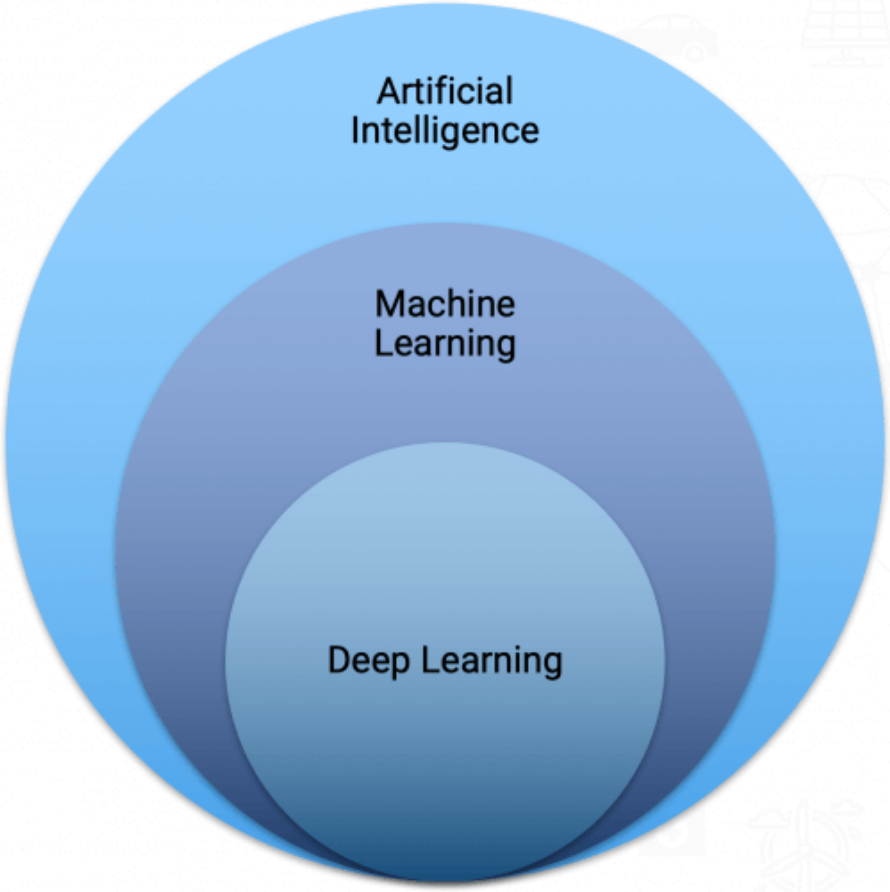
دوره جامع پایتون:

بخش یادگیری ماشین

جلسه نوزدهم

دکتر ذبیح اله ذبیحی

هوش مصنوعی، یادگیری ماشین و یادگیری عمیق



هوش مصنوعی

- به طور کلی هوش مصنوعی مجموعه‌ای از دستورالعمل‌ها است که نحوه اقدام و رفتار مشابه انسان را به کامپیوتر آموزش می‌دهد. نحوه عملکرد کامپیوتر رمزگردانی می‌شود. برای مثال «اگر این اتفاق افتاد، آن کار را انجام بده». طبق یک قاعده سرانگشتی اگر هوش مصنوعی بگوید که چه تصمیمی باید اتخاذ شود، این برنامه در خارج از حوزه هوش مصنوعی قرار می‌گیرد.

یادگیری ماشین

- یادگیری ماشین به عنوان زیر مجموعه هوش مصنوعی به طور خودکار قادر به اقدام است. الگوریتم یادگیری ماشین برخلاف هوش مصنوعی نحوه تفسیر اطلاعات را بیان نمی‌کند.
- یادگیری ماشین مثل یک کودک باید با کمک پایگاه داده‌های طبقه بندی شده یا درون‌داد input، آموزش ببیند. به عبارت دیگر، علاوه بر معرفی داده، باید محتوای آن هم بیان شود. برای مثال این تصویر سگ یا گربه است. شبکه‌های عصبی مصنوعی با کمک این اطلاعات بدون داشتن دستورالعمل‌های مشخص به نتایج و برونداد می‌رسند.
- ماشین لرنینگ یا یادگیری ماشین دانشی است که کمک می‌کند رایانه‌ها بدون برنامه‌ریزی مشخص و با الگو گرفتن از رفتار خودشان کارهای جدید انجام دهند.

یادگیری عمیق

- یادگیری عمیق Deep Learningشامل مفاهیم و الگوریتم‌هایی است که از شبکه‌های عصبی مصنوعی که ساختار مغز انسان را تشکیل می‌دهند الهام گرفته شده است.

- به عبارتی دیپ لرنینگ زیرمجموعه یادگیری ماشین است و خودِ یادگیری ماشین هم به عنوان زیرمجموعه هوش مصنوعی در نظر گرفته می‌شود.

یادگیری ماشین

- یادگیری ماشین یا Machine Learning، توانایی یادگیری مستقل را برای ماشین‌ها فراهم می‌کند. به بیان دیگر یک ماشین می‌تواند از مشاهدات، تجربیات و الگوهایی که طبق یک مجموعه داده تجزیه و تحلیل می‌کند، آموزش ببیند. البته برای انجام این کار لازم نیست به شکل اختصاصی برنامه ریزی شده باشد.

- در شروع یادگیری ماشین، ما مجموعه‌ای از داده‌ها را وارد می‌کنیم تا از این طریق دستگاه بتواند با شناسایی و تجزیه الگوهای موجود در داده‌ها، یادگیری داشته باشد و بر اساس این یادگیری بتواند از مشاهدات و اطلاعات خود نتیجه بگیرد و تصمیم گیری کند.

- این موارد در نهایت باعث شکل گیری یک سیستم هوشمند و دارای قدرت تولید می‌شود که می‌تواند کارهای بسیاری انجام دهد. کارهایی که با یادگیری ماشین می‌تواند انجام شود، بسیار متنوع است. برای نمونه بسیاری از تکنولوژی‌هایی که امروزه باعث شگفتی شما می‌شوند، مانند سیستم تشخیص چهره، سیستم تشخیص هویت و... از یادگیری ماشین ناشی می‌شوند.

- به طور کلی ، یک مسئله یادگیری مجموعه ای از nنمونه داده را در نظر می گیرد و سپس سعی می کند خصوصیات داده های ناشناخته را پیش بینی کند. اگر هر نمونه بیش از یک عدد منفرد باشد و به عنوان مثال یک ورودی چند بعدی باشد (یا داده های چند متغیره) ، گفته می شود که دارای چندین ویژگی یا ویژگی است.

کاربردهای یادگیری ماشین

- پیش بینی آب و هوا: با استفاده از علم ماشین لرنینگ و تجزیه تحلیل داده‌ها می‌توان عمل پیش بینی آب و هوا برای یک بازه‌ی زمانی مشخص را انجام داد.
- تشخیص پزشکی: یکی از مهم‌ترین استفاده‌ها و کاربردهای علم هوش مصنوعی و یادگیری ماشین توانایی تشخیص پزشکی است. به این ترتیب کامپیوتر یا ماشین می‌تواند بیمار بودن یا نبودن یک فرد را از روی داده‌ها و علائم وی تشخیص دهد.
- تجزیه و تحلیل داده‌ها در حجم زیاد: در حال حاضر حجم داده‌های موجود در جهان در تصور انسان نمی‌گنجد. انسان‌ها بدون کمک کامپیوترها نمی‌توانند از بهره‌ای از این داده‌ها ببرند. بنابراین ماشین لرنینگ می‌تواند در زمینه انجام پردازش‌های مختلف روی آن‌ها کارآمد باشد.
- تشخیص چهره: شناسایی چهره در یک تصویر (یا تشخیص اینکه آیا چهره‌ای وجود دارد یا خیر).
- فیلتر کردن ایمیل‌ها: دسته‌بندی ایمیل‌ها در دو دسته هرزنامه و غیر هرزنامه.

- اثبات قضیه بطور خودکار
- وبسایت های تطبیقی
- هوش مصنوعی احساسی
- بیوانفوماتیک
- واسط مغز و رایانه
- شیمی انفورماتیک
- طبقه بندی رشته های DNA
- آناتومی محاسباتی
- بینایی ماشین، از جمله شناسایی اشیاء
- شناسایی کارت اعتباری جعلی
- بازی عمومی ((general game playing
- بازیابی اطلاعات

- شناسایی کلاه برداری های اینترنتی
- زبان شناسی
- بازاریابی
- کنترل یادگیری ماشین
- ادراک ماشین
- تشخیص پزشکی
- اقتصاد
- بیمه
- پردازش زبان طبیعی
- استنباط زبان طبیعی
- بهینه سازی و الگوریتم های فرا ابتکاری
- تبلیغات آنلاین
- سیستم های توصیه گر

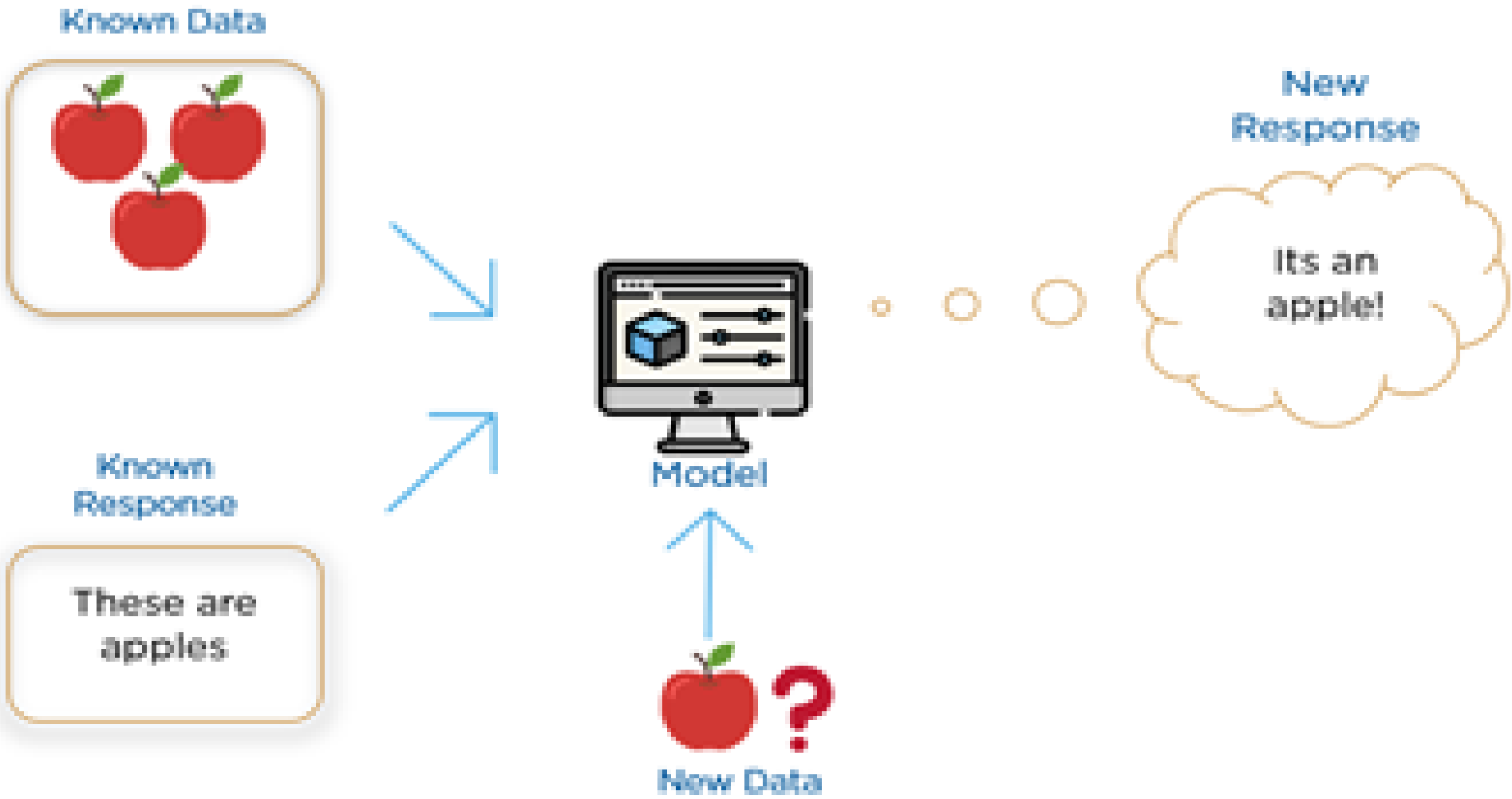
- حرکت ربات
- موتورهای جستجو
- تحلیل احساسات (یا نظر کاوی)
- مهندسی نرم افزار
- شناسایی گفتار و دست نوشته
- تحلیل بازارهای مالی
- نظارت بر درستی ساختار
- الگوشناسی ترکیبی
- پیش بینی سری های زمانی
- تحلیل رفتار کاربر
- ترجمه

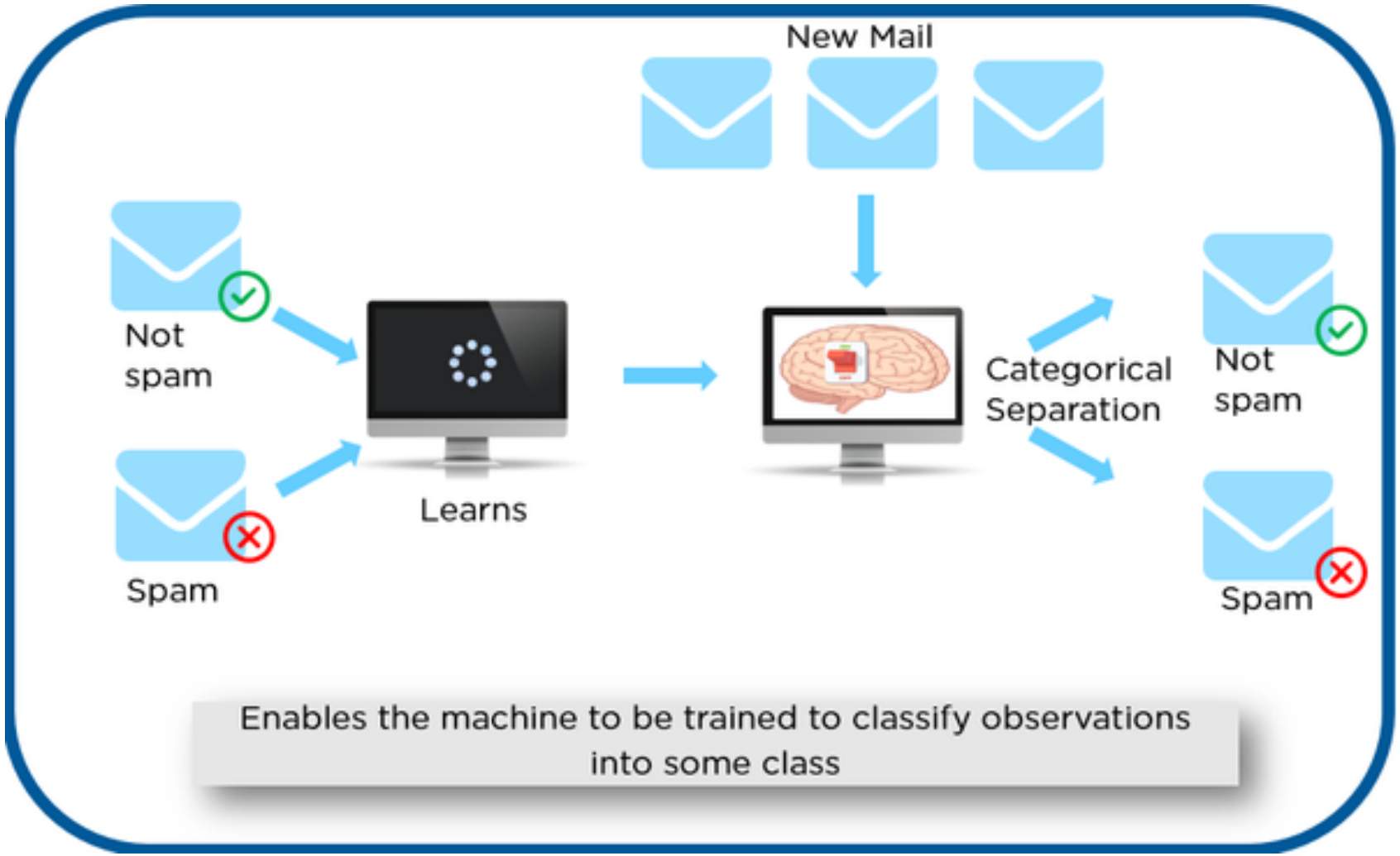
انواع یادگیری ماشین

- یادگیری نظارت شده
- یادگیری نظارت نشده
- یادگیری نیمه نظارتی
- یادگیری تقویتی

یادگیری نظارت شده یا با ناظر (supervised)

- الگوریتم‌های یادگیری ماشین با ناظر یا تحت نظارت، می‌توانند آنچه را که در گذشته آموخته شده است را به آینده تعمیم دهند و از آن‌ها برای پیش بینی استفاده کنند.
- در روش‌های با ناظر، ماشین با شروع از تجزیه و تحلیل مجموعه داده‌های شناخته شده برای آموزش، یک الگوریتم یادگیری و تابع استنباط‌شده را برای پیش‌بینی مقادیر خروجی تولید می‌کند. این سیستم پس از آموزش کافی قادر است الگوریتم ایجاد شده را برای هر ورودی جدید فراهم کند. الگوریتم یادگیری همچنین می‌تواند خروجی خود را با خروجی صحیح در نظر گرفته شده مقایسه کرده و خطاهایی را پیدا کند تا مدل را متناسب با آن اصلاح کند.


Known Data
New Response
Its an apple!
Known Response
These are apples
Model
New Data


New Mail
Not spam
Spam
Learns
Categorical Separation
Not spam
Spam
Enables the machine to be trained to classify observations into some class

- هرچه مجموعه اطلاعاتی که در اختیار ماشین قرار می‌دهید بزرگتر باشد، ماشین بیشتر می‌تواند در مورد موضوع یاد بگیرد.
- به طور خلاصه، در یادگیری با ناظر، ما یک سری ویژگی داریم و یک لیبل؛ مثلا در یک دیتاست قیمت خانه، یک سری ویژگی (تعداد اتاق‌ها، متراژ، فاصله از مرکز شهر و ...) داریم و یک لیبل (قیمت خانه). با داشتن این دیتاست و با استفاده از روش‌های یادگیری با ناظر، می‌توان مدلی ساخت که قیمت یک خانه را با گرفتن ویژگی‌های آن پیش‌بینی کند.

- به عنوان مثال، فرض کنید در این روش الگوریتم یادگیری با تصاویری از ماهی و تصاویری از اقیانوس که به ترتیب تحت عنوان Fishو Oceanبرچسب‌دار شده‌اند، مورد آموزش قرار گیرد. این الگوریتم پس از آموزش دیدن با این تصاویر و برچسب‌ها، قادر خواهد بود تا تصاویر بدون برچسب ماهی و اقیانوس را به ترتیب به عنوان Fishو Oceanمورد شناسایی قرار داده و این تصاویر را با برچسب تصاویری که با آن‌ها آموزش دیده یکسان قلمداد کند.
- یکی از کاربردهای رایج یادگیری نظارت‌شده زمانی است که با استفاده از اطلاعات گذشته قرار است اتفاقات آینده نزدیک مورد پیش‌بینی قرار گیرند. به عنوان مثال، با این روش می‌توان اطلاعات چند ماه یا چند هفته اخیر بازار سهام را برای پیش‌بینی نواسانات بازار در هفته‌ها و ماه‌های آتی مورد استفاده قرار داد. یک نمونه دیگر استفاده از این الگوریتم نیز در تشخیص ایمیل‌های اسپم (هرزنامه) از غیر اسپم است.

- مسائل یادگیری ماشین نظارت شده قابل تقسیم به دو دسته «دسته‌بندی» و «رگرسیون» هستند.

دسته‌بندی: یک مساله، هنگامی دسته‌بندی محسوب می‌شود که متغیر خروجی یک دسته یا گروه باشد. برای مثالی از این امر می‌توان به تعلق یک نمونه به دسته‌های «سیاه» یا «سفید» و یک ایمیل به دسته‌های «هرزنامه» یا «غیر هرزنامه» اشاره کرد.

رگرسیون: یک مساله هنگامی رگرسیون است که متغیر خروجی یک مقدار حقیقی مانند «قد» باشد.

در واقع در دسته‌بندی با متغیرهای گسسته و در رگرسیون با متغیرهای پیوسته کار می‌شود.

دسته بندی

- نمونه ها به دو یا چند کلاس تعلق دارند و ما می خواهیم از داده های برچسب خورده یاد بگیریم که چگونه کلاس داده های غیر برچسب را پیش بینی کنیم. یک مثال از یک مسئله طبقه بندی ، شناسایی رقمی دست نویس است که در آن هدف اختصاص هر بردار ورودی به یکی از تعداد محدودی از دسته های گسسته است. روش دیگر برای تفکر در مورد طبقه بندی ، یک فرم گسسته (بر خلاف مداوم) یادگیری تحت نظارت است که در آن یکی تعداد محدود دسته بندی دارد و برای هر یک از nنمونه های ارائه شده ، یکی از آنها تلاش برای برچسب گذاری آنها با دسته یا کلاس صحیح است. .

رگرسیون

- اگر خروجی مورد نظر از یک یا چند متغیر پیوسته تشکیل شده باشد ، آن کار را رگرسیون می نامند. یک مثال از یک مشکل رگرسیون می تواند پیش بینی طول ماهی قزل آلا به عنوان تابعی از سن و وزن آن باشد.

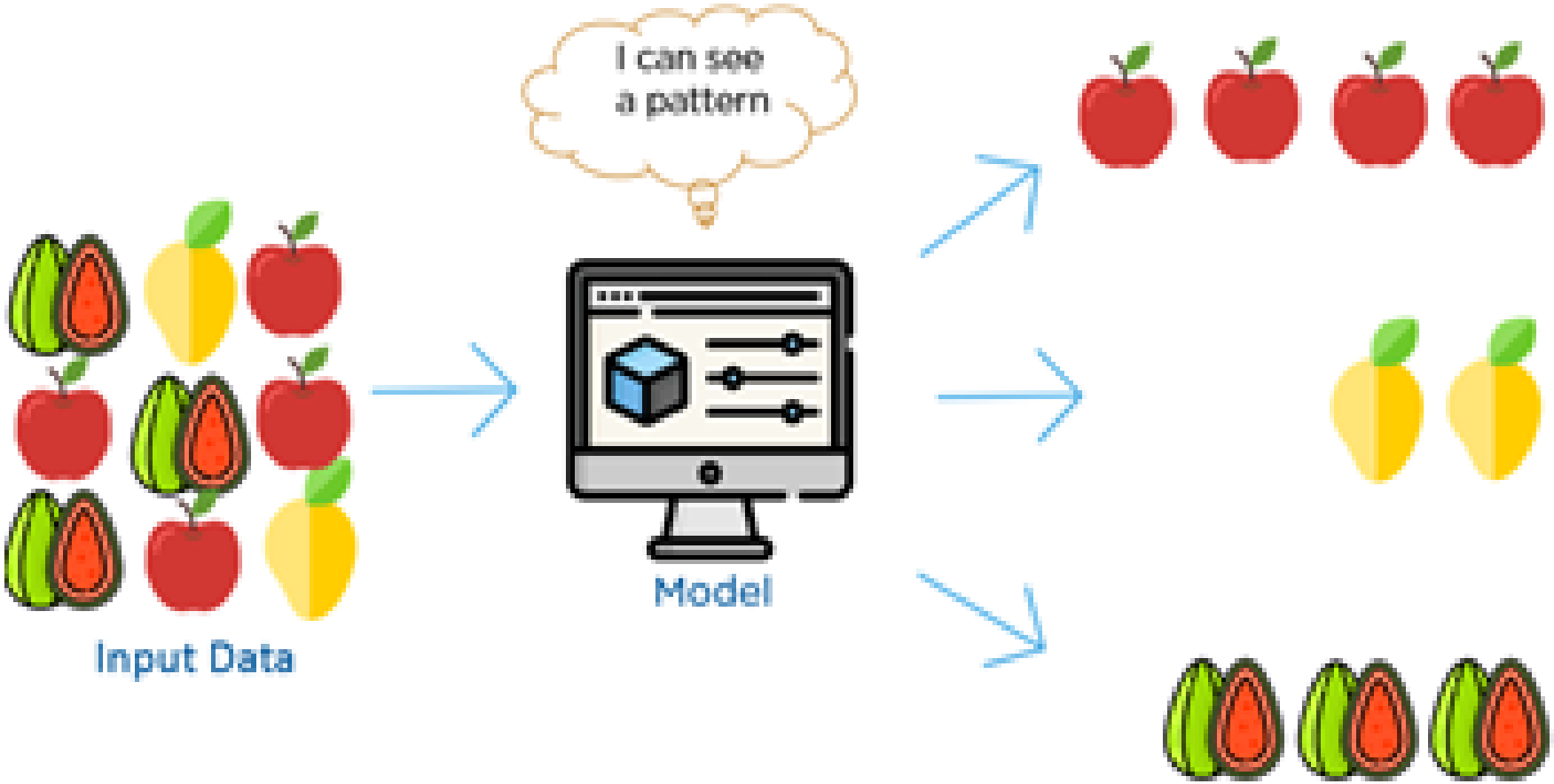
الگوریتم‌ها که در یادگیری نظارتی

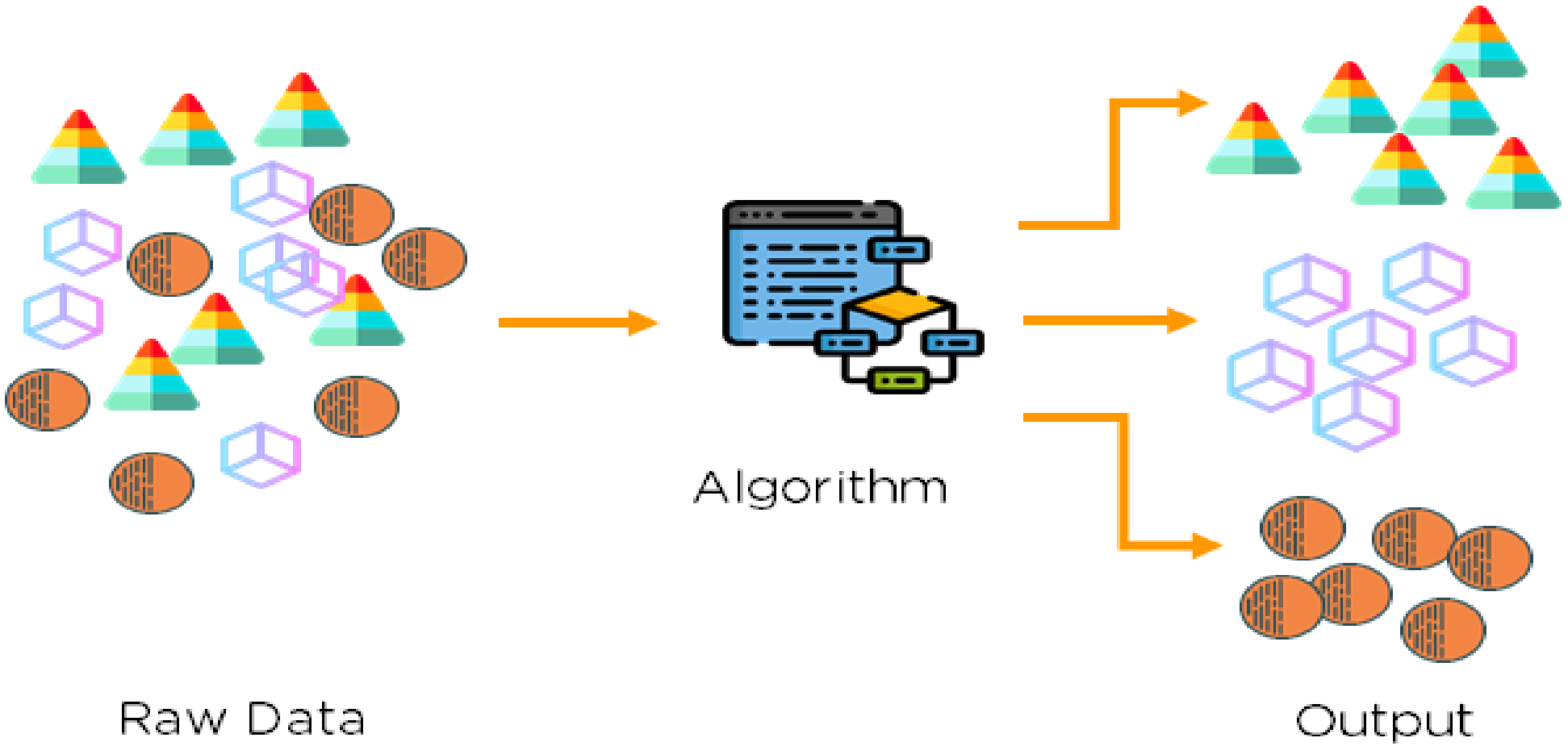
- درخت تصمیمِ Decision Tree
- دسته‌بندی کننده بیز Naive Bayes classifier
- کمینه مربعات
- رگرسیون لجستیک logistic regression

یادگیری نظارت نشده یا بدون ناظر (unsupervised)

- الگوریتم‌های بدون نظارت در مقابل الگوهای با ناظر قرار دارند. آن‌ها هنگامی به کار می‌روند که اطلاعات مورد استفاده برای آموزش نه طبقه‌بندی و نه برچسب‌گذاری شده باشند. به بیان ریاضی، یادگیری نظارت نشده مربوط به زمانی است که در مجموعه داده فقط متغیرهای ورودی Xوجود داشته باشند و هیچ متغیر داده خروجی موجود نباشد. به این نوع یادگیری، نظارت نشده گفته می‌شود زیرا برخلاف یادگیری نظارت شده، هیچ پاسخ صحیح داده شده‌ای وجود ندارد و ماشین خود باید به دنبال پاسخ باشد.
- در یادگیری بدون ناظر، سیستم‌ها می‌توانند تابعی را برای توصیف ساختار پنهان از داده‌های بدون برچسب استنباط کنند.
- روش‌های یادگیری ماشین بدون ناظر خروجی صحیح را تشخیص نمی‌دهند، اما داده‌ها را کاوش می‌کنند و می‌توانند از مجموعه داده‌ها ساختارهای پنهان آن‌ها را استنباط کنند.

- **بطور خلاصه**، یادگیری بدون نظارت ، که در آن داده های آموزش از مجموعه ای از بردارهای ورودی xو بدون هیچ مقادیر متناظر تشکیل شده است. هدف در چنین مسائلی ممکن است کشف گروههایی از نمونه های مشابه در داده ها باشد ، جایی که خوشه بندی نامیده می شود ، یا تعیین توزیع داده ها در فضای ورودی ، که به عنوان تخمین تراکم شناخته می شود ، یا پروژه پردازی داده ها از یک بعد بالا به منظور تجسم به دو یا سه بعد فاصله دهید


I can see
a pattern
Input Data
Model


Raw Data
Algorithm
Output

- یادگیری نظارت نشده مانند گوش دادن به یک فایل صوتی با زبانی ناشناس است. وقتی شما تنها به این فایل صوتی با زبان ناشناس گوش دهید، چیز زیادی دستگیرتان نمی‌شود. اما چنانچه مدت زیادی به این کار ادامه دهید، مغز شما در مورد آن زبان، شروع به ایجاد نوعی الگو می‌کند و کم کم در هنگام گوش دادن به آن پادکست، انتظار شنیدن اصوات خاصی را خواهد داشت.

- با به‌کارگیری الگوریتم یادگیری نظارت‌نشده ممکن است در نهایت مشخص شود که مثلاً خانم‌هایی که در یک بازه سنی خاص قرار دارند و صابون بدون‌بو خریداری نموده‌اند، احتمالاً باردار هستند. بدین ترتیب با شناسایی این الگو و از آنجا که این مشتریان احتمالاً به محصولات مرتبط با دوران بارداری و پس از بارداری نیاز دارند، اگر پیشنهادات مرتبط با بارداری و مراقبت نوزاد نیز در اختیار این دسته از مشتریان قرار گیرد، احتمال خرید این محصولات بالاتر خواهد رفت.

- یادگیری نظارت نشده قابل تقسیم به مسائل خوشه‌بندی و انجمنی است.
- **قوانین انجمنی:** یک مساله یادگیری هنگامی قوانین انجمنی محسوب می‌شود که هدف کشف کردن قواعدی باشد که بخش بزرگی از داده‌ها را توصیف می‌کنند. مثلا، «شخصی که کالای A را خریداری کند، تمایل به خرید کالای B نیز دارد».
- **خوشه‌بندی:** یک مساله هنگامی خوشه‌بندی محسوب می‌شود که قصد کشف گروه‌های ذاتی (داده‌هایی که ذاتا در یک گروه خاص می‌گنجند) در داده‌ها وجود داشته باشد. مثلا، گروه‌بندی مشتریان بر اساس رفتار خرید آن‌ها

الگوریتم‌ها که در یادگیری غیر نظارتی

- الگوریتم خوشه بندی Cluster analysis
- تحلیل مولفه‌های اصلی Principal Component Analysis
- تجزیه مقادیر منفرد Singular Value Decomposition
- تحلیل مولفه‌های مستقل Independent Component Analysis

یادگیری نیمه نظارتی

- الگوریتم های یادگیری ماشین نیمه نظارت شده، ویژگی‌هایی در بین روش یادگیری با ناظر و بدون ناظر قرار می‌گیرند. زیرا در این روش، بخشی از داده‌های ارائه شده برای آموزش دارای برچسب هستند و برخی بدون برچسب و دسته بندی.
- به طور معمول در روش یادگیری نیمه نظارت شده مقدار کمی از داده‌های دارای برچسب و مقدار زیادی از داده ها بدون برچسب هستند. سیستم‌هایی که از این روش استفاده می‌کنند، می‌توانند به میزان قابل توجهی دقت یادگیری را افزایش دهند.
- در این روش‌ها به کامپیوتر تنها یک سیگنال آموزشی ناقص داده می‌شود. منظور از سیگنال آموزشی ناقص، داده‌هایی است که بسیاری از خروجی‌های آن از دسترس خارج هستند.

- در این نوع یادگیری نیز مانند یادگیری نظارت نشده، داده‌های مورد استفاده برای یادگیری، برچسب گذاری نمی‌شوند. زمانی که پرسشی برای داده‌ها مطرح شد، نتیجه آن درجه بندی می‌شود.

یادگیری تقویتی (Reinforcement Learning)

- در روش‌های یادگیری ماشین تقویت شده، ماشین برای رسیدن به هدفی خاص، مثلا برنده شدن در یک مسابقه کامپیوتری تلاش می‌کند. ماشین در این روش یادگیری از آزمون و خطا برای تقویت و بهبود عملکرد خود استفاده می‌کند.
- دلیل اینکه این روش را با نام یادگیری تقویت شده می‌شناسیم، این است که ماشین با استفاده از بازخوردهای مثبت و منفی که دریافت می‌کند، می‌تواند عملکرد خود را ارتقا دهد. بنابراین در نهایت تجربه کافی به دست می‌آورد و می‌تواند به هدف مشخص شده برسد.

- یک مثال مناسب در این زمینه، ترتیب دادن یک بازی است. اگر ماشین برنده بازی شود، می‌تواند از نتیجه کار برای تقویت حرکات آینده خود در حین بازی استفاده کند. البته این نکته مهم را به خاطر بسپارید که اگر کامپیوتر تنها یک یا دو بار بازی را انجام دهد، این روش تاثیری در عملکرد آن نخواهد داشت. اما وقتی هزاران بار بازی را تکرار کند، به تدریج می‌تواند نوعی استراتژی پیروزی را شکل دهد.
- به عنوان یک مثال کاربردی، می‌توان ربات فوتبالیستی را در نظر گرفت که با قرار گرفتن در موقعیت‌های مختلف و اتخاذ تصمیم‌های متناسب با این موقعیت‌ها و رفع تدریجی خطاهای خود، سرانجام می‌آموزد که در هر موقعیتی **درست‌ترین** تصمیم را برای شوت زدن بگیرد.

ریاضیات مورد نیاز

- جبر خطی: ماتریس‌ها و عملیات روی آن‌ها، اتحاد و تجزیه، ماتریس‌های متقارن، متعامدسازی.
- نظریه آمار و احتمالات: قوانین احتمال و اصل (منطق)، نظریه بیزی، متغیرهای تصادفی، واریانس ، انحراف از معیار و امید ریاضی، توزیع‌های آماری، توزیع استاندارد.
- حساب: حساب دیفرانسیل و انتگرال، مشتقات جزئی.
- الگوریتم‌ها و بهینه‌سازی پیچیدگی‌ها: درخت‌های دودویی، هیپ، استک.

همبستگی

- Correlation همبستگی عبارت است از میزان وابستگی دو متغییر نسبت به یکدیگر که با ضریبی به نام ضریب همبستگی نشان داده می‌شود که مقدار آن عددی مابین ۱ و ۱- است. ضریب همبستگی ۰ میان دو پارامتر به این معنا است که این پارامترها هیچ‌گونه وابستگی نسبت به یکدیگر ندارند و هر چه‌قدر ضریب همبستگی دو پارامتر از ۰ دورتر باشد، به معنای این است که تغییرات دو پارامتر وابستگی بیشتری به یکدیگر دارند.
- ثبت بودن ضریب همبستگی هم به این معنا است که اگر یکی از دو پارامتر افزایش یابد، دیگری نیز افزایش خواهد یافت و اگر یکی از آن‌ها کاهش یابد، دیگری نیز کاهش خواهد یافت اما منفی بودن ضریب همبستگی به معنای وابستگی معکوس میان دو پارامتر است. به عبارت دیگر، اگر یکی از آن‌ها کاهش پیدا کند، دیگری افزایش می‌یابد و اگر یکی افزایش پیدا کند، دیگری کاهش خواهد یافت. به عنوان مثال، همبستگی میان قد و وزن انسان‌ها معمولاً یک همبستگی مثبت است و هر چه قد افراد بلندتر باشد، وزن آن‌ها نیز بیشتر است (البته همواره استثناءهایی وجود خواهد داشت.)

رگرسیون

- Regressionدر لغت به معنای «بازگشت» است. هنگامی که دو متغیر با یکدیگر همبستگی بالایی داشته باشند، رگرسیون پیش‌بینی و بیان تغییرات یک متغیر بر اساس تغییرات متغیر دیگر را امکان‌پذیر می‌سازد

رگرسیون: ضریب r2

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

then the variability of the data set can be measured with two sums of squares formulas:

- The total sum of squares (proportional to the variance of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

نصب ماژول sklearn

- pip install sklearn

رگرسیون خطی با sklearn

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import  linear_model
from sklearn.metrics import  r2_score
# data
x=np.array([0,2,4,6,8,10,12])
y=np.array([0,2,4,6,8,10,12])
X = x[:, np.newaxis] # The input data for sklearn is 2D: (samples == 3 x features == 1)
# Create linear regression object
model = linear_model.LinearRegression()
# Train the model using the training sets
model.fit(X, y)
# Make predictions using the testing set
x_test=X
y_test=y
y_pred = model.predict(x_test)
print("r2=",r2_score(y_test,y_pred))
#plot
plt.scatter(X,y,label="data")
plt.plot(x_test,y_pred,label="predict")
plt.legend()
plt.show()
```

رگرسیون چند جمله ای با numpy

```
import numpy as np

x=np.array([0,2,4,5,7,9,15,20,22])
y=np.array([-3,-1,4,3,5,9,5,6,7])
model=np.poly1d(np.polyfit(x,y,7))
x_test=x
y_test=y
y_pred=model(x_test)
import matplotlib.pyplot as plt
plt.plot(x,y,"o",label="data")
plt.plot(x_test,y_pred,label="pred")
plt.legend()
plt.xlabel("x")
plt.ylabel("y")

from sklearn.metrics import r2_score
print("r2=",r2_score(y_test,y_pred))

plt.show()
```

رگرسیون غیر خطی با scipy

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import  linear_model
from sklearn.metrics import  r2_score
from scipy.optimize import curve_fit
# data
x=np.array([0,2,4,5,7,9,15,20,22])
y=np.array([-3,-1,4,3,5,9,5,6,7])
# Create linear regression object
# Train the model using the training sets
def f(x,a,b,c,d):
    return a*x**3+b*x**2+c*x+d
popt,pocv=curve_fit(f,x,y)
a,b,c,d=popt
# Make predictions using the testing set
x_test=x
y_test=y
y_pred = f(x,a,b,c,d)
print("r2=",r2_score(y_test,y_pred))
#plot
plt.scatter(x,y,label="data")
plt.plot(x_test,y_pred,label="predict")
plt.legend()
plt.show()
```

رگرسیون دو متغیره در Sklearn

مثال

Car	Model	Volume	Weight	CO2
Toyota	Aygo	۱۰۰۰	۷۹۰	۹۹
Mitsubishi	Space Star	۱۲۰۰	۱۱۶۰	۹۵
Skoda	Citigo	۱۰۰۰	۹۲۹	۹۵
Fiat	۵۰۰	۹۰۰	۸۶۵	۹۰
Mini	Cooper	۱۵۰۰	۱۱۴۰	۱۰۵
VW	Up!	۱۰۰۰	۹۲۹	۱۰۵
Skoda	Fabia	۱۴۰۰	۱۱۰۹	۹۰
Mercedes	A-Class	۱۵۰۰	۱۳۶۵	۹۲
Ford	Fiesta	۱۵۰۰	۱۱۱۲	۹۸
Audi	A1	۱۶۰۰	۱۱۵۰	۹۹
Hyundai	I20	۱۱۰۰	۹۸۰	۹۹
Suzuki	Swift	۱۳۰۰	۹۹۰	۱۰۱
Ford	Fiesta	۱۰۰۰	۱۱۱۲	۹۹
Honda	Civic	۱۶۰۰	۱۲۵۲	۹۴
Hundai	I30	۱۶۰۰	۱۳۲۶	۹۷
Opel	Astra	۱۶۰۰	۱۳۳۰	۹۷
BMW	۱	۱۶۰۰	۱۳۶۵	۹۹
Mazda	۳	۲۲۰۰	۱۲۸۰	۱۰۴
Skoda	Rapid	۱۶۰۰	۱۱۱۹	۱۰۴

Ford	Focus	۲۰۰۰	۱۳۲۸	۱۰۵
Ford	Mondeo	۱۶۰۰	۱۵۸۴	۹۴
Opel	Insignia	۲۰۰۰	۱۴۲۸	۹۹
Mercedes	C-Class	۲۱۰۰	۱۳۶۵	۹۹
Skoda	Octavia	۱۶۰۰	۱۴۱۵	۹۹
Volvo	S60	۲۰۰۰	۱۴۱۵	۹۹
Mercedes	CLA	۱۵۰۰	۱۴۶۵	۱۰۲
Audi	A4	۲۰۰۰	۱۴۹۰	۱۰۴
Audi	A6	۲۰۰۰	۱۷۲۵	۱۱۴
Volvo	V70	۱۶۰۰	۱۵۲۳	۱۰۹
BMW	۵	۲۰۰۰	۱۷۰۵	۱۱۴
Mercedes	E-Class	۲۱۰۰	۱۶۰۵	۱۱۵
Volvo	XC70	۲۰۰۰	۱۷۴۶	۱۱۷
Ford	B-Max	۱۶۰۰	۱۲۳۵	۱۰۴
BMW	۲	۱۶۰۰	۱۳۹۰	۱۰۸
Opel	Zafira	۱۶۰۰	۱۴۰۵	۱۰۹
Mercedes	SLK	۲۵۰۰	۱۳۹۵	۱۲۰

```
import pandas
from sklearnimport linear_model
df= pandas.read_excel("car.xlsx")
X = df[['Weight', 'Volume']]
y = df['CO2']
model= linear_model.LinearRegression()
model.fit(X, y)
predictedCO2 = model.predict([[2300, 1300]])
print(predictedCO2)
print(model.coef_)
```

- We have already predicted that if a car with a 1300cm3 engine weighs 2300kg, the CO2 emission will be approximately 107g.
- These values tell us that if the weight increase by 1kg, the CO2 emission increases by 0.00755095g.
- And if the engine size (Volume) increases by 1 cm3, the CO2 emission increases by 0.00780526 g

رگرسیون لجستیک

- رگرسیون خطی، متغیر وابسته یک متغیر کمی در سطح فاصله‌ای یا نسبی است و پیش بینی کننده ها از نوع متغیرهای پیوسته، گسسته یا ترکیبی از این دو هستند. اما هنگامی که متغیر وابسته در کمی نباشد، یعنی به صورت دو یا چندمقوله‌ای باشد، از رگرسیون لجستیک استفاده می‌کنیم که امکان پیش‌بینی عضویت گروهی را فراهم می‌کند. این روش موازی روش‌های تحلیل تشخیصی و تحلیل لگاریتمی است. برای مثال، پیش بینی مرگ و میر نوزادان بر اساس جنسیت نوزاد، دوقلو بودن و سن و تحصیلات مادر.
- رأی دادن یا ندادن در انتخابات، مالکیت (مثلاٌ داشتن یا نداشتن کامپیوتر شخصی) و سطح تحصیلات (مانند: داشتن یا نداشتن تحصیلات دانشگاهی) ارزیابی می‌شود. از جمله حالت های پاسخ دوتایی عبارتند از: موافق- مخالف، موفقیت – شکست، حاضر – غایب و جانبداری – عدم جانبداری.

مثال

- هدف: ایجاد یک مدل رگرسیون لجستیک در پایتون تا تعیین کند که آیا داوطلبان در یک دانشگاه معتبر پذیرفته می شوند یا خیر.
- دو نتیجه احتمالی وجود دارد: پذیرفته شده (۱) در مقابل رد شده (۰)
- سپس می توان یک رگرسیون لجستیک در پایتون ایجاد کرد ، جایی که:
- متغیر وابسته نشان می دهد که آیا فرد پذیرفته می شود. و ۳ متغیر مستقل نمره GMAT، معدل (GPA)و سالها سابقه کار هستند

gmat	gpa	work_experience	admitted
780	4	3	1
750	3.9	4	1
690	3.3	3	0
710	3.7	5	1
680	3.9	4	0
730	3.7	6	1
690	2.3	1	0
720	3.3	4	1
740	3.3	5	1
690	1.7	1	0
610	2.7	3	0
690	3.7	5	1
710	3.7	6	1
680	3.3	4	0
770	3.3	3	1
610	3	1	0
580	2.7	4	0
650	3.7	6	1
540	2.7	2	0
590	2.3	3	0
620	3.3	2	1
600	2	1	0
550	2.3	4	0
550	2.7	1	0
570	3	2	0
670	3.3	6	1
660	3.7	4	1
580	2.3	2	0
650	3.7	6	1
660	3.3	5	1
640	3	1	0
620	2.7	2	0
660	4	4	1
660	3.3	6	1
680	3.3	5	1
650	2.3	1	0
670	2.7	2	0
580	3.3	1	0
590	1.7	4	0
690	3.7	5	1

بخش اول: بارگزاری داده ها و متغیرها

```
import pandas as pd
candidates = {'gmat':
[780,750,690,710,680,730,690,720,740,690,610,690,710,680,770,610,580,650,540,590,620,600,55
0,550,570,670,660,580,650,660,640,620,660,660,680,650,670,580,590,690],
'gpa':
[4,3.9,3.3,3.7,3.9,3.7,2.3,3.3,3.3,1.7,2.7,3.7,3.7,3.3,3.3,3,2.7,3.7,2.7,2.3,3.3,2,2.3,2.7,3,3.3,3.7,2.3,3.
7,3.3,3,2.7,4,3.3,3.3,2.3,2.7,3.3,1.7,3.7],
'work_experience': [3,4,3,5,4,6,1,4,5,1,3,5,6,4,3,1,4,6,2,3,2,1,4,1,2,6,4,2,6,5,1,2,4,6,5,1,2,1,4,5],
'admitted': [1,1,0,1,0,1,0,1,1,0,0,1,1,0,1,0,0,1,0,0,1,0,0,1,0,0,0,0,1,1,0,1,1,0,0,1,1,1,0,0,0,0,1]
}
df= pd.DataFrame(candidates,columns= ['gmat', 'gpa','work_experience','admitted'])
#print (df)
X = df[['gmat', 'gpa','work_experience']]
y = df['admitted']
```

بخش دوم: فراخوانی مدل رگرسیون لجستیک

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)
model= LogisticRegression()
model.fit(X_train,y_train)
```

تابع train_test_splitعمل تفکیک داده‌ها را انجام می‌دهیم. در اینجا ۲۵٪ داده‌ها را برای بخش آزمایش در نظر گرفته‌ایم. معمولا نسبت داده‌های آموزشی به آزمایشی به صورت ۸۰ به ۲۰ یا ۳۰ به ۷۰ در نظر گرفته می‌شود. یعنی برای مثال ۸۰ درصد داده‌ها برای مدل‌سازی و ۲۰ درصد باقی‌مانده برای برآورد خطای مدل به کار می‌رود.

بخش سوم: پیش بینی

```
y_pred=model.predict(X_test)
```

بخش چهارم:تعیین صحت

```
from sklearn import metrics
print('Accuracy: ',metrics.accuracy_score(y_test, y_pred))
```