Useful Sources: Section 28.3 of textbook, Wikipedia articles on Gradient Descent and Machine Learning (first half of articles)

Machine learning is concerned with algorithms that can learn from data (or experience) to improve performance.

Assume the performance measure is an error function (or loss, risk, or cost function).

Assume algorithm finds values for parameters and/or constructs a specific kind of data structure such as a neural network or Bayesian network.
A *hypothesis* is a specific assignment to the parameters/data structure.

The typical basic machine learning algorithm performs *incremental improvement*. That is, it has an initial hypothesis, then makes incremental changes to decrease its performance measure.

*Gradient descent* is one approach for implementing incremental improvement.

There are many more sophisticated variations, but they generally follow the above ideas.
As an example of producing a least-squares fit, suppose that we have five data points

\[
(x_1, y_1) = (-1, 2), \\
(x_2, y_2) = (1, 1), \\
(x_3, y_3) = (2, 1), \\
(x_4, y_4) = (3, 0), \\
(x_5, y_5) = (5, 3),
\]

shown as black dots in Figure 28.3. We wish to fit these points with a quadratic polynomial

\[ F(x) = c_1 + c_2 x + c_3 x^2. \]

Much of 28.3 is concerned with an algorithm to directly compute the parameters/coefficients/weights.

Type of hypothesis

Parameters to be filled in

\[
\begin{align*}
(2 - (c_1 + c_2(-1) + c_3(-1)^2))^2 + (1 - (c_1 + c_2(1) + c_3(1)^2))^2 \\
+ (1 - (c_1 + c_2(2) + c_3(2)^2))^2 + (0 - (c_1 + c_2(3) + c_3(3)^2))^2 \\
+ (3 - (c_1 + c_2(5) + c_3(5)^2))^2
\end{align*}
\]

Find values for \(c_1, c_2, c_3\) to minimize.
Matrix Representation

Minimize $\|AC - Y\|^2$

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 5 & 25 \end{pmatrix}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 3 \end{pmatrix}$$

$$c = \begin{pmatrix} 1.200 \\ -0.757 \\ 0.214 \end{pmatrix}$$

Example of a hypothesis

CS 3343 Analysis of Algorithms

Learning by Gradient Descent – 5
Illustration of Solution

Introduction
Introduction 2
Example 1
Example 2
Example 3

PERFORMANCE MEASURE
sum of squared differences

black points - data

hypothesis

\[ F(x) = 1.2 - 0.757x + 0.214x^2 \]

\[ x \]

CS 3343 Analysis of Algorithms

Learning by Gradient Descent – 6
\[ y = 0.5 + 0.5x \]

\[ y = 1.0 + 0.25 \pi \]

\[ 8.25 = 2^2 + 0^2 + (0.5)^2 + 2^2 + 0^2 \]

\[ 16.25 = 0^2 + 0^2 + (0.5)^2 + 1^2 + y^2 \]

\[ 5.25 = (1.25)^2 + (0.25)^2 + (0.75)^2 + (0.75)^2 \]

\[ y = (x - 0)^2 + (y - 0.5)^2 \]

\[ y = -0.5x^2 + 1.5x + 0 \]

\[ y = 0 + 1.5x - 0.5x^2 \]