

# Regularization + Assignments 5, 6

(Neural Networks Implementation and Application Tutorial)

Vilém Zouhar, Noon Pokaratsiri Goldstein

15th, 14th December 2021

# Overview

- Assignment 5
- Regularization
- Data Augmentation
- Ensembling
- Dropout
- Assignment 6

# Assignment 5

## Tips

- Start early
- Ask questions on Piazza
  - ▶ We try to be very active there and lead you to solutions
- What went wrong?
- What was the hardest part?

# Assignment 5

## Tips

- Start early
- Ask questions on Piazza
  - ▶ We try to be very active there and lead you to solutions
- What went wrong?
- What was the hardest part?
- Who wants to present their solution?
- Any questions?

# Regularization

## General

- What is regularization, why do we need it?
- What is the training & generalization error?

# Regularization

## General

- What is regularization, why do we need it?
- What is the training & generalization error?

## Norm Penalty

- What is a norm penalty?
- Where do we add this penalty?
- What's the practical effect of  $L_1$  and  $L_2$  (sparsity)?

# Data Augmentation

- What is it?
- Why & how does it work?
- Examples in computer vision domain?
- Examples in computational linguistics domain?

# Data Augmentation

- What is it?
- Why & how does it work?
- Examples in computer vision domain?
- Examples in computational linguistics domain?

Does data augmentation of the outputs also make sense?



# Data Augmentation

- What is it?
- Why & how does it work?
- Examples in computer vision domain?
- Examples in computational linguistics domain?

Does data augmentation of the outputs also make sense?

## Yes! - Label smoothing

- $(0, 0, 1, 0) \rightarrow (0.1, 0.1, 0.7, 0.1)$
- Still same argmax (predictions)
- Does not punish softmax and  $\sigma$  for not predicting 1 which they can't
- Reduces dependency on individual samples

# When to stop training?

- Why not when the training loss reaches 0?
  - ▶ Can sometimes not do that
  - ▶ We're interested in generalization error

# When to stop training?

- Why not when the training loss reaches 0?
  - ▶ Can sometimes not do that
  - ▶ We're interested in generalization error

## Patience $k$

- Stop if your metric does not improve for at least  $k$  epochs
- Can use validation loss but also extrinsic metric

# Ensembling

- Simple trick: train multiple networks with different seed
- Why does this work?
  - ▶ Some of their errors are not systematic and is caused by noise
  - ▶ Noise cancels out

## Implementation

- Classification:
    - ▶ Hard classification: make prediction by every model and pick most common class
    - ▶ Soft classification: sum probabilities and pick argmax class
    - ▶ Are they always the same? 🤔
  - Regression:
    - ▶ Mean
- 
- Provides systematic improvement of a few points
  - Not that interesting, costs a lot of power
  - SotA papers publish results without ensembling because everyone knows it would help slightly

# Dropout

- Poor man's bagging
- How does it work?
- Part of assignment 6

## What went wrong?

- A model had training loss of 10.1 and development loss of 2.0.
- How could this be?

# Assignment 6

Any questions?