# The Fundamentals of Machine Learning

وقتی بیشتر مردم کلمه یادگیری ماشین را می شنوند ، یک ربات را به خاطر می آورند اما یادگیری ماشین به آینده ای فانتزی مربوط نیست. دهه ها است که در برخی از برنامه های تخصصی مثل نویسه خوان نوری ( Optical Character Recognition or OCR ) وجود دارد. اولین برنامه یادگیری ماشین که زندگی صدها میلیون آدم را بهبود بخشید مربوط می شود به دهه ۱۹۹۰ و برنامه فیلتر اسپم. به دنبال آن صدها برنامه یادگیری ماشین وجود داشت و وجود دارد که اکنون بی سر و صدا به محصولاتی که به طور منظم از آن ها استفاده می کنیم اضافه شده است مثل توصیه های بهتر برای جستجوی صوتی.

## یادگیری ماشین چیست؟

یادگیری ماشین علم و هنر برنامه نویسی کامپیوتر است تا بتواند از داده ها بیاموزد.

یک تعریف کلی تر از آرتور ساموئل سال ۱۹۵۹ ارائه شده است که بیان می کند یادگیری ماشین رشته تحصیلی است که به کامپیوترها توانایی یادگیری را بدون برنامه ریزی صریح می دهد.

یک تعریف مهندسی گرانه وجود دارد از تام میچل که سال ۱۹۹۷ ارائه شده است و بیان می کند یادگیری ماشین به برنامه ای می گویند که از تجربه یا E با توجه به برخی از وظایف یا T و برخی معیارهای عملکرد یا P ، اگر عملکرد آن روی T همانطور که توسط P اندازه گیری می شود ، با تجربه E بهبود می یابد.

به عنوان مثال ، فیلتر اسپم شما یک برنامه یادگیری ماشین است که می تواند با پرچم گذاری اسپم و نمونه هایی از ایمیل معمولی یا غیر اسپم که به آن ها Ham گفته می شود ، یاد بگیرد. به نمونه هایی که سیستم برای یادگیری از آن ها استفاده می کند ، مجموعه آموزشی ( Training Set ) می گویند. به هر مثال آموزش یک نمونه آموزش ( Sample ) می گویند. در این حالت وظیفه یا T پرچم گذاری اسپم برای ایمیل های جدید است ، تجربه یا E داده های آموزشی است و اندازه گیری عملکرد یا P باید تعریف شود.

## چرا باید از یادگیری ماشین استفاده کرد؟

در نظر بگیرید که چگونه می توان از تکنیک های سنتی برای برنامه نویسی فیلتر اسپم استفاده کرد.

ابتدا ببینیم هرزنامه چگونه است. ممکن است متوجه شوید که برخی از کلمات در این موضوع زیاد مطرح می شوند. همچنین ممکن است چند الگو دیگر در نام فرستنده ، متن ایمیل و غیره مشاهده شود.

برای هر یک از الگوهایی که یافتید یک الگوریتم تشخیص می نویسید و اگر تعدادی از این الگوها شناسایی شود برنامه شما ایمیل را به عنوان اسپم پرچم گذاری می کند.

برنامه خود را تست و مراحل قبل را تکرار کنید تا به اندازه کافی خوب شود.

با این روش برنامه شما تبدیل به لیست طولانی از قوانین پیچیده تبدیل می شود.

در مقابل یک فیلتر هرزنامه مبتنی بر تکنیک های یادگیری ماشین با شناسایی الگوهای مکرر غیرمعمول کلمات در نمونه های هرزنامه در مقایسه با نمونه های غیر اسپم ، به طور خودکار می آموزد کدام کلمات و عبارات پیش بینی کننده خوبی برای هرزنامه هستند. برنامه بسیار کوتاه تر و دقیق تر است.

علاوه بر این هرزنامه نویسان اگر متوجه بشوند ، از کلمات مشابه استفاده می کنند که برنامه ما در این شرایط نیاز دارد بروز رسانی شود. اما الگوریتم های یادگیری ماشین به طور خودکار این کار را انجام می دهد.

یکی دیگر از حوزه هایی که یادگیری ماشین استفاده می شود و الگوریتم های سنتی برای تشخیص آن کافی نیست و نمی تواند جوابگو باشد تشخیص صدا هست. بدیهی است که برای یک کلمه که هر انسان ممکن است با میزان متفاوتی از شدت صدا بیان کند یا با زبان و گویش متفاوت ، اگر به صورت سنتی برنامه نویسی کنیم دچار خطا و مشکلات زیادی خواهد شد حال آنکه اگر به صورت یادگیری ماشین چند نمونه آموزشی به الگوریتم بدیم دیگر این مشکلات پیش نخواهد آمد.

به طور خلاصه یادگیری ماشین برای موارد زیر مناسب است :

- ✓ مسائلی که راه حل های موجود برای آن ها نیاز به تنظیم دستی یا لیست طولانی قوانین زیادی دارد که الگوریتم یادگیری ماشین می تواند کد را راحت تر و عملکرد بهتری داشته باشد.
- ✓ مسائل پیچیده ای که با استفاده از یک رویکرد سنتی هیچ راه حل خوبی برای آن ها وجود ندارد.
- ✓ نوسان محیط که الگوریتم های یادگیری ماشین می توانند با داده های جدید سازگار شوند.
- ✓ دریافت بینش درباره مسائل پیچیده و مقدار زیاد داده.

## انواع سیستم های یادگیری ماشین :

انواع مختلفی از سیستم های یادگیری ماشین وجود دارد که طبقه بندی آن ها در دسته های گسترده بر اساس

- ✓ این که آیا آن ها با نظارت انسانی آموزش دیده اند یا خیر ( یادگیری با ناظر ، یادگیری بدون ناظر ، یادگیری نیمه نظارت شده و یادگیری تقویتی )
- ✓ این که آیا آن ها می توانند به صورت تدریجی در حین کار ، یاد بگیرند یا خیر ( یادگیری برخط و دسته ای )

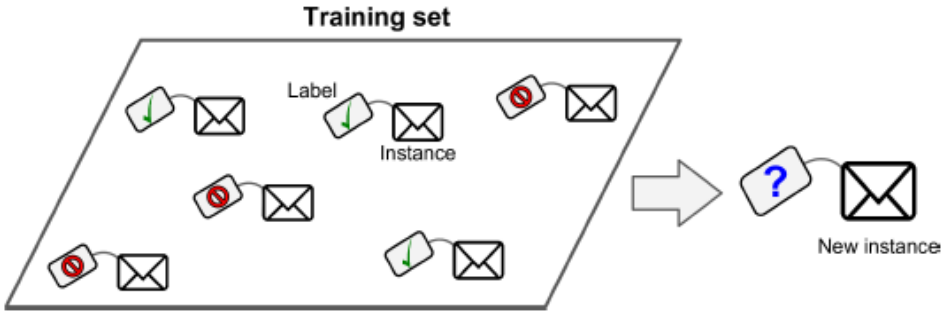
این ها منحصر به فرد نیستند و می شود به صورت ترکیبی آن ها را استفاده کرد.

❖ **یادگیری با ناظر ( Supervised Learning ) :**

در یادگیری با ناظر داده های آموزشی که به الگوریتم می دهیم شامل جواب های مورد نظر به نام برچسب ( Label ) است.

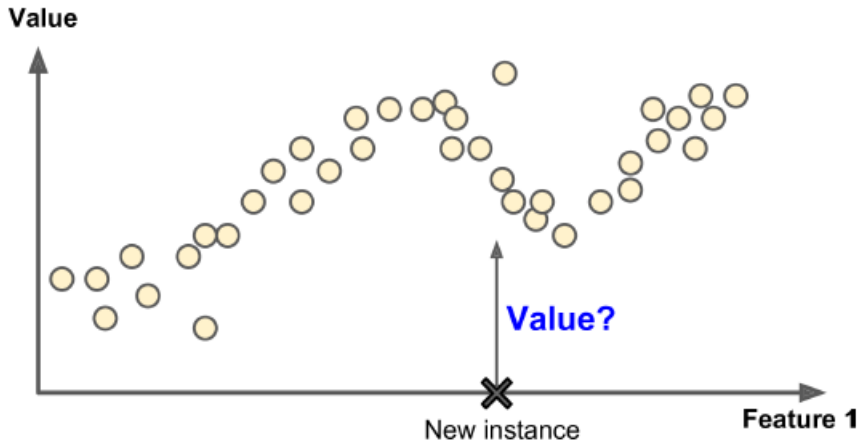
یکی از کاربرد های یادگیری تحت نظارت ، طبقه بندی ( Classification ) است. مثل فیلتر هرزنامه که طبقه بندی می کند یک ایمیل هرز است یا خیر. در این نوع یادگیری ما نمونه های بسیاری را به عنوان نمونه آموزشی می دهیم و در نهایت ایمیل جدیدی که دریافت می کند را تشخیص می دهد که هرز است یا خیر. به این نوع مسائل طبقه بندی می گویند.

در مسائل طبقه بندی خروجی الگوریتم ها مقادیر گسسته ای دارند.




مثال دیگر ، پیش بینی قیمت خودرو که شامل ویژگی هایی از جمله مسافت طی کرده ، سن خودرو ، مدل خودرو و... است. برای پیش بینی باید تعداد زیادی خودرو به عنوان نمونه بدهیم که با توجه به ویژگی ها ، در نهایت قیمت خودرو را پیش بینی می کند. به این نوع مسائل رگرسیون ( Regression ) می گویند.

در مسائل رگرسیون خروجی الگوریتم ها مقادیر پیوسته ای دارند.




می توان برخی از الگوریتم های رگرسیون را برای طبقه بندی استفاده کرد و بالعکس مانند رگرسیون لجستیک.

برخی از مهم ترین الگوریتم های یادگیری با ناظر :

- K – Nearest Neighbors ✓
- Linear Regression ✓
- Logistic Regression ✓
- Support Vector Machines ( SVM ) ✓
- Decision Trees and Random Forests ✓
- Neural Networks ✓

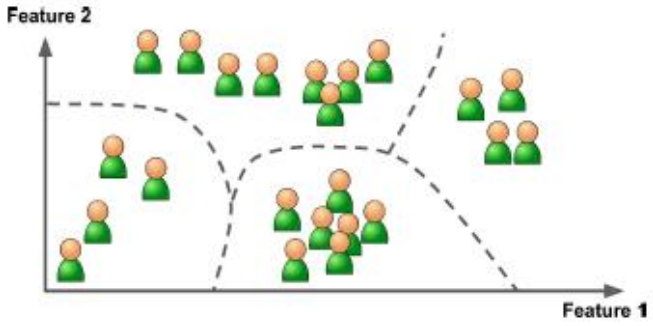
❖ **یادگیری بدون ناظر ( Unsupervised Learning ) :**

در یادگیری بدون ناظر داده های آموزش ، بدون Label هستند. این سیستم سعی می کند بدون معلم یاد بگیرد.

برخی از مهم ترین الگوریتم های یادگیری بدون ناظر به صورت زیر است :

- ✓ Clustering
- ✓ کاهش ابعاد

به عنوان مثال شما اطلاعات زیادی در مورد بازدید کنندگان وبلاگ خود دارید. اگر بخواهید با استفاده از خوشه بندی ، گروهی از بازدیدکنندگان مشابه را شناسایی کنید هیچ وقت به الگوریتم نمی گویید که بازدید کننده از کدام گروه است ، او این ارتباط را بدون کمک شما پیدا می کند.



در کاهش ابعاد هدف ساده سازی داده ها بدون از دست دادن اطلاعات بیش از حد است. یکی از راه های انجام این کار ادغام چندین ویژگی همبسته در یک ویژگی است. مثلا مسافت پیموده شده ماشین و سن ماشین بهم مرتبط هستند بنابراین الگوریتم کاهش ابعاد ، ویژگی ها را باهم ادغام می کند که در این مثال فرسودگی ماشین می شود ، به این عمل استخراج ویژگی ( Feature Extraction ) می گویند.

از کارهای مهم دیگر برای یادگیری بدون ناظر تشخیص ناهنجاری مثلا برای جلوگیری از تقلب در تراکنش های غیر معمول کارت اعتباری است. یکی دیگر از کاربردهای یادگیری بدون ناظر یادگیری قانون وابستگی ( Association Rule Learning ) است برای مثال وابستگی بین خرید چیپس و سس کچاپ که اکثر مشتریان وقتی چیپس خریداری می کنند سپس سس کچاپ نیز خریداری می کنند پس صاحب سوپر مارکت این دو را کنار هم قرار می دهد.

- **یادگیری نیمه نظارتی ( Semi Supervised Learning ) :**

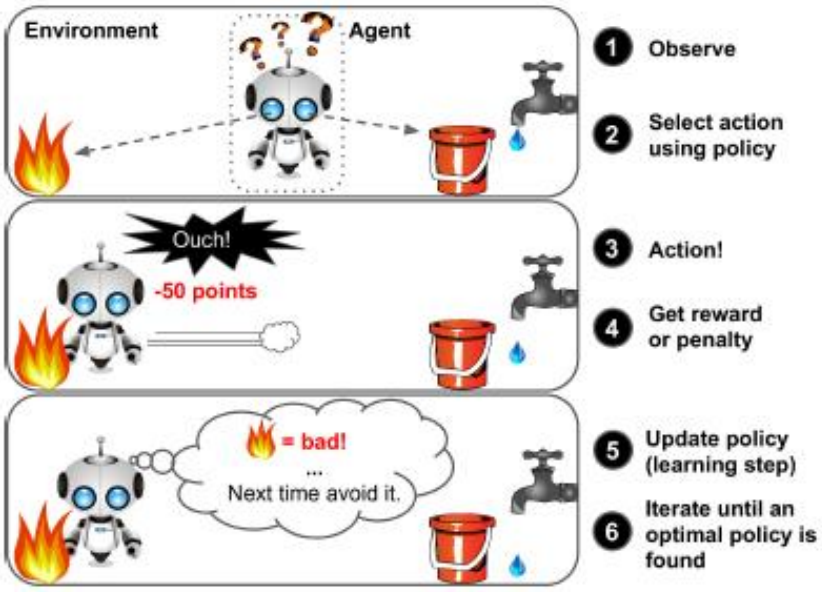
در این مدل یادگیری برخی از داده های آموزشی دارای Label هستند و برخی بدون Label. به الگوریتم با این مجموعه داده آموزشی الگوریتم نیمه نظارتی می گویند.

مثلا در Google Photos تشخیص وجود یک فرد در چند عکس مانند الگوریتم بدون نظارت است هنگامی که شما از برچسب خاصی استفاده کنید الگوریتم با نظارت به حساب می آید.

بیش تر این الگوریتم ها ترکیبی از الگوریتم های باناظر و بدون ناظر هستند.

❖ **یادگیری تقویتی ( Reinforcement Learning ) :**

سیستم یادگیری تقویتی محیط را مشاهده می کند و اقداماتی را انتخاب می کند و عمل به خصوصی را انجام می دهد که در ازای آن پاداش یا مجازات دریافت می کند و سپس باید به خودی خود بیاموزد که بهترین استراتژی برای دریافت بیشترین پاداش در طول زمان چیست. به عنوان مثال بسیاری از ربات ها ، الگوریتم های یادگیری تقویتی را برای یادگیری راه رفتن استفاده می کنند.



❖ **یادگیری دسته ای ( Batch Learning ) :**

در این الگوریتم سیستم قادر به یادگیری تدریجی نیست و باید به صورت یکجا از تمام داده های آموزشی یاد بگیرد.

این کار به زمان و محاسبات زیاد نیاز دارد بنابراین به صورت آفلاین انجام می شود. ابتدا سیستم آموزش داده می شود و سپس بدون یادگیری دیگری از آن استفاده می شود. به آن یادگیری آفلاین هم می گویند.

اگر قصد دارید از داده های آموزشی جدید استفاده کنید باید سیستم قبلی را متوقف کنید و داده های آموزشی جدید و قبلی را به عنوان مجموعه آموزشی جدید استفاده کنید.

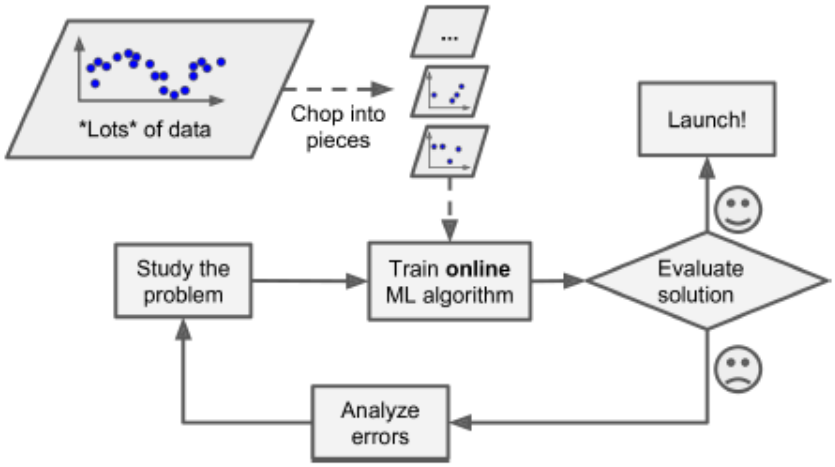
این روش به منابع محاسباتی زیاد احتیاج دارد ( CPU و حافظه و... ).

❖ **یادگیری برخط ( Online Learning ) :**

در سیستم یادگیری آنلاین سیستم را با تغذیه متوالی و تدریجی نمونه داده ها توسط گروه های کوچکی از داده ها به نام Mini-Batches به تدریج آموزش می دهند. هر مرحله سریع و ارزان است و سیستم می تواند به محض رسیدن اطلاعات جدید یاد بگیرد.

این یادگیری مثلا برای قیمت سهام بسیار عالی است زیرا سریعا در حال تغییر است. اگر به منابع محاسباتی محدود هم دسترسی داشته باشید گزینه خوبی است زیرا پس از یادگیری نیازی به داده ها ندارد. چنانچه داده های آموزشی شما زیاد باشد و حافظه اصلی ماشین کم باشد ، می توانید از این الگوریتم استفاده کنید.

یکی از پارامترهای مهم این الگوریتم ها نرخ یادگیری آن ها است که به معنی سرعت سازگاری آن ها با تغییر داده ها است.




## آموزش و تست ( Training and Testing ) :

تنها راه برای دانستن آن که الگوریتم یادگیری ماشین ، چگونه بر روی موارد جدید یا مواردی که از قبل ندیده است ، خوب کار می کند ، ارزیابی و تست آن است. یکی از روش های معمول برای ارزیابی الگوریتم و یا مدل ، تقسیم بندی داده ها به دو قسمت آموزش یا Train و تست یا Test است که بعد از آموزش الگوریتم و یا مدل با استفاده از داده های قسمت آموزش ، داده های قسمت تست را برای ارزیابی به مدل می دهند و سپس خروجی مدل را با Label های داده های تست مقایسه می کنند و به عنوان درصد خطا و یا میزان عملکرد الگوریتم در نظر می گیرند.

**نکته !** به طور معمول ۸۰ الی ۷۵ درصد داده ها را برای آموزش و ۲۰ الی ۲۵ درصد داده ها را برای تست تقسیم بندی می کنند.