DATA 221
Homework 5 (rev 0)

W. Trimble

Due: Tuesday 2023-05-02 - 11:59pm

Kaggle user `hugodarwood` created a dataset with over 20,000 recipes from the website Epicurious, scraped in 2017:

`https://www.kaggle.com/datasets/hugodarwood/epirecipes/code`.

Loosely speaking, there are a few groups of variables in the dataset: nutritional variables (‘calories‘, ‘protein‘, ‘fat‘, ‘sodium‘), Ingredient tags (‘almond‘, ‘amaretto‘, ‘anchovy‘...), Place tags (‘alabama‘, ‘alaska‘, ‘aspen‘,...), and Other tags (‘advance.prep.required‘, ‘anthony.bourdain‘, etc.).

You may want to carry out your own exploratory dataset to a) become familiar with the dataset, and b) carry out some pre-processing on the dataset. You will have to make choices during pre-processing–these choices are up to you. We are happy to offer suggestions during office hours, but as long as you feel that you have made a reasonable choice, we will not take points off.

Once you have pre-processed the data...

1. Logistic Regression with ℓ_1 Regularization

   (a) Find the logistic regression coefficients to predict whether a recipe has the ‘pie‘ tag using ℓ_1 regularization.

   (b) Plot the regression coefficients as a function of the (logarithm of the) regularization parameter t–example on the second page of the assignment.

   (c) Find the optimum regularization parameter t by optimizing for minimum error on a test set.

2. Regression / nutritional information

   (a) Train a (linear regression? neural network?) model to predict the nutrition information fields from everything else.

   (b) Evaluate your model on a testing holdout set, report its accuracy, and predict the nutrition information for the rows where nutrition information is missing.

   (c) Which coefficents are the largest? (Salted? Fried? protein-feed-enhanced?)

3. Principal Component Analysis

   (a) Perform Principal Component Analysis (a.k.a., singular value decomposition) on all the features of the dataset.

   (b) Make a graph showing the total fraction of variance in the first N principal components. How many principal components do you need to retain half of the variance of the data?

   (c) Display scatter plots of the first two principal components, PC1 and PC2, for the two response variable classes. Label the axes with the fraction of the variance explained by PC1 and PC2.

4. Embedding

   (a) Calculate the all-against all distance matrix in Euclidean space for the category labels. Produce a clustering.

   (b) Calculate the all-against all distance matrix with a Minkowski metric for the category labels. Produce a clustering.

   (c) Map the category labels to a word2vec language embedding, calculate an all-against-all distance matrix, and produce a clustering.

   (d) Plot the three clusterings in principal coordinates space with the conventional (PC 1, 1.2%...) labels.