Insights from the US food prices from 2004 to 2010

Wanyi Li

Stanford University Management Science and Engineering Department

Motivation

Food is one of the most basic commodities in everyone's life. Food prices are of great interest to economists because the market for food is very close to a perfectly competitive market. How does the nature of these commodities affect their prices? How does the market vary across geographical locations? What further insights can we gain about the food market through applying machine learning techniques on food price data?

Dataset

The United States Department of Agriculture published the "Quarterly Food-at-Home Price Database"[1] in 2012 which contains prices for 54 food groups at 35 different locations around the US from 2004 to 2010. The food groups are divided into 4 categories: "fruits and vegetables", "grains and dairy", "meats, nuts and egg" and "fats, beverages and prepared foods". Specifically, the dataset contains the prices of each food group, in dollars per 100 grams of food as purchased by consumers, given the year, the quarter and the location. Each location comes with the label for regions (East, Central, South, West).

Methods

Both unsupervised learning and supervised learning are used to analyze the dataset. I am mostly interested in applying the $k-$means clustering algorithm to see if the resulting clusters have any meaning, which would imply similar foods or similar locations share similar pricing patterns. For each sample (a single food group, or a single location), I tested the following three types of features:

1. Unnormalized price as given in the original dataset: p_t ;
2. Normalized prices: $p_t/\bar{p}$;
3. The change of the normalized prices: $(p_t - p_{t-1})/\bar{p}$.

For supervised learning: I slit the data chronologically: the first 4 years as the training set and the last 2 years as the test set. First, I use linear regression predict the prices of each food using prices of other foods at each time and location, so called "horizontal" prediction. The performances of three variations of linear regressions are compared, including feature selection and regularization. Second, I use time series analysis to predict the prices of each food in the future using its prices from the past at a given location, so called "vertical" prediction.

Clustering Food Groups†

Fruits and Vegetables

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

Meats and Egg

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

Grains and Dairies

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

Clustering Locations‡

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

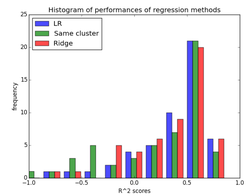
Table: Clustering locations into two groups

• 1: East; • 2: Central; • 3: South; • 4: West.

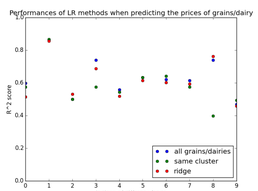
† The third type of feature is used for this clustering results;
‡ The first type of feature is used for this clustering results;
* I choose to use the coefficient of the residual R^2 as the score to measure the performance of linear regression, which is defined as the following (optimal score is 1):

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

Horizontal Prediction



(a) Distribution of the scores of all regressions



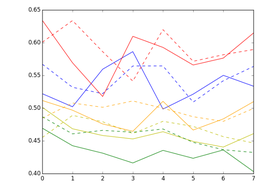
(b) Grains and dairies

Figure: Scores of the linear regression methods*

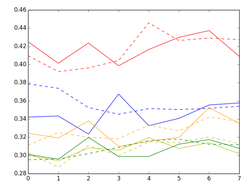
1. Linear regression where X is the prices of all other food groups in a category at each time and location;
2. Linear regression where X is only the prices of food groups in **the same cluster** resulting from $k-$means;
3. 1st method with ridge regularization: $\min_w ||Xw - y||_2^2 + \beta||x||_2^2$.

Vertical Prediction

The prices of each food at each location comes in the form of time series. We can use autoregressive integrated moving average (ARIMA) method to use the past prices to predict the future prices. The parameters chosen for the following predictions are: $(p, d, q) = (3, 1, 0)$



(a) Predicting the Price of Fresh/Frozen Fruit



(b) Predicting the Price of Canned Fruit

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

Future Work

Competitive equilibrium tells us that prices are determined by supply and demand functions. Without knowing them, to better predict food prices, we need to incorporate economic data like inflation rate, GDP and information on agricultural production which are possibly factors that correlate with the food prices.

References

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

Contact Info & Acknowledgments

- Email: wanyili@stanford.com
- Address: Huang Engineering Center 263G, 475 Via Ortega, Stanford, CA.

- I thank the teaching team for their hard work, especially TA Mengwei Liu for providing helpful feedback and suggestions.