Use Machine Learning to Get Insights from the US Food prices from 2004 to 2010

Wanyi Li*

Abstract— **Food market is one of the few markets that have the characteristics of a perfectly competitive market. Using a dataset of seven years of food prices in the US, through unsupervised learning, I discover that similar foods share similar pricing patterns. I also discover that the East and the West have similar food prices, while the Central and the South share similar ones. Then, I use linear regression and time series analysis to predict the food prices at the given location and the time.**

I. Motivation

Everyone's living quality is tied to the prices of food. The food market is one of the few markets that is close to what economists consider a "competitive" market, which means the prices of food items should be determined by their demand and supply functions. In reality, it is very hard to approximate and estimate demand and supply functions. However, we do know that food prices should vary depending on the nature of the food and the location of the market. Given a dataset of seven years of quarterly food prices at different locations in the US, I want to apply machine learning techniques to gain insights about the food market. Often, computational results from machine learning methods can be deemed as a black box by outsiders. Through unsupervised learning, I am interested in deriving results that make sense intuitively and can be interpreted easily. Through supervised learning, I am interested in predicting the food prices in the US.

II. Data

United States Department of Agriculture (USDA) published this data set called "Quarterly Food-At-Home Price Database"[1] in 2012 which includes the prices of 54 different food groups in 35 different geographical locations in the US at a quarterly frequency from 2004 to 2010. Each food groups belongs to one of the following categories: "fruits and vegetables", "grains and dairies", "meats and egg", "fats, beverages and prepared foods". Prices are presented in dollars per 100 grams of food as purchased by consumers. Each of the 26 metropolitan areas and the 9 non-metropolitan areas belongs to one of four regions: East, South, Central, and West. An example row of the data looks like:

Location	Region	Year	Quarter	Price
1	3	2005	3	0.4285299

A very small portion of the prices of certain foods at some locations are missing in the data. So far, I have dealt with it by replacing the empty values with the average of the prices of all time and locations of that food. A better implementation would be to replace empty values with the average price of that food at the same location and in the same year. The difficulty of this method is that, in this dataset, often, the prices of one single food group of all years at the same location are missing altogether. So, we will use the average price to replace empty values for now.

III. Unsupervised Learning

A clustering algorithm is the natural choice for unsupervised learning on this dataset. The dataset includes prices with labels of time, location and food groups. I am particularly interested in using K means clustering algorithm to cluster the locations and the food groups. Explicitly, I want to answer the following questions: do food prices at locations with close proximity follow similar pricing patterns? do similar food groups share similar pricing patterns?

A. Features

The dataset comes with the prices in the unit of dollar per 100 grams of the food. Just by clustering the price vectors of all locations or all food groups, it is possible to see that foods that are more expensive in one cluster, and cheaper the other. So, I am interested in running clustering algorithms on normalized prices or on the

*Wanyi Li is a PhD student at Stanford University Management Science and Engineering Department, Huang Engineering Center 263G, 475 Via Ortega, Stanford, CA. wanyili@stanford.edu

changes of normalized prices over time. Explicitly, I test the following three versions of features:

1) Unnormalized prices as it is given in the data p_t;
2) Normalized prices: $p_t/\bar{p}$ where $\bar{p}$ is the average of the price vector;
3) Changes of normalized prices over time: $(p_t - p_{t-1})/\bar{p}$.

I will display the clustering results in the following section by using the feature and the number of clusters that result in the outcomes that makes the most sense to me, but I do understand that this is a subjective decision.

B. Clustering food groups

For each food category, I cluster the food groups using the changes of normalized prices over time at all locations (third version of features) and the similar food groups fall into the same clusters mostly, though there are a few "anomalies" in each category that does not necessarily match our expectations.

Fresh/frozen low fat meat Fresh/frozen regular fat meat Fresh/frozen poultry Fresh/frozen fish Canned select nutrients
Raw nuts and seeds Processed nuts, seeds and nut butters Canned poultry
Eggs
Canned meat

TABLE I: 4 clusters of meats and egg

In meats and egg category, egg is in its own cluster; most meats are in one cluster; the nuts are in one cluster; canned meat are in its own cluster. It is hard to explain why canned poultry is in the same cluster with nuts, but canned poultry does not have a huge market share in general, so it is not of importance here.

In fruits and vegetables category, most of the fresh/frozen vegetables are in one category and most of the canned vegetables are in one category. The third cluster include fruits with two "anomalies" which are canned starchy vegetables and canned other vegetables. Using two clusters give us similar results but the items in the third clusters are distributed in the first two, so the division between fresh/frozen and canned is not as clear.

Fresh/Frozen dark green vegetables Fresh/Frozen orange vegetable Fresh/Frozen starchy vegetables Fresh/Frozen select nutrient vegetables
Canned dark green vegetables Canned select nutrients Canned orange vegetables Fresh/Frozen other vegetables Canned Legumes Frozen/Dried Legumes Fruit Juice
Fresh/Frozen fruit Canned Fruit Canned starchy vegetables Canned other vegetables

TABLE II: 3 clusters of Fruits and Vegetables

Whole grain bread, rolls, rice, pasta, cereal other bread, rolls, rice, pasta, cereal other frozen/ready to cook grains Low fat yogurt Whole and 2% yogurt Whole and 2% cheese
Low fat milk Whole and 2% milk
other flour and mixes
Low fat cheese

TABLE III: 4 clusters of Grains and Dairies

In the grains and dairy category, milks are in one cluster; grains and other dairies are in one cluster; the last two clusters each containing one item are considered "anomalies" here.

In this last category, all the beverages except caloric drinks, coffee and tea are in one cluster; baked good mixes is in one cluster; the rest are in one cluster.

Though the classifications are not perfect, we can see a basic pattern that follows our intuition: food groups of similar nature or of similar production technologies follow similar pricing patterns. This reinforcements our understanding on food market as a competitive market.

C. Clustering locations

The dataset conveniently provides us the food prices at 26 metropolitan areas and 9 non-metropolitan areas. Each location comes with the label "region" which is the divides all location into East, South, Central, and West. Let's index them into 1,2,3,4 respectively for a brief representation. In this experiment, I will just use the unnormalized prices at different times of different food groups as features and the 35 location as the samples. Naturally, we want to divide the locations into four clusters and check whether each cluster maps into one region. Unfortunately, we did not get the result. However, when we set the number of clusters to be

Water Nonalcoholic nondiet carbonated beverages Diet nonalcoholic carbonated beverages
Baked good mixes
Oils Solid fats Raw sugars Non-carbonated caloric beverages Ice cream and frozen desserts Packaged sweets/baked goods Bakery items, ready to eat Frozen entrees and sides Canned soups, sauces, prepared foods Packaged snacks Ready to cook meals and sides Ready to eat deli items (hot and cold) Unsweetened coffee and tea

TABLE IV: 3 clusters of Fats, Beverages, and Prepared Foods

just two and use the first type of feature which is the unnormalized prices, something interesting came up:

Cluster 1	Cluster 2
1,1,1,3	2,3,3,3,2
4,4,4,3	3,3,1,2,3
1,1,4,4,	3,3,3,4,2
1,4,4,3	1,2,2,3

TABLE V: Clustering locations into two groups

We can see that one cluster is mostly the East and the West, the other is mostly the South and the Central. This meets our expectation though it is possible that the East coast and the West coast have higher food prices, and that the South and the Central has lower food prices because of lowering living cost.

IV. Supervised Learning

Though the dataset is simple and limited, we can still do some predictions on the food prices. Since the prices come with both the location and the food group labels, two types of supervised learning are implemented: first, I use linear regression to predict the prices of each food using the prices of other food given the time and the location, which I call "horizontal prediction"; second, I use time series analysis to predict the prices of each food in the future using their prices in the past at each location, which I call "vertical prediction". For both predictions, I split the data chronologically – the first 5 years of prices as training set, the last 2 years of prices as test set.

A. *Horizontal Prediction*

Three variations of linear regression are implemented:

1) Benchmark linear regression ("LR"): for each category, I fit a linear regression model to predict the price of each food group using prices of all other food groups in that category. For example, if there are n food groups in one category, then I generate n linear regression models. Each model uses a size of $(n-1) \times m$ training data where m is the product of the number of locations and the number of quarters (20) in the training set. Then, I want to compare the performance of linear regression on each food group with the following two methods.
2) Linear regression with the same cluster ("same cluster"): for each category, only prices of the food groups that are in the same cluster of the target food group are used as explanatory variables. The clusters are the results from the above k-means algorithm. For example, if the size of a cluster is n', then each model uses a size of $(n'-1) \times m$ training data, where m is similarly defined with above. This method should work as "feature selection" though we are not explicitly using methods like forward search. If the food groups that are in the same clusters are more correlated, then it is possible to get good predictions by only using features that are in the same clusters.
3) Linear regression with ridge regularization ("ridge"): I add a regularization term to the first method (using all food groups in the same category for prediction), where the objective becomes to minimize the loss function:

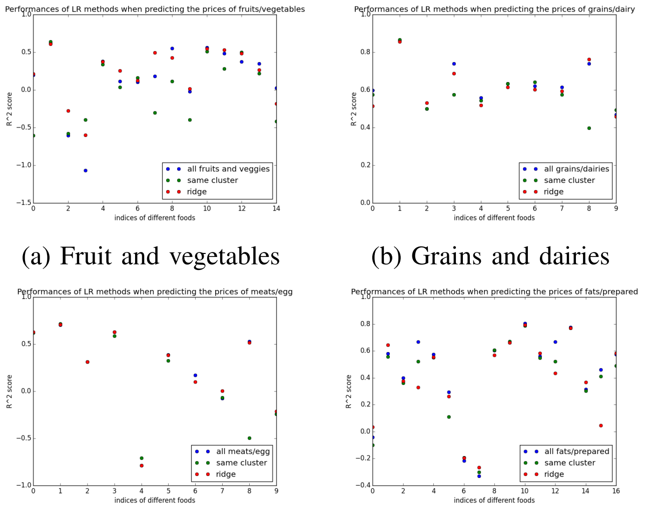
$$\min_{w} ||Xw - y||_2^2 + \beta ||x||_2^2.$$

I measure the performance of the predictions by using the coefficient of the residual term R^2, defined by:

$$s = 1 - \frac{||y - \hat{y}||_2^2}{||y - \bar{y}||_2^2}.$$

y is the true value; $\hat{y}$ is the predicted value; $\bar{y}$ is the mean of the true values. Thus, a perfect prediction means that $s = 1$. It is possible for scores to be below 0 and it means that assigning the average value to the prediction is better than using the linear regression fitted values.

The following graph displays the scores of each method when predicting the prices of food groups in four categories.



(a) Fruit and vegetables

(b) Grains and dairies

(c) Meats and egg

(d) Fats & prepared

Fig. 1: Scores of the linear regression methods

We can see that it is not obvious which method works the best. So, I plotted the following histogram by aggregating the scores of 3 prediction methods of all food groups.

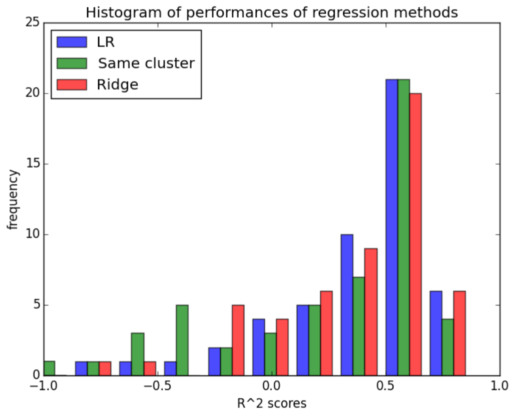
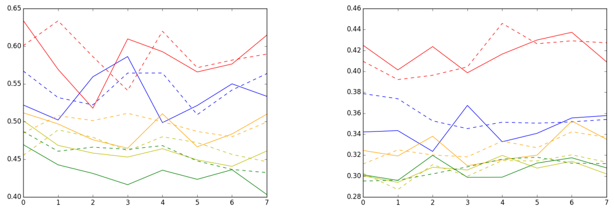


Fig. 2: Distribution of the scores of all regressions

We can see that three methods perform similarly even though the same cluster method and ridge regularization do worse in some instances. One way to explain that the same cluster method does not always do as well is that using the food groups that is not in the same cluster can be negatively correlated with the target food group. When food groups are not correlated with each other, then we cannot use one to predict the other; when food prices are negatively correlated with each other, then we can use this piece of information for prediction. One of the future extension is to look at which food groups are positively correlated and which are negatively correlated, and investigate whether that corresponds to "complement" or "substitute" goods.

B. *Vertical Prediction*

I use time series analysis to predict the prices of each food groups at each location over time. Five years of prices which are 20 data points are used for training. I choose to use autoregressive integrated moving average (ARIMA) method to use the past prices to predict the future prices. After trying out a few combinations of parameters, I chose $(p, d, q) = (3, 1, 0)$ for the following predictions which gave relatively low errors across all food groups and locations. The following to graphs are examples of prediction of two food groups at five different locations. The two years of test data are displayed where the solid lines are the true prices and the dashed likes are the predicted prices. Each color represent a different location.



(a) Predicting the Price of Fresh/Frozen Fruit

(b) Predicting the Price of Canned Fruit

Fig. 3: Predicting the prices at five locations
(Solid lines: true prices; Dashed lines: predicted prices).

Unfortunately, the time series prediction is not very successful due to my limited understanding of ARIMA, and possibly the limited amount of training data.

V. CONCLUSIONS & FUTURE WORK

I use k-means clustering algorithm to do unsupervised learning to find which regions and which food groups resemble similarities in their food pricing patterns. In the future, I want to analyze the relationships between clusters and correlations. More specifically, it will be interesting to derive on results whether the distance between the clusters are related to the correlations within the data.

I use both linear regression and time series analysis to predict the prices of food groups. In the future, it is best to combine the horizontal and the vertical predictions to achieve better results. Due to the limitation and the simplicity of the dataset itself, I should integrate data including the inflation rate, GDP and information about agricultural production with the food prices data, because they are part of mechanisms which determine the food market prices. In addition, having more price data over a longer range of time will also help with better prediction.

ACKNOWLEDGMENT

I thank the teaching team for their hard work: especially TA Mengwei Liu for her helpful suggestions and the two TAs who visited me during the poster session.

REFERENCES

[1] US Department of Agriculture, Quarterly Food-at-Home Price Database, 2012 (accessed November 7, 2017), https://www.ers.usda.gov/data-products/quarterly-food-at-home-price-database/.