1007-final project proposal

Scenario and Motivation: Predicting sales-related quantities is important for retail companies. Our work will focus on the sales data for Walmart. Accurate prediction of sales allows the comany to properly prepare for handling the shocks to product stock and customer support. One challenge of modelling retail data is the need to make decisions based on limited history. Besides, markdowns during special holidays - the Super Bowl, Labor Day, Thanksgiving and Christmas are hard to model in the absence of complete historical data.

Data: Our project is based on three datasets, as following:

stores.csv: contains anonymised information about the 45 stores, indicating the type and size of store.

train.csv: covers data from 02/05/2010 to 11/01/2012 with the following fields:

- Store - the store number
- Dept - the department number
- Date - the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - whether the week is a special holiday week

features.csv: contains additional data to the store, department and regional activity for the given dates with the following fields:

- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymised data related to promotional markdowns that Walmart is running. Only available after Nov 2011, and is not available for all stores all the time.
- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

Data exploration: Preliminary analysis on the dataset is necessary for us to get basic understanding about the data. For example, we may want to see if the data is stationary if we want to build time series models. We may also want to check the correlation between these features before we fit the data into a linear regression model. Checking on the distribution of our features is also useful for us to be aware of the data.

Modelling: This problem can be regarded as a regression task where the target is to forecast the sales of each department in each store. We can try to predict sales with or without using the given features. For instance, the classic ARIMA method and K nearest regression can be used for building time series model without features. In the meanwhile, linear regression and Random Forest can also be used not only to predict sales but also find out the drivers that have impact on sales. We will compare these models and select one with the best performance.

Model evaluation: We will use the loss function given by Kaggle: the weighted mean absolute error(WMAE):

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

where

- n is the number of rows
- $\hat{y}_i$ is the predicted sales
- y_i is the actual sales
- w_i are weights. w = 5 if the week is a holiday week, 1 otherwise

The dataset will be split into training set and test set. We want to use cross-validation on the training set to avoid over-fitting. We may not use the same loss function on both training set and test set but this is to be determined.