SOLUTION for Homework 1

Problem 2.7

The optimal choice for k via leave-one-out cross-validation (using training data, for k = 3, 5, 7, …, 19) is $k=9$, (note that $k=9$ and $k=7$ generate the same LOO error rate, $k=9$ is selected because the the simpler model is expected to generate a better prediction according to Occam's razor – see Table 1) and for 9-NN classifier LOO error rate is 27.45%.

Table 1. The LOO errors of using different k

K	LOO error
3	0.3137
5	0.2941
7	0.2745
9	0.2745
11	0.3725
13	0.3529
15	0.2941
17	0.2941
19	0.3333

The k-NN classifier (with k=9) is tested using 2000 elections data. The test error is 49.02%.

Discussion:

The test error is much higher than the LOO error. This may be due to:

(a) The obesity index changed significantly during 2000~2004. See the trends in Figure 1.

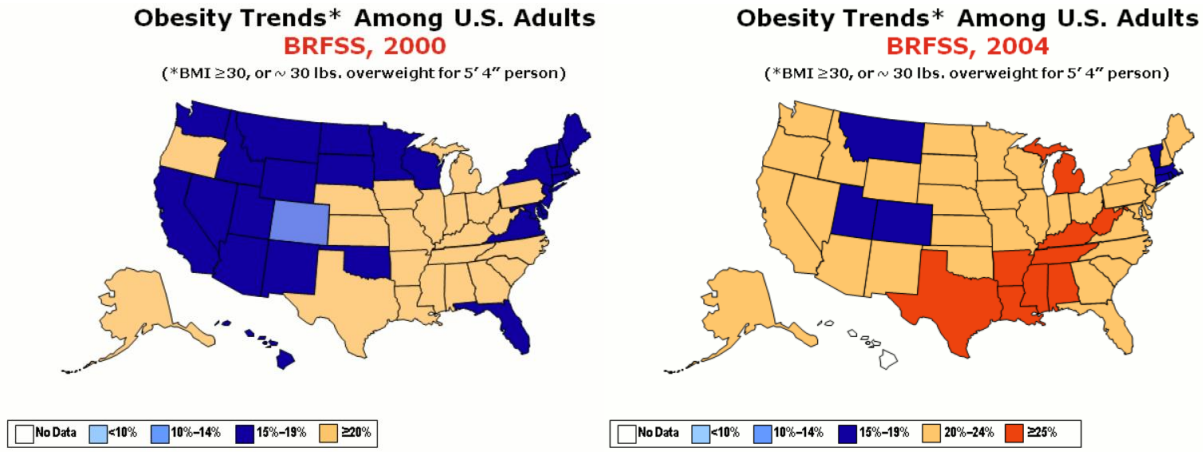

Figure 1. The obesity trends among U.S. adults in 2000 (left) and 2004 (right)

Hence, the model trained using 2004 election data cannot generate good predictions for the election in 2000.

(b) Clearly, the model can be possibly improved by using additional input variables for each state. For example, we may add variables reflecting average educational level, such as % of high school graduates and/or college graduates in each state.

Problem 2.8

Using the year 2000 election results as the training data, the optimal choice for *k* via leave-one-out cross-validation is *k=17*, and for 17-NN classifier LOO error rate is 39.22% – see Table 2 . The corresponding test error rate is 35.29%.

Table 2. The LOO errors of using different k

K	LOO error
3	0.4118
5	0.4118
7	0.4118
9	0.4118
11	0.4510
13	0.3922
15	0.3922
17	0.3922
19	0.4706

Problem 2.11

This problem illustrates the use of analytic methods for model complexity control. The available 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.

Two learning methods are used to estimate the unknown univariate function and to decide which learning method yields better prediction accuracy. For both learning methods, we try m = 0,…,8 in this problem. The model complexity (or the number of free parameters) of the polynomial estimation is m+1. We don't have to try any value greater than 8 for two reasons. First, we have 10 training samples which can be used to construct 10 equations and solve 10 unknowns. For m greater than 9, we would have more unknowns than equations. Second, if m=9, the penalized factor based on the Schwartz Criterion would be infinity and cannot be used to compute the penalized risk.

The modeling results using trigonometric polynomial regression are shown in Table 3. With m=2 we have the optimal model and the corresponding graphical representation with the training data and the true function is shown in Figure 2.

Table 3. Model selection for the trigonometric polynomial regression. The optimal model is m = 2 with complexity (degree of freedom) 3.

m	DOF (m+1)	Remp	Penalized factor (r)	Rpen = r*Remp
0	1	0.1779	1.2558	0.2234
1	2	0.1702	1.5756	0.2683
2	3	0.1011	1.9868	0.2008
3	4	0.0864	2.5351	0.2191
4	5	0.0743	3.3026	0.2453
5	6	0.0740	4.4539	0.3297
6	7	0.0738	6.3727	0.4702
7	8	0.0381	10.2103	0.3895
8	9	0.0117	21.7233	0.2549

9	10	0.0000	Inf	Inf

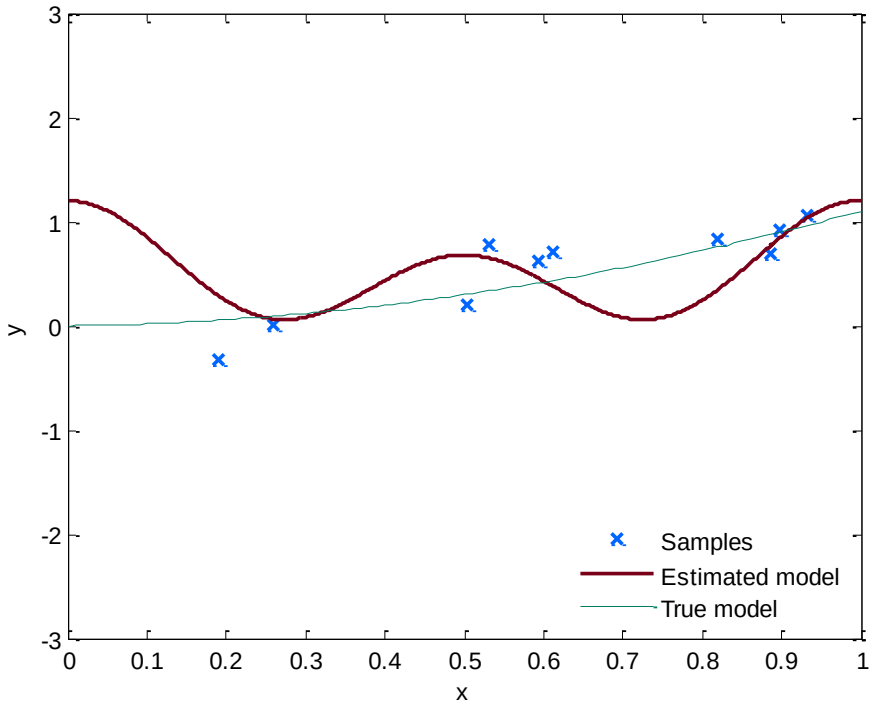

Figure 2. Trigonometric polynomial model (m = 2) estimated from 10 training samples. True target function is shown in green.

The modeling results using algebraic polynomial regression are shown in Table 4. Although Rpen (estimated prediction risk) has the minimum value when m=8 (marked in blue in Table 4), apparently we are overfitting the training data. Therefore, instead of selecting the model according to the global minima of the prediction risks, we should choose a simpler model (smaller m) which yields local minimum prediction risk. In this problem, we select m=2 (marked in red in Table 4). See Fig. 3 showing the empirical and (estimated) prediction risk as a function of model complexity. Figure 4 shows the training data, estimated model and the true function. *Discussion:* Note that when m=8 the fitting error (empirical risk) is very close to zero, as the model with 9 parameters is used to fit/explain perfectly 10 data points. This may suggest that Schwartz criterion does not provide sufficiently high penalization for large values of DoF (m). In contrast, VC penalty factor is larger than Schwarz penalty. That is, condition DoF < 0.5*n (where n ~sample size) should hold for *any estimator* (see p. 140 in the text book).

Table 4. Model selection for the algebraic polynomial regression. The optimal model is m = 2 with complexity (degree of freedom) 3.

m	DOF (m+1)	Remp	Penalized factor (r)	Rpen = r*Remp
0	1	0.1779	1.2558	0.2234
1	2	0.0329	1.5756	0.0518
2	3	0.0224	1.9868	0.0446
3	4	0.0214	2.5351	0.0543
4	5	0.0209	3.3026	0.0689
5	6	0.0154	4.4539	0.0684
6	7	0.0154	6.3727	0.0979
7	8	0.0076	10.2103	0.0775

8	9	0.0006	21.7233	0.0129
9	10	0.0000	Inf	Inf

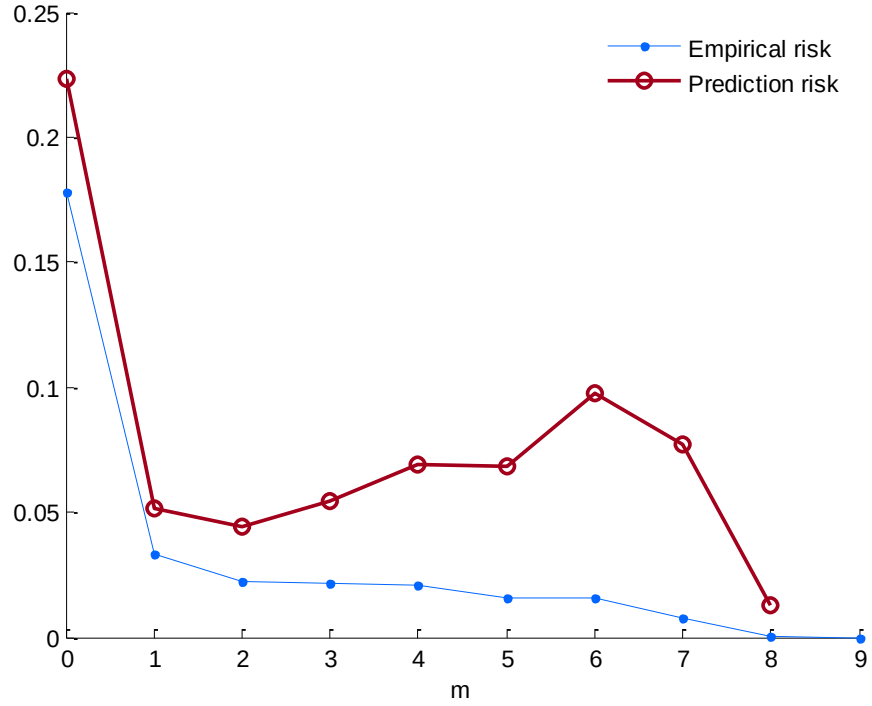

Figure 3. Empirical and (estimated) prediction risk as a function of m (DoF-1).

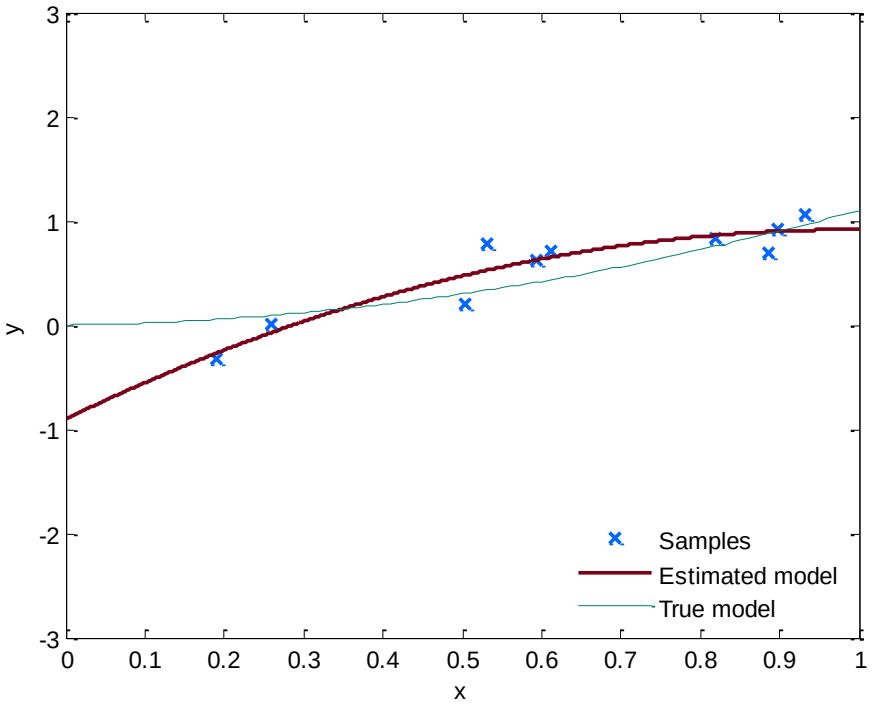

Figure 4. Algebraic polynomial model (m = 2) estimated from 10 training samples. True target function is shown in green.

Problem 2.12

The prediction errors (Normalized Root Mean Squared Error, NRMS) from the two modeling methods are summarized in Table 5.

Discussion: Results in Table 5 indicate

(a) High variability of optimal model complexity selected in different folds (due to variability of small training sample used to estimate a model in each fold). This variability also results in high variability of the test error (NRMS).

(b) Poorly chosen model complexity (large m-values) has relatively minor impact on the quality of trigonometric polynomial models but quite high impact on the quality of algebraic polynomial models. This is due to the bounded nature of trigonometric models, where the basis functions are in the [-1, +1] range. In contrast, algebraic polynomials (of high degree) may exhibit very large values near the boundary of the input (x) domain.

Table 5. Prediction errors (NRMS) of trigonometric and algebraic polynomial models in different folds.

Fold	Trigonometric		Algebraic	
	Optimal m	NRMS	Optimal m	NRMS
1	2	0.8332	1	0.3532
2	4	1.4694	5	10.0803
3	2	1.5177	3	1.0646
4	0	3.8584	5	3.4262
5	0	1.3467	2	0.5682
Average		1.8051		3.0985

Summary: As evident from Table 5, regression estimation using trigonometric polynomials yields better prediction accuracy (lower NRMS) than modeling using algebraic polynomials. However, results in Problem 2.11 indicate that (optimal) trigonometric model has *larger* penalized training error than (optimal) polynomial model. The explanation of this inconsistency is that penalized training errors (reported in Problem 2.11) have been used to select optimal model complexity (for a given pre-specified class of functions). So these penalized training errors cannot be used to measure the true test error (of a learning method), as required in Problem 2.12.