D200, Problem Set 2: Discrete Choice Models

Due: 18 February 2025 here in groups of 4.

Stefan Bucher

This problem set will review classification as discussed in the lecture through the lens of discrete choice modeling, a classically used method in economics.

You will work with the Expedia Dataset using the choice-learn package. In order to load the Expedia dataset, you first need to download it from https://www.kaggle.com/c/expedia-personalized-sort and save the `train.csv` file in your python package's `./lib/python3.12/site-packages/choice_learn/datasets/data/expedia.csv` (otherwise `choice_learn` will tell you the exact location in a `FileNotFoundError`).

Problem 1: Conditional Logit

(1a) Reading up if necessary, provide a brief explanation of the Conditional Logit model and its use in discrete choice modeling.

(1b) Load the Expedia dataset using `choice_learn.datasets.load_expedia(preprocessing="rumnet")`, discard all but the first 5000 choices (for computing efficiency), and split the data into a training and test set. Look at the dataset and its documentation online and describe the dataset's structure and the variables it contains.

(1c) Write down a sensible model specification for the Conditional Logit model for the Expedia dataset, for examples using the hotel features

- log(price)
- star rating
- review
- whether the hotel is a brand
- location desirability scores

You may also want to include hotel fixed effects.

Feel free to play around with alternative model specifications.

(1d) Fit your conditional logit model to the Expedia data and report the cross-entropy loss on the test data using TensorFlow's `tf.keras.losses.CategoricalCrossentropy` loss function.

(1e) Display the resulting parameter estimates and interpret them.

Problem 2: RUMnet

This problem uses the "RUMnet" model, a neural network-based model specifically designed for discrete choice modeling.

(2a) Fit the "RUMnet" model shipped with `choice_learn` to the Expedia dataset and again report the cross-entropy loss on the test data.

(2b) Compare the cross-entropy loss of the Conditional Logit model and the RUMnet model. What do you observe?

(2c) Discuss the advantages and disadvantages of the Conditional Logit model and the RUMnet model.