Machine Learning

Pawel Wocjan

University of Central Florida

Fall 2020

Sources for Slides

- I have extensively used the machine learning materials that have been prepared by Google.

  `https://developers.google.com/machine-learning/crash-course/`

- Google has licensed these materials under the Creative Commons Attribution 3.0 License.

  `https://creativecommons.org/licenses/by/3.0/`

Basic ML Terminology

- What is (supervised) machine learning? Concisely put, it is the following:

  ML systems learn how to combine input to produce useful predictions on never-before-seen data.

- Let's explore basic ML terminology machine learning terminology.

Labels and features

- A **label** is the thing we seek to predict.

  For instance, the label could be the future price of a stock or the type of animal shown in a picture.

- A **feature** x_i is an input variable that that is used to predict the label where $i \in \{1, \ldots, n\}$ and n is the number of features.

  For instance, the feature could be the previous day's closing price of a stock or a pixel of an image.

- A simple machine learning project might use a single or few features, while a more sophisticated machine learning project could use millions of features.

Features for spam detection

- For spam detection, for instance, the feature could include the following features:
  - words in the email text
  - sender's address
  - time of day the email was sent
  - email contains the phrase "one weird trick."

Labeled and unlabeled example

- An **example** is a particular instance of data that is fed into a model.
- We break examples into two categories:
  - labeled examples
  - unlabeled examples

Labeled example

- A **labeled example** includes both feature(s) and the label.

  `labeled example:` {`features, label`}: (x, y)

- We use labeled examples to train the model.

  In our spam detector example, the labeled examples would be individual emails that users have explicitly marked as "spam" or "not spam."

Unlabeled example

- An **unlabeled example** contains features but not the label:

  `unlabeled example:` {`features, ?`}: $(x, ?)$

- Once we've trained our model with labeled examples, we use that model to predict the label on unlabeled examples.

  In the spam detector, unlabeled examples are new emails that humans haven't yet labeled.

Training and inference

- A model defines the relationship between features and labels.
- Let's highlight two phases of a model's life:
  - **Training** means creating or learning the model.

    You show the model labeled examples (x, y) and enable it to gradually learn the relationships between features and label so that its predictions $\hat{y}$ are close to the labels y.
  - **Inference** means applying the trained model to unlabeled examples.

    You use the trained model to make useful predictions $\hat{y}$.

Regression

- A **regression model** predicts continuous values.

  For example, regression models make predictions that answer questions like the following:

  - What is the value of a house in California?
  - What is the probability that a user will click on this ad?

Classification

- A **classification model** predicts discrete values. For example, classification models make predictions that answer questions like the following:
  - Is a given email message spam or not spam?
  - Is this an image of a dog, a cat, or a hamster?

Key Terms

- classification model
- example
- feature
- inference
- label
- model
- regression model
- training