Machine Learning Engineer Nanodegree

Capstone Proposal

Sai Marasani
August 20, 2018

Domain Background:

Picture yourself strolling through your local, open-air market... What do you see? What do you smell? What will you make for dinner tonight?

If you're in Northern California, you'll be walking past the inevitable bushels of leafy greens, spiked with dark purple kale and the bright pinks and yellows of chard. Across the world in South Korea, mounds of bright red kimchi greet you, while the smell of the sea draws your attention to squids squirming nearby. India's market is perhaps the most colorful, awash in the rich hues and aromas of dozens of spices: turmeric, star anise, poppy seeds, and garam masala as far as the eye can see. Some of our strongest geographic and cultural associations are tied to a region's local foods. Particularly I like to cook my food and to explore different varieties of food so this project caught my attention while I was exploring for projects on Kaggle.

Problem Statement:

This is Multi Class Classification problem where there are 20 possible responses from the features we pass as input. Final goal of this project is to predict cuisine type from the ingredients list provided. For out of sample performances I have considered Cross-validation scores. The ultimate goal is to accurately predict the cuisine type provided with features as input. We will use accuracy measurement as a model performance metrics which gives us a percentage based on all correct and incorrect classifications of our model on the test data.

Datasets and Inputs:

"Yummly" has provided the data set which consists of 39774 samples which consists of list of ingredients which are going to be our features and cuisine column which is going to be our response variable. Data looks like below picture

	cuisine	id	ingredients
0	greek	10259	[romaine lettuce, black olives, grape tomatoes...
1	southern_us	25693	[plain flour, ground pepper, salt, tomatoes, g...
2	filipino	20130	[eggs, pepper, salt, mayonaise, cooking oil, g...
3	indian	22213	[water, vegetable oil, wheat, salt]
4	indian	13162	[black pepper, shallots, cornflour, cayenne pe...

The data is in json format. I have started with very naïve approach to decide what features can influence the outcome which is cuisine type like Smell, ingredients; location (sometimes), Spiciness which I think can be used in part of designing feature engineering process.

Data set link: https://www.kaggle.com/c/whats-cooking/data

Solution Statement:

This is a multi-class classification problem as there is an label attached to every row of data which is Cuisine in our case and we are going to predict the cuisine based in ingredients provided. I am planning to try multiple supervised learning algorithms (ensemble, svc..) mostly based on the cheat sheet of algorithms provided in scikit-learn website which shows the flow of algorithms which we should try for particular type of problem . I will use cross validation score on training data as a metric to check accuracy.

Benchmark Model:

I will use Null model as the benchmark model , null model always predicts the most frequent class from the training data. I will be using sklearn accuracy score as metric to check accuracy score of null model.

Evaluation Metrics:

As I am going to use supervised learning approach, I will use accuracy score as evaluation metric.

Project Design:

As I will be dealing with text data, I will start first exploring the data like checking for null values and then I will check possibilities to apply feature engineering techniques to get new features like for example deriving number of ingredients from ingredients list. My goal in feature engineering will be to create signals from data rather than adding noise. Then I will use Term Frequency-Inverse Document Frequency (Tf-idf) to compute the relative frequency that a word appears in a document to its frequency

across all documents in the corpus. Next step would be to use pipline for proper cross validation, by passing pipe line to cross validation score features will be created from X within each fold of cross validation which simulates the real world scenarios where features will not be seen during training which we can use in testing.

Next step will be combining GridSearchCV with pipeline to locate optimal tuning parameters by performing exhaustive grid search with different parameter combinations in order to search for best cross validated accuracy. Another option for tuning parameters would be using randomized searchCV instead of searching all possible combinations. Final step would be predicting on test data and calculate accuracy score.