
بسم الله الرحمن الرحيم

دانشکده مهندسی برق

عنوان:

پیش‌بینی قیمت خودرو

نام استاد : جناب آقای دکتر عبادالهی

نام نگارنده: ارشیا گشتاسبی

آبان ۱۴۰۰

چکیده

افزایش بی‌سابقه حجم داده‌های موجود در ارتباط با یک پدیده، موجب این شده است که به سادگی گذشته نتوان ارزش آن را معلوم کرد. در گذشته‌ی نه چندان دور که رایانه‌ها قدرت پردازشی کمتری داشتند، امکان استفاده از الگوریتم‌های یادگیری ماشین امکان‌پذیر نبود. درحالی که امروزه، با افزایش قدرت پردازش رایانه‌ها و همچنین افزایش داده‌های موجود، این امکان فراهم است تا بتوان ارزش یک پدیده را با استفاده از هوش مصنوعی پیش‌بینی کرد. از این رو هدف این پروژه بررسی الگوریتم‌های یادگیری ماشین و انتخاب بهترین آن‌ها برای پیش‌بینی قیمت و ارزش یک پدیده می‌باشد. در انتها نیز با استفاده از کتابخانه‌های پایتون از صفحه‌های وب، داده‌ها دریافت می‌شود و قیمت یک خودرو را با توجه به ویژگی‌هایش پیش‌بینی می‌شود.

واژه‌های کلیدی: یادگیری ماشین/ الگوریتم پیش‌بینی/ رگرسیون خطی/ شبکه عصبی/ مدل‌های خطی/ پایتون/ استخراج داده

Machine Learning/ Predictive Algorithms/ Linear Regression/ Neural Network/ Linear Models/ Python/ Data Mining

.

فهرست مطالب

عنوان صفحه

فهرست تصاویر و نمودارها

عنوان صفحه

فهرست جدول‌ها

عنوان صفحه

-

فصل ۱: مقدمه

۱.۱ مقدمه

با افزایش چشـــم‌گیر قدرت پردازش رایانه‌ها، بهره بردن از الگوریتم‌های یادگیری ماشـــین که عموما از محاسبات زیادی برخوردار هستند، عملی‌تر شده است و به طور روزافزون استفاده‌ی بیشتری از آن صورت می‌پذیرد. نکته‌ی قابل توجه آن است که هر یک از این الگوریتم‌ها کاربرد و استفاده‌ی مخصوصی دارند و نمی‌توان یک الگوریتم کلی را برای همه‌ی مســـائل به کار برد. از این‌رو، بررســـی دقیق هر یک از این الگوریتم‌ها و توجه به اینکه هر کدام در چه حوزه‌ای کاربرد دارند بسیار اهمیت دارد.

یکی از کاربردهای این الگوریتم‌ها در پیش‌بینی است. این پیش‌بینی می‌تواند پیش‌بینی قیمت، پیش‌بینی فاصله، پیش‌بینی ساعت باشد که پیش‌بینی یک عدد حقیقی است. همچنین می‌تواند پیش‌بینی کلاس یک پدیده باشد، مانند پیش‌بینی محتوایات یک تصویر، پیش‌بینی اسپم بودن یا نبودن ایمیل که یک امر نسبی است باشد.

یکی از موضـــوعاتی که می‌تواند مورد بررســـی قرار گیرد، پیش‌بینی قیمت خودرو دســـت دوم اســـت. سایت‌های متفاوتی به دغدغه اختصاص داده شده‌اند که فروشندگان و خریداران می‌توانند با مراجعه به آن‌ها اطلاعات زیادی راجب خودرو خود به دست آورند. یکی از راهکارهایی که برای استفاده از این داده‌ها موجود است، استفاده از این الگوریتم‌های پیش‌بینی است که توسط آن بتوان ارزش یک خودرو دست دوم را پیش‌بینی کرد. این اتفاق موجب این می‌شـــود که بتوان ارزش عادلانه‌ی یک خودرو را جداگلنه از هر گونه ویژگی‌های منطقی مانند احساسات، پیدا کرد.

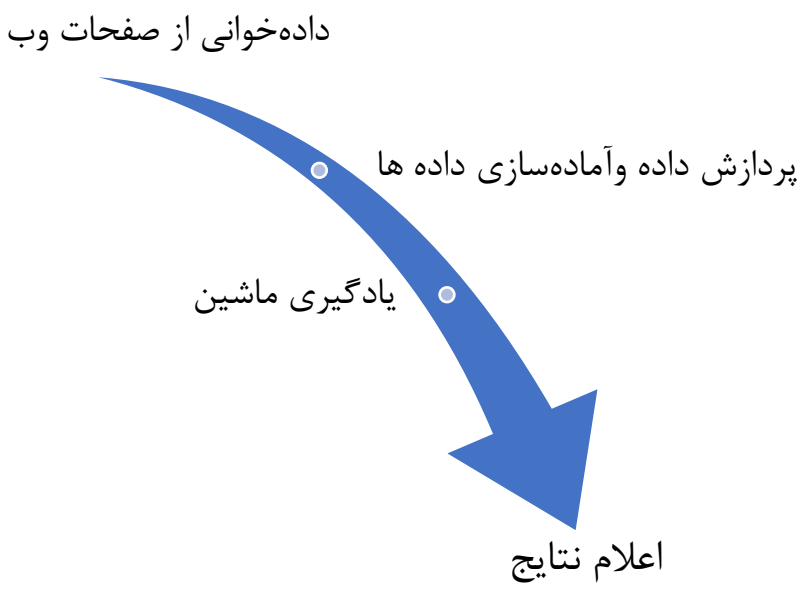
دســـته‌بندی‌های مختلفی برای الگوریتم‌های یادگیری ماشـــین موجود اســـت. یکی از رایج‌ترین آن‌ها تقسیم‌بندی آن‌ها به دو دسته‌ی نظارت شده و نظارت نشده است. در دسته‌ی اول، ماشین پاسخ داده‌هایی که به آن تحویل داده شـده اسـت را در اختیار دارد و آن را بررسـی می‌کند. در مقابل، در دسـته‌ی دوم، چیزی تحت عنوان پاسخ تعریف نشده است و ماشین باید اطلاعات جدیدی را به کاربر ارائه دهد.

۱.۲ ساختار گزارش

در فصل اول، ابتدا به دسته‌بندی الگوریتم‌های ماشین اشاره خواهد شد. در ادامه‌ی این فصل هر الگوریتم و کاربرد آن به صورت کوتاه توضیح داده خواهد شد و در نهایت الگوریتمی که مناسب پروژه هستند انتخاب می‌شوند و در فصل بعدی به بررسی ریزتر آن پرداخته می‌شود. در نهایت در فصل چهارم نیم نگاهی به ساختار صفحات وب انداخته می‌شود تا بتوان وقتی از کتاب‌خانه‌های پایتون استفاده می‌شود، درک بهتری داشت.

نکته‌ای که قابل توجه است این است که پروژه از دو بخش مختلف تشکیل شده است: از یک بخش یادگیری ماشین که ثابت است و یک الگوریتم ثابت دارد و دیگری بخش داده‌خوانی از سطح وب، که وابسته به هر وبسایتی متفاوت است.

می‌توان ساختار پروژه را با دیاگرام زیر نشان داد:



نمودار ۱-۱ - مراحل پروژه

فصل ۲: الگوریتم‌های یادگیری ماشین

۲.۱ مقدمه

هدف از این فصل، بررسی کلی فرآیندهای یادگیری ماشین و مقایسه‌ی اجمالی آن‌هاست. در ابتدا به تعریف یادگیری ماشین و سپس دسته‌بندی این الگوریتم‌ها می‌پردازیم. سپس الگوریتم‌های محبوب و پر کاربرد آن را معرفی کرده و در انتهای فصل بهترین آن‌ها را انتخاب می‌کنیم.

۲.۲ یادگیری ماشین چیست؟

وقتی که اینباکس ایمیل خود را باز می‌کنید، وقتی که در یوتوب به دنبال موضوعی می‌گردید، وقتی که از برنامه‌ی نقشه در تلفن‌های همراهمتان استفاده می‌کنید، ناخواسته از تکلونوژی به نام یادگیری ماشین بهره می‌برید. یادگیری ماشین ایده‌ای قدیمی است که کاربرد آن در دو دهه‌ی اخیر به اوج خود رسیده است.

یادگیری ماشین، یکی از شاخه‌های هوش مصنوعی محسوب می‌شود که در دو دهه‌ی اخیر به سبب افزایش فراوان قدرت پردازنده‌ها، رشد فراوانی داشته است.

آرتور ساموئل در سال ۱۹۵۹، یادگیری ماشین را اینگونه تعریف کرد:

- حوزه‌ای که در آن کامپیوتر می‌تواند بدون اینکه دقیقا برنامه‌ریزی شده باشد، یاد بگیرد.

تام میچل یکی از دانشمندان حوزه‌ی کامپیوتر، در بازی شطرنج مهارت زیادی نداشت. در طی یک پروژه و با استفاده از یادگیری ماشین، به کامپیوتر شخصی خود تصمیم گرفت که شطرنج را آموزش دهد. برای این‌کار، کامپیوتر خود را مجبورکرد که هزاران بازی را با خودش انجام دهد و تجربه بدست بیاورد. این اتفاق سبب شد که بعد از مدتی، کامپیوتر شخصی او بتواند از خود او بهتر شطرنج بازی کند. عنصری که موجب این اتفاق شد این است که کامپیوترها به دلیل اینکه احساسات ندارند و خستگی ناپذیرند، می‌توانند مدت‌های طولانی را به انجام یک کار تکراری بپردازند و ادامه دهند.

این امر موجب شده است تا رایانه‌ها بهترین موجودات برای انجام کارهای پردازشی سنگین باشند و از آن‌ها در علوم مختلف برای این امر استفاده شود.

چندین سال بعد، تام میچل در سال ۱۹۹۸، برای یادگیری ماشین این تعریف را ارائه داد:

- گوییم که یک برنامه کامپیوتری از تجربه E با توجه به کار T و معیار سنجش P، یاد می‌گیرد اگر عملکرد آن در T، که توسط P اندازه‌گیری می‌شود، با تجربه‌ی E بهبود می‌یابد.

شاید تعریف آخر کمی گیج‌کننده باشد. با یک مثال به توضیح آن پرداخته می‌شود:

در دسته‌بندی ایمیل‌ها توسط یادگیری ماشین، دسته‌بندی ایمیل‌ها به دو دسته‌ی اسپم و غیر اسپم کار T، درصد موفقیت در تشخیص این امر به درستی P و تجربه‌ی E وقتی حاصل شده است که سیستم از کاربر یادگرفته است که براساس چه معیارهایی ایمیل اسپم را از غیر اسپم تشخیص دهد.

از دیگر مثال‌های یادگیری ماشین می‌توان به کاربرد آن در پیش‌بینی قیمت ارزهای دیجیتال، تشخیص دست خط، کلاسه‌بندی اخبار، تشخیص گفتار و .... اشاره کرد.

نکته‌ی قابل توجه در این حوزه این می‌باشد که همه‌ی الگوریتم‌ها و عملیات به طور کامل بر پایه ریاضیات می‌باشد و از منطق پیروی می‌کند. ابزارهایی که همواره در کنار این داشمندان بوده، ابزار ماتریس و ابزار محاسبات عددی بوده است.

۲.۳ دسته‌بندی الگوریتم‌ها

دسته‌بندی‌ها متعددی برای الگوریتم‌های یادگیری ماشین وجود دارد. در این بخش به بررسی مشهورترین آن‌ها می‌پردازیم.الگوریتم‌های یادگیری ماشین به طور کلی به سه دسته‌ی «نظارت شده» ، «نظارت نشده» و «یادگیری تقویتی» تقسیم می‌شوند.

در دسته‌ی نخست، کاربر حجمی از داده‌ها را به ماشین ارائه می‌دهد. هر داده مجموعه‌ای از اطلاعات راجع به ویژگی‌های یک حالت از آن پدیده است. از هر داده، یک ویژگی را به عنوان پاسخ در نظر می‌گیرد. در یادگیری نظارت شده، وظیفه‌ی ماشین این است که از آن مجموعه داده یاد بگیرد و در نهایت برای داده‌های جدیدی که قبلا آن‌ها را ندیده است، مقداری را پیش‌بینی کند. پردازش تصویر، پیش‌بینی قیمت ارز، طبقه‌بندی ایمیل‌ها، تشخیص سرطانی بودن یا نبودن تومور با توجه به ویژگی‌های ظاهری آن مثال‌هایی از این دسته از الگوریتم‌ها می‌باشند. از الگوریتم‌های معروف در این حوزه می‌توان به «رگرسیون خطی» ، «رگرسیون لوجستیک» ، «درخت تصمیم» و ... اشاره کرد.

در دسته‌ی دوم، برخلاف حالت اول، کاربر ویژگی از داده‌ها را به عنوان پاسخ انتخاب نمی‌کند، بلکه از ماشین انتظار دارد که اطلاعات جدیدی را درباره‌ی همان مجموعه داده ارائه دهد.تشخیص گفتار، خوشه‌بندی داده‌ها، آنالیز ارتباطات در شبکه‌های مجازی از مثال‌های کاربرد الگوریتم‌های نظارت نشده‌اند. از الگوریتم‌های معروف در این حوزه می‌توان به «SVM»، «KNN»، «خوشه بندی K-means» و ... اشاره کرد.

در یادگیری تقویتی، الگوریتم‌ها براساس تصمیم‌گیری آموزش دیده‌اند. بنابراین بسته به تصمیمات، الگوریتم‌ها خود را برای تولید خروجی موفقیت آمیز و یا شکست آموزش می‌دهند. در نهایت این الگوریتم دارای تجربه‌ای است که قادر به ارائه پیش‌بینی‌های خوب در یک موضوع خواهد بود. بهینه‌سازی کنترلرها و خودروهای خودران مثال‌هایی از پروژه‌هایی هستند که از الگوریتم‌های یادگیری تقویتی بهره برده‌اند. از الگوریتم‌های این حوزه می‌توان به «Q-Learning»، «DQN»، «SARSA» و .... اشاره کرد.

۲.٤ یادگیری نظارت‌شده – دسته‌ی مناسب برای پیش‌بینی

همانطور که گفته شد، در یادگیری نظارت شده پاسخ برای مجموعه‌ای از داده‌ها در ارتباط با آن پدیده موجود است. به طور مثال، اگر قیمت خودرو را پاسخ در نظر گرفته شود، تعداد تصادف‌ها، سال تولید، مسافت طی کرده و .... می‌توانند «ویژگی»هایی باشند که رابطه‌ای با پاسخ دارند. مسئولیت ماشین پیدا کردن بهترین رابطه‌ای است که با کمترین مقدار خطا بتواند این اعداد را پیش‌بینی کند. وقتی که دقت ماشین به حد کافی رسید، از آن برای پیش‌بینی حالت‌های جدید استفاده می‌شود. این ویژگی الگوریتم‌های نظارت شده برای اهداف این پروژه مناسب است چرا که برای خودرو می‌توان قیمت را پاسخ، و سایر ویژگی‌ها را به عنوان ویژگی در نظر گرفت. دو دسته‌ی دیگر چون که چیزی به عنوان پاسخ را نمی‌شناسند، مناسب نیستند. پس در ادامه به بررسی چند الگوریتم نظارت شده پرداخته می‌شود تا در نهایت بتوان بهترین آن‌ها را برای این پروژه برگزید.

۲.۴.۱ رگرسیون خطی

ابتدایی ترین الگویتم این حوزه «رگرسیون خطی» نام دارد. هدف آن این است که ماشین یک معادله براساس ویژگی‌های مجموعه‌ای از داده‌ها پیش‌بینی کند که توسط آن بتواند با کمترین خطای ممکن

مقادیر پاســخ را برای داده‌های جدید پیش‌بینی کند. این معادله با h_θ نمایش داده می‌شــود. البته نکته اینجاســت که این ویژگی‌ها لزوما نباید با مرتبه‌ی یک در این معادله حاضــر شــوند. می‌توانند به صــورت تابعی از این ویژگی‌ها نیز باشــند. مثلا به صــورت ضــرب ویژگی‌ها، لگاریتم یک ویژگی و ... باشــد. از این منظور از «خطی» بودن این اســت که ضــرایب این معادله به صــورت یک عدد بیان می‌شــوند و نه تابعی متغیر.

برای درک بهتر، این مثال را مورد بررسی قرار داده می‌شود:

پیش‌بینی قیمت یک خانه با توجه به ویژگی‌های آن (متراژ، تعداد اتاق، ســال ســاخت) که هر یک از این ویژگی‌ها می‌تواند در تابع h_θ حضور پیدا کند. به طور مثال:

$$h_\theta(x) = 1000\left(\text{متراژ}\right)^2 + 1500(\text{تعداد اتاق}) - 1500\ log(\text{سال ساخت})$$

حالت کلی این معادله به صورت زیر می‌تواند بیان شود:

$$h_\theta(x) = \sum_{i=1}^{n} \theta_i x_i$$

در این رابطه، θ_i ضریب ویژگی i ام است که x_i یک تابع از ویژگی i ام است. n نیز تعداد تابع‌هایی است که مهندس انتخاب کرده است.

در ابتدا مهندس یادگیری ماشین، ویژگی‌هایی که فکر می‌کند مناسب پیش‌بینی می‌باشد، انتخاب می‌کند و سـپس توانی از آن که می‌تواند مناسـب باشـد را اختیار می‌کند.مقدار اولیه‌ای برای ضـرایب این ویژگی‌ها در نظر می‌گیرد و یک سـری عملیات جبری رو داده‌ها اعمال می‌شـود. در نهایت به یک بردار θ می‌رسـد که ضرایب ویژگی‌هایی است که از قبل انتخاب شده است.

در این الگوریتم، یک تابع هزینه به صورت زیر تعریف می‌شود که هدف آن نسبت دادن یک عدد به مقدار خطای ماشین نسبت به داده‌های تست می‌باشد:

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

در این رابطه، m تعداد داده‌هایی است که ماشین از آن‌ها یاد می‌گیرد. $x^{(i)}$ داده‌ی i ام است و $y^{(i)}$ نیز، پاسخ به داده‌ی i ام است.

از مزایای این الگوریتم می‌توان به موارد زیر اشاره کرد:

۱. پیاده سازی آسان‌تری نسبت به سایر الگوریتم‌ها دارد.

۲. نسبت به سایر الگوریتم‌ها داده‌های کمتری را جهت یادگیری نیاز دارد.

از معایب این الگوریتم می‌توان به موارد زیر اشاره کرد:

۱. امکان دارد که معادله‌ی بدست آمده نسبت به داده‌های ورودی به شدت حساس باشد.

۲. ممکن است در جهت کاهش خطا، ضرایس بدست آمده بزرگ شوند. البته این مورد قابل حل است.

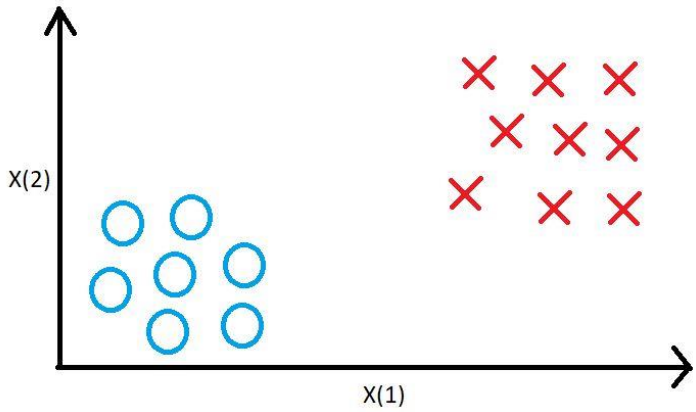
با توجه به ویژگی‌های گفته شده، می‌توان بیان کرد که این الگوریتم می‌تواند گزینه‌ی خوبی برای پیش‌بینی قیمت خودرو باشد.

این الگوریتم به‌طور مفصل‌تر در فصل ۳ مورد بررسی قرار می‌گیرد.

۲.۴.۲ رگرسیون لوجستیک

در این الگوریتم، داده‌های ورودی در دسته‌بندی‌ها مختلف قرار دارند. مقصود از این الگوریتم این است که ماشین بتواند با توجه به ویژگی‌های ورودی بتواند تشخیص دهد که این داده‌ی جدید در کدام دسته قرار می‌گیرد. به طور مثال: تصویری را دریافت کند و تشخیص دهد که این تصویر گل، درخت و یا اجسام محدود دیگر است. برای این کار، تعدادی عکس به ماشین داده می‌شود و به آن اعلام می‌شود که هر کدام از این تصاویر بیانگر کدام یک از موردهای بالاست. سپس تصویری جدید که ماشین آن را ندیده است را به عنوان ورودی به ماشین داده می‌شود و خروجی آن، عنوان دسته‌ای است که ماشین آن را پیش‌بینی کرده است.

از کاربردهای این الگوریتم در حوزه‌ی پزشکی می‌توان به تشخیص بدخیم بودن یا خوش‌خیم بودن یک تومور اشاره کرد.



نمودار ۲-۱ تومور بدخیم یا خوش‌خیم

اگر در نمودار بالا، دایره‌ها نماد خوش‌خیم بودن و ضـربدرها نماد بدخیم بودن تومور براسـاس دو ویژگی x(1) و x(2) باشند، در این صورت هدف الگوریتم این است که بتواند مرزی را بین این دو تشخیص دهد و طبق آن وضعیت تومورهای جدید را پیش‌بینی کند و اعلام کند که آیا خوش‌خیم هستند یا بدخیم.

رابطه‌ی زیر احتمال حضـور داده‌ای را در دسـته‌ی شـماره‌ی i محاسـبه می‌کند.بدیهی اسـت که $g_i(z)$ همواره بین یک و صـفر می‌باشـد. برای ورودی جدید، احتمال حضـور در هر دسـته را محاسـبه کرده و دسته‌ای که بیشترین مقدار را پیش‌بینی کند، به عنوان دسته‌ی پیش‌بینی شده اعلام می‌شود:

$$g_i(z) = \frac{1}{1 + e^{-z}}$$

مقدار z، از همان رابطه‌ی رگرسـیون خطی که در زیربخش قبل معرفی شـد، محاسـبه می‌شـود. اگر مقدار $g_j(z)$ از 0.5 بیشتر باشد، مقدار آن را یک و در غیر این صورت، صفر اعلام می‌کند.

همانند الگوریتم قبلی، مقصـود نهایی الگوریتم آن اسـت که ضـرایب بردار θ را بتواند پیش‌بینی کند که مقدار خطای آن ،که طبق رابطه‌ی زیر محاسبه می‌شود، به حداقل برسد:

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

تعداد داده‌های یادگیری m، پاسخ واقعی سیستم به ورودی i ام $y^{(i)}$ و $h_\theta(x^{(i)})$ برابر با همان z است که توضیح داده شد. دلیل اسـتفاده از لگاریتم آن است که در صورتی که مقدار را درست پیش‌بینی کند خطایی محاسبه نمی‌شود اما اگر برعکس پیش‌بینی کند، خطای بزرگی را به سیستم وارد می‌کند.

بطور کلی این الگوریتم در دسته‌بندی داده‌ها استفاده می‌شود و از آن نمی‌توان انتظار پیش‌بینی قیمت یک پدیده را که یک امر پیوسته است را داشت. از همین رو این الگوریتم مناسب این پروژه نیست.

۲.۴.۳ شبکه‌های عصبی

مغز انسان، همواره یکی از عجیب‌ترین پدیده‌های شناخته شده در طبیعت می‌باشد. دانشمندان علوم کامپیوتر همواره در تلاش بوده‌اند که عملکرد این دستگاه منظم را توسط کامپیوتر شبیه‌سازی کنند. شبکه‌های عصبی حاصل تلاش این دانشمندان در این حوزه است.

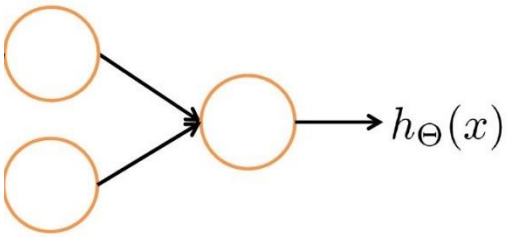
کاربرد این الگوریتم در پیش‌بینی بسیار متعدد است. از آن می‌توان در پردازش تصاویر، پردازش گفتار، پیش‌بینی قیمت ارز و ... اشاره کرد.

قبل از توضیح راجع به عملکرد این الگوریتم لازم است چند تعریف ارائه شود:

سلول عصبی: کوچک‌ترین واحد شبکه‌های عصبی محسوب می‌شود. هر سلول چند ورودی و یک خروجی دارد. به ازای ورودی‌ها و θ معادل با آن سلول، خروجی که براساس رابطه‌ی زیر محاسبه می‌شود:

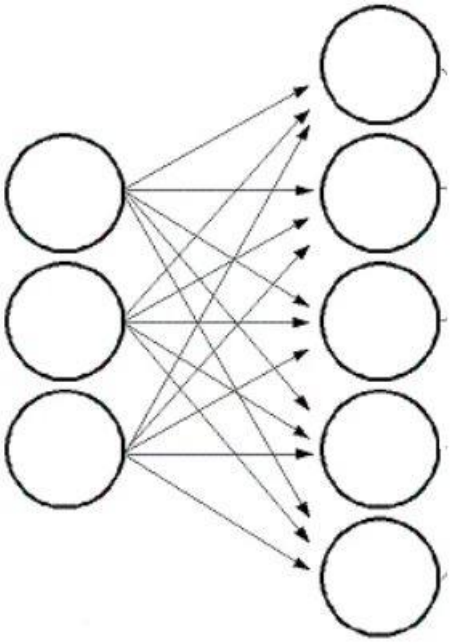
$$g(z) = \frac{1}{1 + e^{-z}}$$

در رابطه‌ی بالا z، همانند رگرسیون لوجستیک، از رابطه‌ی $z = \sum_{i=1}^{n} \theta_i x_i$ محاسبه می‌شود. و مقدار $g(z)$ را به عنوان خروجی بر می‌گرداند.



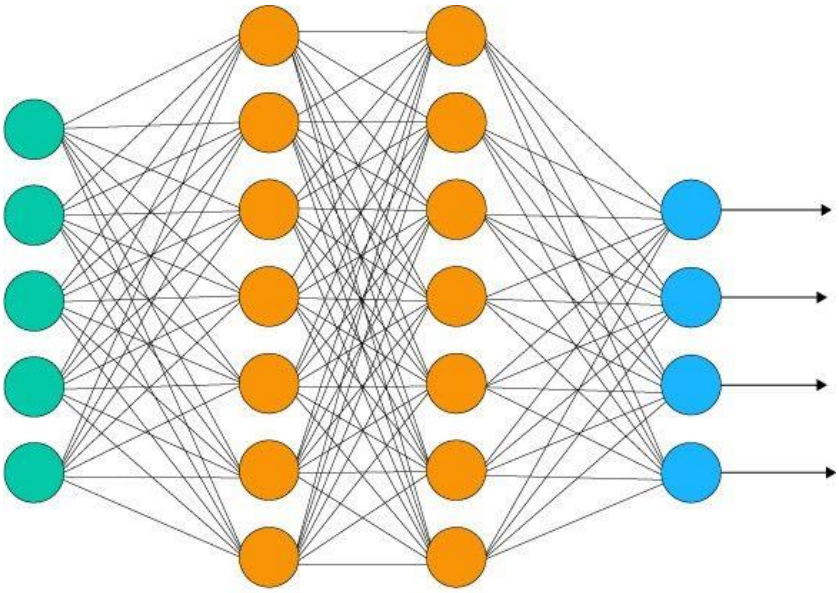
شکل ۲-۱ ساختار یک سلول عصبی

لایه: مجموعه‌ای از سلول‌های عصبی که نسبت به یک دسته از ورودی‌ها محاسبه می‌شوند، لایه گفته می‌شود. برای ارتباط بین هر دو لایه، یک ماتریس $\Theta^{(j)}$ تعریف می‌شود که ضرایب $h_\theta(x)$ برای هر سلول از لایه j به هر سلول $j+1$ است.



شکل ۲-۲ دو لایه شبکه‌ی عصبی

شـبکه عصـبی: به مجموعه‌ای که بعد از در کنارهم قرار گرفتن لایه‌ها شـکل می‌گیرد، شـبکه‌ی عصـبی می‌گویند. هر شبکه‌ی عصبی را می‌توان با در کنار هم قرار گرفتن $\Theta^{(j)}$ ها تشکیل داد.



شکل ۲-۳ شبکه عصبی

پردازش در این الگوریتم این‌گونه است:

ابتدا مهندس یادگیری ماشین، ساختاری برای شبکه‌ی عصبی خود تعیین می‌کند. منظور از این ساختار، تعداد ورودی‌ها،تعداد لایه‌ها و ارتباطات بین لایه‌هاست.این ساختار براساس نیازها و شرایط متفاوت است.سپس با مقادیر اولیه‌ای شروع به محاسبه‌ی مقادیر می‌کند و خطا را می‌سنجد. سپس با روش‌هایی که مبتنی بر رفت و بازگشت می‌باشند شروع به اصلاح ضرایب کرده و تا وقتی که مقدار خطا به مقدار دلخواه نرسیده است، متوقف نمی‌شود.

حوزه‌ی یادگیری عمیق، که به تازگی از محبوبیت زیادی برخوردار است به بررسی این لایه‌ها و اینکه چگونه باید طراحی شوند می‌پردازد. از کاربردهای این الگوریتم می‌توان به پردازش تصویر (مانند تشخیص دست خط) اشاره کرد که در آن هر پیکسل یک تصویر یک ورودی برای شبکه عصبی خواهد بود.خروجی نیز می‌تواند تشخیص این باشد که چه چیزی در تصویر دیده می‌شود. این اتفاق همانند الگوریتم رگرسیون لاجستیک می‌باشد.

از مزایای این الگوریتم می‌توان به موارد زیر اشاره کرد:

۱- قدرتمند بودن آن

۲- امکان پیش‌بینی راحت‌تر ویژگی‌های ترکیبی

از معایب این الگوریتم می‌توان به موارد زیر اشاره کرد:

۱- پیاده سازی دشوار

۲- نیاز به داده‌های زیاد

۳- نیاز به تخصص در طراحی ساختار شبکه

با توجه به نکات بالا، شبکه‌های عصبی نیز می‌تواند یکی از گزینه‌هایی باشد که با آن پیش‌بینی قیمت خودرو صورت پذیرد.

۲.۴.۴ ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان، از الگوریتم‌های نظارت شده می‌باشد که نسبت به سایر الگوریتم‌ها تازه‌تر است. از الگوریتم SVM، در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در کلاس‌های خاص باشد می‌توان استفاده کرد. در ادامه به کاربردهای این الگوریتم به صورت موردی اشاره می‌شود:

سیستم آنالیز ریسک، کنترل هواپیما بدون خلبان، ردیابی انحراف هواپیما، شبیه‌سازی مسیر، سیستم راهنمایی اتوماتیک اتومبیل، سیستم‌های بازرسی کیفیت، آنالیز کیفیت جوشکاری

مزایای این الگوریتم به شرح زیر است:

۱- پیاده‌سازی نسبتاً ساده

۲- برخلاف شبکه‌های عصبی در ماکزیمم محلی گیر نمی‌افتد.

معایب به شرح زیر است:

۱- در مواردی که تعداد ویژگی‌ها از تعداد داده‌ها بیشتر است عملکرد خوبی ندارد.

۲- وقتی نویز زیادی داشته باشد، پیش‌بینی معقولی ارائه نمی‌دهد.

با توجه به موارد بالا و اینکه از الگوریتم تنها در دسته‌بندی استفاده می‌شود نمی‌تواند گزینه‌ی مناسبی برای این پروژه باشد.

۲.۵ جمع‌بندی و انتخاب الگوریتم

با توجه به اینکه نیاز به الگوریتمی می‌باشد که

۱) پاسخ را درک کند و بتواند آن را پیش‌بینی کند

۲) پاسخ بدست آمده باید فرم عددی داشته باشد و دسته‌بندی نباشد.

دو الگوریتم «رگرسیون خطی» و «شبکه‌های عصبی» امکان پذیرند. با توجه به اینکه تعداد داده‌های موجود در ارتباط به یک نوع خودروی خاص محدود است، استفاده از «رگرسیون خطی» منطقی‌تر است.

از این‌رو، فصل بعد به توضیح این الگوریتم اختصاص داده شده است.

فصل ۳: رگرسیون خطی

۳.۱ مقدمه

یکی از ساده‌ترین و پرکاربردترین الگوریتم‌های یادگیری ماشین، «رگرسیون خطی» است. از این الگوریتم در پیش‌بینی امور مختلف می‌توان بهره برد. ساده‌بودن پیاده‌سازی، سریع‌بودن و انتخاب دقت از مزایای این روش می‌باشند. در ادامه‌ی این فصل به بررسی دقیق‌تر این الگوریتم پرداخته می‌شود.

۳.۲ داده‌ها و ویژگی‌ها

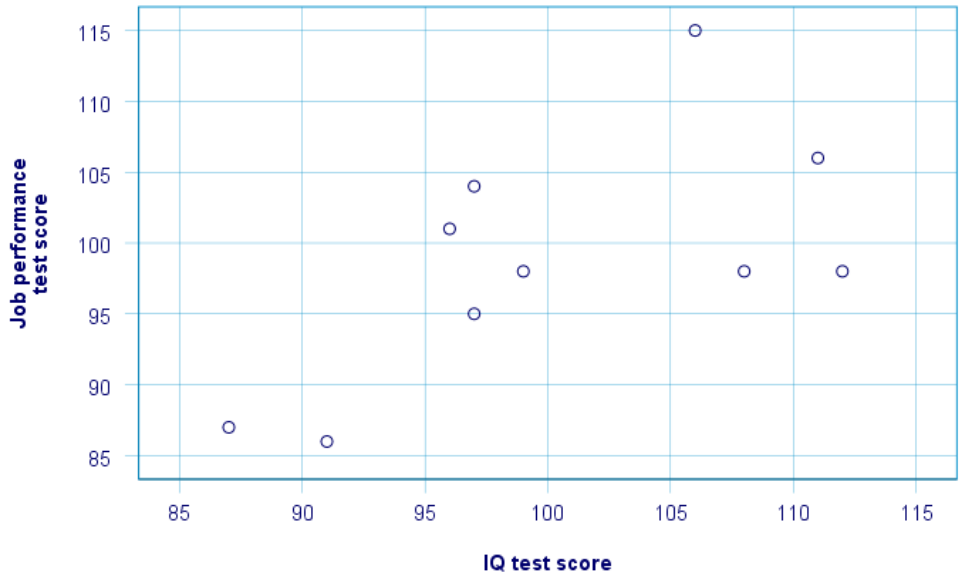
در هر الگوریتم نظارت‌شده، لازم است که داده‌های ورودی به سیستم، پردازش شوند. پس از این رو نیاز به ارائه چند تعریف و نحوه‌ی علامت‌گذاری می‌باشد.

بخشی از هر داده که هدف آن پیش‌بینی بخش دیگر، یا همان پاسخ، است بخش «ورودی» داده گفته می‌شود. در این گزارش با علامت x نمایش داده می‌شود. هر ورودی از چند «ویژگی» تشکیل شده‌است.

بخش دیگر داده که هدف ماشین، پیش‌بینی مقادیر آن است، «پاسخ» نامیده می‌شود و با علامت y نمایش داده می‌شود.

در هر ماشین، تعدادی داده به ماشین وارد می‌شود که شامل دو بخش ورودی و پاسخ می‌باشد. برای این‌که اشاره به داده‌ی i ام راحت باشد، ورودی آن را با $x^{(i)}$ و خروجی را با $y^{(i)}$ نمایش داده می‌شوند.

نکته‌ی قابل توجه این است که $x^{(i)}$ لزوما یک اسکالر نیست و در صورتی که چند ویژگی را شامل باشد، تبدیل به بردار خواهد شد. البته این موضوع راجع به $y^{(i)}$ صحیح نمی‌باشد و همواره یک اسکالر خواهد بود.



نمودار ۳-۱ ورودی و خروجی در رگرسیون خطی

در شکل ۳-۱، داده‌ها به صورت یک دایره نمایش داده شده‌اند. برای هر داده، مقدار آن در راستای افقی، برابر ورودی یا همان $x^{(i)}$ است. همچنین مقدار آن در راستای عمودی برابر پاسخ آن داده است.

۳.۳ تابع فرضیه

همان‌گونه که قبل‌تر توضیح داده‌شد، هدف از این روش یادگیری ماشین این است که به یک تابع برسد که بتولند با کمترین مقدار خطای ممکن، پاسخ ورودی‌های جدید را پیش‌بینی کند. به این تابع، تابع فرضیه گفته می‌شود.

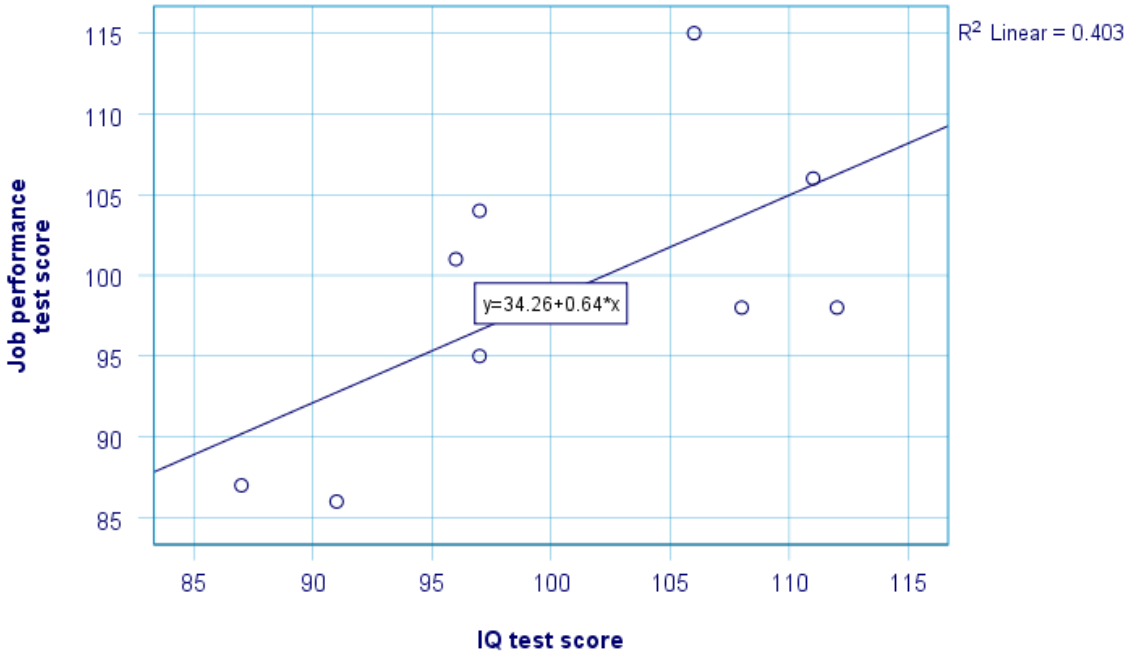
این تابع به‌صورت ترکیب خطی توابعی از ویژگی‌ها بیان می‌شود:

$$h_\theta(x) = \sum_{i=1}^{n} \theta_i x_i$$

$h_\theta(x)$ همان تابع فرضیه می‌باشد. ضرایب این رابطه، θ_i ها، فاکتورهایی هستند که تابع فرضیه را توصیف می‌کنند. هدف «رگرسیون خطی» پیداکردن بهترین ترکیب ممکن برای این تابع می‌باشد. نکته‌ی جالب اینجاست که می‌توان این تابع را به صورت ضرب برداری دو بردار x و θ نوشت:

$$h_\theta(x) = \theta^T x$$

البته می‌توان مقادیر x_i ها را با تابعی از جنس ویژگی‌های ورودی جایگزین کرد.



نمودار ۳-۲ تابع فرضیه

در شکل ۳-۲، خطی که رسم شده است، همان تابع فرضیه است که تلاش دارد با کمترین خطای ممکن، مقادیر پاسخ را پیش‌بینی کند.

۳.٤ تابع خطا

همانند هر الگوریتم نظارت‌شده‌ای برای بررسی مقدار خطای سیستم نیاز به یک تابع خطا می‌باشد. در روش رگرسیون خطی، همانطور که هدف پیداکردن ضرایب بردار θ است، تابع خطا نیز براساس این بردار تعریف می‌شود:

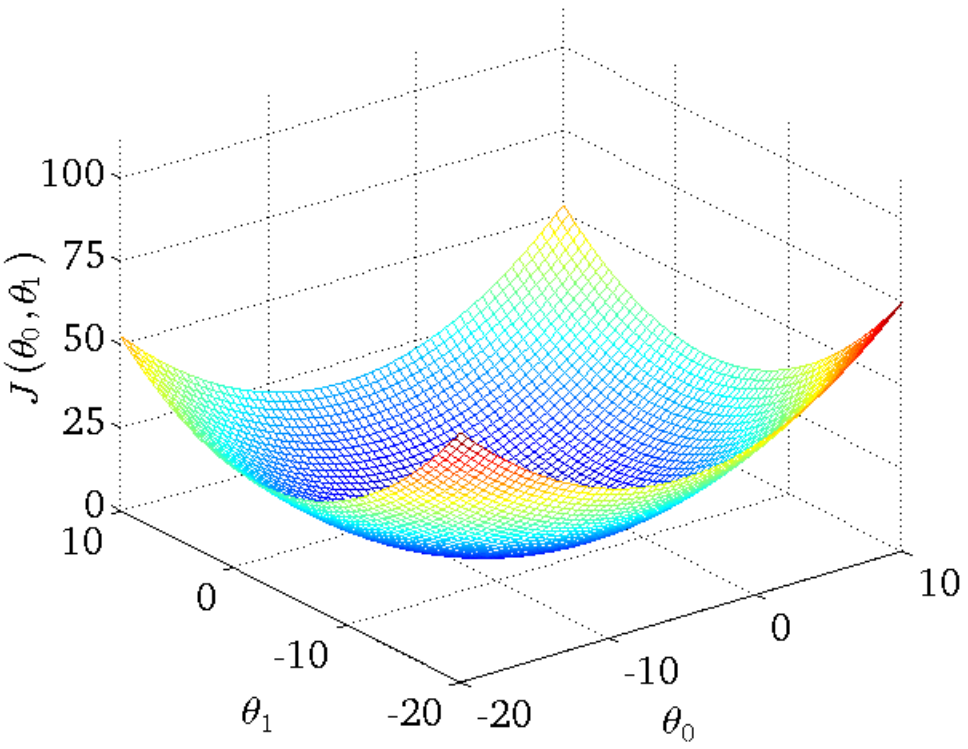
$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

در این رابطه، m تعداد داده‌های یادگیری‌شده، می‌باشد. هرچه این تابع مقدار کمتری داشته باشد، ضرایب بهتری برای معادله یافت شده است.

برای شهود از این تابع، تابع فرضیه به‌صورت زیر در نظر گرفته می‌شود:

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

تابع هزینه‌ی این تابع به شکل زیر خواهد بود:



نمودار ۳-۳ تابع هزینه برای یک تابع فرضیه با دو ضریب

در نمودار ۳-۳، دو محور به ضرایب ویژگی‌ها اختصاص داده شده‌است. محور سوم نیز بیانگر تابع خطا براساس این دو ضرایب می‌باشد.

۳.۵ یافتن بهترین بردار ضرایب

یکی از روش‌هایی که بواسطه‌ی آن کمینه‌ی تابع خطا، محاسبه می‌شود، روش «گرادیان کاهشی» است. در این روش، با استفاده از مشتق جزئی، مقدار بردار ضرایب را مرحله به مرحله بروزرسانی می‌شوند و این فرآیند تا جایی ادامه دارد که تابع خطا از حد موردنظر کمتر شود.

در این روش هر درایه بردار ضرایب به شکل زیر بروزرسانی می‌شود:

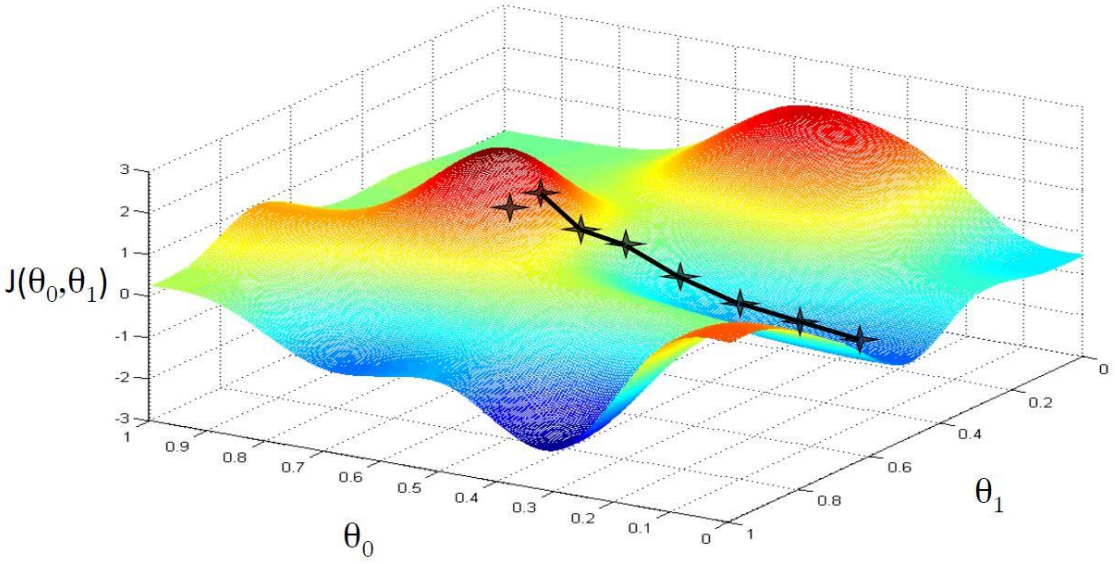
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

در این رابطه، α به عنوان نرخ یادگیری شناخته می‌شود.

این رابطه وقتی در فضای رگرسیون خطی وجود داشته باشد، به حالت ساده‌تر زیر تبدیل می‌شود:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}{}_j$$

برای درک بهتر از شهود این روش، مثال زیر آورده شده است:



نمودار ۳-۴ گرادیان کاهشی

در نمودار ۳-۴، وظیفه‌ی گرادیان کاهشی این است که بتواند از نقطه‌ی قرمز رنگ به نقطه‌ی آبی رنگ برسد. برای این اتفاق، از مشتق جزئی و نرخ یادگیری استفاده می‌شود.

نکته‌ی قابل توجه این است که همواره این روش بهترین بردار ضرایب را پیدا نمی‌کند، اما تا حد خیلی خوبی باعث بهتر شدن عملکرد سیستم می‌شود. همچنین به نقطه‌ی شروع اولیه بردار θ وابسته است.

۳.٦ جمع‌بندی

بعد از آشـنایی با اجزای مختلف الگوریتم «رگرسـیون خطی» می‌توان روند کلی الگوریتم را به این شـکل بیان کرد:

۱. ویژگی‌هایی که می‌توانند مناسب باشند انتخاب می‌شوند.

۲. فرم ویژگی‌ها انتخاب می‌شوند.(توان دوم، لگاریتم، ضرب چند ویژگی و ....)

۳. بردار ضرایب با مقادیر اولیه آماده می‌شوند.

٤. تابع فرضیه تشکیل می‌شود و پاسخ داده‌ها محاسبه می‌شوند.

٥. این پاسخ‌ها را با جواب واقعی مقایسه و مقدار تابع خطا را به ازای بردار ضرایب محاسبه می‌شود.

٦. اگر خطا بیشـتر از حد انتظار بود، گرادیان کاهشـی را بر روی بر روی بردار ضرایب اجرا کرده و به دستور ۴ باز می‌گردد.

۷. بعد از اینکه خطا به حد کافی کم شد، داده‌های جدیدی وارد سیستم کرده و از مقادیر پیش‌بینی شده استفاده می‌شود.

فصل ۴: داده کاوی

۴.۱ مقدمه

داده کاوی به فرآیند استخراج اطلاعات و داده از پایگاه داده گفته می‌شود. این اطلاعات ممکن است در ارتباط با موضوعات مختلف باشد. می‌توان از جنس عدد، کلمه و ... باشد.

با توجه به افزایش حجم داده‌ها، داده‌کاوی از حوزهای فعال و داغ دنیای تکلونوژی است. در بخش اول این پروژه که باید داده از صفحات وب برداشت شوند، استفاده از این راهکار می‌تواند کمک فراوانی بکند.

در ادامه‌ی فصل به بررسی ساختار و ابزارهایی که می‌توانند به ما کمک کنند پرداخته می‌شوند.

۴.۲ ساختار صفحات وب

همه‌ی صفحات وب که امروزه در اختیار عموم قرار دارند، از سه زبان HTML، CSS و JS استفاده می‌کنند. خطوط کد که با استفاده از این سه عنصر نوشته شده‌اند، توسط مرورگر خوانده می‌شوند و به صورتی که دیده می‌شوند به کاربر نمایش داده می‌شوند. نکته اینجاست که در استخراج داده از این صفحات، باید با ساختار این صفحات آشنا بود و از حجیم بودن این صفحات در قالب اصلی‌شان(کد) نترسید.

۴.۳ ابزار

زبان برنامه‌نویسی پایتون از کابردهای متفاوتی برخوردار است. از یادگیری ماشین تا بازی سازی. از استخراج داده تا رسم نمودارهای ریاضی. دلیل این ویژگی، وجود کتاب‌خانه‌های بی‌شماری است که در هر زمینه‌ای موجود است.

در زمینه‌ی استخراج داده از صفحات وب، کتاب‌خانه‌ی Beautiful Soup، یکی از پرکاربردترین‌ها می‌باشد. توسط این کتاب‌خانه می‌توان صفحات وب را در قالب کد دریافت و داده‌ها را از آن استخراج کرد.

در این پروژه از این زبان و این کتاب‌خانه برای داده‌کاوی از سطح وب استفاده خواهد شد.

٤.٤ جمع‌بندی

با توجه به این‌که در این پروژه نیاز به دریافت داده از سطح اینترنت وجود دارد، از زبان برنامه‌نویسی پایتون و کتاب‌خانه‌های آن استفاده خواهد گردید.

فصل ۵: جمع‌بندی

۵.۱ جمع‌بندی

استفاده از هوش مصنوعی می‌تولند همواره به انسان در جهت زیستن بهتر کمک کند. یکی از زیرشاخه‌های آن، یادگیری ماشین، در این زمینه می‌تواند در حوزه‌های پیش‌بینی کمک‌رسانی کند. یکی از موضوعاتی که داده‌های فراوانی در ارتباط با آن موجود است، قیمت خودروهای دست دوم می‌باشد. با استفاده از پایتون، داده‌ها از سایت‌های خرید و فروش این کالا دریافت می‌شود. توسط الگوریتم رگرسیون خطی، ارزیابی می‌شود و در نهایت با توجه به شرایط خودرو کاربر، ارزش آن را تعیین می‌کند.

۵.۲ نتیجه‌گیری

با توجه به این نکته که گزارش شامل معرفی الگوریتم‌های یادگیری ماشین و توضیح یکی از پرکاربردترین آن‌ها بود، امکان نتیجه‌گیری وجود ندارد و باید بعد از نوشته‌شدن برنامه‌ی اصلی و تست کردن آن، نتیجه‌گیری کرد.

منابع و مراجع

[۱] میرمیرانی، جادی؛ دوره‌ی آموزش برنامه‌نویس با پایتون (پیشرفته)، سایت مکتب‌خونه
https://maktabkhooneh.org/course/%D8%A2%D9%85%D9%88%D8%B2%D8%B4-%D8%A8%D8%B1%D9%86%D8%A7%D9%85%D9%87-%D9%86%D9%88%DB%8C%D8%B3%DB%8C-%D8%A8%D8%A7-%D9%BE%D8%A7%DB%8C%D8%AA%D9%88%D9%86-%D9%BE%DB%8C%D8%B4%D8%B1%D9%81%D8%AA%D9%87-mk387/

[2] blog.faradars.org/best-predictive-algorithms

[3] https://towardsdatascience.com/introduction-to-various-reinforcement-learning-algorithms-i-q-learning-sarsa-dqn-ddpg-72a5e0cb6287

[4] https://jonathan-hui.medium.com/rl-introduction-to-deep-reinforcement-learning-35c25e04c199

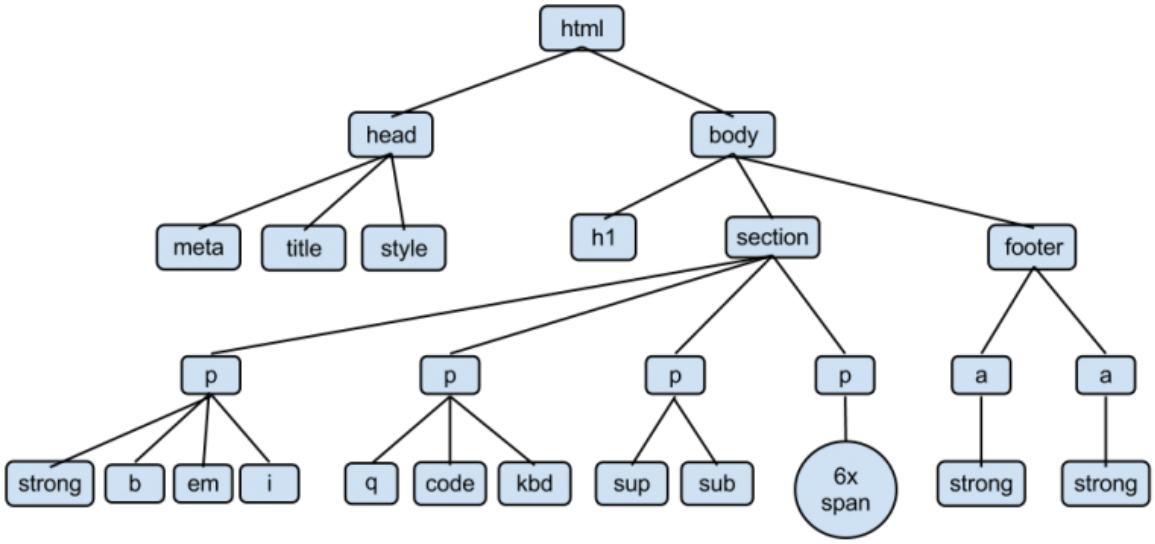
[5] Andrew Ng, Machine Learning Course; Coursera.org
https://www.coursera.org/learn/machine-learning
Slides/ MATLAB Projects.

[6] Youtube.com

پیوست‌ها

پیوست الف – آشنایی با کتابخانه‌ی BeautifulSoup

کتابخانه BeautifulSoup یک کتابخانه پایتون است که به‌منظور استخراج داده از فایل‌های html و xml مورد استفاده قرار می‌گیرد. این کتابخانه صفحات مورد نظر خود را بصورت یک درخت تجزیه می‌کند. درخت تجزیه این امکان را برای برنامه ایجاد می‌کند، که هرگونه دسترسی به عناصر صفحه html با سرعت بیشتری امکان‌پذیر گردد. با این روش شرایط مناسبی برای جستجوی اطلاعات مورد نظر فراهم می‌شود. در زیر نحوه تجزیه عناصر صفحه xml در قالب درخت نمایش داده شده است.



برای اطلاعات بیشتر به لینک زیر مراجعه کنید:

https://bigdata-ir.com/%D9%BE%D8%A7%D8%B1%D8%B3-%D8%B5%D9%81%D8%AD%D8%A7%D8%AA-%D9%88%D8%A8-%D8%A8%D8%A7-%DA%A9%D8%AA%D8%A7%D8%A8%D8%AE%D8%A7%D9%86%D9%87-/beautifulsoup-%D9%BE%D8%A7%DB%8C%D8%AA%D9%88