Lecture 4: Introduction to Optimization

COMP90049
Introduction to Machine Learning

Semester 2, 2020

Hadi Khorshidi, CIS

The slides prepared by Lea Frermann, CIS

Roadmap

Last time... Probability

- estimate the (conditional, joint) probability of observations
- Bayes rule
- Marginalization
- Probabilistic models
- Maximum likelihood estimation (taster)
- Maximum aposteriori estimation (taster)

Roadmap

Last time... Probability

- estimate the (conditional, joint) probability of observations
- Bayes rule
- Marginalization
- Probabilistic models
- Maximum likelihood estimation (taster)
- Maximum aposteriori estimation (taster)

Today... Optimization

- Curves, minima
- Derivatives
- Recipe for numerical optimization
- Maximum likelihood of the Binomial (from scratch!)

Optimization

Introduction I

We are all here to **learn** about Machine **Learning**.

- What is learning?

Introduction I

We are all here to **learn** about Machine **Learning**.

- What is learning?
- It probably has something to do with **change** or **mastering** or **optimizing** performance on a specific task

Introduction I

We are all here to **learn** about Machine **Learning**.

- What is learning?
- It probably has something to do with **change** or **mastering** or **optimizing** performance on a specific task
- Machine learning typically involves to build models (like seen last time), and learning boils down to **finding model parameters that optimize some measure of performance**

Introduction I

We are all here to **learn** about Machine **Learning**.

- What is learning?
- It probably has something to do with **change** or **mastering** or **optimizing** performance on a specific task
- Machine learning typically involves to build models (like seen last time), and learning boils down to **finding model parameters that optimize some measure of performance**

But, how do we know what is optimal?

Finding Optimal Points I

Finding the parameters that optimize a **target**

Ex1: Estimate the study time which leads to the **best grade** in COMP90049.

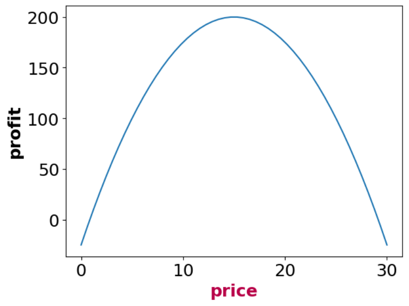
Ex2: Find the shoe price which leads to **maximum profit** of our shoe shop.

Finding Optimal Points I

Finding the parameters that optimize a **target**

Ex1: Estimate the study time which leads to the **best grade** in COMP90049.

Ex2: Find the shoe price which leads to **maximum profit** of our shoe shop.



Finding Optimal Points I

Finding the parameters that optimize a **target**

Ex1: Estimate the study time which leads to the **best grade** in COMP90049.

Ex2: Find the shoe price which leads to **maximum profit** of our shoe shop.

Ex3: Predicting **housing prices** from a weighted combination of house age and house location

Ex4: Find the parameters θ of a spam classifier which lead to the **lowest error**

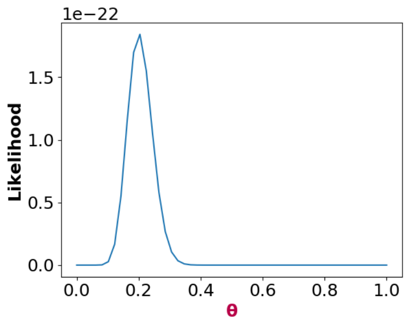
Ex5: Find the parameters θ of a spam classifier which lead to the **highest data log likelihood**

Finding Optimal Points I

Finding the parameters that optimize a **target**

Ex1: Estimate the study time which leads to the **best grade** in COMP90049.

Ex2: Find the shoe price which leads to **maximum profit** of our shoe shop.



Ex3 eighted combination of house age

Ex4 ssifier which lead to the **lowest error**

Ex5 ssifier which lead to the **highest**

Objective functions

Find parameter values θ that maximize (or minimize) the value of a function $f(\theta)$

- we want to find the **extreme** points of the **objective function**. Depending on our **target**, this could be
- ...the **maximum**
  E.g., the **maximum** profit of our shoe shop
  E.g., the **largest** possible (log) likelihood of the data

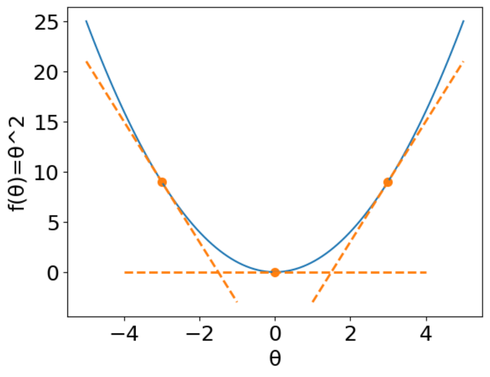
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} f(\theta)$$

- ...or the **minimum** (in which case we often call f a **loss function**)
  E.g., the **smallest** possible classification error

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}} f(\theta)$$

Finding extreme points of a function

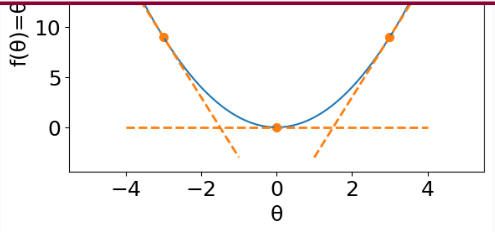
- At its **extreme point**, $f(\theta)$ is 'flat': its **slope** is equal to **zero**.
- We can measure the **slope** of a function at any point through its first **derivative** at that point
- The derivative measures the change of the output $f(\theta)$ given a change in the input θ
- We write the derivative of f with respect to θ as $\frac{\partial f}{\partial \theta}$



- At its **extreme point**, $f(\theta)$ is 'flat': its **slope** is equal to **zero**.
- We can measure the **slope** of a function at any point through its first **derivative** at that point
- The derivative measures the change of the output $f(\theta)$ given a change in the input θ
- We write the derivative of f with respect to θ as $\frac{\partial f}{\partial \theta}$

In order to find the parameters that maximize / minimize an objective function, we find those inputs at which the derivative of the function evaluates to zero.

That's it!



Example

- For our function, with a single 1-dimensional parameter θ

  $$f(\theta) = \theta^2$$

  Take the derivative

  $$\frac{\partial f}{\partial \theta} = 2\theta$$

  We want to find the point where this derivative is zero, so

  $$2\theta = 0$$

  and solve for θ

  $$\theta = 0$$

Finding a Minimum / Maximum

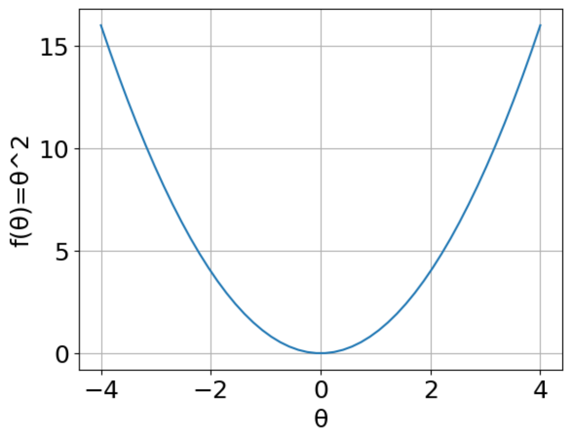
Example

- For our function, w

  Take the derivative

  We want to find the

  and solve for θ



The global minimum of $f(\theta) = \theta^2$ occurs at the point where θ=0.

Recipe for finding Minima / Maxima

1. Define your function of interest $f(\theta)$ (e.g., data log likelihood)
2. Compute its first derivative with respect to its input θ
3. Set the derivative equal to zero
4. Solve for θ

Recipe for finding Minima / Maxima

1. Define your function of interest $f(\theta)$ (e.g., data log likelihood)
2. Compute its first derivative with respect to its input θ
3. Set the derivative equal to zero
4. Solve for θ

Let's do this for a more interesting problem. Recall our binomial spam model from the last lecture?

Maximum Likelihood Optimization of the Binomial Spam Model

1. Problem setup / identifying the function of interest

- Consider a data set of emails, where each email is an observation x which is labeled either as `spam` or `not spam`
- We have N observations, each with 2 possible outcomes. The data consequently follows a **binomial distribution** and the data likelihood is

$$\mathcal{L}(\theta) = p(X; N, \theta) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x}$$

- So the parameter $\theta = P(spam)$

1. Problem setup / identifying the function of interest

- Consider a data set of emails, where each email is an observation x which is labeled either as `spam` or `not spam`
- We have N observations, each with 2 possible outcomes. The data consequently follows a **binomial distribution** and the data likelihood is

$$\mathcal{L}(\theta) = p(X; N, \theta) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x}$$

- So the parameter $\theta = P(spam)$
- Imagine we have a data set of 100 emails: 20 are `spam` (and consequently 80 emails are `not spam`).
- In the last lecture, we agreed intuitively that $P(spam) = \theta = 20/100 = \frac{x}{N}$.
- We will now derive the same result mathematically, and show that $\theta = \frac{x}{N}$ is the $\hat{\theta}$ that maximizes the likelihood of the observed data

Maximum Likelihood Optimization of the Binomial Spam Model

2. Computing its first derivative

$$\mathcal{L}(\theta) = p(X; N, \theta) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x}$$
$$\approx \theta^x(1-\theta)^{N-x}$$

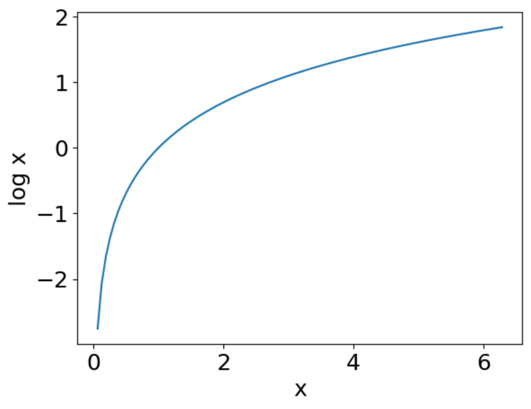
Move to log space (makes our life easier)

$$log\mathcal{L}(\theta) = xlog\theta + (N-x)log(1-\theta)$$

Maximum Likelihood Optimization of the Binomial Spam Model

(Log transformation aside)

- Log is a monotonic transformation: The same θ will maximize both $p(x, y)$ and $log\ p(x, y)$
- Log values are less extreme (cf. x scale vs y scale)
- Products become sums (avoid under/overflow)



Maximum Likelihood Optimization of the Binomial Spam Model

2. Computing its first derivative

$$\mathcal{L}(\theta) = p(X; N, \theta) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x} \approx \theta^x(1-\theta)^{N-x}$$

Move to log space (makes or life easier)

$$log\mathcal{L}(\theta) = xlog\theta + (N-x)log(1-\theta)$$

Maximum Likelihood Optimization of the Binomial Spam Model

2. Computing its first derivative

$$\mathcal{L}(\theta) = p(X; N, \theta) = \frac{N!}{x!(N-x)!}\theta^x(1-\theta)^{N-x} \approx \theta^x(1-\theta)^{N-x}$$

Move to log space (makes or life easier)

$$log\mathcal{L}(\theta) = xlog\theta + (N-x)log(1-\theta)$$

Take the derivative of $\mathcal{L}$ wrt the parameters θ

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{x}{\theta} - \frac{N-x}{1-\theta}$$

THE UNIVERSITY OF MELBOURNE

Maximum Likelihood Optimization of the Binomial Spam Model

3. Set the derivative to zero

$$0 = \frac{x}{\theta} - \frac{N - x}{1 - \theta}$$

4. Solve for θ

$$\begin{aligned}
\frac{x}{\theta} &= \frac{N-x}{1-\theta} && [\times(1-\theta)] \\
\frac{x \times (1-\theta)}{\theta} &= N - x && [\times \frac{1}{x}] \\
\frac{1-\theta}{\theta} &= \frac{N-x}{x} && [\text{rearrange}] \\
\frac{1}{\theta} - 1 &= \frac{N}{x} - 1 && [+1] \\
\frac{1}{\theta} &= \frac{N}{x} && [\textit{flip}] \\
\hat{\theta} &= \frac{x}{N}
\end{aligned}$$

Which corresponds to our estimate of $\frac{x}{N} = \frac{20}{100} = 0.2$ for our spam classification problem.

Possible Complications

Can you think of scenarios where this approach breaks down?

Possible Complications

Can you think of scenarios where this approach breaks down?

- Our loss function is not differentiable
- It is mathematically impossible to set the derivative to 0 and solve for the parameters θ. “No closed-form solution”.
- Our function has multiple ‘extreme points’ where the slope equals zero. Which one is the correct one?

to be continued...

Summary

- What is optimization?
- Objective function / loss function
- Gradients, derivatives, and slopes

Next: Naive Bayes

Optional: Solution subject to Constraints

Constrained Optimization

Finding the parameters that optimize a **target** subject to one or more constraints.

- Find the shoe price which leads to maximum **profit** of our shoe shop. But we can't chare more than 60$ (for tax reasons).

- I want to estimate the parameters of a Categorical distribution to maximize the **data log likelihood** and I know that the parameters must sum to 1.

Constrained Optimization

In mathematical optimization

- it often happens that the parameters we want to learn have to obey constraints

$$\underset{\theta}{\mathrm{argmin}} f(\theta)$$

$$\text{subject to } g_c(\theta) \geq 0,$$

- ideally, we would like to incorporate such constraints and still be able to follow the general recipe for optimization discussed before
- **Lagrangians** allow us to do exactly that
- we combine our target functions with (sets of) constraints multiplied through **Lagrange multipliers** λ

$$\mathcal{L}(\theta, \lambda) = f(\theta) + \sum_c \lambda_c g_c(\theta)$$

- proceed as before: derivative, set to zero, solve for θ

Constrained Optimization

Example

- Find an optimal parameter vector θ such that each all θ_i sum up to a certain constant b.
- Formalize the constraint:

$$\sum_i \theta_i = b$$

- Set the constraint to zero

$$0 = \sum_i \theta_i - b = -b + \sum_i \theta_i$$

- set the constraint and write the Lagrangian

$$g_c(\theta) = -b + \sum_i \theta_i$$

$$\begin{aligned}\mathcal{L}(\theta, \lambda) &= f(\theta) + \lambda g_c(\theta) \\ &= f(\theta) + \lambda(-b + \sum_i \theta_i)\end{aligned}$$

- proceed as before: derivative, set to zero, solve for θ