Recipe Rating Prediction

Sarah Hensley
hensleys@stanford.edu
Stanford | ENGINEERING
Electrical Engineering

Summary

Objective:
Predict the star rating of a recipe given its ingredients

Methods:
Compare linear regression and kernel methods and compare different ingredient embeddings to predict rating

Results:
Best model achieves average mean square error of 0.172

Dataset

- Online recipes from Food.com of any cuisine
- Each recipe has a variable-length list of ingredients and an average star rating out of five from users
- 5000 examples in training dataset, 1000 examples in validation dataset, 1000 examples in test dataset

[“canned chicken”, “green chilies”, “cream cheese”, “green onions”, “flour tortillas”, “cooking oil”] ⇒ 5.0

Models

Loss function:
Mean squared error

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Linear Regression:
Gradient descent to learn θ for $h(x) = \theta^T x$

Kernel Methods:
Gradient descent to learn β for Square and Gaussian kernels

$$h(x) = \sum_{i=1}^{n}\beta_i K(x^{(i)}, x)$$

$$K(x, z) = (x^T z)^2 \qquad K(x, z) = \exp\left(-\frac{||x - z||^2}{2\sigma^2}\right)$$

Features

One-Hot Encoding

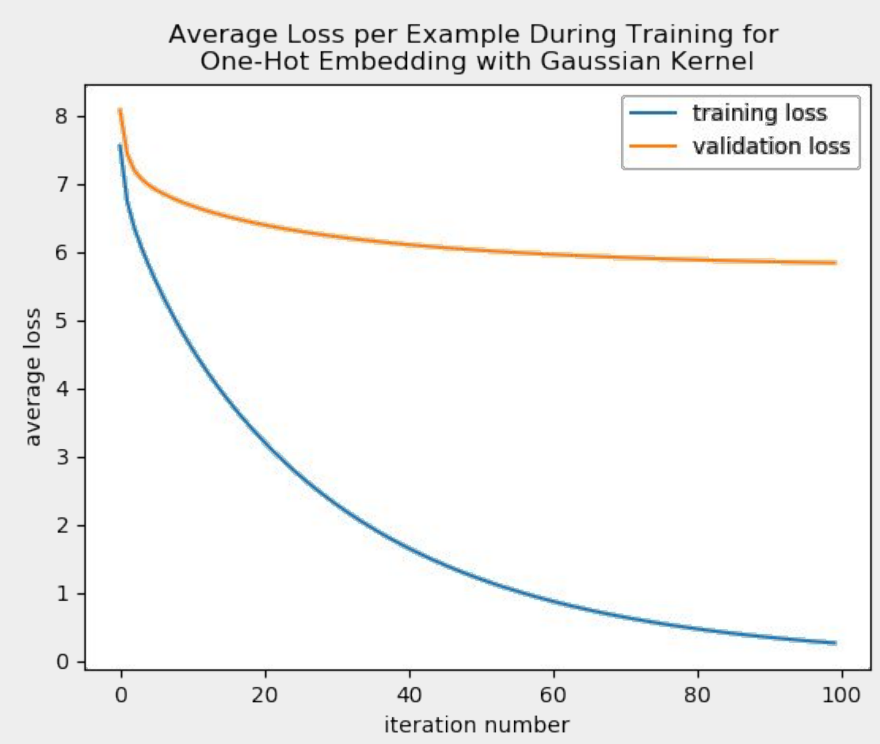
- Let each unique ingredient map to one-hot vector
- Recipe ingredient list as sum of one-hot vectors
- Total embedded vector length 3866

BERT Embeddings

- Embedding from BERT, a pre-trained NLP model
- Each ingredient list is represented as a “sentence”, then embedded as a length 768 vector
- Capture similarities between, for example, “olive oil”, “extra virgin olive oil”, and “cooking oil”

Results

	Linear Regression, One-Hot	Linear Regression, BERT	Square Kernel, One-Hot	Gaussian Kernel, One-Hot	Square Kernel, BERT	Gaussian Kernel, BERT
Training Loss	0.1958	0.2114	0.1181	0.266	0.2338	0.0002
Validation Loss	0.1893	0.1998	0.2028	5.843	0.2262	0.3886
Test Loss	0.1719	0.1840	0.1866	6.962	0.2125	0.3989



Discussion

- Conducted hyperparameter sweep over step size and number of iterations, using validation set to choose best values
- Linear regression with one-hot embedding performs best
  - L1 distance between predictions and true values is 0.436
  - Within one half of a star (out of five) on average
- Gaussian kernel generalizes poorly to unseen data
  - One-hot embedding is extremely sparse, making the “distance” between recipes not very informative
  - Improved performance on BERT embeddings supports this
  - Expect better performance on larger dataset

Future Work

- Apply work to larger dataset
  - Total of 226,000 recipes available, but dataset cannot fit in memory
- Compare to other text embeddings
  - BERT assumes sentence structure
  - Other text embeddings may better capture ingredient relationships
- Explore pathological performance of Gaussian kernel
- Compare to other kernels

[1] Ng, A. & Ma, T. (2019) Linear Regression. From *CS 229 Lecture Notes*.
[2] Ng, A. & Ma, T. (2019) Kernel Methods. From *CS 229 Lecture Notes*.
[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint.