Linear Regression with Gradient Descent

Outline

Introduction

Linear Regression

Introduction

What is Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed." – Arthur Samuel, IBM, Stanford University, 1959

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." – Tom Mitchell [Mit97]

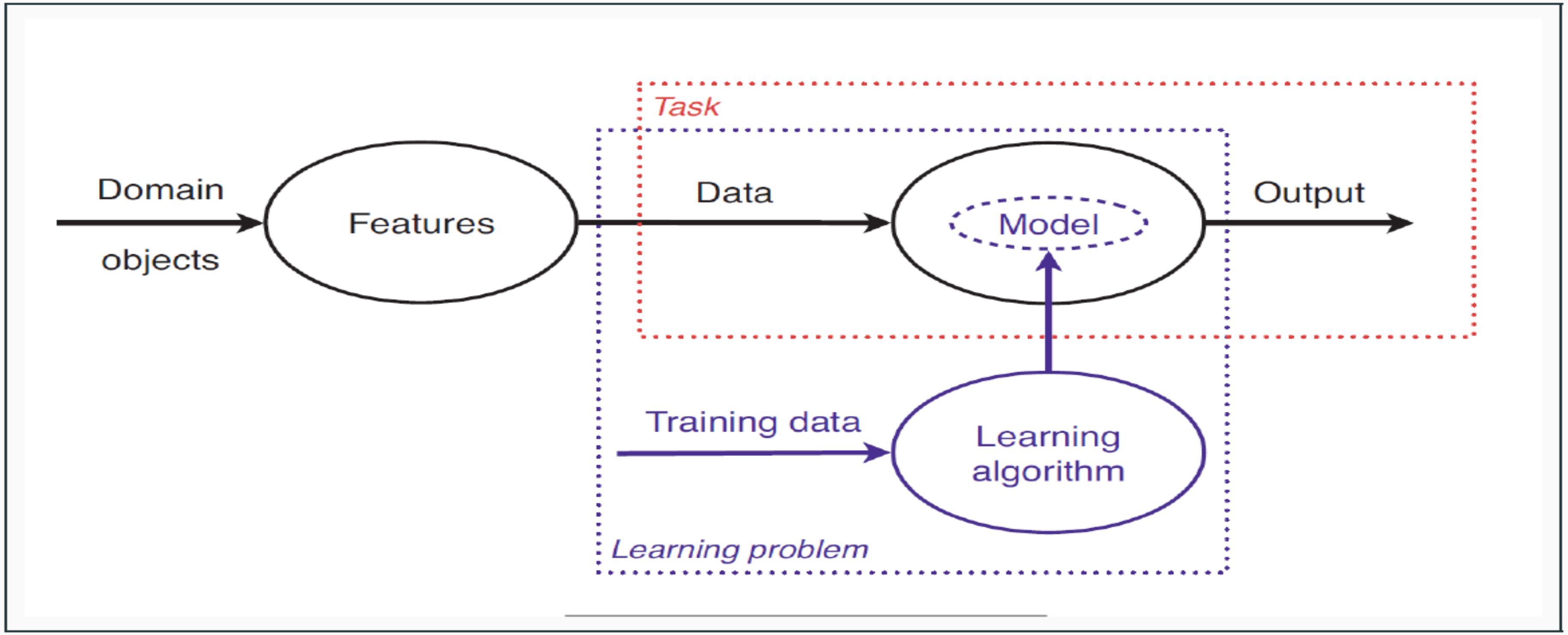
"Instead of writing a program by hand, we collect lots of examples that specify the correct output for a given input. A machine learning algorithm then takes these examples and produces a program that does the job." – Geoffrey Hinton [Hin14]

Applications

- Pattern recognition
  - Facial identities, medical images
  - Handwritten text
- Prediction
  - Stock prices
  - Marketing campaign outcomes
- Classification
  - Spam detection
  - Find similar content

Addressing a Task

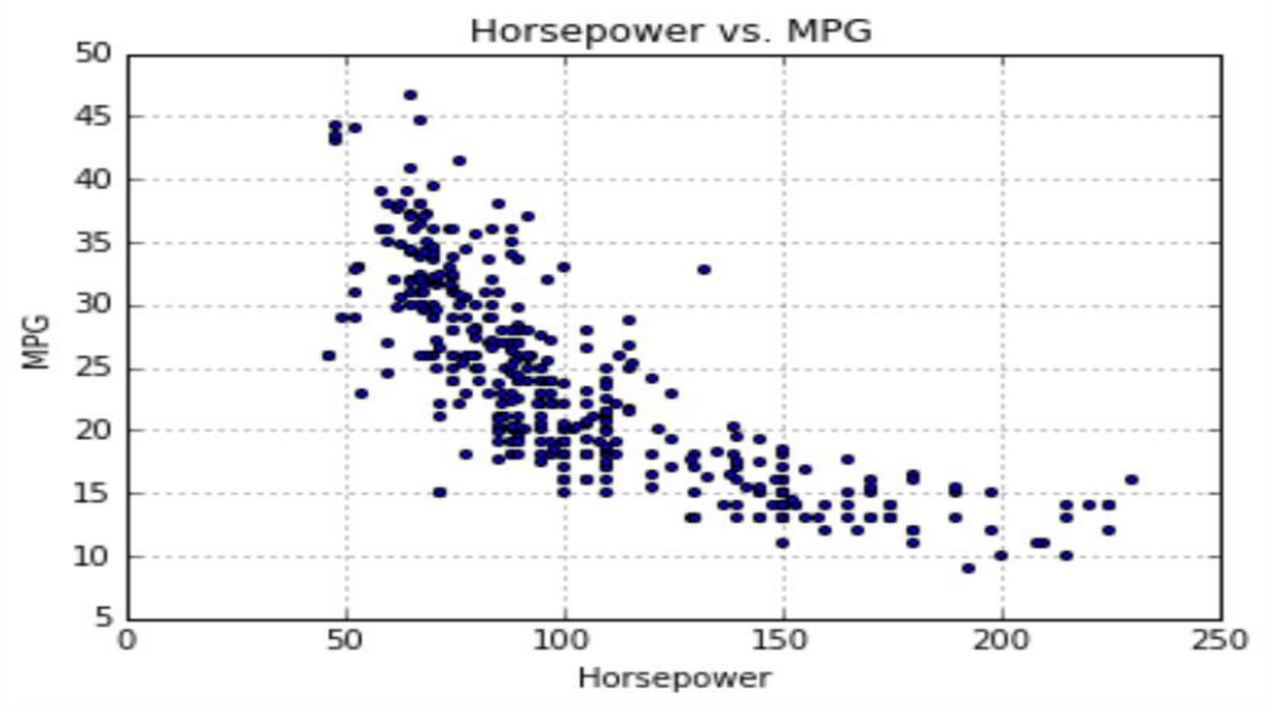
"...tasks are addressed by models, whereas learning problems are solved by learning algorithms that produce models." [Fla12]



MPG vs Horsepower

Simple use-case

- Auto MPG Data Set
- Predicting *MPG* based on *Horsepower*



Linear Regression

Univariate Linear Regression

- Model a single **response** (dependent, outcome) variable based on one or more **input** (independent, predictor) variables
- Assume **linear relationship** between input and response variables

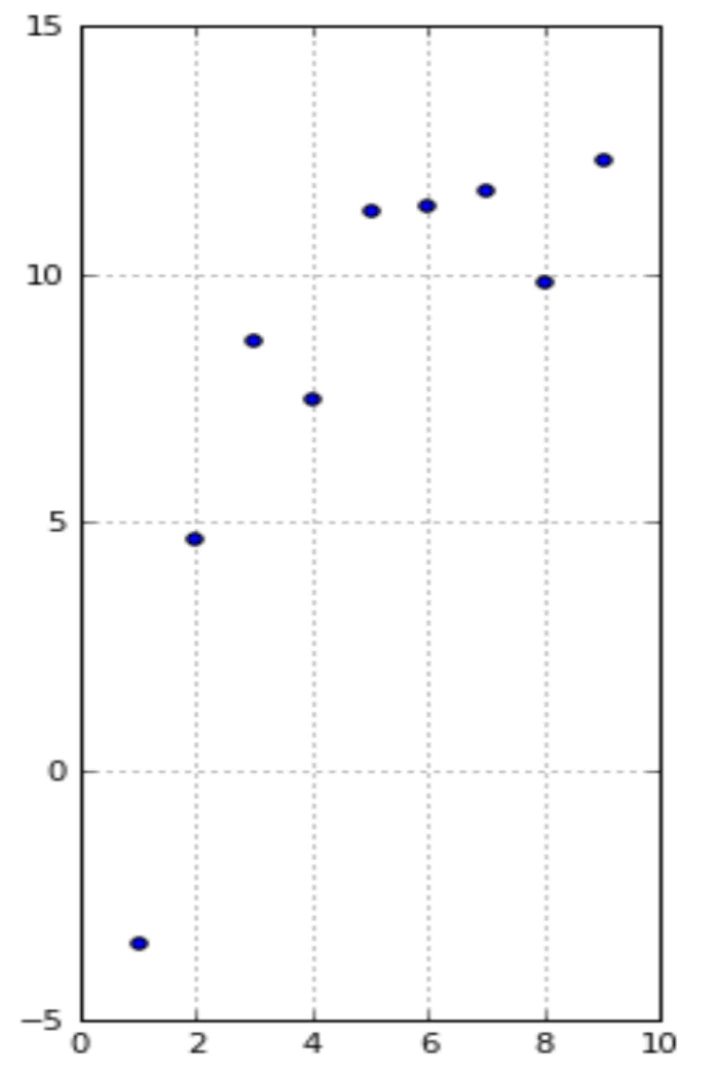
Fitting a linear regression model

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}^T$$

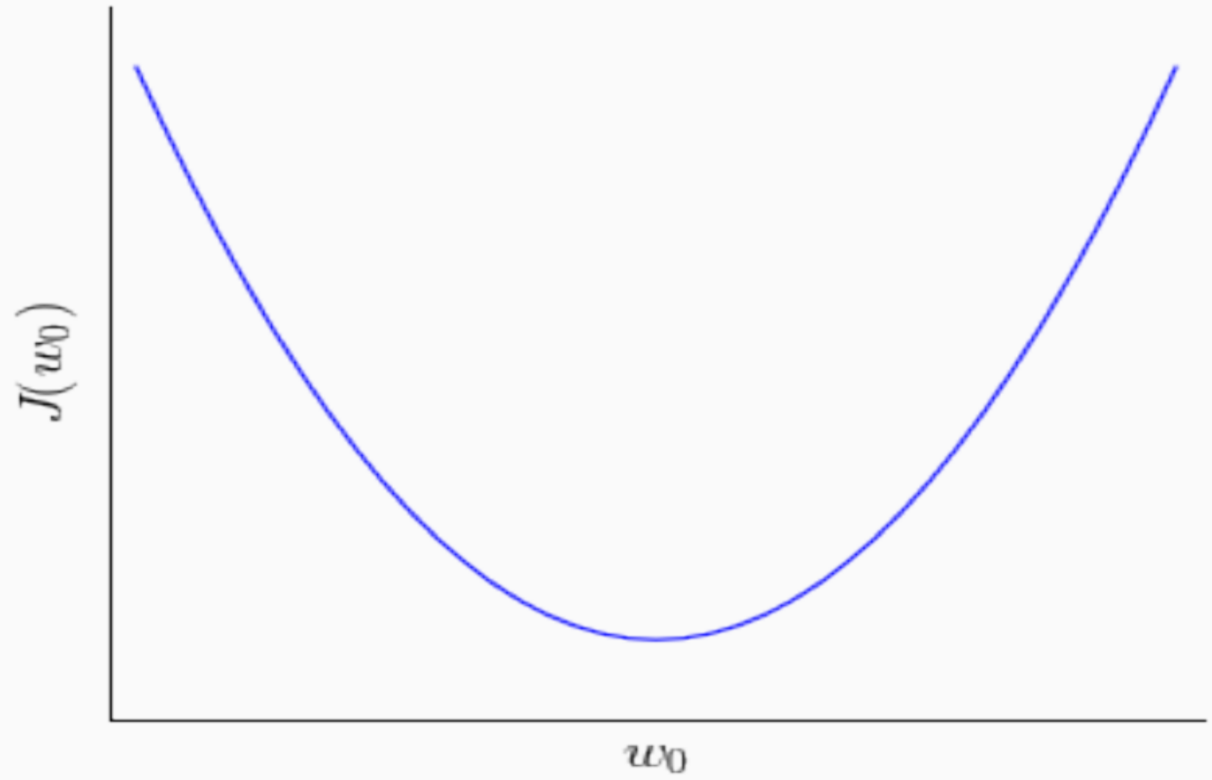
$$\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}^T$$

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

$$J(\boldsymbol{w}) = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$



Gradient Descent



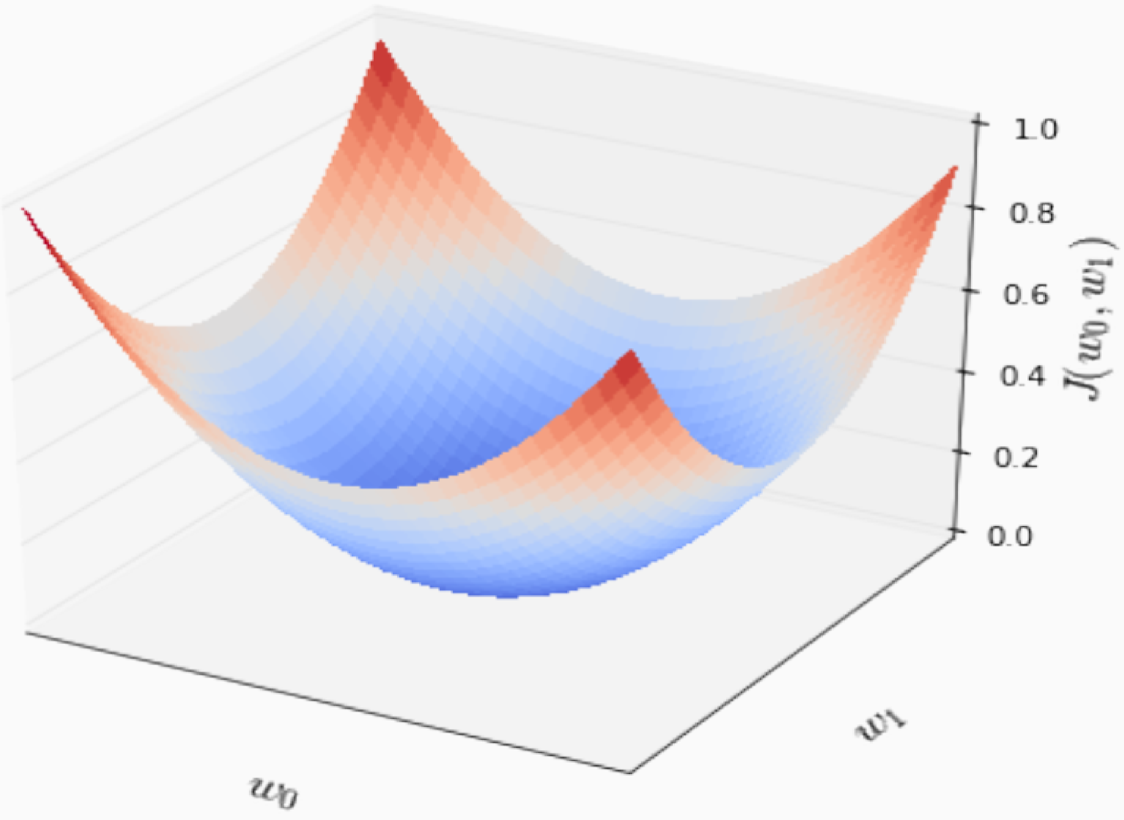
Update rule

$$\frac{d}{dw_0}J(w_0)$$

$$w_0 := w_0 - \alpha\frac{d}{dw_0}J(w_0)$$

- A positive $\alpha\frac{d}{dw_0}J(w_0)$ moves w_0 to the left
- A negative $\alpha\frac{d}{dw_0}J(w_0)$ moves w_0 to the right

Two parameters cost function



Solution

$$\frac{\partial}{\partial w_j} J(w_0, w_1)$$

repeat until convergence $\{$

Simultaneously update for every j:

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(\boldsymbol{w})$$

$\}$

Gradient Descent – Multiple Regression

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}^T$$

$$\boldsymbol{y} = \{y_1, y_2, \dots, y_N\}^T$$

$$\boldsymbol{w} = \{w_0, w_1, \dots, w_D\}^T$$

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(\boldsymbol{w})$$

$$\frac{\partial}{\partial w_j} J(\boldsymbol{w}) = \frac{\partial}{\partial w_j} \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \frac{1}{2N} \sum_{i=1}^{N} \frac{\partial}{\partial w_j} (\hat{y}_i - y_i)^2 =$$

$$\frac{1}{2N} \sum_{i=1}^{N} 2(\hat{y}_i - y_i)^2 \frac{\partial}{\partial w_j} (\hat{y}_i - y_i) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i) x_i$$

Matrix Notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1D} \\ 1 & x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{N1} & x_{N2} & \dots & x_{ND} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$$

Hypothesis: $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}$

Cost function: $J(\boldsymbol{w}) = \frac{1}{2N}\sum_{i=1}^{N}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$

Update rule:

repeat until convergence $\{$

$$\boldsymbol{w} := \boldsymbol{w} - \alpha\frac{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})}{N}$$

$\}$

$J(\theta)$ Solution by Differentiation

As given in Note 1, CS229 Lecture notes [Ng12]

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta - y)^T(X\theta - y) \\
&= \frac{1}{2}\nabla_\theta(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y) \\
&= \frac{1}{2}\nabla_\theta \mathrm{tr}(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y) \\
&= \frac{1}{2}\nabla_\theta(\mathrm{tr}\theta^T X^T X\theta - 2\mathrm{tr}y^T X\theta) \\
&= \frac{1}{2}(X^T X\theta + X^T X\theta - 2X^T y) \\
&= X^T X\theta - X^T y
\end{aligned}
$$