Datasets, generalization, and overfitting

Send feedback

Estimated module length: 105 minutes

Learning objectives

- Identify four different characteristics of data and datasets.
- Identify at least four different causes of data unreliability.
- Determine when to discard missing data and when to impute it.
- Differentiate between direct and derived labels.
- Identify two different ways to improve the quality of human-rated labels.
- Explain why to subdivide a dataset into a training set, validation set, and test set; identify a potential problem in data splits.
- Explain overfitting and identify three possible causes for it.
- Explain the concept of regularization. In particular, explain the following:
  - Bias versus variance (adaptation to outliers…)
  - L_2 regularization, including Lambda (regularization rate)
  - Early stopping
- Interpret different kinds of loss curves; detect convergence and overfitting in loss curves.

Prerequisites:

This module assumes you are familiar with the concepts covered in the following modules:

- Introduction to Machine Learning
- Linear regression
- Working with numerical data
- Working with categorical data

Introduction

This module begins with a leading question. Choose one of the following answers:

If you had to prioritize improving one of the following areas in your machine learning project, which would have the most impact?

- [x] Improving the quality of your dataset

  Data trumps all. The quality and size of the dataset matters much more than which shiny algorithm you use to build your model.

  Correct answer.

- [ ] Applying a more clever loss function to training your model

And here's an even more leading question:

Take a guess: In your machine learning project, how much time do you typically spend on data preparation and transformation?

- [ ] Less than half of the project time
- [x] More than half of the project time

  Yes, ML practitioners spend the majority of their time constructing datasets and doing feature engineering.

  Correct answer.

In this module, you'll learn more about the characteristics of machine learning datasets, and how to prepare your data to ensure high-quality results when training and evaluating your model.

Help Center

Previous
Test your knowledge (10 min)

Next
Data characteristics (10 min)

Was this helpful?
Send feedback