**Dissertation Topic Notes**

Thesis will be on a “Machine Learning” topic on how to use *data and prior knowledge* to build a regression and/or classification function that is optimal for a given cost function.

Directions:

* Parametric or Nonparametric
* Linear or Non-linear (in parameters)
* “Classical” or Bayesian
  + Probabilistic treatment of model (function, parameters)
  + Probabilistic treatment of input data???
* Expected or Empirical cost function
* Batch or Sequential (“Online”) learning
* Deterministic or Stochastic (e.g. particle filter) learning function
* Regression and/or Classification applications
* Supervised and/or Unsupervised

Concepts:

* Uniform prior is implied by Classical estimation. Bayesian framework always applicable?
  + ML equates to MAP with a uniform prior
* Use of hierarchical priors
* MAP cost function (always?) equates with a regularized empirical cost function
* Use of Empirical cost equates to the Expected cost when an estimated joint PDF is used that simply places a Dirac function at the training data
* The unrestricted (or “unweighted”) ML estimate of a joint PDF simply has Dirac functions at the training data locations
  + Prior knowledge may suggest that such an estimate is “overfit” – Empirical costs may be inappropriate
  + ML estimates to ill-posed PDF estimation problem may be poor without restricting search to a low-dimensional manifold (e.g. parametric estimation)
* Sparse representations: selection of prior PDF or search region is tied to selection of dictionary functions (linear expansions) – together they define a union of linear subspaces in which the data is assumed to reside
  + Are sparse priors appropriate for learning functions that are non-linear in the parameters? Can sparsity be used to learn low-dimensional non-linear manifolds?
  + Are there better ways to select dictionary functions (or kernels for RKHS methods) beyond ad-hoc methods like cross-validation?
* If true prior is unknown, what choice of prior is “best” for function selection?
  + Does the uniform (non-informative) prior result in the min-max error?
  + Does maximum entropy minimize max error?
* Are subjective priors appropriate for the completely general learning problem? “No Free Lunch” Theorem suggests that assuming a prior can only improve performance for the given problem at the expense of performance on others
* Classes are extrinsic to data
* How do unsupervised learning methods (esp. in humans) restrict the resultant data-class joint distributions? What role does non-linear dimensionality reduction play?
  + Selection of classes (and the joint PDF) in humans seems to allow very low classification error – is there little class conditional PDF overlap (cross-entropy?)
  + Use and understanding of this effect should inform which functions parametric learning algorithms favor (NFLT motivates such preferential treatment)
* Asymptotic performance: Estimated joint PDF of input-output data converges to the true PDF – minimum Bayesian Risk estimates are achievable
  + Priors can still be used? Estimates should naturally weight data over prior as size of training set increases (e.g. Kay, Bayesian chapter examples)
* Nonparametric learning in RKHS uses *data-dependent* functions and learns a finite number of coefficients (superiority of SVM/SVR vs RBF)
  + Do these methods thus achieve functions in a higher-dimensional space than the usual generalized linear models? Are there connections with density estimation (especially for localized kernels?)
* Modelling classification problem using MSE leads to real vector output (infinite set) whose optimal value is the posterior class PMF
  + Does estimation of full class posterior have any additional value over my formulation?
* Humans classify data using high-dimensional, redundant hypothesis space (fruit/vegetable AND red/orange). IMAGE-NET!!!
  + Classifiers seem distinct and don’t seem to be applied jointly. Rather, a hierarchical approach is used, with vertical and horizontal dimension.
    - Ex: Red/Orange is not a subclass of Fruit/Vegetable, but Apple/Banana is

Questions:

* Why do regularized costs (deterministic) and prior distributions (stochastic) tend to favor parameters with small norms??

Tasks:

* Use Gaussian distributions for linear regression MSE application. Investigate asymptotic results as variance increases/decreases and compare to non-informative model priors, deterministic input assumption, etc.
* Attempt to refine results for classification application with hit/miss cost
  + Generate a simplifying case (like linear regression for MSE) to ease investigation
* Research the mathematics necessary to generalize existing results to a non-parametric treatment of the generating in/out PDF model
  + Random processes? Does increasing model parameter dimensionality provide an asymptotic result?
* Create a framework for mismatch between model prior and true generating model (different PDF or a single value?)
  + Optimize selected prior with respect to min-max cost? Entropy?
* Investigate bias/variance tradeoff
* Investigate No-Free-Lunch theorem
* Attempt classification application using GMM?