



Introduction to Machine Learning

Linear Regression and Gradient Descent

Nikolay Manchev

May 26, 2016

London Machine Learning Study Group

Special thanks to our sponsors

- IBM
- Skills Matter

Next events

<http://www.meetup.com/London-Machine-Learning-Study-Group>

<https://twitter.com/nikolaymanchev>

Slides and code

Available at <https://github.com/nmanchev/MachineLearningStudyGroup>

Introduction

Linear Regression

Introduction

What is Machine Learning

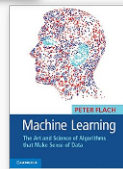
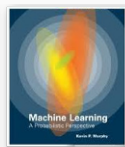
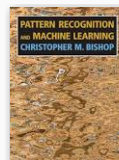
“Field of study that gives computers the ability to learn without being explicitly programmed.” – Arthur Samuel, IBM, Stanford University, 1959

“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]

Recommended books


- Pattern Recognition and Machine Learning [Bis06]
- Machine Learning: A Probabilistic Perspective [Mur12]
- Machine Learning: The Art and Science of Algorithms That Make Sense of Data [Fla12]



UCI Machine Learning Repository –

archive.ics.uci.edu/ml


- Great resource for Machine Learning data sets
- Over 330 freely available sets
- Auto MPG Data Set
 - Fuel consumption in MPG
 - Attributes: mpg, cylinders, displacement, horsepower, weight, acceleration etc.



Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Auto MPG Data Set

Download: [Data Folder](#), [Data Set Description](#)



Abstract: Revised from CMU StatLib library, data concerns city-cycle fuel consumption

Data Set Characteristics:	Multivariate	Number of Instances:	398	Area:	N/A
Attribute Characteristics:	Categorical, Real	Number of Attributes:	8	Date Donated	1993-07-07
Associated Tasks:	Regression	Missing Values?	Yes	Number of Web Hits:	167833

- Pattern recognition
 - Facial identities, medical images
 - Handwritten text

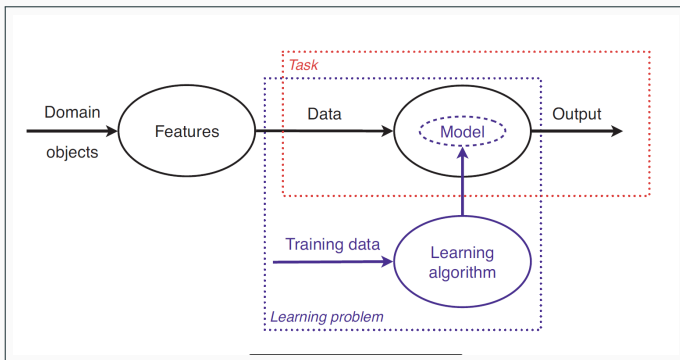
- Pattern recognition
 - Facial identities, medical images
 - Handwritten text
- Prediction
 - Stock prices
 - Marketing campaign outcomes

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

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

Addressing a Task

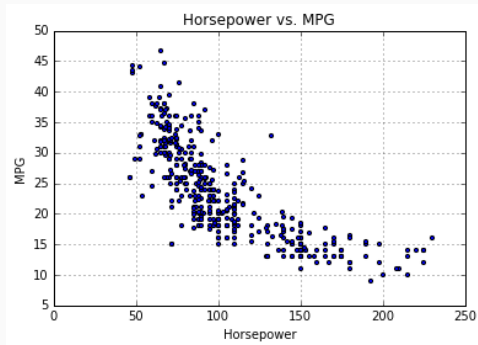
“...tasks are addressed by models, whereas learning problems are solved by learning algorithms that produce models.” [Fla12]



MPG vs Horsepower

Simple use-case

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



Linear Regression

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

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

Univariate Linear Regression

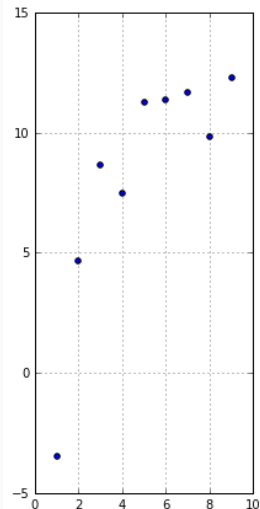
Fitting a linear regression model

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

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

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

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



Gradient Descent



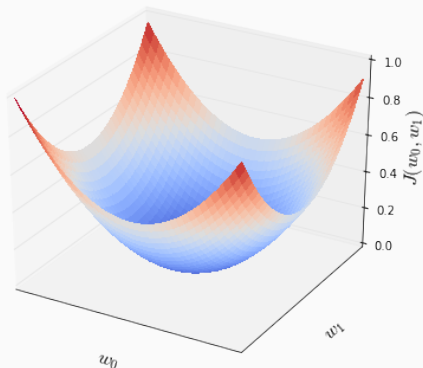
Update rule

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

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

- 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 :

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

}

Gradient Descent – Multiple Regression

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

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

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

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

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

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

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: $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$

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

Update rule:





repeat until convergence {




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

$J(\theta)$ Solution by Differentiation

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

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

-  Olivier Bousquet and Léon Bottou, *The tradeoffs of large scale learning*, Advances in Neural Information Processing Systems 20 (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), Curran Associates, Inc., 2008, pp. 161–168.
-  Christopher M. Bishop, *Pattern recognition and machine learning (information science and statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
-  Peter Flach, *Machine learning: The art and science of algorithms that make sense of data*, Cambridge University Press, New York, NY, USA, 2012.
-  Geoffrey Hinton, *Csc321: Introduction to neural networks and machine learning*, University Lecture, 2014.

-  Thomas M. Mitchell, *Machine learning*, 1 ed., McGraw-Hill, Inc., New York, NY, USA, 1997.
-  Kevin P. Murphy, *Machine learning: A probabilistic perspective*, The MIT Press, 2012.
-  Andrew Ng, *Note 1, Lecture notes in CS 229 Machine Learning*, 2012.