Fundamentals of Linear Algebra and Optimization
Motivations: Fitting Data (Regression)

Jean Gallier and Jocelyn Quaintance

CIS Department
University of Pennsylvania

jean@cis.upenn.edu

August 18, 2023

Motivations

Two central problems in machine learning are

Motivations

Two central problems in machine learning are

(1.) *Data fitting* (or *learning a function*).

Motivations

Two central problems in machine learning are

(1.) *Data fitting* (or *learning a function*).
(2.) *Data classification*.

Motivations

Two central problems in machine learning are

(1.) *Data fitting* (or *learning a function*).
(2.) *Data classification*.

For this introduction we focus on the more classical problem of data fitting.

Fitting Points in the Plane

Assume we have some data points in the plane given as a list of m coordinates

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i, y_i \in \mathbb{R}.$$

Fitting Points in the Plane

Assume we have some data points in the plane given as a list of m coordinates

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i, y_i \in \mathbb{R}.$$

The figure on the next slide shows an example of 100 points in the plane.

Fitting Points in the Plane

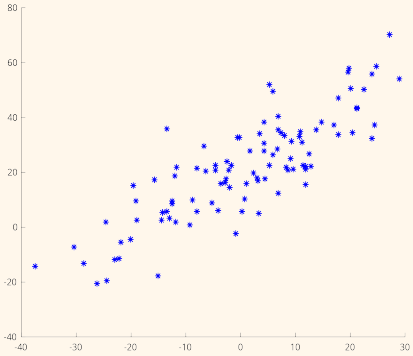

Figure 1: A data set of 100 points in the plane.

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

The simplest kind of function is an *affine map*, that is, a map of the form

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

The simplest kind of function is an *affine map*, that is, a map of the form

$$f(x) = wx + b,$$

for some real numbers w, b. The number w is called a *weight*.

Learning an Affine Map

We are looking for a function $f\colon \mathbb{R} \to \mathbb{R}$ such that $f(x_i) = y_i$ for $i = 1, \ldots, 100$.

The simplest kind of function is an *affine map*, that is, a map of the form

$$f(x) = wx + b,$$

for some real numbers w, b. The number w is called a *weight*.

The numbers w and b must satisfy the 100 (affine) equations

$$y_i = f(x_i) = wx_i + b.$$

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

We are asking for too much. A more promising approach is to *minimize the error*.

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

We are asking for too much. A more promising approach is to *minimize the error*.

But what is the error?

Learning an Affine Map

In general, unless all the points lie on the same line, the above linear system has no solution.

We are asking for too much. A more promising approach is to *minimize the error*.

But what is the error?

Gauss and Legendre proposed a method over 200 years ago: the *least squares method*.

What is the Error?

Every equation $y_i = wx_i + b$ can be written as

$$y_i - wx_i - b = 0.$$

Think of $y_i - wx_i - b$ as an *error*.

What is the Error?

Every equation $y_i = wx_i + b$ can be written as

$$y_i - wx_i - b = 0.$$

Think of $y_i - wx_i - b$ as an *error*.

In the method of least squares, *the error (or loss) is the sum of the squares of the errors*:

$$\sum_{i=1}^{100} (y_i - wx_i - b)^2.$$

Least Squares Solution

Here the least squares solution for our data set of 100 points.

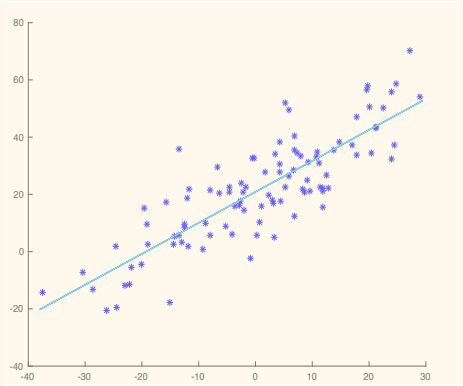

Figure 2: The least squares best fit.

Fitting Points in $\mathbb{R}^n$

We can generalize the problem to data in $\mathbb{R}^n$.

Fitting Points in $\mathbb{R}^n$

We can generalize the problem to data in $\mathbb{R}^n$.
Assume we have some data given as a list of m pairs

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i \in \mathbb{R}^n, y_i \in \mathbb{R}.$$

Fitting Points in $\mathbb{R}^n$

We can generalize the problem to data in $\mathbb{R}^n$.
Assume we have some data given as a list of m pairs

$$((x_1, y_1), \ldots, (x_m, y_m)), \quad x_i \in \mathbb{R}^n, y_i \in \mathbb{R}.$$

We wish to learn an affine map $f\colon \mathbb{R}^n \to \mathbb{R}$ of the form

$$f(z) = w_1 z_1 + \cdots + w_n z_n + b,$$

with $z = (z_1, \ldots, z_n)$ and where $w_1, \ldots, w_n \in \mathbb{R}$ are *weights*.

It is convenient to denote the quantity $w_1 z_1 + \cdots + w_n z_n$ (an inner product) as $z^\top w$.

The Euclidean Norm (or ℓ^2-Norm)

The *Euclidean norm* (or *ℓ^2-norm*) of a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is defined as

$$\|z\|_2 = (z_1^2 + \cdots + z_n^2)^{1/2} = (z^\top z)^{1/2}.$$

The Euclidean Norm (or ℓ^2-Norm)

The *Euclidean norm* (or *ℓ^2-norm*) of a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is defined as

$$\|z\|_2 = (z_1^2 + \cdots + z_n^2)^{1/2} = (z^\top z)^{1/2}.$$

The *least squares problem* is find $w \in \mathbb{R}^n$ that minimizes

$$\|\xi\|_2^2,$$

where $\xi = (\xi_1, \ldots, \xi_m)$ is the vector given by

$$\xi_i = y_i - x_i^\top w - b.$$

Pseudo-Inverse

It turns out that there is a *unique solution* $\begin{pmatrix} w \\ b \end{pmatrix}^+$ *of least* ℓ^2*-norm*.

Pseudo-Inverse

It turns out that there is a *unique solution* $\begin{pmatrix} w \\ b \end{pmatrix}^+$ *of least* ℓ^2*-norm*.

Furthermore, this solution $\begin{pmatrix} w \\ b \end{pmatrix}^+$ is expressed in terms of something called a *pseudo-inverse*.

Pseudo-Inverse

It turns out that there is a *unique solution* $\binom{w}{b}^+$ *of least* ℓ^2*-norm*.

Furthermore, this solution $\binom{w}{b}^+$ is expressed in terms of something called a *pseudo-inverse*.

In our case

$$\binom{w}{b}^+ = A^+ y,$$

where A^+ is the pseudo-inverse of the matrix

$$A = \begin{pmatrix} x_1^\top & 1 \\ \vdots & \vdots \\ x_m^\top & 1 \end{pmatrix}.$$

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*).

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*).

The SVD and the pseudo-inverse will be discussed extensively later.

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*).

The SVD and the pseudo-inverse will be discussed extensively later.

The solution given by the pseudo-inverse is not always desirable or too expensive to compute.

Pseudo-Inverse

The pseudo-inverse of a matrix A can be computed in terms of its *singular value decomposition* (or *SVD*).

The SVD and the pseudo-inverse will be discussed extensively later.

The solution given by the pseudo-inverse is not always desirable or too expensive to compute.

Another method is to *penalize the ℓ^2-norm of w*.

Ridge Regression

The problem to solve is the following minimization problem known as *ridge regression*:

Ridge Regression

The problem to solve is the following minimization problem known as *ridge regression*:

$$\begin{aligned} &\text{minimize} && \|\xi\|_2^2 + K\|w\|_2^2 \\ &\text{subject to} \\ & && y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where K is positive constant.

Ridge Regression

The problem to solve is the following minimization problem known as *ridge regression*:

$$\begin{aligned} &\text{minimize} && \|\xi\|_2^2 + K\|w\|_2^2 \\ &\text{subject to} && \\ & && y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where K is positive constant.

This time there is a unique solution given in terms of the matrix X whose rows are the (row) vectors $x_i^\top$. For simplicity assume $b = 0$.

Ridge Regression

The unique minimizer is given by the expression

$$w = X^\top (XX^\top + KI_m)^{-1} y.$$

Ridge Regression

The unique minimizer is given by the expression

$$w = X^\top (XX^\top + KI_m)^{-1} y.$$

The matrix

$$XX^\top + KI_m$$

is particularly nice because it is *symmetric positive definite*. There are more efficient methods for solving linear system involving SPD matrices. We will study such matrices extensively.

ℓ^1-Norm and Lasso Regression

One of the weak points of ridge regression is that when the dimension n of the data is relatively large, the weight vector w is *not sparse*, which means that very few weights w_i are close to zero.

ℓ^1-Norm and Lasso Regression

One of the weak points of ridge regression is that when the dimension n of the data is relatively large, the weight vector w is *not sparse*, which means that very few weights w_i are close to zero.

A remedy to this problem is to *penalize the ℓ^1-norm $\|w\|_1$ of w* instead of its ℓ^2-norm $\|w\|_2^2$.

ℓ^1-Norm and Lasso Regression

One of the weak points of ridge regression is that when the dimension n of the data is relatively large, the weight vector w is *not sparse*, which means that very few weights w_i are close to zero.

A remedy to this problem is to *penalize the ℓ^1-norm $\|w\|_1$ of w* instead of its ℓ^2-norm $\|w\|_2^2$.

The *ℓ^1-norm* of a vector $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is defined as

$$\|z\|_1 = |z_1| + \cdots + |z_n|.$$

Lasso Regression

Lasso regression is the following minimization problem:

Lasso Regression

Lasso regression is the following minimization problem:

$$\begin{array}{ll} \text{minimize} & \|\xi\|_2^2 + \tau \|w\|_1 \\ \text{subject to} & \\ & y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{array}$$

where τ is positive constant.

Lasso Regression

Lasso regression is the following minimization problem:

$$\begin{aligned} &\text{minimize} && \|\xi\|_2^2 + \tau \|w\|_1 \\ &\text{subject to} \\ & && y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where τ is positive constant.

This time, there is no closed-form solution. However a solution can be computed using an iterative process (ADMM) which solves a sequence of linear systems involving SPD matrices.

Elastic Net Regression

There are still undesirable features of lasso, especially when the dimension n of the data is much larger than the number m of data.

Elastic Net Regression

There are still undesirable features of lasso, especially when the dimension n of the data is much larger than the number m of data.

A way to retain the best features of ridge regression and lasso is to *penalize both the ℓ^1-norm and the ℓ^2-norm of w*.

Elastic Net Regression

There are still undesirable features of lasso, especially when the dimension n of the data is much larger than the number m of data.

A way to retain the best features of ridge regression and lasso is to *penalize both the ℓ^1-norm and the ℓ^2-norm of w.*

Elastic net regression is the following minimization problem:

$$\begin{aligned} &\text{minimize} && \|\xi\|_2^2 + K\|w\|_2^2 + \tau\|w\|_1 \\ &\text{subject to} \\ & && y_i - x_i^\top w - b = \xi_i, \quad i = 1, \ldots, m \end{aligned}$$

where K and τ are positive constants.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

When m is much larger than n, elastic net is much slower than lasso, especially for small K.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

When m is much larger than n, elastic net is much slower than lasso, especially for small K.

Remarkably, least squares, ridge regression, lasso, and elastic net, all rely on *solving linear systems involving SPD matrices*.

Elastic Net Regression

Elastic net can also be solved using an iterative process (ADMM) which solves linear systems involving SPD matrices.

When m is much larger than n, elastic net is much slower than lasso, especially for small K.

Remarkably, least squares, ridge regression, lasso, and elastic net, all rely on *solving linear systems involving SPD matrices*.

This is why most of this course will be devoted to these topics! The notion of *orthogonality* also play a crucial role.