یادگیری ماشینی چیست؟

یادگیری ماشینی، زیرمجموعه ای از علم داده، مطالعه علمی الگوریتم های محاسباتی و مدل های آماری برای انجام وظایف خاص از طریق الگوها و استنتاج به جای دستورالعمل های صریح است. یادگیری ماشینی را می توان مجموعه ای از ابزارها برای ساخت مدل ها بر روی داده ها توصیف کرد. دانشمندان داده داده‌ها را کاوش می‌کنند، مدل‌هایی را انتخاب می‌کنند و می‌سازند (ماشین)، پارامترهایی را تنظیم می‌کنند که یک مدل با مشاهدات (یادگیری) مطابقت داشته باشد، سپس از مدل برای پیش‌بینی و درک جنبه‌های داده‌های نادیده جدید استفاده می‌کنند.

یادگیری تحت نظارت و بدون نظارت

در یادگیری ماشینی، ما در مورد یادگیری تحت نظارت و بدون نظارت صحبت می کنیم. یادگیری تحت نظارت زمانی است که ما یک هدف شناخته شده داریم (که برچسب نیز نامیده می شود) بر اساس داده های گذشته (به عنوان مثال، پیش بینی قیمت یک خانه به چه قیمتی به فروش می رسد) و یادگیری بدون نظارت زمانی است که یک پاسخ قبلی شناخته شده وجود ندارد (مثلاً تعیین موضوعات مورد بحث در بررسی رستوران).

در این ماژول، رگرسیون خطی، یک الگوریتم یادگیری ماشینی تحت نظارت را بررسی خواهیم کرد. در ماژول های آینده، ما همچنین یک الگوریتم یادگیری ماشین نظارت شده دیگر، طبقه بندی، و همچنین یک الگوریتم یادگیری ماشین بدون نظارت، خوشه بندی را بررسی خواهیم کرد.
مشکلات رگرسیون و طبقه بندی هر دو مسائل یادگیری تحت نظارت هستند.

Scikit-learn

Scikit-learn، یکی از شناخته‌شده‌ترین کتابخانه‌های یادگیری ماشین در پایتون برای یادگیری ماشین، تعداد زیادی از الگوریتم‌های رایج را پیاده‌سازی می‌کند. صرف نظر از نوع الگوریتم، سینتکس همان گردش کار را دنبال می کند: import > instantiate > fit > predict. هنگامی که کاربرد و نحو اصلی Scikit-learn برای یک مدل درک شد، تغییر به یک الگوریتم جدید ساده است. بنابراین برای بقیه دوره ما با scikit-learn کار خواهیم کرد تا مدل های یادگیری ماشین را در موارد استفاده مختلف بسازیم.

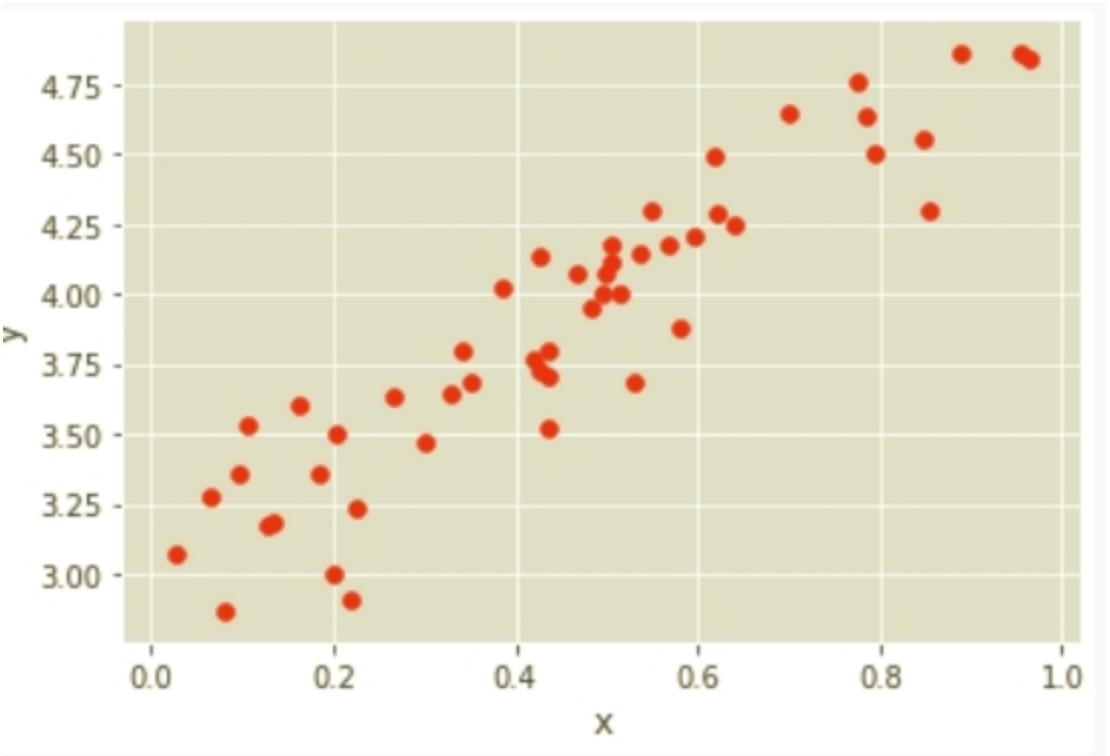
در این ماژول، نحوه پیش بینی قیمت مسکن در بوستون، ایالات متحده آمریکا با استفاده از رگرسیون خطی را یاد می گیریم.

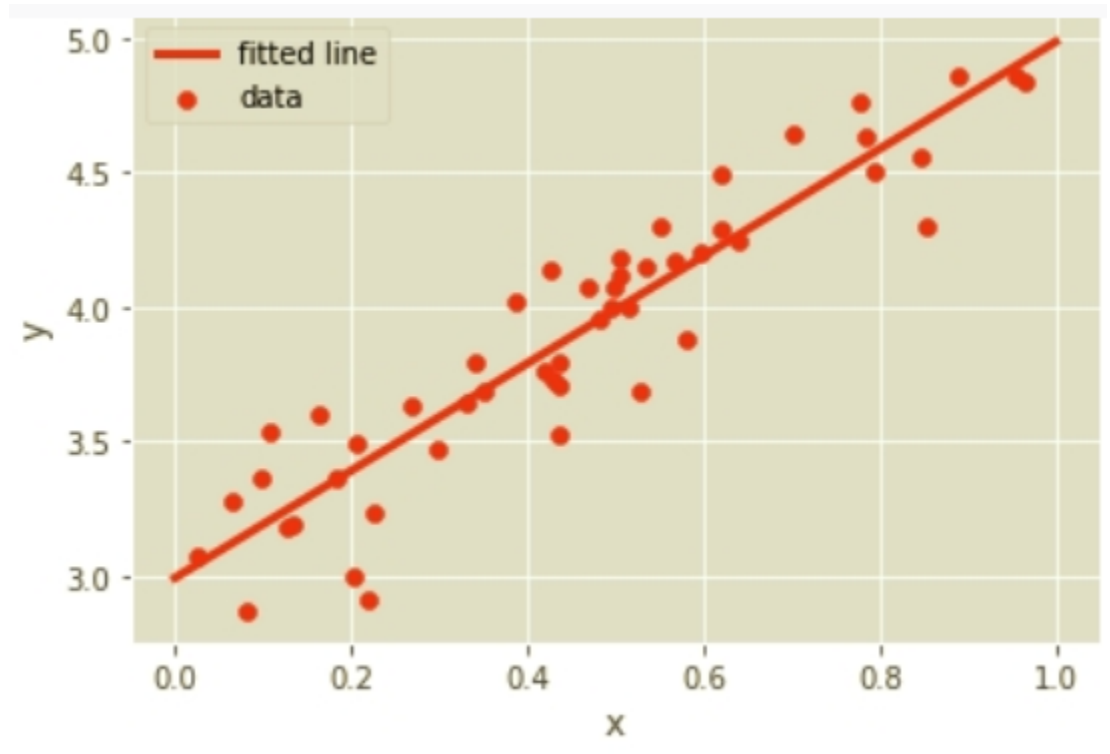
Linear Regression

ما با رگرسیون خطی، یک مدل یادگیری نظارت شده ساده شروع می کنیم. رگرسیون خطی با یک خط مستقیم به داده ها مطابقت دارد، از نظر ریاضی:

$$y = b + m*x$$

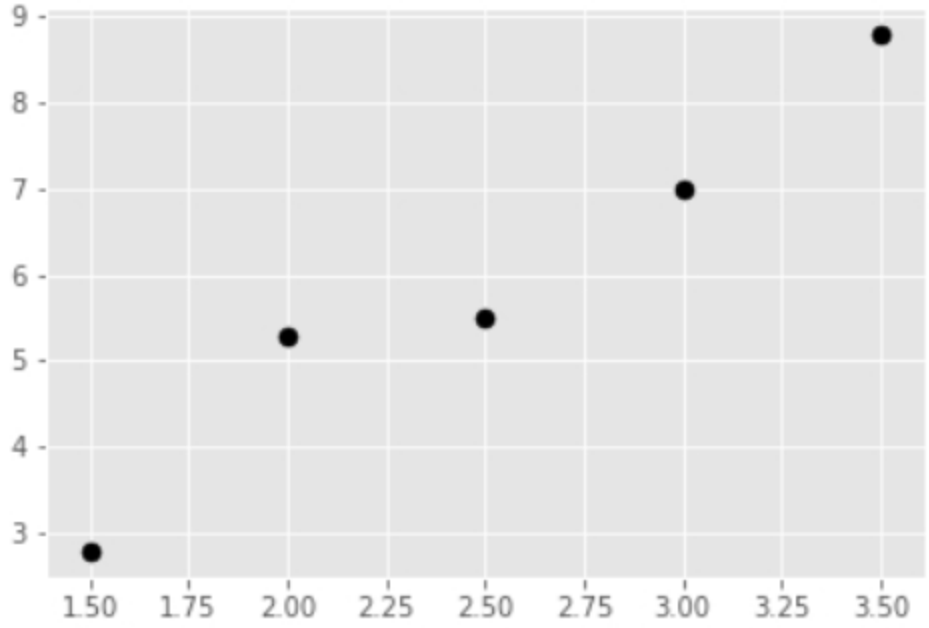
جایی که b نقطه قطع و m شیب است، x یک ویژگی یا یک ورودی است، در حالی که y یک برچسب یا یک خروجی است. وظیفه ما یافتن m و b است تا خطاها به حداقل برسد.



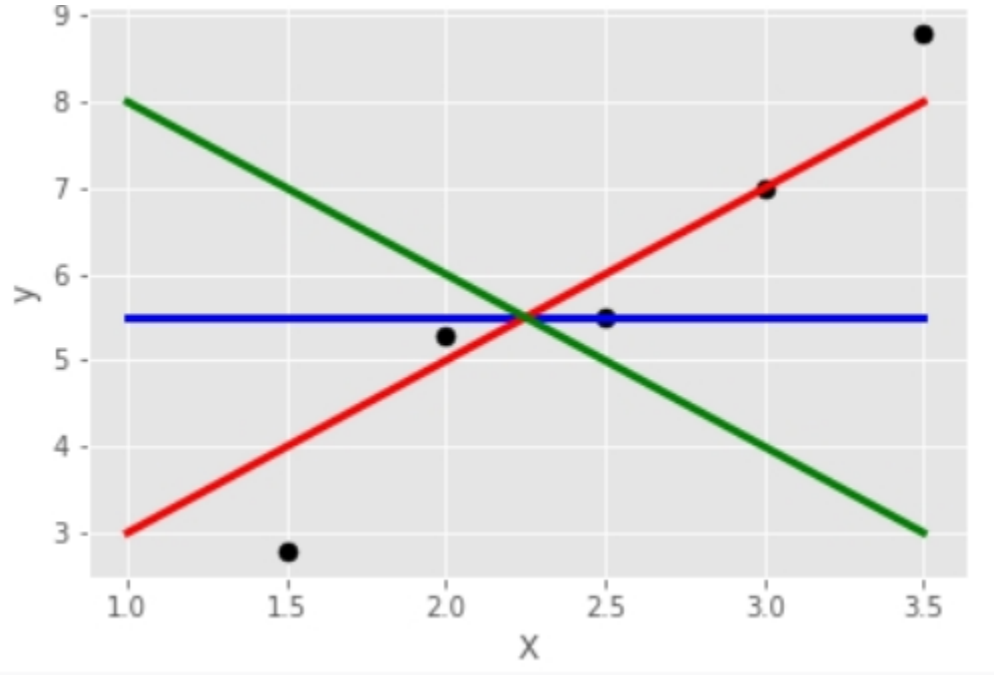


برای تجسم

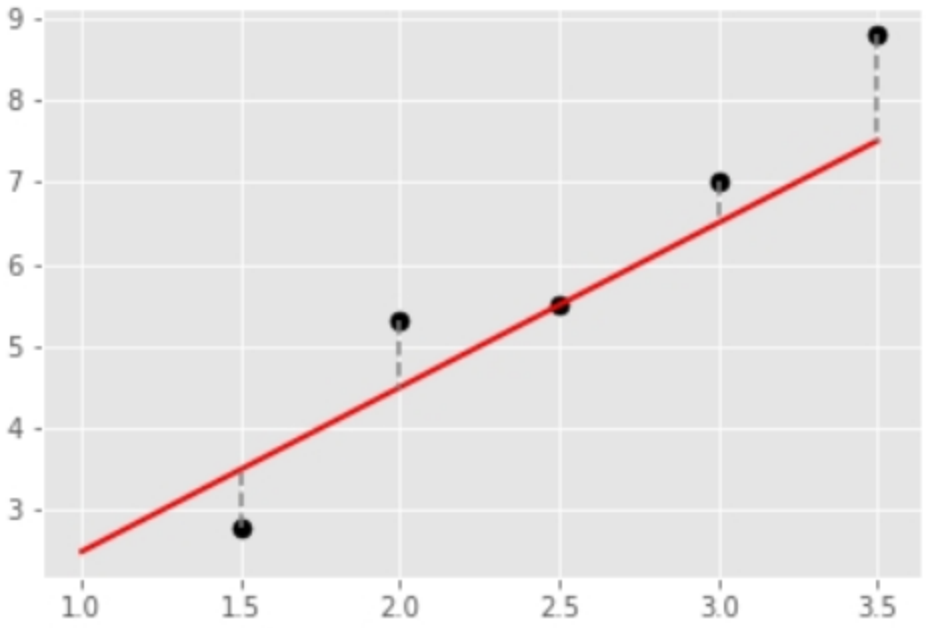
مفهوم، اجازه دهید با پنج نقطه (1.5، 2.8)، (2، 5.3)، (2.5، 5.5)، (3، 7)، (3.5، 8.8) شروع کنیم:



ما می‌خواهیم خطی را در میان این نقاط داده قرار دهیم، با این حال حتی با نگاه کردن به آن، خطی وجود ندارد که از هر پنج نقطه عبور کند، بنابراین بهترین کار را انجام خواهیم داد. این یعنی چی؟ از بین سه خط نشان داده شده در زیر، به نظر شما کدام یک با داده ها مناسب تر است؟ خط سبز y =10 + (-2)*X، خط آبی y = 5.5 + 0*X و خط قرمز y = 1 + 2*X است:



خط قرمز! چرا؟ زیرا رابطه خطی بین X و y را به بهترین نحو نشان می دهد و به نقاط نزدیکتر است. از نظر ریاضی، فاصله بین خط برازش و نقاط داده با باقیمانده ها محاسبه می شود که با خط عمودی سیاه چین در نمودار زیر نشان داده شده است:



بنابراین رگرسیون خطی اساساً یافتن خطی است که مجموع مجذور باقیمانده را به حداقل می رساند که بعداً به آن خواهیم پرداخت.