Homework 1 (10 pts)

Topics: *risk minimization + model selection*
All problems are taken from the textbook *Predictive Learning*, Cherkassky 2013.

Warning: all homework solutions should represent individual student's effort. Collective (group) solutions will be regarded as **cheating**.
Presentation of HW solution results: should be well-presented (in a readable form). All solutions should be typed (not hand-written) and all Tables/ Figs presented using consistent notation (e.g., terminology introduced in this course).

1. Problem 2.7 (2.5 pts**)**

2.7 Consider predicting the results of US presidential election using the obesity index (for each state) as a single input variable. The obesity index is the percentage of adult population that has BMI index greater than 30. You can find this obesity index for each US state from the public domain sources. This problem can be formalized as a binary classification problem, because there are only two voting outcomes (Republican or Democrat). That is, each data sample (x,y) represents a state, where single input x is the obesity index (for that state) and y is a binary label denoting the presidential election voting result. Historical data for this problem can be found from public-domain sources. For example, see US obesity rates and election results at:
http://en.wikipedia.org/wiki/Obesity_in_the_United_States
http://en.wikipedia.org/wiki/United_States_presidential_election,_2004
The results of 2004 presidential election are used as the training data for estimating a binary classifier. The results of 2000 elections are used as test data. Estimate a k-nearest neighbor classifier to predict the election results (for each state) using the obesity index. An optimal k-value is selected using leave-one-out cross-validation on the training data. Then use this k-nearest neighbor model to predict the results of 2000 elections. Clearly present all modeling results, including resampling error, optimal k-value and test error. Also, *briefly* discuss your results. Your discussion may include: the quality of your predictive model, comparison of the resampling error and test error, and critique of modeling assumptions.

The two obesity datasets (for 2000 and 2004) are available on Moodle.

2. Problem 2.8 (2.5 pts**)**

2.8 Repeat Problem 2.7 using election results for year 2000 as training data, and election results for year 2004 as test data.

3. Problem 2.11 (2.5 pts**)**

2.11 For the data set used in Example 2.6, estimate an optimal regression model using:

(a) Trigonometric polynomial estimators, i.e.

$$f_m(x, \mathbf{w}) = \sum_{i=1}^{m} w_i \cos(2\pi i x) + w_0$$ where $m+1$ is the model complexity parameter. Use analytic Schwartz criterion for model selection.

(b) Algebraic polynomial estimators. Use analytic Schwartz criterion for model selection.

Is it possible to choose the best predictive model, trigonometric vs. algebraic polynomial, using *only* the results of model selection for each method?

About the data used in Example 2.6: The data consists of n = 10 samples, (x, y), where x is uniformly distributed in [0,1] and y = x^2 +0.1x+noise and the noise has Gaussian distribution N(0, 0.25). Note that the noise has variance 0.25 or standard deviation 0.5.

4. Problem 2.12 (2.5 pts**)**

2.12 Repeat application of trigonometric and algebraic polynomial estimators to the data set in Problem 2.11. However, now the goal is to compare the prediction accuracies of these methods using 5-fold cross validation. Use Schwartz criterion for model selection.