CPSC 330 Lecture 14: Feature Engineering and Selection

Announcements

- HW5 is due next week Monday. Make use of office hours and tutorials this week.
- CPSC 330 final exam window (in the CBTF) will be Dec. 17-19
- Midterm 1 results will be out tonight (as soon as I get home!)
  - Viewing sessions (in the CBTF) will be next week

Continue with Lecture 13 demo

- SHAP!

Break

Let’s take a 10-min break

iClicker Exercise 14.0

iClicker cloud join link: https://join.iclicker.com/VYFJ

Suppose you are working on a machine learning project. If you have to prioritize one of the following in your project which of the following would it be?

a. The quality and size of the data

b. Most recent deep neural network model

c. Most recent optimization algorithm

Feature engineering motivation

Discussion question

- Suppose we want to predict whether a flight will arrive on time or be delayed. We have a dataset with the following information about flights:
  - Departure Time
  - Expected Duration of Flight (in minutes)

Upon analyzing the data, you notice a pattern: flights tend to be delayed more often during the evening rush hours. What feature could be valuable to add for this prediction task?

Garbage in, garbage out.

- Model building is interesting. But in your machine learning projects, you'll be spending more than half of your time on data preparation, feature engineering, and transformations.
- The *quality* of the data is important. Your model is only as good as your data.

Activity: Measuring quality of the data

- Discuss some attributes of good- and bad-quality data

What is feature engineering?

> **Feature engineering** is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.
> - Jason Brownlee

- Better features: more flexibility, higher score, we can get by with simple and more interpretable models.
- If your features, i.e., representation is bad, whatever fancier model you build is not going to help.

Some quotes on feature engineering

A quote by Pedro Domingos A Few Useful Things to Know About Machine Learning

> … At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.

Some quotes on feature engineering

A quote by Andrew Ng, Machine Learning and AI via Brain simulations

> Coming up with features is difficult, time-consuming, requires expert knowledge. “Applied machine learning” is basically feature engineering.

Better features usually help more than a better model

- Good features would ideally:
  - capture most important aspects of the problem
  - allow learning with few examples
  - generalize to new scenarios.
- There is a trade-off between simple and expressive features:
  - With simple features overfitting risk is low, but scores might be low.
  - With complicated features scores can be high, but so is overfitting risk.

The best features may be dependent on the model you use

- Examples:
  - For counting-based methods like decision trees separate relevant groups of variable values
    - Discretization makes sense
  - For distance-based methods like KNN, we want different class labels to be “far”.
    - Standardization
  - For regression-based methods like linear regression, we want targets to have a linear dependency on features.

Motivating Feature Engineering

Questions:

- What are two possible ways we could “engineer” features?
  - Think broadly and philosophically rathern than an implementation…