Exploring Data

Learn how to describe data, and some of the challenges of analyzing imperfect data.

Instructors: Any notes to add here?

Materials Needed

- Puzzles
- Survey on Google
- Worksheets on observations about data Some worksheets?

Learning Objectives

Introduction to Machine Learning

- Set the stage for the camp, and get students excited about Machine Learning!

Set-up and Surveys

- Get everyone logged in to their iPad.
- Gather data from students to be used in later activities

Imperfect Data

- Understand the types of problems that can occur with imperfect data.
- Understand why cleaning data is important, but also why it can be problematic.

Summary Statistics

- Understand the different types of data: numerical, categorical,...
- Describe data using the mean, median, mode, standard deviation, etc.
- Understand the strengths and limitations of these descriptors.
- Students will practice using these descriptors with their own data, from the survey.

Machine Learning Prequel

- Understand our goals, what we're going to do to get there.

Exploring Data

The primary goals of this session are to introduce the camp and to get students some immediate experience playing with data. By the end of the session, students should have a basic understanding of what machine learning is and how it can be used, the challenges of working with data, and how to describe data using summary statistics.

Introduction to Machine Learning

This is a short introductory talk on Machine Learning, given by some nonempty subset of {Kaisa, Vern}. The goal is to give students an idea of what machine learning is, what it can do, and what types of things we'll need to learn in order to be able to use it effectively.

Set-up and Surveys

Get students logged in to lab computers and ipads, and have them fill out surveys.

Survey draft: `https://forms.gle/ZGqRbCxyFmTM15K37`

Imperfect Data

While students are doing this activity, someone will work on preparing the survey data for the next activity.

Each group will be given a puzzle to assemble, but each puzzle will have several pieces missing, several duplicate pieces, and several pieces from a different puzzle. Students will be told that the puzzle pieces represent "data." After students do their best to assemble the puzzle, we'll have a group discussion about the problems they ran into when trying to assemble their puzzles, and what this represents for data.

We'll make sure to talk about the following points:

- Sometimes data is incomplete (missing piece)
- Sometimes data is duplicated (duplicate piece)
- Sometimes data is incorrect (piece from a different puzzle)
- Even if we have imperfect data, we can still figure out the big picture from the data that we do have.

Transition to a discussion about cleaning data, asking students what kinds of things we can do to fix imperfect data. Some possible suggestions:

- Just omit data that is missing, incomplete, or obviously wrong.
- Think about what happens if missing data takes a range of values.
- Change data to what we think it should be.

Then, talk about why we need to be careful when cleaning data. What kinds of issues can we introduce if we are overzealous with our cleaning? Why is changing data potentially a problem? How do we decide when data is "obviously wrong"?

Summary Statistics

We'll talk about various summary statistics, and students will practice using these with their own data. Need to develop worksheets/notebooks for these activities, decide what data we need for the surveys. How do we want them to do computations?

Machine Learning Prequel

Some kind of short wrap-up discussion that brings us back to the big picture. Maybe map out what they'll do for the next day and a half?