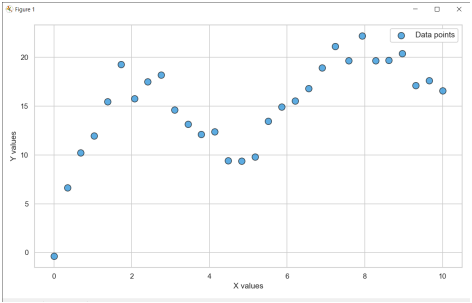
یادگیری با نظارت

رگرسیون

المپیاد هوش مصنوعی - یادگیری ماشین

یک مسئله

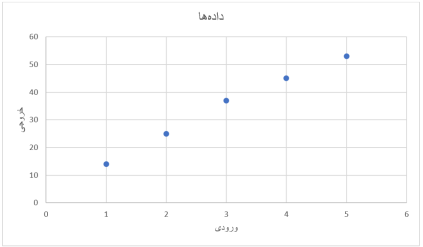
داده‌هایی به شکل (x_i, y_i) به شما داده شده که در آن x_i ویژگی ورودی است و y_i متغیر خروجی. شما می‌دانید که بین x_i و y_i یک رابطه وجود دارد، امّا نمی‌دانید چه رابطه‌ای. هدف شما این است که به ازای هر ویژگی ورودی جدید که آن را تا به حال ندیده‌اید، بتوانید مقدار متغیر خروجی آن را پیشبینی کنید. چطور این کار را انجام می‌دهید؟



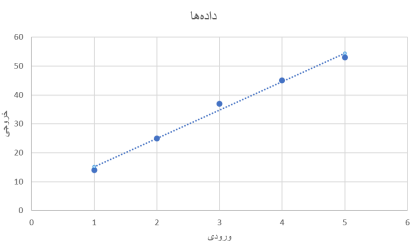
شکل: نمودار داده‌هایی که داریم.

رگرسیون خطی

رگرسیون یک روش یادگیری ماشین نظارت‌شده است که برای پیش‌بینی مقادیر پیوسته به کار می‌رود. برای شروع، فرض کنید به شما تعدادی نقطه در یک صفحه داده شده و شما می‌خواهید یک خط از روی آنها رد کنید :



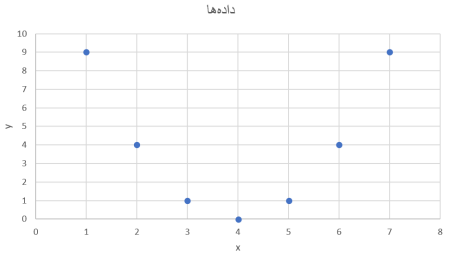
شکل: داده‌ها.



شکل: از روی داده‌ها یک خط رد کردیم.

رگرسیون غیرخطی

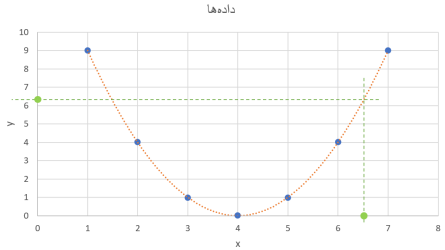
البته گاهی بهتر است از روی داده‌ها یک منحنی غیرخطی رد کنیم. فرض کنید ۷ نقطه در فضای دو بعدی به شما داده شده است که نمودار آنها مانند شکل زیر است. حالا اگر نقطه‌ای با $x = 6.5$ داشته باشیم، حدس می‌زنید مقدار y آن چه قدر باشد؟



شکل: داده‌ها.

رگرسیون غیرخطی

برای پاسخ به سوال صفحه‌ی قبل، یک روش معقول این است که یک منحنی از روی داده‌ها رد کنیم و سپس مقدار y را به ازای $x = 6.5$ از روی آن بخوانیم.



شکل: از روی داده‌ها یک منحنی رد کردیم.

چندجمله‌ای‌ها

احتمالاً می‌دانید که به عبارت‌هایی با شکل زیر، یک چند جمله‌ای درجه n می‌گویند:

$$\sum_{i=0}^{n} a_i x^i = a_0 + a_1 x^1 + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n \qquad (a_n \neq 0)$$

برای مثال خطی که از دو نقطه‌ی $(0,0)$ و $(2,3)$ رد می‌شود، می‌تواند با چندجمله‌ای $f(x) = \frac{3}{2}x$ نشان داده شود (یعنی $a_0 = 0, a_1 = \frac{3}{2}$) زیرا $f(0) = 0, f(2) = 3$.

آیا می‌توانید یک چندجمله‌ای درجه ۱ دیگر مثال بزنید که از دو نقطه‌ای که گفتیم بگذرد؟

چندجمله‌ای‌ها

جواب سوال صفحه‌ی قبل، خیر است.

حال فرض کنید همان دو نقطه‌ی $(0,0)$ و $(2,3)$ را داریم ولی این بار می‌خواهیم یک چندجمله‌ای درجه ۲ از روی آنها رد کنیم. سه چندجمله‌ای درجه ۲ متفاوت پیدا کنید که از روی این دو نقطه بگذرند.

چندجمله‌ای‌ها

$$f(x) = -x^2 + \frac{7}{2}x + 0$$
$$g(x) = \frac{3}{4}x^2 + 0x + 0$$
$$h(x) = x^2 - \frac{1}{2}x + 0$$

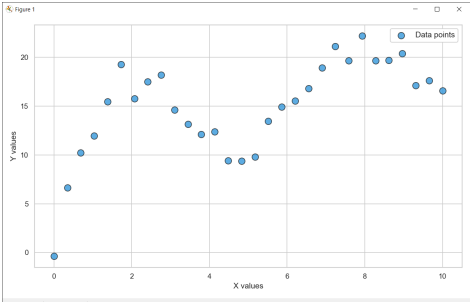
تمام چندجمله‌ای‌های درجه ۲ی بالا از نقاط $(0,0)$ و $(2,3)$ می‌گذرند! در واقع، تمام معادلاتی که شروط زیر در آنها برقرار است از این دو نقطه می‌گذرند:

$$P(x) = ax^2 + bx + c$$

$$\begin{cases} P(0) = 0 \implies a \cdot 0 + b \cdot 0 + c = 0 \\ P(2) = 3 \implies a \cdot 4 + b \cdot 2 + c = 3 \end{cases} \implies \begin{cases} c = 0 \\ 4a + 2b = 3 \end{cases}$$

رگرسیون

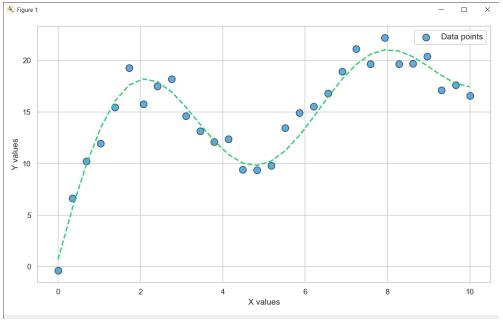
یک مسئله: داده‌هایی به شکل (x_i, y_i) به شما داده شده که در آن x_i ویژگی ورودی است و y_i متغیر خروجی. شما می‌دانید که بین x_i و y_i یک رابطه وجود دارد، امّا نمی‌دانید چه رابطه‌ای. هدف شما این است که به ازای هر ویژگی ورودی جدید که آن را تا به حال ندیده‌اید، بتوانید مقدار متغیر خروجی آن را پیشبینی کنید. چطور این کار را انجام می‌دهید؟



شکل: نمودار داده‌هایی که داریم.

رگرسیون

یک روش معقول این است که منحنی‌ای رسم کنیم که از نقاطی که داریم بگذرد، و حدس بزنیم رابطه‌ی میان x و y توسط آن منحنی مشخص می‌شود.

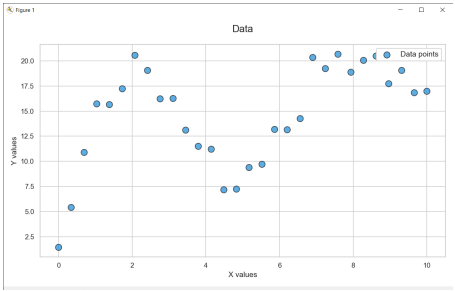


شکل: منحنی‌ای که حدس می‌زنیم رابطه‌ی میان x و y را مشخص می‌کند.

کدام منحنی؟

امّا بی‌نهایت منحنی مختلف وجود دارند که همگی از داده‌های ما عبور می‌کنند! این سوال پیش می‌آید که کدام یک را از میانشان انتخاب کنیم؟ به دلایلی همچون سادگی، ما معمولاً ترجیح می‌دهیم که فرض کنیم میان x و y یک رابطه‌ی چندجمله‌ای وجود دارد. امّا باز هم باید به یک سوال جواب دهیم: درجه‌ی چند جمله‌ای را چند در نظر بگیریم؟

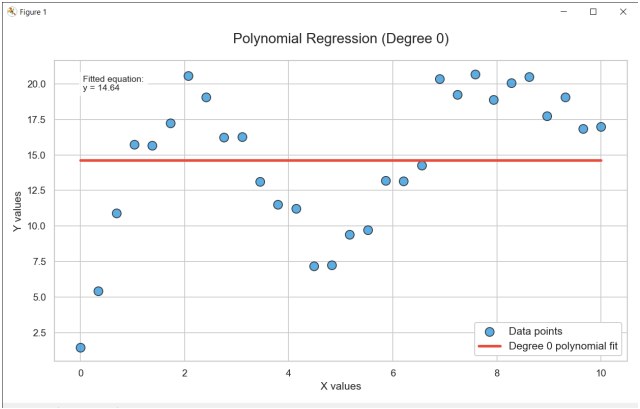
برای مثال به داده‌های زیر توجه کنید:



شکل: داده‌ها.

کدام چندجمله‌ای؟

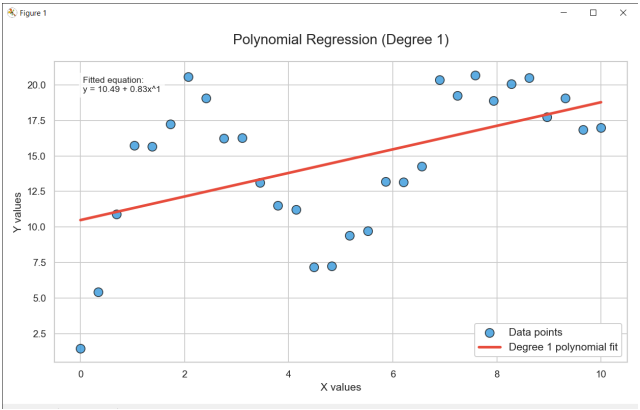
آیا یک مدل ثابت (چندجمله‌ای درجه صفر) رابطه‌ی بین x و y را به طور مناسب توصیف می‌کند؟



شکل: حدس ما از رابطه‌ی میان x و y.

کدام چندجمله‌ای؟

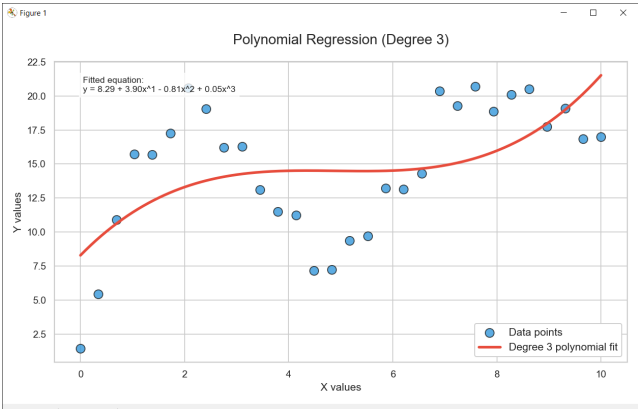
آیا یک مدل خطی (چندجمله‌ای درجه یک) رابطه‌ی بین x و y را به طور مناسب توصیف می‌کند؟



شکل: حدس ما از رابطه‌ی میان x و y.

کدام چندجمله‌ای؟

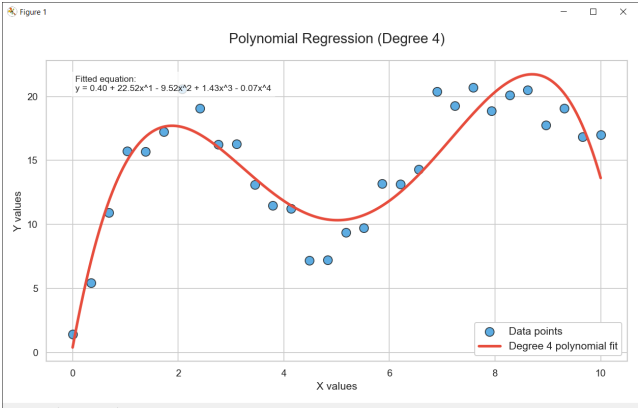
آیا یک چندجمله‌ای درجه سه رابطه‌ی بین x و y را به طور مناسب توصیف می‌کند؟



شکل: حدس ما از رابطه‌ی میان x و y.

کدام چندجمله‌ای؟

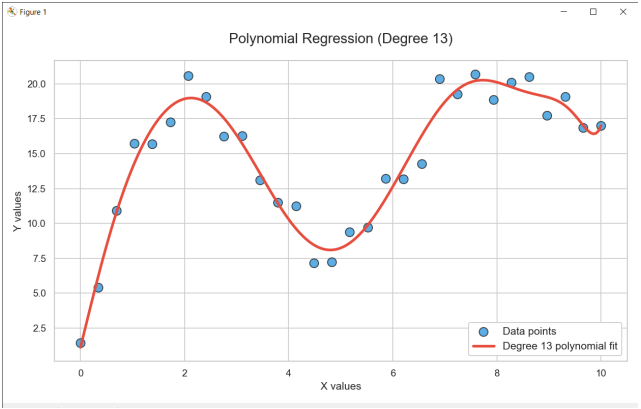
آیا یک چندجمله‌ای درجه چهار رابطه‌ی بین x و y را به طور مناسب توصیف می‌کند؟



شکل: حدس ما از رابطه‌ی میان x و y.

کدام چندجمله‌ای؟

آیا یک چندجمله‌ای درجه سیزده رابطه‌ی بین x و y را به طور مناسب توصیف می‌کند؟



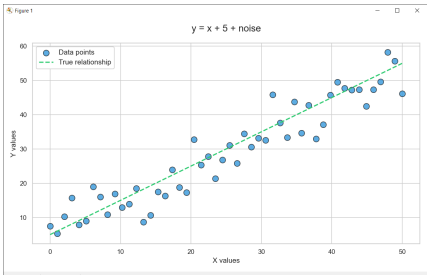
شکل: حدس ما از رابطه‌ی میان x و y.

کدام چندجمله‌ای؟

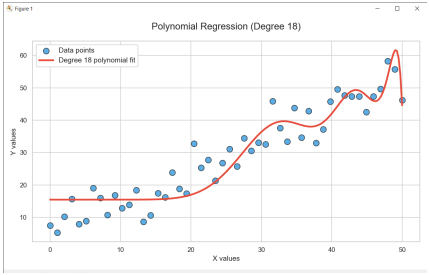
همانطور که در صفحات قبل دیدید، هر چه قدر درجه‌ی چندجمله‌ای بالاتر باشد، می‌تواند روابط پیچیده‌تری را توصیف کند (یعنی منحنی پر پیچ و خم‌تری داشته باشد). پس آیا می‌توان نتیجه گرفت هر چه درجه را بیشتر بگیریم بهتر است؟

کدام چندجمله‌ای؟

خیر! به مثال زیر توجّه کنید. فرض کنید میان ورودی‌ها و خروجی‌ها رابطه‌ی $y = x + 5$ برقرار است، امّا به دلایلی مثل خطای اندازه‌گیری یا قطعی نبودن رابطه و وجود عناصر تصادفی، داده‌هایی که جمع‌آوری کردیم مقداری با خط $y = x + 5$ فاصله دارند و دقیقاً روی آن نمی‌افتند. حال اگر سعی کنیم از روی این داده‌ها یک خط رد کنیم بهتر است یا یک چندجمله‌ای با درجه‌ی زیاد؟



شکل: داده‌ها و رابطه‌ی واقعی میان ورودی و خروجی.



شکل: چندجمله‌ای درجه ۱۸ از روی داده‌ها رد کردیم.

کدام چندجمله‌ای؟

همان‌طور که دیدید، باید سعی کنیم پیچیدگی مدل منطبق بر پیچیدگی رابطه‌ی واقعی باشد. نه بیشتر، نه کمتر.
هر چه درجه‌ی چندجمله‌ای بیشتر باشد، می‌تواند مدل‌های پیچیده‌تری را توصیف کند؛ یعنی منحنی پر پیچ و خم‌تری خواهد داشت.

کدام چندجمله‌ای درجه k؟

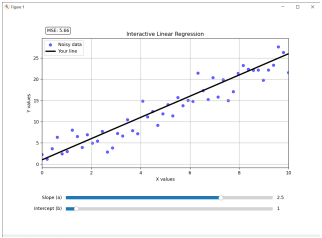
تا اینجای کار گفتیم که دوست داریم رابطه‌ی میان ورودی (ویژگی‌ها) و خروجی را با یک چندجمله‌ای مدل کنیم. سپس کمی در مورد انتخاب درجه‌ی چندجمله‌ای بحث کردیم. حال فرض کنید تصمیم گرفتیم که رابطه‌ی بین ورودی و خروجی را با یک چندجمله‌ای درجه k مدل کنیم. یعنی:

$$\hat{y} = f(x) = \sum_{i=0}^{k} a_i x^i$$

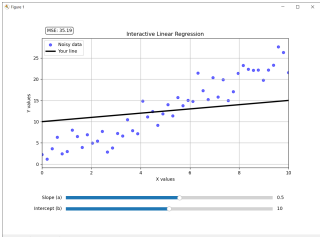
که در آن $\hat{y}$ نشان‌دهنده‌ی پیش‌بینی ما از مقدار واقعی یا همان y است. گام بعدی این است که ضرایب را بیابیم، یعنی a_iها. یعنی تصمیم بگیریم دقیقاً کدام چندجمله‌ای درجه k را برای پیش‌بینی استفاده کنیم. به بیان دیگر، مجموعه‌ی فرض ما مجموعه‌ی تمام چندجمله‌ای‌های درجه k است و ما می‌خواهیم در این مجموعه برای یافتن بهترین گزینه جست‌وجو کنیم.

کدام خط؟

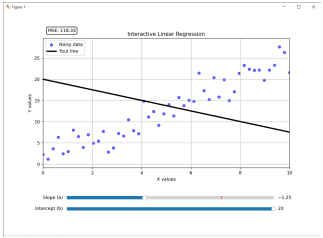
فرض کنید داده‌های به شکل (x_i, y_i) داریم و تصمیم گرفتیم که می‌خواهیم رابطه‌ی میان ورودی و خروجی را با یک خط مدل‌سازی کنیم. حال باید از فضای تمام خطوط، مناسب‌ترین خط را پیدا کنیم. چطور؟



شکل: خط ۳.



شکل: خط ۲.



شکل: خط ۱.

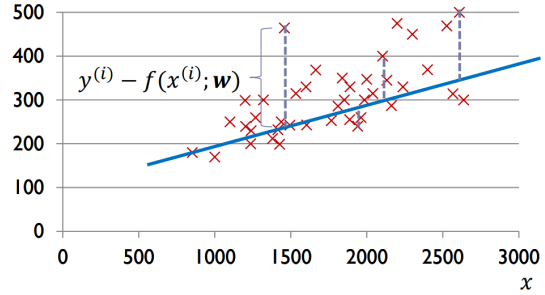
کدام خط؟

ما دوست داریم که داده‌ها تا جای ممکن به خطمان نزدیک باشند. برای مثال اگر داده‌ی $(1500, 480)$ را داشته باشیم ولی خط ما برای $x = 1500$ مقدار $\hat{y} = 250$ را پیش‌بینی کند، پیش‌بینی مدل ما به مقدار $480 - 250 = 230$ از مقدار واقعی فاصله داشته است. پس اگر مقدار واقعی خروجی را y و مقدار پیش‌بینی شده توسط مدلمان را $\hat{y}$ بنامیم، می‌توانیم خطای مدل برای این داده را به صورت $y - \hat{y}$ تعریف کنیم. امّا به دلایلی (از جمله این که دوست داریم فاصله همیشه نامنفی باشد)، ترجیح می‌دهیم خطای پیش‌بینی یک داده را به شکل $(y - \hat{y})^2$ تعریف کنیم. به خطای مدل، "هزینه"ی مدل نیز گفته می‌شود.

تابع هزینه

اگر بخواهیم تابع هزینه‌ی مدل برای کل داده‌ها را تعریف کنیم، یک روش این است که بین خطاهای مدل برای داده‌ها میانگین بگیریم. یعنی (اگر n داده داشته باشیم):

$$Mean\ Squared\ Error\ (MSE) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



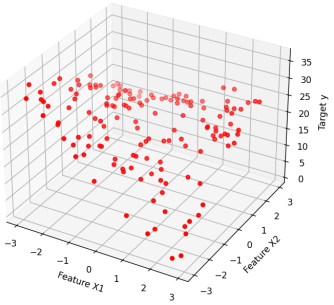
تا اینجا

فهمیدیم که برای مدل‌سازی یک رابطه در دنیای واقعی، باید ابتدا مجموعه‌ی فرضمان را مشخص کنیم و سپس در بین اعضای آن مجموعه جست‌وجو کنیم تا عضوی را پیدا کنیم که مقدار تابع هزینه برای آن کمترین مقدار را دارد. این عملیّات جست‌وجو به کمک روش‌هایی مثل $Gradient\ Descent$ انجام می‌شود و به لطف کتابخانه‌های پایتون، شما لازم نیست از جزئیات ریاضیات آن سر در بیاورید و خود کتابخانه برای شما کار جست‌وجو را انجام می‌دهد! امّا انتخاب دو چیز با شماست:

1. مجموعه‌ی فرض.
2. تابع هزینه.

ابعاد بالاتر

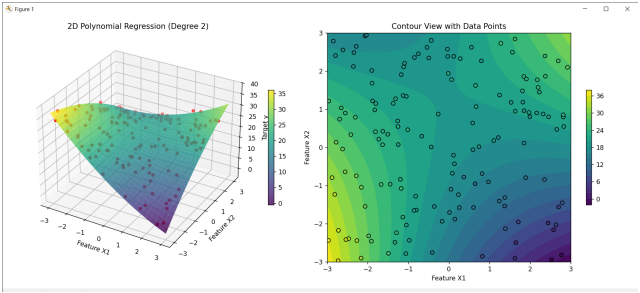
این بار فرض کنید ورودی به جای یک بعد، دو بعد دارد. یعنی هر داده، دو ویژگی دارد که می‌خواهیم به کمک آنها، مقدار متغیر خروجی آن را پیش‌بینی کنیم. فرض کنید داده‌هایمان را به شکل زیر در فضا نشان داده‌ایم.



شکل: نمودار Y بر حسب $X1$ و $X2$.

ابعاد بالاتر

همانطور که در صفحه (فضای دو بعدی) از روی داده‌ها یک منحنی رد می‌کردیم، اینجا کافیست از روی داده‌ها یک رویه (صفحه‌ای با پیچ و خم) رد کنیم. برای قشنگی به نقاطی که Y بزرگ دارند رنگ زرد، و به نقاطی که Y کوچک دارند رنگ بنفش اختصاص دادیم. تصویر سمت راست، همان فضای سمت چپ است که از بالا به آن نگاه شده.



شکل: از روی داده‌ها یک رویه رد کردیم.

ابعاد بالاتر

رگرسیون را می‌توان برای داده‌هایی با هر ابعادی انجام داد. البته دیگر نمی‌توانیم برای نمایش آن از نموداری مثل مثال‌های قبلی استفاده کنیم و باید به نمایشش با عبارات ریاضی بسنده کنیم، زیرا رسم فضای دو بعدی و فضای سه بعدی را بلدیم ولی رسم فضایی با ابعاد بالاتر را نه :)