Framing: اصطلاحات کلیدیML

یادگیری ماشینی (با نظارت) چیست؟ به طور خلاصه به شرح زیر است

- سیستم‌های ML یاد می‌گیرند که چگونه ورودی‌ها را برای تولید پیش‌بینی‌های مفید روی داده‌هایی که قبلاً دیده نشده‌اند، ترکیب کنند

بیایید اصطلاحات اساسی یادگیری ماشین را بررسی کنیم.

Labels

یک **Labels** چیزی است که ما پیش‌بینی می‌کنیم - متغیر y در رگرسیون خطی ساده. این label می تواند قیمت آتی گندم، نوع حیوان نشان داده شده در تصویر، معنای یک کلیپ صوتی یا تقریباً هر چیزی باشد.

Features

یک **feature** یک متغیر ورودی است - متغیر x در رگرسیون خطی ساده. یک پروژه یادگیری ماشینی ساده ممکن است از یک ویژگی استفاده کند، در حالی که یک پروژه یادگیری ماشینی پیچیده‌تر می‌تواند از میلیون‌ها ویژگی استفاده کند که به شرح زیر است:

$$x_1, x_2, \ldots, x_N$$

در مثال آشکارساز spam، ویژگی ها می تواند شامل موارد زیر باشد:

- کلمات در متن ایمیل
- آدرس فرستنده
- ساعت از روز ایمیل ارسال شد
- ایمیل حاوی عبارت "one weird trick" است.

Examples

یک example یک نمونه خاص از داده ها، x است. (ما x را با خط پررنگ قرار می دهیم تا نشان دهیم که بردار است.) مثال ها را به دو دسته تقسیم می کنیم

- labeled examples ----- نمونه های برچسب گذاری شده
- unlabeled examples ----- نمونه های بدون برچسب

یک **labeled example** شامل هر دو ویژگی (ها) و برچسب است. به این معنا که:

```
-- labeled examples: {features, label}: (x, y)
```

برای آموزش مدل(train) از labeled examples استفاده کنید. در مثال spam ما، نمونه‌های برچسب‌گذاری‌شده ایمیل‌های فردی هستند که کاربران به‌صراحت آن‌ها را به‌عنوان «spam» یا «not spam » علامت‌گذاری کرده‌اند.

به عنوان مثال، جدول زیر 5 نمونه برچسب گذاری شده از مجموعه داده حاوی اطلاعاتی در مورد قیمت مسکن در کالیفرنیا را نشان می دهد:

housingMedianAge (feature)	totalRooms (feature)	totalBedrooms (feature)	medianHouseValue (label)
15	5612	1283	66900
19	7650	1901	80100
17	720	174	85700
14	1501	337	73400
20	1454	326	65500

یک **unlabeled example**حاوی ویژگی‌ها است اما label ندارد. به این معنا که:

```
unlabeled examples: {features, ?}: (x, ?)
```

در اینجا 3 unlabeled examples از همان مجموعه داده مسکن وجود دارد که مستثنی هستند medianHouseValue :

housingMedianAge (feature)	totalRooms (feature)	totalBedrooms (feature)
42	168	361
34	1226	180
33	1077	271

هنگامی که مدل خود را با labeled examples آموزش دادیم، از آن مدل برای پیش بینی برچسب روی unlabeled examples استفاده می کنیم. در آشکارساز spam، نمونه‌های unlabeled examples ایمیل‌های جدیدی هستند که انسان‌ها هنوز برچسب‌گذاری نکرده‌اند.

Models

یک مدل رابطه بین ویژگی ها و برچسب را تعریف می کند. به عنوان مثال، یک مدل تشخیص spam ممکن است ویژگی های خاصی را به شدت با "spam" مرتبط کند. بیایید دو مرحله از زندگی یک مدل را برجسته کنیم:

- **Training** (آموزش) به معنای ایجاد یا **learning** (یادگیری) مدل است. به این معنا که شما نمونه هایی با برچسب مدل را نشان می دهید و مدل را قادر می سازید تا به تدریج روابط بین ویژگی ها و برچسب را یاد بگیرد.
- **Inference** استنتاج به معنای به کارگیری مدل آموزش دیده برای نمونه های بدون برچسب است. یعنی از مدل آموزش دیده برای پیش بینی های مفید .(y') استفاده می کنید. به عنوان مثال، در طول استنتاج، می توانید medianHouseValue را برای نمونه های جدید بدون برچسب پیش بینی کنید.

Regression vs. classification

یک مدل رگرسیون مقادیر پیوسته را پیش بینی می کند. به عنوان مثال، مدل های رگرسیون پیش بینی هایی را انجام می دهند که به سؤالاتی مانند زیر پاسخ می دهند:

- ارزش خانه در کالیفرنیا چقدر است؟
- احتمال اینکه کاربر روی این تبلیغ کلیک کند چقدر است؟

یک مدل طبقه بندی (classification) مقادیر گسسته را پیش بینی می کند. به عنوان مثال، مدل های طبقه بندی پیش بینی هایی را انجام می دهند که به سؤالاتی مانند زیر پاسخ می دهند:

- آیا یک پیام ایمیل داده شده spam است یا not spam ؟
- آیا این تصویر یک سگ، گربه یا همستر است؟