یادگیری ماشین : Machine Learning

یادگیری ماشینی (ML) رویکردی در برنامه‌نویسی است که در آن رایانه‌ها به جای برنامه‌ریزی صریح راه‌حل ، از داده‌ها برای حل یک کار یاد می‌گیرند .

مثال : (email spam detector)

اگر سعی کنید آن را با استفاده از رویکرد برنامه نویسی سنتی (بدون ML) بسازید، برای نوشتن منطق برنامه خود مشکل خواهید داشت، متی به صورت دستی یک لیست کلمات هرزنامه ایجاد کنید.

از طرف دیگر، می‌توانید نمونه‌های زیادی از نامه‌های هرزنامه و نامه‌های هم‌زمان را به یک مدل یادگیری ماشینی که خود به خود یاد می‌گیرد، بدهید.

داده‌هایی که به مدل ML می‌دهیم تا یاد گیرد ، training set نامیده می‌شود.

در مثال بالا، training set مجموعه ای از ایمیل ها است که قبلاً به عنوان هرزنامه یا سالم برچسب گذاری شده اند.

انواع یادگیری ماشین :

Supervised Learning : یادگیری نظارت شده یک تکنیک یادگیری ماشینی است که در آن مدل بر روی یک مجموعه آموزشی برچسب‌گذاری شده آموزش داده می‌شود.

محبوب ترین وظایف یادگیری تحت نظارت عبارتند از:

Regression (مثلا پیش بینی قیمت خانه . برای این کار به یک مجموعه آموزشی با برچسب قیمت های دیگر خانه نیاز دارید.)

Classification (مثلا طبقه بندی ایمیل به عنوان هرزنامه/سالم . برای این کار به یک training set با عنوان spam/ham نیاز دارید.)

Unsupervised Learning : یادگیری بدون نظارت یک تکنیک یادگیری ماشین است که در آن مدل بر روی یک training set بدون برچسب آموزش داده می شود.

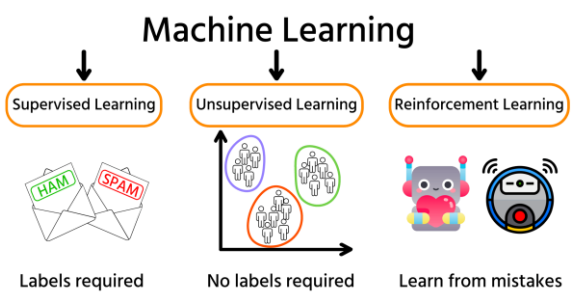
محبوب ترین کارهای یادگیری بدون نظارت عبارتند از:

Clusterization (این فرآیند گروه بندی نقاط داده مشابه در خوشه ها است. شما نیازی به برچسب گذاری داده ها برای آن ندارید. به عنوان مثال، مجموعه ای آموزشی از ایمیل های بدون برچسب اسپم/سالم انجام خواهد شد.)

Anomaly Detection (این فرآیند شناسایی انحرافات از رفتار عادی داده است . به عنوان مثال، کشف تقلب در معاملات کارت اعتباری. نیازی به برچسب نیست . به سادگی اطلاعات تراکنش را به یک مدل بدهید، که مشخص می کند آیا تراکنش صحیح است یا خیر.)

Dimensionality Reduction (این فرآیند کاهش تعداد ابعاد و در عین حال حفظ تا حد امکان اطلاعات مرتبط است . همچنین به هیچ برچسبی نیاز ندارد.)

Reinforcement Learning : یادگیری تقویتی با دو نوع قبلی تفاوت زیادی دارد. این تکنیکی است که برای آموزش وسایل نقلیه خودران، ربات ها، هوش مصنوعی در بازی و غیره استفاده میشود. یادگیری تقویتی یک تکنیک یادگیری ماشین است که در آن عامل (به عنوان مثال، ربات جاروبرقی) با تصمیم گیری و دریافت پاداش در صورت صحیح بودن تصمیم و جریمه در صورت اشتباه بودن تصمیم می آموزد. در مورد ربات جاروبرقی، اگر به یک منطقه کثیف حرکت کند، پاداش و اگر به منطقه ای که قبلا تمیز شده است، جریمه دریافت می کند. همچنین، هنگامی که کل منطقه تمیز شود، پاداش زیادی دریافت می کند.



مجموعه آموزشی (Training Set) :

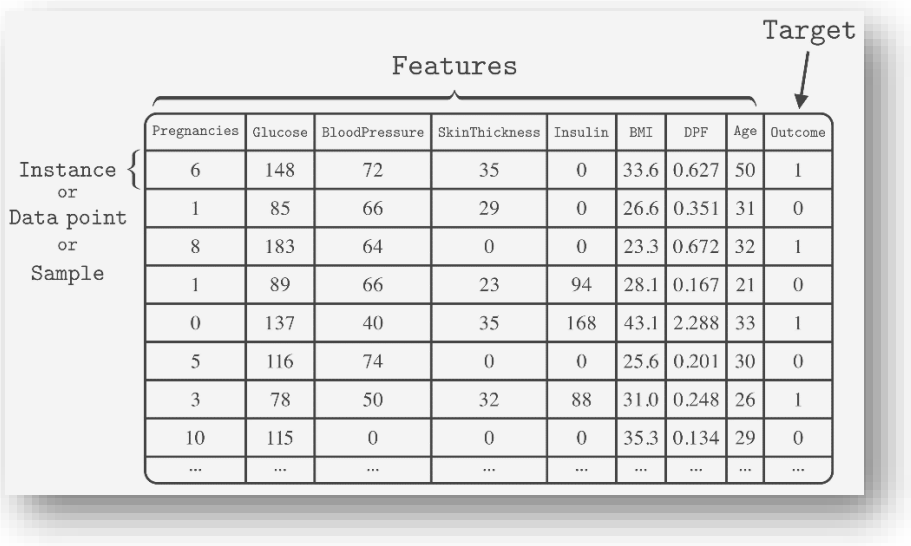
اگر در مورد یادگیری تحت نظارت یا بدون نظارت صحبت کنیم، مجموعه آموزشی معمولاً به شکل جدول خواهد بود.

مجموعه داده دیابت را در نظر بگیرید که وظیفه آن پیش بینی دیابت است یا خیر.

اطلاعات مربوط به 768 زن با پارامترهایی مانند سن، شاخص توده بدن، فشار خون و غیره را در خود دارد. این پارامترها **ویژگی** نامیده می شوند.

مجموعه داده همچنین حاوی اطلاعاتی در مورد اینکه آیا فرد مبتلا به دیابت است یا خیر، در ستون "نتیجه" وجود دارد، چیزی که ما می خواهیم پیش بینی کنیم. **هدف** نامیده می شود.

هر سطر در یک جدول، **نمونه** (یا نقطه داده یا نمونه) نامیده می شود. در این مورد، اطلاعات مربوط به یک زن است.



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31.0	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
...	...	...	...	...	...	...	...	...

جدول (مجموعه آموزشی) دارای یک ستون هدف در آن است، به این معنی که برچسب گذاری شده است.

وظیفه آموزش مدل ML بر روی این مجموعه آموزشی است، و پس از آموزش، می‌تواند برای افراد دیگر (نمونه‌های جدید) پیش‌بینی کند که آیا آنها دیابت دارند یا خیر.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age	Outcome
2	122	70	27	0	36.8	0.34	27	?
5	121	72	23	112	26.2	0.245	30	?
1	126	60	0	0	30.1	0.349	47	?
1	93	70	31	0	30.4	0.315	23	?

توجه داشته باشید :

مجموعه آموزشی باید تا حد امکان با موارد جدید مرتبط باشد. به عنوان مثال، این مجموعه داده دیابت حاوی اطلاعاتی در مورد زنان حداقل 21 ساله است، بنابراین این مدل می‌تواند پیش‌بینی‌های بدتری را در مورد نمونه‌های جدید مرد در مقایسه با زنان انجام دهد.

X

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30
3	78	50	32	88	31.0	0.248	26
10	115	0	0	0	35.3	0.134	29
...	...	...	...	...	...	...	...

y

Outcome
1
0
1
0
1
0
1
0
...

هنگام کدنویسی، ستون های ویژگی معمولاً به ***X*** و ستون های هدف به عنوان ***y*** اختصاص داده می شوند و ویژگی های نمونه های جدید به عنوان ***X_new*** اختصاص داده می شود.

X_new

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DPF	Age
2.0	122.0	70.0	27.0	0.0	36.8	0.34	27.0
5.0	121.0	72.0	23.0	112.0	26.2	0.245	30.0
1.0	126.0	60.0	0.0	0.0	30.1	0.349	47.0
1.0	93.0	70.0	31.0	0.0	30.4	0.315	23.0

انواع داده ها :

هر ستون (ویژگی) در یک مجموعه آموزشی دارای یک نوع داده مرتبط با آن است. این نوع داده ها را می توان به عددی، دسته بندی، و تاریخ و (یا) زمان گروه بندی کرد.

Numerical

Age
50
31
32
21
33
30
26
29
...

Dates

Date
2023-04-01
2023-04-05
2023-04-10
2023-04-15
2023-04-20
2023-04-25
2023-04-28
2023-05-03
...

Categorical

Sex
MALE
FEMALE
MALE
FEMALE
FEMALE
MALE
FEMALE
MALE
...

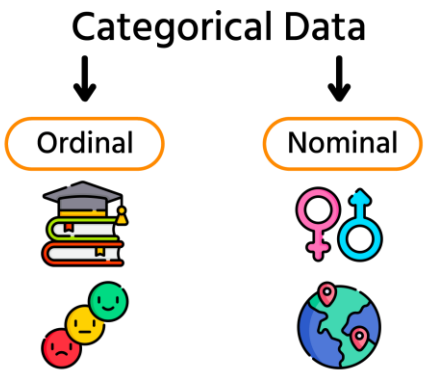
متأسفانه اکثر الگوریتم های ML فقط با اعداد ، خوب کار می کنند. بنابراین ما به راهی برای تبدیل داده های categorical و داده های تاریخ به اعداد نیاز داریم.

با توجه به تاریخ و زمان، می توانید بر اساس وظیفه خود از ویژگی هایی مانند 'سال'، 'ماه' و غیره استفاده کنید. این ویژگی ها مقادیر عددی هستند، بنابراین مشکلی با آنها وجود ندارد. پردافتن به داده های categorical کمی چالش برانگیزتر است.

Types of categorical data :

Ordinal data : داده‌های ترتیبی نوعی از داده‌های طبقه‌بندی هستند که در آن دسته‌ها از نظم طبیعی پیروی می‌کنند. مثلاً سطح تحصیلات (از دبستان تا دکتری) یا میزان (از خیلی بد به خیلی خوب) و غیره.

Nominal data : داده‌های اسمی نوعی از داده‌های طبقه‌بندی هستند که از ترتیب طبیعی پیروی نمی‌کنند. به عنوان مثال، نام، جنسیت، کشور مبدا و غیره.



Machine Learning Workflow :

Step 1. Get the data : برای این مرحله باید مشکل را تعریف کنید و اینکه چه داده هایی مورد نیاز است. سپس، یک معیار را انتخاب کنید و مشخص کنید که چه نتیجه ای رضایت بخش خواهد بود.

در مرحله بعد، باید این داده ها را با هم جمع آوری کنید، معمولاً از چندین منبع (پایگاه داده) در قالبی مناسب برای پردازش بیشتر در پایتون. گاهی اوقات داده ها از قبل در قالب csv. هستند و آماده برای پیش پردازش هستند و می توان از این مرحله صرف نظر کرد.

Step 2. Preprocess the data :

این مرحله شامل:

پاکسازی داده ها – برخورد با مقادیر از دست رفته، داده های غیر عددی و غیره.

تجزیه و تحلیل داده های اکتشافی (EDA) – (Exploratory data analysis) تجزیه و تحلیل و تجسم مجموعه داده ها برای یافتن الگوها و روابط بین ویژگی ها و به طور کلی، به دست آوردن بینشی در مورد چگونگی بهبود مجموعه آموزشی.

مهندسی ویژگی – انتخاب، تبدیل، یا ایجاد ویژگی های جدید بر اساس بینش های EDA برای بهبود عملکرد مدل.

Step 3. Modeling :

این مرحله شامل:

انتخاب مدل – در این مرحله، شما یک یا چند مدل را انتخاب می کنید که بهترین عملکرد را در مورد مشکل شما دارند. درک الگوریتم و آزمایش‌ها را با مدل‌ها ترکیب می‌کند تا مدل‌های مناسب برای مشکل شما را پیدا کند.

Hyperparameter Tuning – فرآیندی برای یافتن هایپرپارامترهایی که به بهترین عملکرد منجر می شوند.

فراپارامترها را به‌عنوان دستگیره‌ها و صفحه‌های روی یک دستگاه در نظر بگیرید که می‌توانید برای کنترل نحوه عملکرد آن تنظیم کنید. در یادگیری ماشین ، این شستی ها و شماره گیری ها تنظیمات (مقادیر) هستند که یک دانشمند داده قبل از شروع آموزش مدل خود آنها را تنظیم می کند. برای مثال، فراپارامترها ممکن است شامل مدت زمان آموزش مدل یا جزئیات آموزش باشد.

ارزیابی مدل – اندازه گیری عملکرد مدل بر روی داده های دیده نشده.

Step 4. Deployment :

هنگامی که یک مدل تنظیم شده دارید که عملکرد خوبی را نشان می دهد، می توانید آن را اجرا کنید. اما این جایی نیست که کار شما تمام می شود. بیشتر اوقات، شما همچنین می خواهید عملکرد مدل مستقر شده را زیر نظر داشته باشید، راه هایی برای بهبود آن بیابید و داده های جدید را در حین جمع آوری تغذیه کنید که این اتفاق شما را به مرحله 1 برمی گرداند.