Some of the most promising and still unfinished 2002 Projects

1. Hypothesis Testing with Small Sets (tp and Dradulov)

2. Feature Selection for SVMs: Theory and Experiments (Sayan)

3. Reusing the Test Set: Dataminining Bounds (Sayan)

4. Large-Scale Nonlinear Least Square Regularization (Rif)

5. Local vs Global classifiers: experiments and theory (conjecture: unlike Vapnik's statement are local and

global subsumed under the same formulation?). Alternatively, critically review Vapnik and Bottou two papers on local algorithms (Alex)

6. RKHS invariance to measure: historical math(tp)

7. Kernel synthesis and selection (and theoretical foundations for kernel alignment): write paper for a conference (tp+sayan)

8. Bayesian Interpretation of regularization and in particular of SVMs: finish proofs of limits theorem and integral approximations; connect to Gaussian Processes literature (tp)

9. Phylosophy project: history of induction from Kant to Popper and current state (Alex)

10. “Religious” project: Bayesian Priorhood (Sayan)

A few new 2003 projects

1. Review techniques to transform a variable length input vector into a fixed length one. What is an acceptable set of measurements? Consider in particular time series.

2. Sparsity of representation and learning: what is the connection? Is sparsity – in the sense of sparsity of an overcomplete dictionary – "good" for learning?

3. Study Girosi's (Girosi, F. An Equivalence between Sparse Approximation and Support Vector Machines, Neural Computation, Vol. 10, 1455-1480, 1998; see link on Publication page in CBCL Web site) result

about "equivalence" of BPD and SVM for $K(x, x_i) = xx_i$. What does it say? can it be generalized?

4. (suggested by steve smale) Approximate indicator functions with kernels from a RKHS with very little smoothness. Calculate approx and sample error using bounds such as Cucker Smale etc.. Verify with computer simulations.

5. (also suggested by steve smale) Do careful proof – mimicking theorem 4 in CS p. 37 – that the RKHS defined for unbounded domains through the Mercer-like Fourier representation (Girosi) is the same as the RKHS define through the r.k. without Fourier.

6. (suggested by M. Bertero) Use L_2 compactness of monotonic functions for regularizing density estimation ?

7. Summarize critically results of Ding and Micchelli (and us) on density of RKHS in L_2 and C

8. Critically review Nature Neuroscience paper (June 2002) by Weiss, Simoncelli and Adelson: it is just regularization! using bayesian view of regularization.

9. Review recent approaches to prediction of time series (advice: avoid financial time series). Review approaches based on combination of classifiers for time

series prediction – such as mixture of Gaussians (see Gerschefeld Nature paper, January 28, 1999)

10. Do a review on non i.i.d. Brownian bridge processes

11. Multiple hypothesis testing: review critically a paper by Benjamini and Hochberg that controls false discovery rates

12. Relate statistical complexity concepts to computational complexity concepts. The underlying thesis is that more complex solutions, such as solutions with small margin (eg solutions with large RKHS norm) correspond to solutions that require more time to compute (eg because the condition number of the linear system of equations in the case of square loss is poorer)

New April 2003 projects

1. "Challenge project": set of data of which 70 percent are noise and 30 percent contain signal enough for binary classification. Challenge is to find a good learning algorithm in this situation. There is a real blind test set hidden until partecipants are ready for the test. (Alex, tp)

2. Bayes noise model for hinge loss function (for binary classification). The idea is to look at the noise model of logistic regression (see for instance Applied Logistic Regression by Hosmer and Lemeshow) of which the SVM loss function is an approximation. The probability distribution underlying logistic regression is a binomial instead of a Gaussian as in quadratic regression. The challenge is to obtain for the SVMC loss function a similar derivation as we have for the SVMR loss function (Gaussian is to SVMR noise model as Binomial is to the to-be-found SVMC noise model). (tp)

3. Intron recognition. Building a SVM classifier for sequence motifs to recognize introns in human genes. Building a Bayes Net classifier and comparing the performances of SVM and Bayesian Network. Data will be provided. (Gene)

4. Exploring Marginalized Kernels for RNA sequence data analysis. Can we derive sensible SVM formulations that are similar to SCFG to model RNA secondary structure ? Data will be provided. (Gene)