From Calculus to Machine Learning

Paul Siegel

July 2, 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

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.

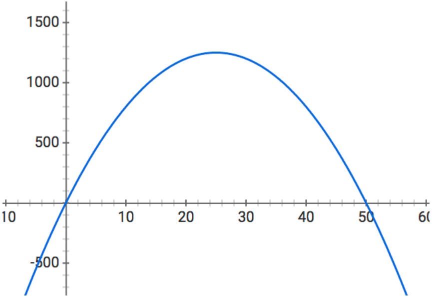
The objective function is quadratic and the constraint is linear, so we can hope to solve it analytically.

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 $x = 25$

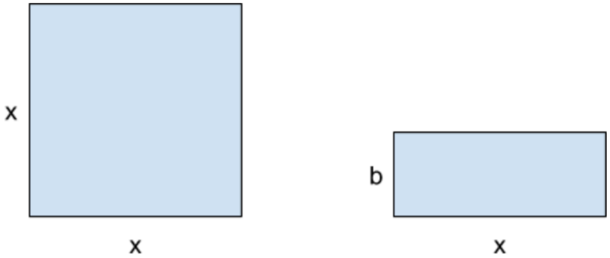
How does the algebraic magic work?

How does the algebraic magic work?

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

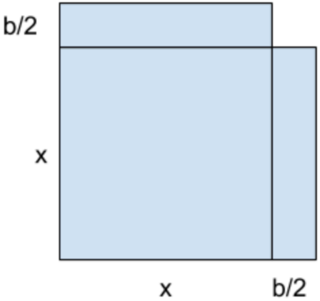
How does the algebraic magic work?

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



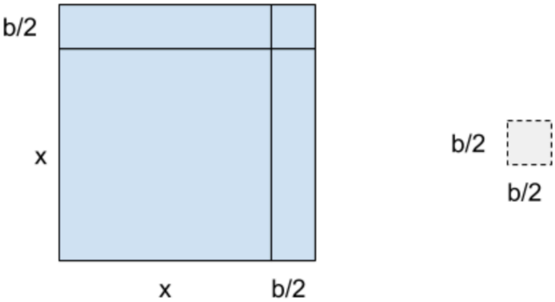
Cut the bx rectangle in half and rearrange:

Cut the bx rectangle in half and rearrange:

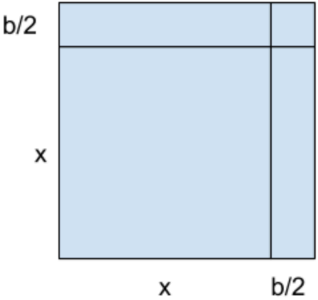


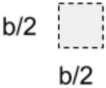
Complete the square!

Complete the square!



Complete the square!





Conclusion:

$$x^2 + bx = \left(x + \frac{b}{2}\right)^2 - \frac{b^2}{4}$$

In the example above, it looks like this:

In the example above, it looks like this:

$$\begin{aligned} A(x) &= x(100-2x) \\ &= -2x^2 + 100x \\ &= -2(x^2 - 50x) \\ &= -2((x-25)^2 - 625) \quad \text{magic!} \\ &= -2(x-25)^2 + 1250 \end{aligned}$$

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:

We want to minimize the surface area $S(x, y) = x^2 + 4xy$

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 $x^2 y = 25$.

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

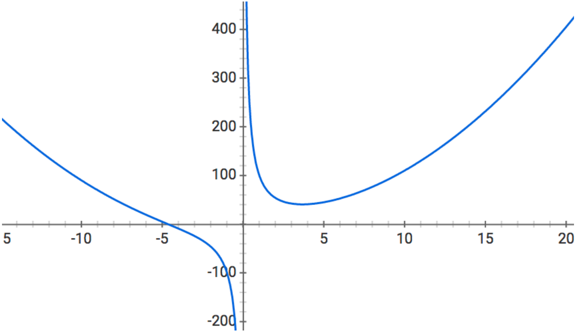
Use the constraint to eliminate y and get:

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 a quadratic function.