Homework 2 (10 pts)

Topics: VC-theory and its applications

Problem 1 (4 pts):
Perform model selection comparisons for univariate regression using trigonometric polynomials

$$f_m(x,\ w, v, b) = b + \sum_{i=1}^{m} w_i \sin(ix) + v_i \cos(ix)$$

Follow the experimental procedure as described in Section 4.4, and use the same regression data sets i.e., pure noise and sine-squared. Show your comparison results in graphical form, similar to Figs. 4.10-4.11 (The section and figures are in direct reference to the course material, *Predictive Learning' by V. Cherkassky (2013)*).

Problem 2 (6 pts): Trading international mutual funds
Background: this HW applies predictive data-analytic methods to frequent trading of international mutual funds. This practice called 'timing of mutual funds' attempts to profit from daily price fluctuations, under the assumption that the next-day price changes may be statistically 'predictable' from today's market data. For background reading on the difficulty of market timing, see http://en.wikipedia.org/wiki/Market_timing . In early 2000's, this practice of 'timing of mutual funds' by hedge funds and speculators, has resulted in financial scandals in the mutual fund industry. See http://en.wikipedia.org/wiki/Mutual-fund_scandal_(2003)
and http://financialservices.house.gov/media/pdf/110603ez.pdf
In this HW, you will investigate the effectiveness of such an approach via simple trading strategies estimated using classification and regression methods.
Problem statement: An objective is to design effective (i.e., profitable and safe) trading strategy for the international mutual fund, Fidelity International Discovery (symbol FIGRX). All US mutual funds accept buy/sell orders once a day (before US stock market close at 3pm), so your strategy needs to generate a BUY or SELL signal at the end of each trading day, based on today's input indicators. Effectively, by placing the BUY (or SELL) order *today*, you are betting that the price of this mutual price will go UP (or DOWN) *tomorrow*. Two input indicators (for making trading decisions) are the daily closing prices of the following indices:

- SP 500 stock index (symbol ^GSPC on http://finance.yahoo.com);
- Euro-to-dollar exchange rate (symbol EURUSD=X).

These two inputs may be good predictors international stocks, because (a) foreign markets closely follow US market; (b) US mutual funds price foreign securities in US dollars.

Trading strategy and data encoding: your total account (say $1 million) should stay 100% invested either in FIGRX or in cash. You can switch between cash position and FIGRX at the end of each trading day, with no transaction costs. The inputs for making trading decisions represent the % daily change of each input variable, i.e.

% daily change = 100% * [GSPC_close(t) – GSPC_close(t-1)] / GSPC_close(t-1)

where GSPC_close (t) is today's market closing price, and GSPC_close (t-1) is yesterday's (previous business day) closing price. The output (response) value, FIGRX, should be similarly encoded as % daily change. Note that in your model the inputs correspond to **today's** daily changes, whereas the output is **tomorrow's** daily change.

Historical data of daily closing prices of ^GSPC and FIGRX can be obtained from Yahoo! Finance web site: http://finance.yahoo.com**.** Daily closing prices of euro-to-dollar exchange rate can be obtained at the University of British Columbia web site http://fx.sauder.ubc.ca/data.html

Experimental Set-up
General approach is to estimate 'good' trading strategy during the training period, and then apply this strategy and evaluate its performance during the test period. Here a trading strategy is a predictive model estimated classification or regression modeling (as explained below). The *training period* is specified as year 2004; and the *test period* is year 2005.

Learning Methods
You need to implement 2 learning approaches for classification, as outlined next. All approaches estimate a model from the training data (x1, x2, y) which represents % daily changes of input and output (y) variables.
(a) Linear classifier estimated via linear least-squares regression. Under this approach, the training data (x1,x2, y), with real-valued y-values, is used to estimate *linear regression* model:

$$t(\mathbf{w}) = w_1 x_1 + w_2 x_2 + w_0$$

This model is estimated by minimizing the mean squared error (on the training data):

$$R(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}\cdot\mathbf{x}_i + w_0 - y_i\right)^2$$

Estimated regression model $t(x_1, x_2) = w_0^* + w_1^* x_1 + w_2^* x_2$ is used to predict the next-day price changes (UP or DOWN) according to decision rule $Sign(w_0^* + w_1^* x_1 + w_2^* x_2)$.

(b) Quadratic decision boundary classifier. Under this approach, real-valued response values of the training data are initially transformed into class labels, i.e. y = Sign (y), so that class label -1 encodes DOWN, and class label +1 encodes UP. So the training data can be used for standard binary classification formulation. This method estimates a quadratic (nonlinear) decision boundary in the input space (x1, x2): $g(\mathbf{x}, \mathbf{w}) = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + w_0$.

Suggested Plan of work and presentation of results:

(1) Download the data from Yahoo and UBC web sites, encode both inputs and an output as daily % changes. (pre-processed data (x1, x2, y) is available for your convenience on the course web site)

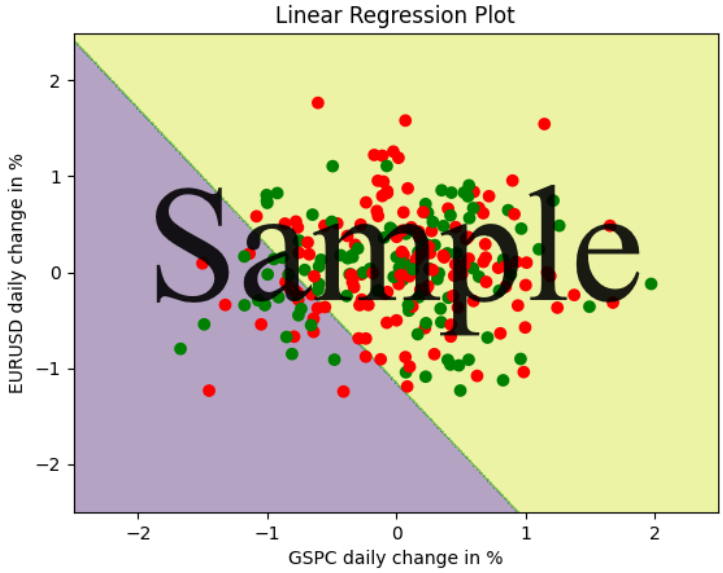
(2) Apply two learning methods to year 2004 training data, and obtain 2 models (decision boundaries). Display each of these models along with actual training data on a 2D plot (see Example attached below). For consistency, on each plot show input variables in (-2%, +2%) range. Also, show the plots of each model overlapped on the test data. (total 4 plots).

(3) Test performance of your predictive model using independent test data (year 2005). Please compare the performance of the two classifiers (linear and quadratic) and *Buy-and-Hold* (B-H) using two indices, *GAIN* and *EXPOSURE*.
Note that the *GAIN* is the cumulative gain (or loss) in percentage at the end of the year. The *EXPOSURE* is the proportion of days when the account is fully invested in **FIGRX**.

(4) Discussion and interpretation of results in (2) and (3). In particular, discuss the following
(a) can this trading strategy be used in practice? Would you bet your own money using this strategy?

(b) Explain your results using VC-theoretical arguments. Why the learning methods in this HW *do not* use model complexity control?

Appendix: example plot showing the training data and estimated (linear) decision boundary



Appendix: example plot showing the effectiveness of the trading strategy using linear classifier