Final project

Description
The final project is to apply machine learning in one of the most challenging areas in finance: price prediction. In the course of the project, you will get familiar with (1) the procedure of machine learning prediction on time series; (2) extracting informative features and constructing useful labels; (3) selecting features and classifiers to optimize the performance and optionally (4) building an algorithmic trading strategy based on machine learning.
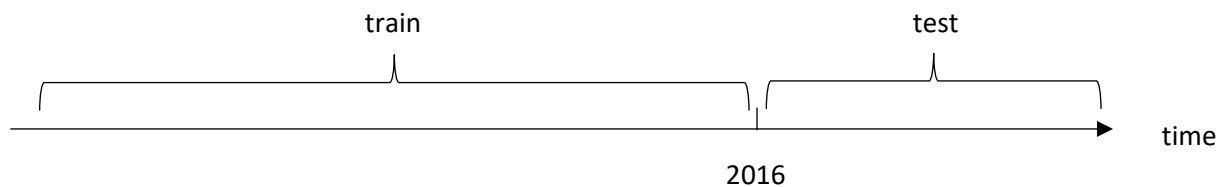
Data
We focus on the S&P stock universe and use daily price data. You can search for other types of data, if (1) they are publicly available and (2) there is full document about their description and quality.

Format
The final submission should include a jupyter notebook that contains code and description. The result needs to be reproducible. Additionally, you should also include data, reference and other supporting materials that are not provided initially.

Prediction
The simplest way to apply out-of-sample prediction is the following. Suppose you predict on a single stock using its price data, and you split the data set at 2016, the train/test sets are

A better way (optional) is to re-train the model after a certain period, such that the training data always contain the most seasonal information. Graphically, suppose we still start in 2016, this looks the following:
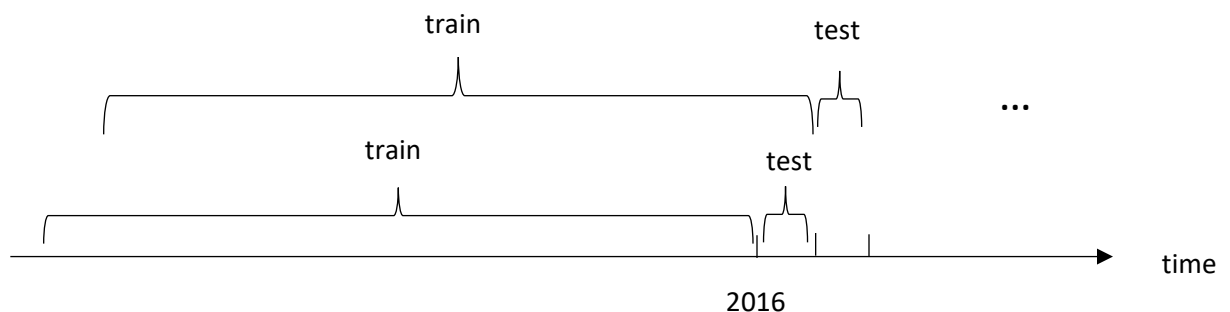
Features

Generating features is the main task in stock prediction. The most basic ones are past returns

$$r_t = \frac{p_t - p_{t-n}}{p_{t-n}}$$

You need to construct features that you think will be useful for predicting stock prices.

Label

A commonly used label for stock prediction is sign of return:

$$l_t = sign(p_{t+h} - p_t)$$

You can try other labels as well.

Evaluation

At least 2500 out-of-sample predictions are required. This can be done by either doing 10 year prediction on S&P index (assuming that you are trading an ETF that tracks S&P), or doing prediction on multiple stocks (e.g. half year prediction 20 individual stocks). Though out-of-sample accuracy and other prediction metrics are an important part of the evaluation, you will also get scores for

- Finding useful features and labels. For both, you can use the materials that are on the course website, search the literature for interesting papers or come up with ideas on your own.
- Applying appropriate parameter optimization. You should report all tuned parameters and how the performance changes.
- Selection of algorithms. Using multiple classifiers is a plus.
- Carefully analyzing features and finding financial and economical reasons for choosing or neglecting them.
- Building a strategy is a plus.