را تحلیل نمایید. (۵ نمره)

ج) در مسئله ی دسته بندی دو کلاسه با توجه به شکل 2 ، کلاس مربوط به نقطه ی $x = [2 , 2]^T$ ، تحت روشهای نزدیکترین همسایه و نزدیکترین میانگین و با استفاده از معیارهای فاصله ی ذکر شده در جدول زیر، کدام یک از کلاس های آبی یا قرمز خواهد بود؟ (۷ نمره)

	1- Nearest Neighbor	Nearest Centroid
$d_1(x,y) = \max \lvert x_i - y_i \rvert$		
$d_2(x,y) = \sum_{i=1}^{d} \lvert x_i - y_i \rvert$		
$d_3(x,y) = \sum_{i=1}^{d} (x_i - y_i)^2$		



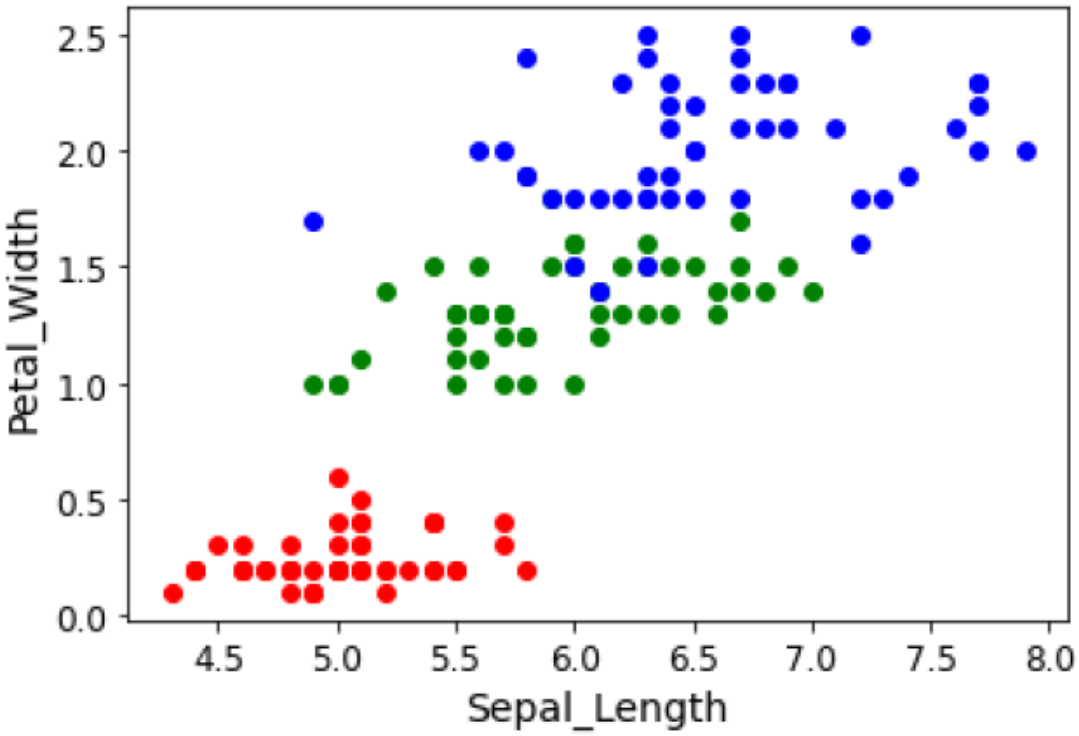
شکل ۲) فضای ویژگی

د) راجع به متد Generalized Linear Regression و ارتباط آن با الگوریتم های Linear regression و logistic regression توضیح دهید.(۴ نمره)

6. (شبیه‌سازی) (۲۰ نمره)

در این سوال بر روی دیتاست Iris ضمیمه شده کار خواهید کرد.. برای دو بخش ابتدایی امکان استفاده از پکیج های یادگیری ماشین را ندارید.

الف) ابتدا برای درک بهتر این دیتاست نمودار نقاط آن را بر حسب هر دو تایی از ویژگی ها رسم کنید. حتما اسامی ویژگی ها را یر روی نمودار مشخص کنید(مطابق شکل ۳) . حال از بین این نمودار ها مشخص کنید که یک طبقه بند خطی بر حسب کدام ویژگی می تواند با دقت بیشتری کلاس ها را جدا نماید.



شکل ۳) نمونه ی یکی از دوتایی های دیتاست Iris

ب) داده ها را با نسبت مشخص به تست و ترین تقسیم کنید. یک طبقه بند نزدیک ترین به میانگین را پیاده سازی کرده و داده ی تست را کلاس بندی نمایید. دقت طبقه بند و confusion matrix را برای آن گزارش نمایید

ج) تک تک گام های قبل (جداسازی داده ، پیاده سازی طبقه بند و ...) را توسط پکیج های آماده ی scikit-learn انجام دهید و نتایج را گزارش نمایید. هم چنین به کمک این پکیج ها نمودار ROC (برای هر کلاس در یک نمودار) رسم کرده و مساحت سطح زیر آن را نیز گزارش نمایید.

7. (شبیه‌سازی) .(۲۰ نمره)

در این سوال بر روی دیتا ست Breast_cancer_data کار خواهید کرد. در بخش الف سوال از پکیج های آماده یادگیری ماشین استفاده نکنید و الگوریتم را پیاده سازی کنید.

الف) ابتدا مختصرا راجع به الگوریتم های naïve bayes و optimal bayes توضیح دهید. در ادامه الگوریتم naïve bayes را با فرض گاوسی بودن داده پیاده سازی کنید.

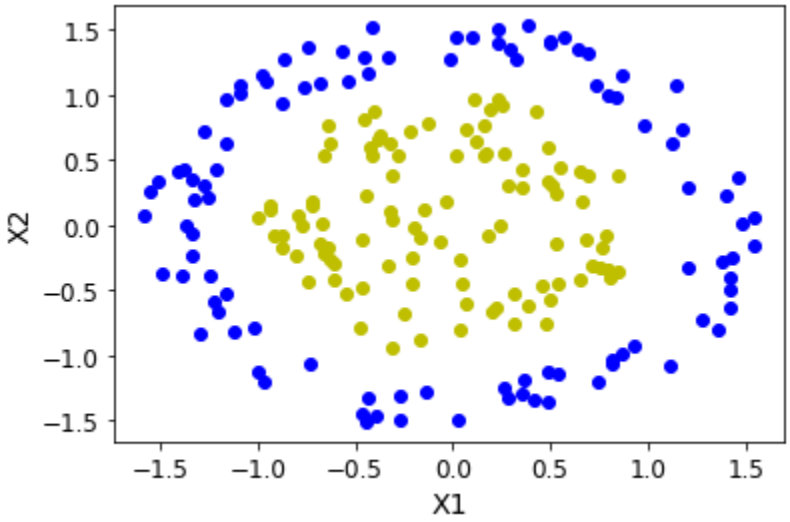
ب) حال الگوریتم پیاده سازی شده را بر روی مجموعه ی داده (پس از پیش پردازش های معمول) تست کرده و نتایج را گزارش کنید.(دقت ، کانفیوژن ماتریکس و ...)

ج) در این قسمت به کمک کتابخانه scikit-learn دو الگوریتم ذکر شده را بر روی داده اعمال کرده و نتایج را با بخش قبلی مقایسه نمایید.

8. (شبیه‌سازی) .(۲۰ نمره)

در این سوال بر روی دیتاست Class_A کار خواهید کرد. همانطور که در شکل ۴ مشخص هست جداسازی خطی این داده در دو بعد با یک خط ، خطای زیادی خواهد داشت. بنابرین از Nonlinear کلاسیفایر ها استفاده می کنیم.

(در این سوال در صورت استفاده از پکیج های آماده یادگیری ماشین **نصف نمره** را دریافت می نمایید.)



شکل ۴) تصویر داده در دو بعد

الف) با استفاده از الگوریتم Logistic Regression دو کلاس این مجموعه داده را تحت f زیر جداسازی نموده و نتایج این کلاسیفایر(دقت ، ماتریس آشفتگی و ...) را گزارش نمایید.(فرض کنید که برای داده ی X داشته باشیم $X: (x,y)^T$)

$$f(x, y; c) = c_0 + c_1 x + c_2 y + c_3 x^2 + c_4 xy + c_5 y^2$$

ب) برای تابع f غیر خطی فوق مرز تصمیم گیری را در فضای دو بعدی مشخص کرده و رسم نمایید