۱. رگرسیون (Regression) :

۱.۱. رگرسیون خطی (Linear regression):

در این بخش می‌خواهیم روی یک مجموعه داده که مربوط به ویژگی‌های تعدادی خانه و قیمت آن‌ها است عمل رگرسیون خطی انجام دهیم. ابتدا برای دریافت این مجموعه داده به لینک زیر مراجعه کنید:

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

این مجموعه داده از ۱۴ ستون تشکیل شده‌است که ستون آخر بیانگر قیمت‌خانه و ۱۳ ستون دیگر بیانگر ویژگی‌هایی مانند تعداد اتاق، تعداد جرایم رخ داده در منطقه و ... است. (برای اطلاعات بیشتر در مورد مجموعه داده می‌توانید فایل housing.names را در لینک بالا مطالعه کنید.) هدف ما در در این مسئله این است که با توجه به ویژگی‌های موجود برای یک خانه از طریق fit کردن یک خط روی داده‌های موجود، قیمت آن را با تقریب خوبی محاسبه کنیم.

۱.۱.۱. رگرسیون تک متغیره :(univariate regression)

در این بخش برای انجام پیش‌بینی، قیمت خانه را تنها به یکی از ویژگی‌های مربوط به آن وابسته در نظر می‌گیریم. در اینجا هدف ما مانند سایر مسائل بهینه سازی، کمینه کردن تابع هزینه است. در تمامی بخش‌های این سوال تابع هزینه را تابع MSE(mean square error) در نظر بگیرید.

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

که در مدل خطی در این مسئله $h_\theta(x)$ مطابق زیر خواهد بود:

$$h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x_1$$

حال برای بهینه سازی و یافتن بهترین مقدار برای ضرایب θ_0 و θ_1 از روش گرادیان کاهشی (Gradient descent) که در مباحث مربوط به شبکه‌های عصبی با آن آشنا شدید استفاده کنید. در هر گام گرادیان کاهشی، پارامترهای Θ به مقادیر بهینه نزدیکتر میشوند تا جایی که میزان هزینه $J(\theta)$ به کمترین مقدار خود برسد.

اولین و سومین ویژگی موجود در مجموعه داده‌ها که به ترتیب بیانگر تعداد جرایم رخ داده‌شده در منطقه خانه مورد نظر و مقدار مالیات مربوط به ملک مورد نظر است را در نظر بگیرید:

۱- ابتدا نمودار قیمت خانه را برحسب هر یک از این دو ویژگی به صورت جداگانه رسم کنید.

۲- با استفاده از روش گرادیان کاهشی و تابع هزینه‌ای که در بالا توضیح داده شد برای هر یک از این دو ویژگی پارامترهای θ را طوری پیدا کنید که با استفاده از خط پیدا شده بتوان قیمت ملک را با تقریب خوبی پیش بینی کرد. خط محاسبه شده را نیز در کنار داده‌هایی که در بخش ۱ رسم کرده‌اید قرار دهید. (فرمول‌های به دست آمده از روش گرادیان نزولی را نیز در گزارش خود ذکر کنید.)

۳-برای هر دو این ویژگی‌ها تغییرات تابع هزینه در طول مراحل الگوریتم بهینه‌سازی رسم کنید.

۴-مقدار خطا میانگین را برای تمامی نمونه‌ها به ازای هر دو ویژگی تعیین شده محاسبه کنید.

۱.۱.۲ رگرسیون خطی چندمتغیره : (multivariate linear regression)

در این بخش قصد داریم تا با توجه به تمام ویژگی‌ها مربوط به یک نمونه از داده، قیمت مربوط به آن داده را پیش‌بینی کنیم. تابع خطا و الگوریتم مورد نظر همانند قسمت بالا است با این تفاوت که در این بخش داریم :

$$h_\Theta(x) = \Theta^T x = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \cdots + \Theta_{13} x_{13}$$

که در اینجا x1 به معنای مقدار ویژگی اول برای نمونه ورودی x است.
۱-با استفاده از روش گرادیان نزولی یک عمل رگرسیون خطی روی داده‌های موجود انجام دهید. (فرمول‌های به دست آمده از روش گرادیان نزولی را نیز در گزارش خود ذکر کنید.)
۲-تغییرات تابع هزینه را در طول مراحل الگوریتم بهینه‌سازی رسم کنید.
۳-نمودار مقدار پیش‌بینی شده برای قیمت ها را برحسب قیمت واقعی خانه رسم کنید.
۴-مقدار خطا میانگین را برای تمامی نمونه‌ها به ازای هر دو ویژگی تعیین شده محاسبه کنید.
۵-با استفاده از روش L2 norm به روش بهینه سازی خود regularization را نیز اضافه‌کنید. مقدار میانگین خطا روی داده‌های موجود را در این روش و در بخش قبل با یکدیگر مقایسه کرده و نتیجه این مقایسه را تحلیل کنید.

۱.۲. رگرسیون لجستیک (Logistic regression):

در این بخش می‌خواهیم با استفاده از روش Logistic Regression یک طبقه‌بند طراحی کنیم. داده‌ی مورد استفاده در این بخش در کنار صورت پروژه قرارداده شده‌است. در این مجموعه داده مقدار ۲ فاکتور A و B در خون افراد مختلف سنجیده شده‌است (ستون اول و دوم) و در ستون سوم مبتلا بودن فرد به نوعی خاص از بیماری آمده است. مقدار ۱ در این ستون به این معنا است که فرد مورد نظر این بیماری را دارد. در این بخش هدف ما این است که با استفاده از این مجموعه داده یک مدل طبقه‌بند طراحی کنیم که با دریافت مقدار این فاکتورها در خون بتواند پیش‌بینی کند که آیا فرد بیمار است یا خیر.
۱- ابتدا می‌خواهیم داده را به تصویر بکشیم تا درک بهتری از آن داشته باشیم. محورهای نمودار باید میزان فاکتورهای موجود در خون افراد باشند. بیمار بودن یا نبودن یک فرد را با نشانگرهای مختلف روی نمودار مشخص کنید.
۲-حال با توجه به توضیحات زیر مدل طبقه‌بند را پیاده‌سازی کنید:
می‌دانیم که فرضیه رگرسیون لجستیک مطابق زیر تعریف می‌شود:

$$h_\theta(x) = \text{sigmoid}(\theta^T x)$$

در پیاده‌سازی خود باید تابعی بنویسید تا مقادیر هزینه و گرادیان را بازگرداند. همان‌طور که می‌دانید تابع هزینه در رگرسیون لجستیک به شکل زیر تعریف می‌شود:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

حال باید با استفاده از روش گرادیان نزولی مقدار بردار θ را به گونه‌ای پیدا کنیم که مقدار تابع هزینه به ازای آن کمینه شود. فرمول‌های محاسبه شده بر اساس روش گرادیان نزولی را در گزارش خود بیاورید.

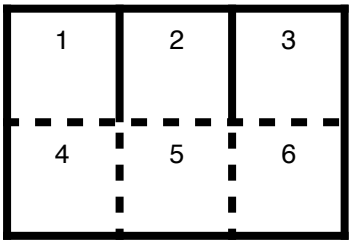
۳- با استفاده از روش L2 norm با تغییر تابع هزینه regularization را نیز به طبقه‌بند خود اضافه کنید.
۴-تابعی بنویسید که یک مجموعه داده و یک بردار پارامتر θ را به عنوان ورودی بگیرد و بیمار یا بیمار نبودن افراد را به ازای این مجموعه داده تعیین کند.

۱.۳. SVM (Support Vector Machine) :

در کلاس درس با SVM و مفاهیم آن آشنا شدیم. در این بخش از شما خواسته شده تا با استفاده از این روش یک بار دیگر یک مدل طبقه‌بند روی داده‌ها مربوط به بیماری که در بخش قبل از آن استفاده کردید آموزش دهید. در ابتدا ۱۰٪ از داده موجود را به عنوان داده تست انتخاب کنید. فاز آموزش را روی ۹۰٪ بقیه داده‌ها انجام داده و سپس از این ۱۰٪ برای سنجش دقت مدلتان استفاده کنید. روش کار خود در این بخش را به طور کامل در گزارش خود توضیح داده وسپس این روش را با Logistic Regression که پیش‌تر با آن آشنا شدید و هر دو برای classification کاربرد دارند مقایسه کرده و مزایا و معایب آن‌ها را نسبت به هم بیان کنید.

۲. Q-Learning

تمرین کامپیوتری اول را به یاد بیاورید، هدف از انجام آن تمرین پیاده‌سازی الگوریتم هاي جستجویی که در درس یاد گرفته اید را براي حل کردن یک ماز (Maze) بود. در این تمرین نیز هدف ما حل یک ماز است اما تفاوتی که وجود دارد این است که این بار از پیش ساختار شکل ماز را نمی‌دانیم. به عنوان مثال در شکل زیر اگر عامل(agent) در خانه شمار ۱ قرار داشته باشد، تا وقتی که حرکت به راست را انجام ندهد (اقدام (action) مربوط به حرکت سمت راست را انتخاب نکند) از وجود دیوار بین دو خانه ماز بی‌اطلاع خواهد بود.



برای حل این مسئله قصد داریم از الگوریتم Q-Learning استفاده کنیم. در اینجا مدل مسئله تشکیل شده از یک عامل، وضعیت ها S و مجموعه از اقدامات A برای هر وضعیت.با انجام یک اقدام $a \in A$، عامل از یک وضعیت به وضعیت بعدی حرکت کرده و هر وضعیت پاداشی به عامل می دهد. هدف عامل حداکثر کردن پاداش دریافتی کل خود است. این کار با یادگیری اقدام بهینه برای هر وضعیت انجام می گردد. الگوریتم دارای تابعی است که ترکیب <state, action> را محاسبه می نماید :

$$Q : S \times A \rightarrow \mathbb{R}$$

قبل از شروع یادگیری، Q مقدار ثابتی را که توسط شما انتخاب شده برمی گرداند. سپس هر بار که به عامل پاداش داده می شود، مقادیر جدیدی برای هر ترکیب <state, action> محاسبه می گردد. هسته الگوریتم از یک بروز رسانی تکراری ساده تشکیل شده است. به این ترتیب که بر اساس اطلاعات جدید مقادیر قبلی اصلاح می شود.

$$Q(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha_t(s_t, a_t)}_{\text{learning rate}} \times \left[\overbrace{\underbrace{R(s_t)}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \underbrace{\max_{a_{t+1}} Q(s_{t+1}, a_{t+1})}_{\text{max future value}}}^{\text{learned value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right]$$

از فایل‌های مربوط به تمرین کامپیوتری اول برای تولید ماز و دیدن رفتار agent استفاده کنید. انتخاب reward function به عهده خودتان است حداقل دو مورد را به همراه نتایج بدست آمده در گزارش بیاورید و علت استفاده از هرکدام را توضیح دهید. درمورد این که reward function های مختلف چه تاثیری روی رفتار agent می‌گذارند خوب فکر کنید، هنگام تحویل سوالاتی در این رابطه از شما پرسیده خواهد شد. همچنین در گزارش خود درباره تاثیر این توابع و پارامترهای موجود در الگوریتم یادگیری بر روی عملکرد agent توضیح دهید.

۳. خوشه‌بندی (Clustering)

۳.۱. خوشه بندی با الگوریتم K-means:

در این بخش قصد داریم الگوریتم خوشه‌بندی k-means را پیاده‌سازی کنیم. همان‌طور که می‌دانید الگوریتم‌های خوشه‌بندی از دسته الگوریتم‌های یادگیری بدون نظارت هستند و سعی دارند با مینیمم کردن تابع هزینه تعریف‌شده برای آن‌ها به یک خوشه‌بندی بهینه دست پیدا کنند. در اینجا تابع هزینه‌ای که برای این الگوریتم تعریف می‌شود به‌صورت مجموع فاصله نمونه‌های متعلق به هر خوشه تا مرکز آن خوشه است.

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

رابطه ۱:تابع هزینه برای الگوریتم k-means: m تعداد نمونه‌ها بوده و $\mu_{c(i)}$ مرکز خوشه‌ی است که داده i ام به آن تعلق دارد

مجموعه داده‌ای که قرار است در این بخش عمل خوشه‌بندی را روی آن انجام دهید اطلاعات مربوط به تعدادی ماشین است که ویژگی‌های مختلفی از آن‌ها مانند سال ساخت، وزن، قدرت موتور و ... در اختیار شما قرار گرفته است. شما باید با توجه به این اطلاعات روی این داده‌ها الگوریتم k-means را انجام دهید و این ماشین‌ها را به تعدادی خوشه تقسیم کنید.

۱) تعداد تکرار برای الگوریتم k-means را برابر با ۲۰۰ در نظر بگیرید و مقادیر مختلف {2 , 3 , 5 ,6} = k را برای برای تعداد خوشه‌ها در نظر بگیرید. با استفاده از معیار شباهت درونی و بیرونی بهترین عدد برای تعداد خوشه‌ها روی این داده را محاسبه کرده و نوع محاسبه شباهت درونی و بیرونی و نتایج محاسبات برای مقادیر مختلف k را در گزارش خود بیاورید.

۲)الگوریتم k-means را ۳ مرتبه با شرایط اولیه تصادفی متفاوت اجرا کرده و مقدار تابع هزینه برای مقادیر مختلف k که در سوال بالا گفته شد رسم کنید.

۳.۲. خوشه بندی فازی با الگوریتم Fuzzy C-means:

در کلاس با مفاهیم منطق فازی آشنا شدید. همان‌طور که می‌دانید این مفهوم در خوشه‌بندی نیز قابل پیاده‌سازی است. با این مفهوم که به‌جای اینکه در خوشه بندی هر نمونه را تنها به یک خوشه نسبت بدهیم به عضویت هر یک از داده‌ها به هر خوشه احتمالی نسبت میدهیم. الگوریتم fuzzy c-means یکی از روش‌های پیاده‌سازی خوشه‌بندی فازی است.

ابتدا الگوریتم fuzzy c-means را روی مجموعه داده cars با بهترین تعداد خوشه‌ای که در بخش قبلی به دست آورید اجرا کنید و نتیجه را در قالب یک ماتریس ارائه کنید سپس در مورد خوشه بندی فازی به سوالات زیر پاسخ بدهید:

۱-فرض کنید در بین داده‌های موجود برای خوشه‌بندی تعداد بسیار محدودی از داده‌ها از پیش label داشته باشند. به نظر شما توجه به این مسئله چگونه می‌تواند در پیاده‌سازی الگوریتم fuzzy c-means و تابع هدف آن تاثیر داشته‌باشد؟

۲-فرض کنید می‌خواهیم فقط تعداد محدودی از داده‌ها در خوشه‌بندی شرکت داده‌شوند. مثلا در مثال بالا میخواهیم فقط داده‌هایی که سال ساخت آن‌ها نزدیک به 1985 است خوشه بندی شوند. با توجه به این که "نزدیک بودن سال ساخت به ۱۹۸۵" هم یک جمله فازی است، به نظر شما چگونه می‌توانیم این مسئله را در پیاده‌سازی الگوریتم fuzzy c-means تاثیر دهیم؟

۳-فرض کنید از پیش از روی دامنه مسئله مورد بررسی، در مورد میزان شباهت یا تفاوت بعضی از نمونه‌ها اطلاعاتی داریم. به نظر شما توجه به این مسئله چگونه می‌تواند در پیاده‌سازی این الگوریتم و تابع هدف آن تاثیر داشته باشد؟

نکات پایانی:

- در این پروژه مجاز به استفاده از کتابخانه های آماده (Tensorflow, PyTorch, Scikit-learn,...)نیستید.
- موارد ذکر شده در صورت پروژه و جواب سوالات تئوری را حتما در فایل گزارش بیاورید.
- جواب سوالات تئوری را در فایل گزارش به همراه پروژه آپلود کنید.
- این تمرن باید در قالب گروه‌های دو نفره انجام شود
- با هر گونه تقلب با جدیت برخورد خواهد شد. استفاده از پیاده‌سازی‌های موجود در اینترنت نیز تقلب محسوب می‌شود.
- از شما انتظار می‌رود به مفاهیم مطرح شده مسلط باشید، هنگام تحویل سوالاتی از این مفاهیم از شما پرسیده می‌شود.
- ایمیل طراحان:

saghar.talebipoor@gmail.com

omrani.ali.96@gmail.com