From Calculus to Machine Learning

Paul Siegel

July 28, 2018

What do linear regression, maximum likelihood estimation, support vector machines, and neural networks have in common?

What do linear regression, maximum likelihood estimation, support vector machines, and neural networks have in common?

All fit into the following framework:

What do linear regression, maximum likelihood estimation, support vector machines, and neural networks have in common?

All fit into the following framework:

- Identify a space of “reasonable” models

What do linear regression, maximum likelihood estimation, support vector machines, and neural networks have in common?

All fit into the following framework:

- Identify a space of "reasonable" models
- Construct a function which computes how well a given model fits the data

What do linear regression, maximum likelihood estimation, support vector machines, and neural networks have in common?

All fit into the following framework:

- Identify a space of "reasonable" models
- Construct a function which computes how well a given model fits the data
- Find the model(s) which maximize (or minimize) the function

Data scientists play a significant role in this process:

Data scientists play a significant role in this process:

- Identifying reasonable models uses domain expertise

Data scientists play a significant role in this process:

- Identifying reasonable models uses domain expertise
- Constructing a good objective function usually uses statistics

Data scientists play a significant role in this process:

- Identifying reasonable models uses domain expertise
- Constructing a good objective function usually uses statistics
- Finding the extremal model(s) uses math and/or engineering

Data scientists play a significant role in this process:

- Identifying reasonable models uses domain expertise
- Constructing a good objective function usually uses statistics
- Finding the extremal model(s) uses math and/or engineering

In this seminar we'll try to understand some of the theory behind all three steps.

The plan:

1. Optimization for functions of one variable
2. Linear algebra and PCA
3. Optimization for functions of several variables
4. Conditional probability and Bayesian statistics
5. Linear regression
6. Perceptrons
7. Back propagation and gradient descent

1.1 Optimizing quadratic functions of one variable

You wish to build a rectangular fence next to a river. You have 100m of fence to work with and you want to enclose as much area as possible. How do you do it?

What is the space of models?

What is the space of models?

- Each possible fence is determined by its height x and its width y

What is the space of models?

- Each possible fence is determined by its height x and its width y
- Constraints: $x \geq 0$, $y \geq 0$, and $2x + y = 100$

What is the objective function?

What is the objective function?

We want to maximize the area $A(x, y) = xy$

This is now just a math problem: maximize $A(x, y) = xy$ subject to the constraints $x \geq 0$, $y \geq 0$, and $2x + y = 100$

This is now just a math problem: maximize $A(x, y) = xy$ subject to the constraints $x \geq 0$, $y \geq 0$, and $2x + y = 100$

This is an example of a *constrained optimization* problem.

This is now just a math problem: maximize $A(x, y) = xy$ subject to the constraints $x \geq 0$, $y \geq 0$, and $2x + y = 100$

This is an example of a *constrained optimization* problem.

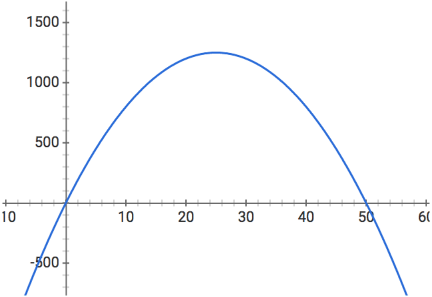
The objective function is quadratic and the constraint is linear, so we can hope to solve it analytically.

Using the constraint, eliminate y to get:

$$A(x) = x(100 - 2x)$$

Using the constraint, eliminate y to get:

$$A(x) = x(100 - 2x)$$



Algebraic magic:

$$A(x) = x(100 - 2x) = -2(x - 25)^2 + 1250$$

Algebraic magic:

$$A(x) = x(100 - 2x) = -2(x - 25)^2 + 1250$$

Since

$$-2(x - 25)^2 \leq 0$$

Algebraic magic:

$$A(x) = x(100 - 2x) = -2(x - 25)^2 + 1250$$

Since

$$-2(x - 25)^2 \leq 0$$

we get

$$A(x) \leq 1250$$

with equality if and only if $x = 25$

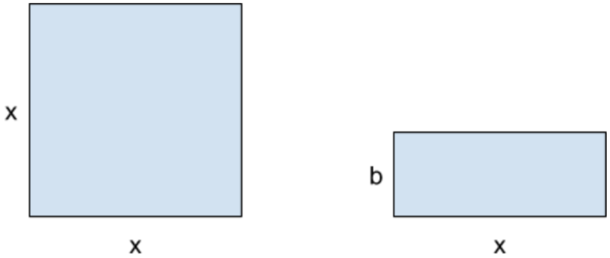
How does the algebraic magic work?

How does the algebraic magic work?

Start with an expression of the form $x^2 + bx$.

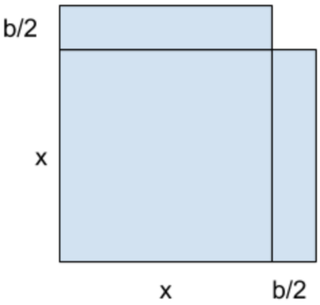
How does the algebraic magic work?

Start with an expression of the form $x^2 + bx$.



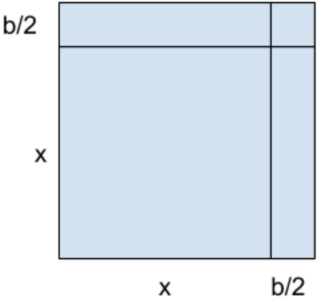
Cut the bx rectangle in half and rearrange:

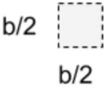
Cut the bx rectangle in half and rearrange:



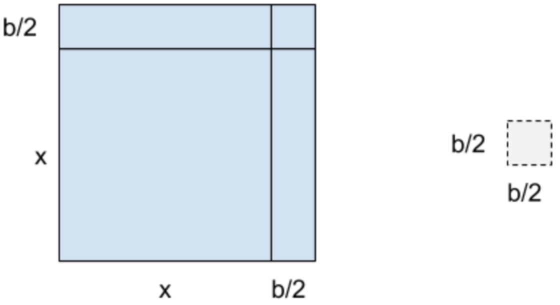
Complete the square!

Complete the square!





Complete the square!



Conclusion:

$$x^2 + bx = \left(x + \frac{b}{2}\right)^2 - \left(\frac{b}{2}\right)^2$$

Let's apply the algebraic magic to the function $A(x) = x(100 - 2x)$ which computes the area of our fence.

Let's apply the algebraic magic to the function $A(x) = x(100 - 2x)$ which computes the area of our fence.

To do so, we need to find an expression of the form $x^2 + bx$; the first step is to expand the parentheses:

Let's apply the algebraic magic to the function $A(x) = x(100 - 2x)$ which computes the area of our fence.

To do so, we need to find an expression of the form $x^2 + bx$; the first step is to expand the parentheses:

$$A(x) = x(100 - 2x) = 100x - 2x^2 = -2x^2 + 100x$$

Let's apply the algebraic magic to the function $A(x) = x(100 - 2x)$ which computes the area of our fence.

To do so, we need to find an expression of the form $x^2 + bx$; the first step is to expand the parentheses:

$$A(x) = x(100 - 2x) = 100x - 2x^2 = -2x^2 + 100x$$

Now factor out a -2:

Let's apply the algebraic magic to the function $A(x) = x(100 - 2x)$ which computes the area of our fence.

To do so, we need to find an expression of the form $x^2 + bx$; the first step is to expand the parentheses:

$$A(x) = x(100 - 2x) = 100x - 2x^2 = -2x^2 + 100x$$

Now factor out a -2:

$$A(x) = -2x^2 + 100x = -2(x^2 - 50x)$$

Finally, apply the algebraic magic to the expression $x^2 + 50x$, leaving the -2 on the outside:

Finally, apply the algebraic magic to the expression $x^2 + 50x$, leaving the -2 on the outside:

$$A(x) = -2(x^2 - 50x) = -2((x - 25)^2 - 25^2)$$

Finally, apply the algebraic magic to the expression $x^2 + 50x$, leaving the -2 on the outside:

$$A(x) = -2(x^2 - 50x) = -2((x - 25)^2 - 25^2)$$

Conclude with a little bit of simplifying:

$$A(x) = -2(x - 25)^2 + 2 \cdot 25^2 = -2(x - 25)^2 + 1250$$

Let's look at another optimization problem:

You want to build an open box with a square base which holds $25m^3$ of water. How much material do you need?

Space of models:

Space of models:

Each box is determined by the width x of the base and the height y

Space of models:

Each box is determined by the width x of the base and the height y

Constraints: $x \geq 0$, $y \geq 0$, Volume $= x^2 y = 25$

Objective function:

Objective function:

The amount of material needed is determined by the surface $S(x, y) = x^2 + 4xy$ of the box.

Constrained optimization problem: minimize $S(x, y) = x^2 + 4xy$ subject to the constraints $x \geq 0$, $y \geq 0$, and $x^2 y = 25$.

Constrained optimization problem: minimize $S(x, y) = x^2 + 4xy$ subject to the constraints $x \geq 0$, $y \geq 0$, and $x^2y = 25$.

The objective function is quadratic, but the constraint $x^2y = 25$ is cubic, so we expect this to be harder.

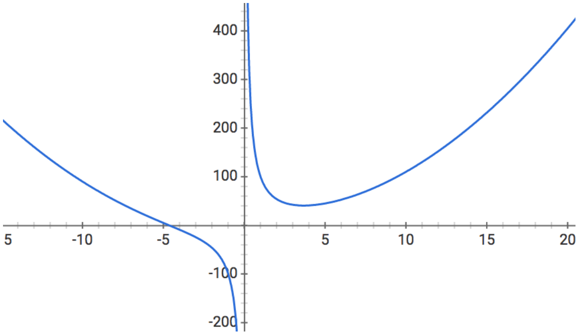
Use the constraint to eliminate y and get:

Use the constraint to eliminate y and get:

$$S(x) = x^2 + \frac{100}{x}$$

Use the constraint to eliminate y and get:

$$S(x) = x^2 + \frac{100}{x}$$



No algebraic magic; we'll need some tools.

No algebraic magic; we'll need some tools.

Main idea: approximate a general function with linear and quadratic functions.

1.2 Optimization via linear approximation

Our objective now is to solve optimization problems of the following form:

Find the maximum and/or minimum value of a function $f : A \to \mathbb{R}$ where $\mathbb{R}$ is the set of all real numbers and A is a subset of $\mathbb{R}$.

1.2 Optimization via linear approximation

Our objective now is to solve optimization problems of the following form:

Find the maximum and/or minimum value of a function $f : A \to \mathbb{R}$ where $\mathbb{R}$ is the set of all real numbers and A is a subset of $\mathbb{R}$.

This problem is completely hopeless in general, but there are lots of techniques which work in special cases which come up in applications.

In this seminar we'll solve optimization problems by using calculus to understand the geometry of the graph of f.

In this seminar we'll solve optimization problems by using calculus to understand the geometry of the graph of f.

Assume that f is defined on an interval (or finite collection of intervals) in $\mathbb{R}$; we will look at:

- The *local behavior* of f near points of interest in the interior of its domain

In this seminar we'll solve optimization problems by using calculus to understand the geometry of the graph of f.

Assume that f is defined on an interval (or finite collection of intervals) in $\mathbb{R}$; we will look at:

- The *local behavior* of f near points of interest in the interior of its domain
- The *boundary behavior* of f near the endpoints of its domain

Definition

A point $x_{\max \in A}$ is said to be a *local maximum* for f if

$$f(x_{\max}) \geq f(x)$$

for all x sufficiently close to $x_{\max}$.
Similarly, $x_{\min}$ is said to be a local minimum if $f(x_{\min}) \leq f(x)$ for all x near $x_{\min}$.

To find local extrema, we will use *local linear approximations*.

To find local extrema, we will use *local linear approximations*.

Main idea: if x_0 is a local extremum of f and f can be well approximated by a line near x_0 then that line is flat (has slope zero).

1.3 Lines

Before approximating functions by lines, let us review some basic facts about them.

Definition

A function L is said to be *linear* if it satisfies the following conditions:

$$L(x+y) = L(x) + L(y), \quad L(ax) = aL(x)$$

for every x, y, and a.

Definition

A function L is said to be *linear* if it satisfies the following conditions:

$$L(x+y) = L(x) + L(y), \quad L(ax) = aL(x)$$

for every x, y, and a.

Fact: every linear function $L\colon \mathbb{R} \to \mathbb{R}$ has the form $L(x) = mx$ for some constant m called the *slope* of L.

Definition

A function ℓ is said to be *affine* if the $\ell(x) - b$ is a linear function for some constant b, called the *intercept* of ℓ.

Definition

A function ℓ is said to be *affine* if the $\ell(x) - b$ is a linear function for some constant b, called the *intercept* of ℓ.

Every affine function $\ell\colon \mathbb{R} \to \mathbb{R}$ has the form $\ell(x) = mx + b$ for some constants m and b.

Definition

A function ℓ is said to be *affine* if the $\ell(x) - b$ is a linear function for some constant b, called the *intercept* of ℓ.

Every affine function $\ell \colon \mathbb{R} \to \mathbb{R}$ has the form $\ell(x) = mx + b$ for some constants m and b.

Fact: a function $\mathbb{R} \to \mathbb{R}$ is affine if and only if its graph is a line, and any non-vertical line is the graph of some affine function.

Definition

A function ℓ is said to be *affine* if the $\ell(x) - b$ is a linear function for some constant b, called the *intercept* of ℓ.

Every affine function $\ell\colon \mathbb{R} \to \mathbb{R}$ has the form $\ell(x) = mx + b$ for some constants m and b.

Fact: a function $\mathbb{R} \to \mathbb{R}$ is affine if and only if its graph is a line, and any non-vertical line is the graph of some affine function.

It is standard to simply refer to an affine function as a line and write its equation as $y = mx + b$.

Given a point (x_0, y_0) in the plane and a slope m, one can construct a line which passes through (x_0, y_0) with slope m by solving the following equation for y:

$$y - y_0 = m(x - x_0)$$

Given a point (x_0, y_0) in the plane and a slope m, one can construct a line which passes through (x_0, y_0) with slope m by solving the following equation for y:

$$y - y_0 = m(x - x_0)$$

Example

Find a line with slope -2 which passes through the point $(3, 5)$.

Given two points (x_0, y_0) and (x_1, y_1) in the plane (not on the same vertical line), one can construct a line which passes through both points by using the slope:

$$m = \frac{y_1 - y_0}{x_1 - x_0}$$

Given two points (x_0, y_0) and (x_1, y_1) in the plane (not on the same vertical line), one can construct a line which passes through both points by using the slope:

$$m = \frac{y_1 - y_0}{x_1 - x_0}$$

Example

Find a line which passes through the points $(1, 4)$ and $(7, 5)$.

1.4 Linear approximation and limits

When we speak of a linear approximation to a function $f(x)$ at a point a, our hope is to find an affine function $\ell(x)$ whose values are as close as possible to the values of f near a.

1.4 Linear approximation and limits

When we speak of a linear approximation to a function $f(x)$ at a point a, our hope is to find an affine function $\ell(x)$ whose values are as close as possible to the values of f near a.

Such a linear approximation does not necessarily exist:

Assuming it does exist, how would we find it?

Assuming it does exist, how would we find it?

Certainly it should pass through the point $(a, f(a))$, so if we can find the slope m then the line has the form

$$y = f(a) + m(x - a)$$

Look at the slope of the line passing through $(a, f(a))$ and $(x, f(x))$ for x close to a:

$$m = \frac{f(x) - f(a)}{x - a}$$

Look at the slope of the line passing through $(a, f(a))$ and $(x, f(x))$ for x close to a:

$$m = \frac{f(x) - f(a)}{x - a}$$

In some cases it is clear what happens as x gets closer and closer to a; for instance, take $f(x) = x^2$:

Look at the slope of the line passing through $(a, f(a))$ and $(x, f(x))$ for x close to a:

$$m = \frac{f(x) - f(a)}{x - a}$$

In some cases it is clear what happens as x gets closer and closer to a; for instance, take $f(x) = x^2$:

$$\begin{aligned}\frac{f(x) - f(a)}{x - a} &= \frac{x^2 - a^2}{x - a} \\ &= \frac{(x - a)(x + a)}{x - a} \\ &= x + a\end{aligned}$$

Look at the slope of the line passing through $(a, f(a))$ and $(x, f(x))$ for x close to a:

$$m = \frac{f(x) - f(a)}{x - a}$$

In some cases it is clear what happens as x gets closer and closer to a; for instance, take $f(x) = x^2$:

$$\begin{aligned} \frac{f(x) - f(a)}{x - a} &= \frac{x^2 - a^2}{x - a} \\ &= \frac{(x - a)(x + a)}{x - a} \\ &= x + a \end{aligned}$$

So as x approaches a the slope approaches $a + a = 2a$.

Thus part of the fundamentals of linear approximation lies in the notion of a *limit*.

Thus part of the fundamentals of linear approximation lies in the notion of a *limit*.

Informally, we say that a function g approaches a number m as x approaches a point a, written

$$\lim_{x \to a} g(x) = m$$

provided that the values of g can be made arbitrarily close to m by restricting the inputs x to some neighborhood of a.

This definition is admittedly a little vague, but it is possible to make it completely rigorous and to organize all of the theory in this course around the precise definition of a limit.

This definition is admittedly a little vague, but it is possible to make it completely rigorous and to organize all of the theory in this course around the precise definition of a limit.

However, scientists and mathematicians happily used calculus to solve hard problems for over two centuries before the rigorous definition was discovered, so we will follow their lead in this seminar and reason from intuition when it is necessary to work with limits.

Definition

- The *derivative* of a function f at a point a is the number

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

if the limit exists. In this case f is said to be *differentiable* at a.

Definition

- The *derivative* of a function f at a point a is the number

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

  if the limit exists. In this case f is said to be *differentiable* at a.

- If f is differentiable at a then the *local linear approximation* of f at a is the function

$$\ell(x) = f(a) + f'(a)(x - a)$$

Intuition: $f'(a)$ represents the factor by which f stretches tiny intervals centered at a:

Intuition: $f'(a)$ represents the factor by which f stretches tiny intervals centered at a:

Example

Use the previous definition to compute the local linear approximation of the given function at the given point.

- $f(x) = x^2$, $a = 2$

Example

Use the previous definition to compute the local linear approximation of the given function at the given point.

- $f(x) = x^2$, $a = 2$
- $f(x) = \sqrt{x}$, $a = 4$

Example

Use the previous definition to compute the local linear approximation of the given function at the given point.

- $f(x) = x^2$, $a = 2$
- $f(x) = \sqrt{x}$, $a = 4$
- $f(x) = \frac{1}{x}$, $a = 1$

1.5 Computing derivatives

Did you struggle a little with $\sqrt{x}$ and $\frac{1}{x}$?
Imagine trying $\frac{x^5}{\sqrt{4x^2-7}}$!

1.5 Computing derivatives

Did you struggle a little with $\sqrt{x}$ and $\frac{1}{x}$?
Imagine trying $\frac{x^5}{\sqrt{4x^2-7}}$!

There are a number of handy rules for recovering the derivative of a complicated function from the derivatives of simpler pieces, and this is normally how one handles functions like the above.

Power functions

For $p \neq 1$, the derivative of $f(x) = x^p$ is given by

$$f'(x) = px^{p-1}$$

Power functions

For $p \neq 1$, the derivative of $f(x) = x^p$ is given by

$$f'(x) = px^{p-1}$$

Example

- Differentiate $f(x) = x^3$

Power functions

For $p \neq 1$, the derivative of $f(x) = x^p$ is given by

$$f'(x) = px^{p-1}$$

Example

- Differentiate $f(x) = x^3$
- Differentiate $f(x) = \frac{1}{x^3}$

Power functions

For $p \neq 1$, the derivative of $f(x) = x^p$ is given by

$$f'(x) = px^{p-1}$$

Example

- Differentiate $f(x) = x^3$
- Differentiate $f(x) = \frac{1}{x^3}$
- Differentiate $f(x) = \sqrt[3]{x}$

Linearity

Given $f(x) = ag(x) + bh(x)$ we have

$$f'(x) = ag'(x) + bh'(x)$$

if g and h are differentiable at x.

Linearity

Given $f(x) = ag(x) + bh(x)$ we have

$$f'(x) = ag'(x) + bh'(x)$$

if g and h are differentiable at x.

Example

Differentiate $f(x) = x^3 + \frac{2}{x^3} - 4\sqrt[3]{x}$

Chain rule

Let $f = h \circ g$, meaning $f(x) = h(g(x))$. Then:

$$f'(x) = g'(x)h'(g(x))$$

if g is differentiable at x and h is differentiable at $g(x)$.

Chain rule

Let $f = h \circ g$, meaning $f(x) = h(g(x))$. Then:

$$f'(x) = g'(x)h'(g(x))$$

if g is differentiable at x and h is differentiable at $g(x)$.

Intuition: if g stretches by a factor of m near x and h stretches by a factor of n near $h(x)$ then $g \circ h$ stretches by a factor of mn.

Example

- Differentiate $f(x) = (x - 5)^2$

Example

- Differentiate $f(x) = (x-5)^2$
- Differentiate $f(x) = \frac{1}{x^3+2x}$

Example

- Differentiate $f(x) = (x - 5)^2$
- Differentiate $f(x) = \frac{1}{x^3+2x}$
- Differentiate $f(x) = \sqrt[3]{x^2+1}$

Product rule

Given $f(x) = g(x)h(x)$, we have:

$$f'(x) = g(x)h'(x) + g'(x)h(x)$$

if g and h are differentiable at x.

Product rule

Given $f(x) = g(x)h(x)$, we have:

$$f'(x) = g(x)h'(x) + g'(x)h(x)$$

if g and h are differentiable at x.

Intiuition:

Example

- Differentiate $f(x) = (x-5)^2(x-2)^5$

Example

- Differentiate $f(x) = (x-5)^2(x-2)^5$
- Differentiate $f(x) = \frac{\sqrt{x}}{x+2}$

Quotient rule

Given $f(x) = \frac{g(x)}{h(x)}$, we have:

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$$

if g and h are differentiable at x and $h(x) \neq 0$.

Quotient rule

Given $f(x) = \frac{g(x)}{h(x)}$, we have:

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$$

if g and h are differentiable at x and $h(x) \neq 0$.

Example

- Differentiate $f(x) = \frac{x^2}{\sqrt{x-1}}$.

Quotient rule

Given $f(x) = \frac{g(x)}{h(x)}$, we have:

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$$

if g and h are differentiable at x and $h(x) \neq 0$.

Example

- Differentiate $f(x) = \frac{x^2}{\sqrt{x-1}}$.
- Differentiate $f(x) = \frac{x^5}{\sqrt{4x^2-7}}$.