Day 4 - Linear Regression

Agenda:

- Review: Supervised vs Unsupervised Learning
- Regression vs Classification
- Parametric Regression, Linear Regression
- Least Squares Formulation
- Solving Least Squares Formulation
- Issues to Pay Attention To with Linear Regression

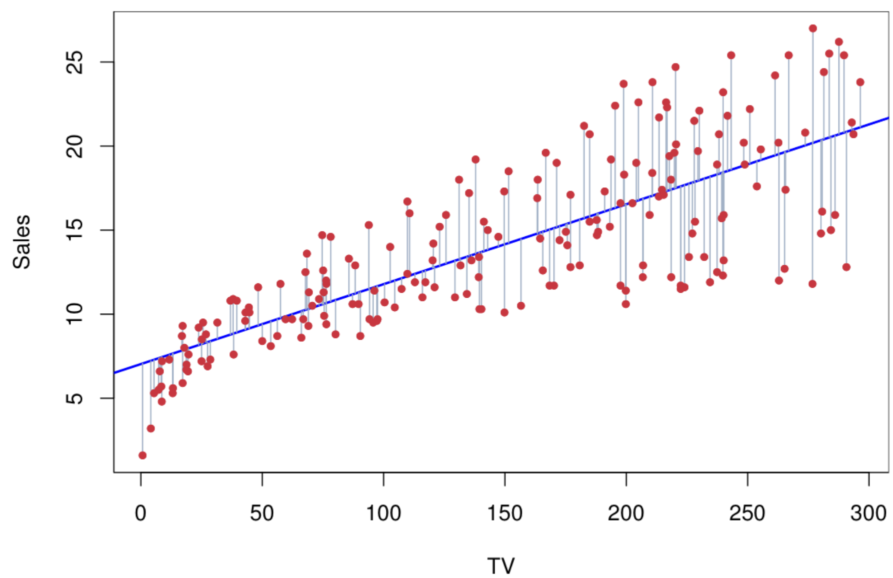


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend in the left of the plot.*

Regression:

Given $\{(x^{(i)}, y_i)\}_{i=1}^{n}$ where $y_i \approx f(x^{(i)})$,

Find f

continuously valued

Supervised Learning

Supervised vs Unsupervised Learning

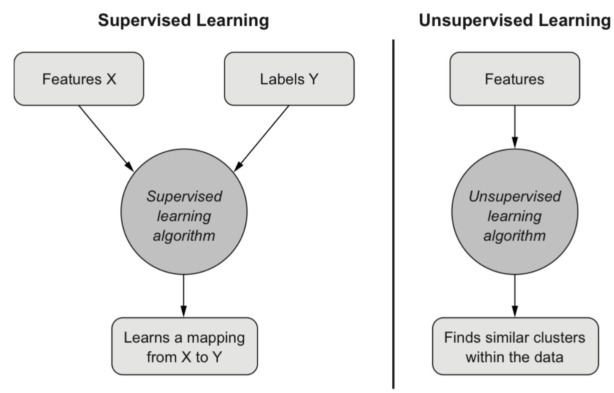


Figure 3.6 The differences in input and output of supervised and unsupervised algorithms

Supervised Learning Pipeline

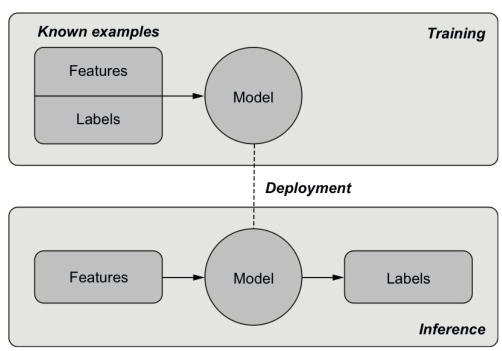
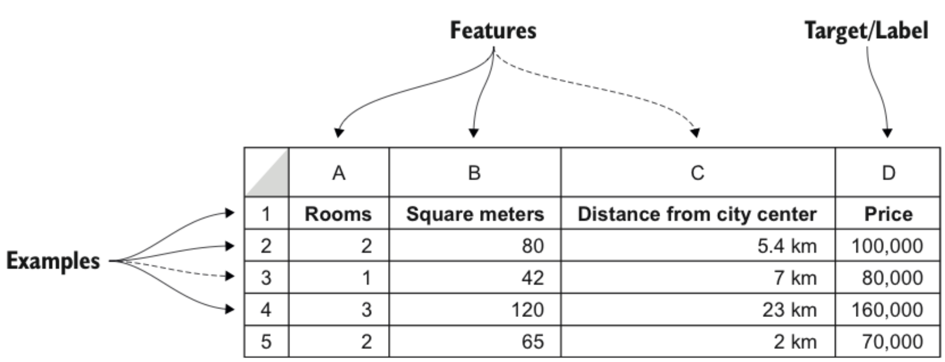


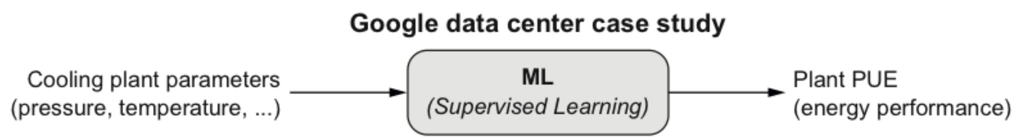
Figure 2.3 The two phases of machine learning: training and inference

Example of features and labels from the context of home prices:



	A	B	C	D
1	Rooms	Square meters	Distance from city center	Price
2	2	80	5.4 km	100,000
3	1	42	7 km	80,000
4	3	120	23 km	160,000
5	2	65	2 km	70,000

Figure 2.2 An Excel sheet with features and labels for several examples



Regression and Classification

Regression: Predict a **CONTINUOUS** value/label/output based on features/inputs

Classification: Predict a **DISCRETE/CATEGORICAL** label based on features

Examples:

Regression:
- Determine the energy efficiency of a data center given cooling operational data

- Determine the value a home will sell for given features of the home

Classification:
- Given an image of a telephone phone, determine if there is rust on it

- Given an image of a dog, determine what breed of dog it is.

Activity: Decide if these are regression problems or classification problems

- You are have a conveyer belt of cucumbers. A photograph is taken and the cucumber is identified as either high, medium, or low quality.

- You are given a resume of a person, and you predict the salary they will be offered for a job if hired.

Activity: Would you approach the following task as a regression or a classification problem?

The company Square has access to the financial transactions (on its platform) for a given company. The company asks for a $2000 loan. Square needs to decide if they will approve or deny the loan request.

Figure 2.7 How the house-prediction example and the Square and Google case studies used supervised learning

Regression:

Features: Account balances, recent financial transactions, other market conditions
Training Labels: Market valuation OR account balance in two weeks

Classification:

Features: Account balances, recent financial transactions, other market conditions
Training Labels: Did they repay loan (requested in the past)

Predict:

Mathematically

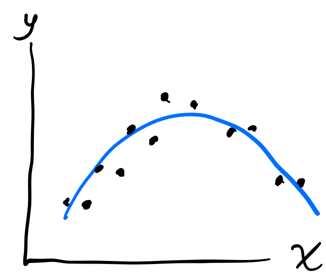
Regression: predict a continuous value

Let $f: \mathbb{R}^d \to \mathbb{R}$

$y = f(x) + \text{noise}$

Given: $\{(x^{(i)}, y_i)\}_{i=1\cdots n}$

Find: f



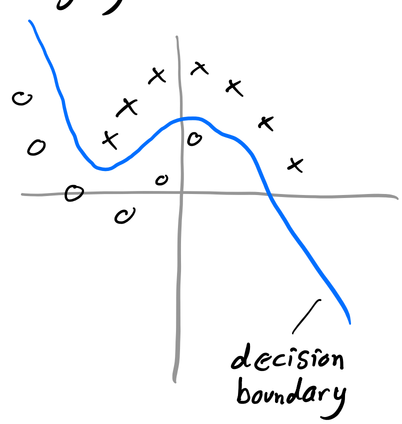
Classification: predict membership in a category

Let $f: \mathbb{R}^d \to \left\{\begin{array}{c}\text{cat } 1\\ \vdots \\ \text{cat } m\end{array}\right\}$

$y = f(x) + \text{noise}$

Given: $\{(x^{(i)}, y_i)\}_{i=1\cdots n}$

Find: f



Terminology:

- x – input variables, predictors, independent vars, features
- y – response, dependent variable, output variable
- f – model, predictor, hypothesis

The parametric approach to regression

With a parametric approach to regression, you **assume the relationship of the input to output is given by a function, f, with unknown parameters**. This requires making **modeling assumptions** on f.

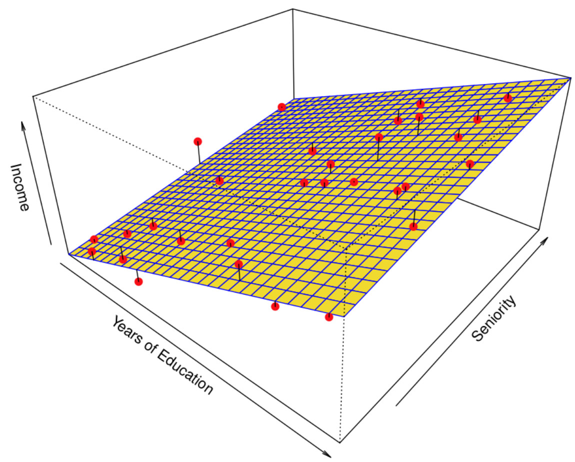


FIGURE 2.4. *A linear model fit by least squares to the* `Income` *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

Example of data model

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$

unknown parameters

Mathematically,

Consider $X \in \mathbb{R}^2$, $y = f(X) \in \mathbb{R}$ of form

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{error}$$

for unknown $\beta_0, \beta_1, \beta_2$

Or perhaps we want a model that is not linear in its input:

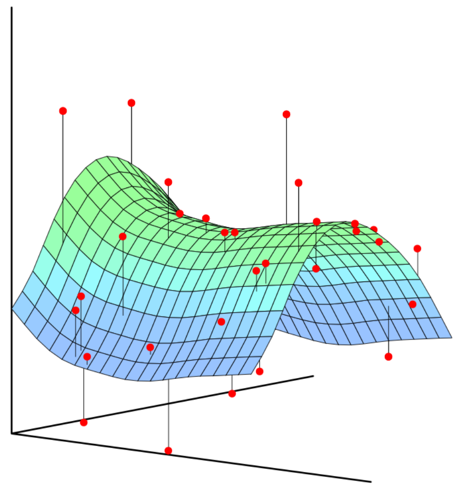


FIGURE 2.10. *Least squares fitting of a function of two inputs. The parameters of $f_\theta(x)$ are chosen so as to minimize the sum-of-squared vertical errors.*

$$y = \beta_{00} + \beta_{10} X_1 + \beta_{20} X_1^2 + \beta_{01} X_2 + \beta_{02} X_2^2 + \beta_{11} X_1 X_2 + \text{error}$$

for unknown $\beta_{00}, \beta_{10}, \beta_{20}, \beta_{01}, \beta_{02}, \beta_{11}$

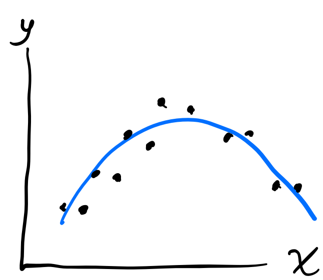
Parametric Regression: predict a continuous value

Model $f_\theta : \mathbb{R}^d \to \mathbb{R}$

$y = f_\theta(x) + \text{noise}$

Given: $\{(x^{(i)}, y_i)\}_{i=1 \dots n}$

Find: θ



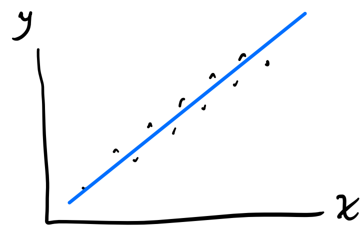
What is linear regression?

A linear regression problem is one where the **response is linear in the model parameters.**

Examples

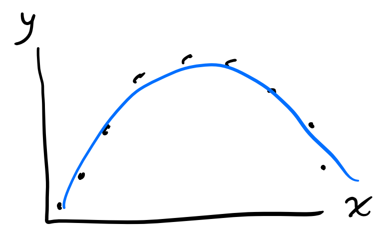
$$f: \mathbb{R} \to \mathbb{R}$$
$$y = \beta_0 + \beta_1 x$$

linear in β_0, β_1



$$f: \mathbb{R} \to \mathbb{R}$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

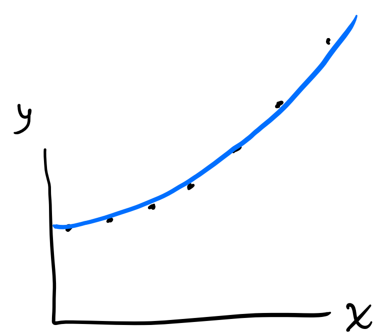
linear in $\beta_0, \beta_1, \beta_2$



NOT LINEAR REGRESSION

$$f: \mathbb{R} \to \mathbb{R}$$
$$y = \beta_1 + x^{\beta_2}$$

not linear in β_1, β_2



Are the following models linear in their parameters?

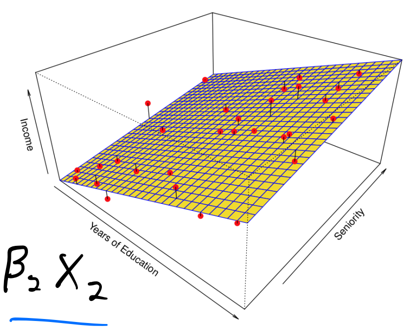
$$f: \mathbb{R} \to \mathbb{R}$$
$$y = \beta_1 e^x + \beta_2 \sin x$$

YES

$$f: \mathbb{R} \to \mathbb{R}$$
$$y = \sin(\beta_1 x + \beta_2 x^2) + \beta_2$$

No

Higher dimensional examples



$$f: \mathbb{R}^2 \to \mathbb{R}$$
$$y = \underline{\beta_0} + \underline{\beta_1 x_1} + \underline{\beta_2 x_2}$$

YES

$$f: \mathbb{R}^2 \to \mathbb{R}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

yes

$$f: \mathbb{R}^2 \to \mathbb{R}$$

$$y = \beta_0 + X_1 X_2^{\beta_2}$$

No

Can you use linear regression to solve for the parameters of the following models?

$$f: \mathbb{R} \to \mathbb{R}$$

$$y = \sqrt{\beta_0 + \beta_1 X}$$

$$y^2 = \beta_0 + \beta_1 X$$

yes

$$f: \mathbb{R}^2 \to \mathbb{R}$$

$$y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}$$

$$\log y = \log \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2$$

Least Squares Formulation for Linear Regression (for models that are linear wrt input)

Given: $D = \{(X^{(i)}, y_i)\}_{i=1 \cdots n}$, $X^{(i)} \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Model: $y = \underbrace{\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d}_{f_\theta(x)} + \text{Error}$

Want θ such that $y_i \approx f_\theta(X^{(i)})$

Idea: minimize square deviation of y_i from $f_\theta(X^{(i)})$

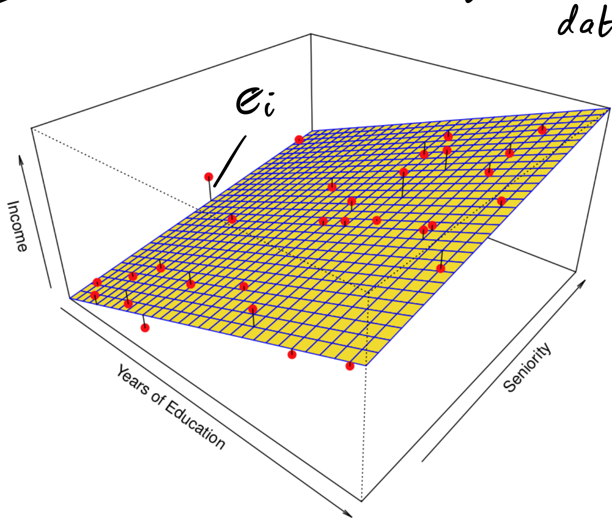
$$\min_\theta \frac{1}{2} \sum_{i=1}^{n} \Big(\underbrace{y_i - (\theta_1 x_1^{(i)} + \cdots + \theta_d x_d^{(i)})}_{e_i \text{ - error of } i^{th} \text{ data point}} \Big)^2$$

funct $\mathbb{R}^d \to \mathbb{R}$

minimize

Take ∇_θ

set $= 0$



Rewrite using vectors and matrices

Let $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

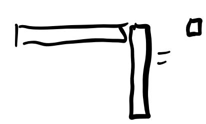
$\mathbb{R}^{n \times d}$ $\quad X = \begin{pmatrix} — x^{(1)} — \\ — x^{(2)} — \\ \vdots \\ — x^{(n)} — \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ \vdots & & \ddots & \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{pmatrix}$

Want $\quad y \approx X\theta$

$$\min_{\theta} \frac{1}{2} \| y - X\theta \|^2$$

Recall: if $a \in \mathbb{R}^n, b \in \mathbb{R}^n$

- $\langle a, b \rangle = \sum_{i=1}^{n} a_i b_i$

- $\| a \|^2 = \sum_{i=1} a_i^2$
  $= a^t a$
  
  $= \langle a, a \rangle$

Least squares formulation for linear regression (with models linear in their input)

$$\min_{\theta} \frac{1}{2} \| y - X\theta \|^2$$

where $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ $\quad X = \begin{pmatrix} — x^{(1)} — \\ — x^{(2)} — \\ \vdots \\ — x^{(n)} — \end{pmatrix}$

Solving the least squares formulation using vector calculus

Let $\underbrace{y}_{\text{known}} \in \mathbb{R}^n$, $\underbrace{X}_{\text{known}} \in \mathbb{R}^{n \times d}$, $\underbrace{\theta}_{\text{unknown}} \in \mathbb{R}^d$

Consider $\min_{\theta} \frac{1}{2} \| y - X\theta \|^2$

If X has rank d,

the solution is given by $\theta = (X^t X)^{-1} X^t y$

Why? $\nabla_\theta \frac{1}{2}\|y - X\theta\|^2 = -X^t(y - X\theta)$

set this gradient equal to 0,

$$X^t(y - X\theta) = 0$$

normal eqns

$$X^t X\theta = X^t y$$

$$\theta = (X^t X)^{-1} X^t y \quad (\text{if } X^t X \text{ is invertible})$$

If X has rank d, $X^t X$ has rank d

Note $X^t X$ is $d \times d$, so $X^t X$ is invertible ∎

Why is $\nabla_\theta \frac{1}{2}\|y - X\theta\|^2 = -X^t(y - X\theta)$?

Recall Taylors thm for multivariate functions

If $f: \mathbb{R}^d \to \mathbb{R}$

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + O(\|h\|^2)$$

You can use this to read off a gradient after applying a perturbation h

Let $f(\theta) = \frac{1}{2}\|y - X\theta\|^2 = \frac{1}{2}\|X\theta - y\|^2$

$$= \frac{1}{2}\langle X\theta - y, X\theta - y \rangle$$

So

$$f(\theta + h) = \frac{1}{2}\langle X\theta + Xh - y, X\theta + Xh - y \rangle$$

$$= \frac{1}{2}\langle X\theta - y + Xh, X\theta - y + Xh \rangle$$

$$= \frac{1}{2}\langle X\theta - y, X\theta - y \rangle + \frac{1}{2}\langle Xh, X\theta - y \rangle$$

$$+ \tfrac{1}{2}\langle X\theta - y, Xh\rangle + \tfrac{1}{2}\langle Xh, Xh\rangle$$

$$= f(\theta) + \langle X\theta - y, Xh\rangle + O(\|h\|^2)$$

$$= f(\theta) + \langle \underbrace{X^t(X\theta - y)}_{\nabla f(\theta)}, h\rangle + O(\|h\|^2)$$

So $\nabla f(\theta) = X^t(X\theta - y) = -X^t(y - X\theta)$ ∎

When is X of rank d?

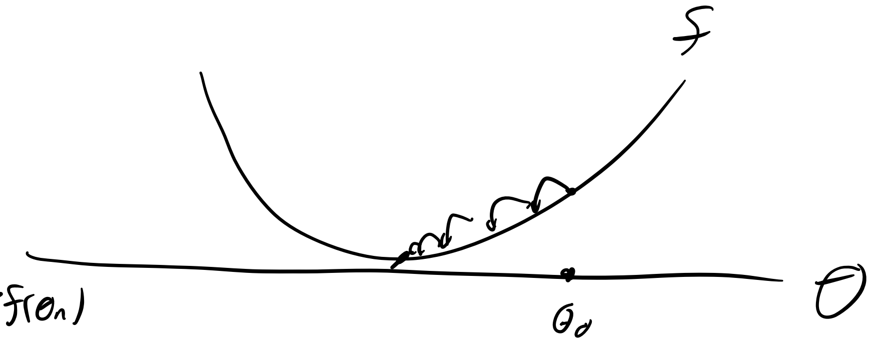
- If $n<d$, X is of rank $\leq n$
  $\Rightarrow$ Need more data points than parameters to get a unique θ.

- If any features are duplicates (or linear combinations of each other) X is of rank $<d$.
  $\Rightarrow$ Need to remove dependent features to use formula above

Other ways to solve least squares formulation for linear regression

Use a computer package such as TensorFlow or PyTorch to run Gradient Descent down the objective.

$$\min_{\theta} f(\theta)$$



$$\theta_{n+1} = \theta_n - \varepsilon \nabla f(\theta_n)$$

Grad Desc.

Least Squares Formulation for Linear Regression (for a general model)

Given: $D = \{(X^{(i)}, y_i)\}_{i=1\cdots n}$, $X^{(i)} \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Model: $y = \underbrace{\theta_1 g_1(X^{(i)}) + \cdots \theta_k g_k(X^{(i)})}_{f_\theta(x)} + \text{Error}$

Want $y \approx \mathbb{X}\theta$

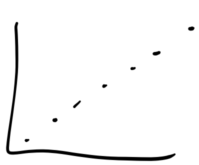
$$\text{w/ } \mathbb{X} = \begin{pmatrix} g_1(X^{(1)}) & g_2(X^{(1)}) \cdots & g_k(X^{(1)}) \\ \vdots & \ddots & \vdots \\ g_n(X^{(n)}) & \cdots & g_k(X^{(n)}) \end{pmatrix} = \begin{pmatrix} - & g(X^{(1)}) & - \\ - & g(X^{(2)}) & - \\ & \vdots & \\ - & g(X^{(n)}) & - \end{pmatrix}$$

So, do same process as above but w/ features

$(g_1(X), g_2(X), \cdots g_k(X))$

instead of X.

Ex: $f: \mathbb{R} \to \mathbb{R}$ $\{(x^{(i)}, y_i)\}_{i=1}^n$

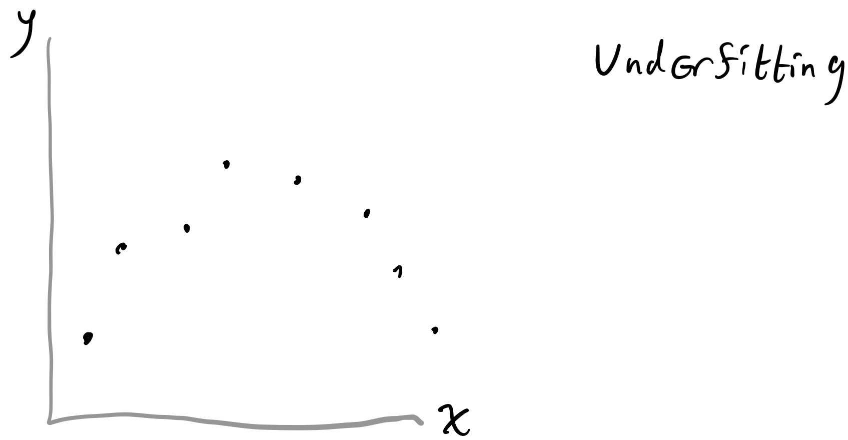


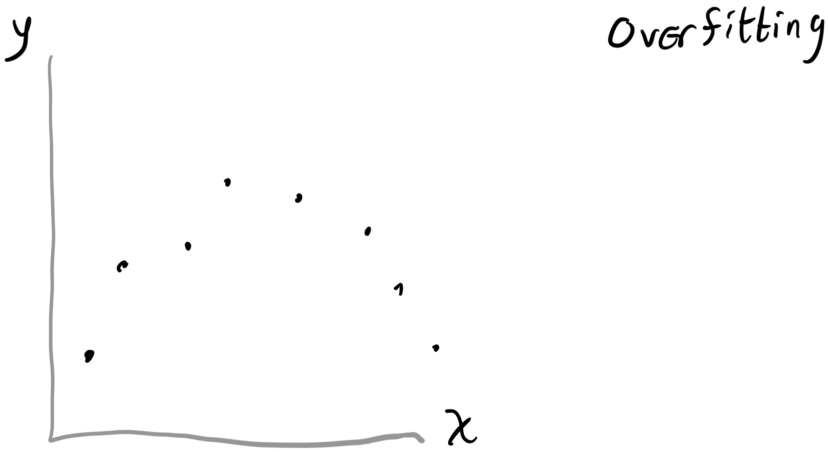
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\min \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x^{(i)} + \beta_2 x^{(i)2}))^2$$

$$\min \frac{1}{2} \| y - X\theta \|^2 \quad \text{w/ } \theta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad X = \begin{pmatrix} 1 & x^{(1)} & x^{(1)2} \\ 1 & x^{(2)} & x^{(2)2} \\ \vdots & \vdots & \vdots \\ 1 & x^{(n)} & x^{(n)2} \end{pmatrix}$$

Things that can go wrong: Underfitting and Overfitting





Other topics:

What happens when there is fewer data than features?

How do you deal with categorical features?

Be careful about whether you want to view your problem as a prediction task