Introducetion to Ensembles of Experts and Hybrid Methods

Nathan Intrator

Tel-Aviv University and Brown University

`www.physics.brown.edu/users/faculty/intrator`

Collaboration:

Yair Shimshoni, Yuval Raviv, Inna Stainvas, Shimon Cohen

Outline

- Basic definitions
- Averaging formulation
- Combining models of different nature
- Assessing the goodness of an expert
- Predicting large ensemble performance from a small ensemble
- Regularization revisited

General Setup

- **Problem:** Small training set, large number of dependent variables

- **Best Solution:** Detailed modeling of the data with very few free parameters to estimate

- **Second best:** Use a more flexible model, estimate many parameters

General Setup

Trade off between:

- number of free parameters
- data complexity
- reliability of the estimation

What are Ensembles and Hybrid architectures

What are Ensembles and Hybrid architectures

Definition:

Combining different models where each is capable of modeling the observations separately

Reasons for Ensembles and Hybrid Methods

Reasons for Ensembles and Hybrid Methods

- Uncertainty about the desired model

Reasons for Ensembles and Hybrid Methods

- Uncertainty about the desired model
  - Uncertainty about model parameters
  - Uncertainty about model capacity
  - Uncertainty about model complexity
  - Uncertainty about model type and architecture

Uncertainty about model parameters

- When the optimization solution is unique, uncertainty results from the choice of the training set

- When the solution is non-unique, additional uncertainty results from the initial choice of model parameters

Uncertainty about model parameters (continued)

Usually addressed by:

- Imposing a prior $\phi(W)$ on the distribution of parameters

- Integration over the distribution:

$$\int \phi(W)W(x)dx.$$

This may be problematic due to multiple local minima.

Uncertainty about model parameters (continued)

A better approach: Integrate (average) over the prediction M_W of all these models

$$\int \phi(W) M_W dW.$$

Leads to ensemble of experts as an approximation to a model posterior

Combining models of different nature

Sequential methods of a Hybrid flavor

- Additive and Generalized additive models (GAM)
- Projection pursuit regression
- Matching pursuit: Choose from a (nonorthogonal) collection of basis functions

Combining models of different nature

Reasons for combination

- Efficiency difference between models, training methodology
- Sequential modeling
- Divide and conquer
- Model interpretation

Actual combination

Similar to the combination of models with different parameter values:

- Construct (or empirically estimate) a posterior to the models $\phi(M_i(W))$, where i represents the different models

- Integrate over the prior

$$\int \phi(M_i(W)) M_i(W)$$

Pros and cons of combining models using a posterior distribution

Pros

- Appears to model the data better, fit the more appropriate models

- Removes naturally very unrelated models

- Smaller ensemble size works fine

Pros and cons of combining models using a posterior distribution

Cons

- Regularization is simpler
- Sensitivity to wrong models is reduced
- Training for optimal ensemble performance is simpler

Main caveat for ”smart” averaging: Construct useful Model Assessment

- A ”good” model assessment could be useful for model averaging.

- When two models have similar predictions should we give them same importance?

Main caveat for "smart" averaging: Construct useful Model Assessment

- Simply put, if a 40 hidden unit architecture performs as well as a 5 hidden unit architecture, which one should we prefer?

Main caveat for "smart" averaging: Construct useful Model Assessment

- Simply put, if a 40 hidden unit architecture performs as well as a 5 hidden unit architecture, which one should we prefer?

Information theory may surprise us here..

Model Assessment (Hinton & van Camp, 1993)

Basic idea

- The performance of an expert is a function of its error (residual) and a function of its complexity.

- The complexity of a model is a function of the number of parameters and the required accuracy for the parameters

Model Assessment (Hinton & van Camp, 1993)

- To use the same scale, we measure the code-length of the residual and of the model parameters
- The code-length of a model is obtained using the posterior probability of the parameters
- Model assessment is thus inversely proportional to the sum of the code-lengths

Pros and cons of combining models using a posterior distribution

Cons

- Regularization is simpler and sensitivity reduced
- Variance of the ensemble can be reduced
- Training for optimal ensemble performance
- Predict large ensemble performance from a small set

Variance/Bias Decomposition for Ensembles

$$\bar{f}(x) = \frac{1}{Q} \sum_{i=1}^{Q} f_i(x).$$

$$\begin{aligned} E[(\bar{f} - E[\bar{f}])^2] &= E[(\frac{1}{Q} \sum f_i - E[\frac{1}{Q} \sum f_i])^2] \\ &= E[(\frac{1}{Q} \sum f_i)^2] - (E[\frac{1}{Q} \sum f_i])^2. \end{aligned} \quad (1)$$

The first RHS term can we rewritten as

$$E[(\frac{1}{Q} \sum f_i)^2] = \frac{1}{Q^2} \sum E[f_i^2] + \frac{2}{Q^2} \sum_{i<j} E[f_i f_j],$$

Variance/Bias Decomposition for Ensembles

and the second term gives,

$$(E[\frac{1}{Q}\sum f_i])^2 = \frac{1}{Q^2}\sum(E[f_i^2])^2 + \frac{2}{Q^2}\sum_{i<j} E[f_i]E[f_j].$$

Plugging these equalities into (1) gives

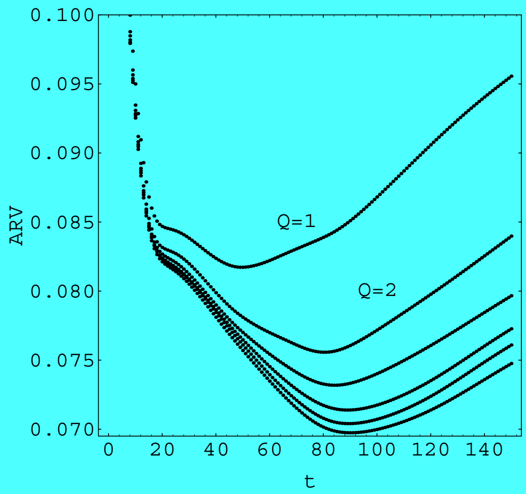
$$E[(\bar{f}-E[\bar{f}])^2] = \frac{1}{Q^2}\sum\{E[f_i^2]-(E[f_i])^2\}+\frac{2}{Q^2}\sum_{i<j}\{E[f_if_j]-E[f_i]E[f_j]\}.$$

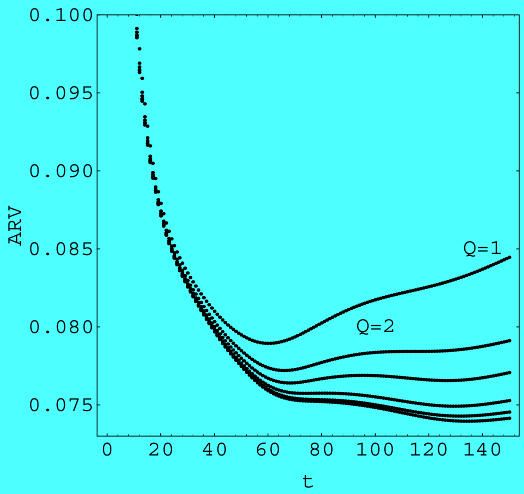
Set $\gamma = \text{Var}(f_i) + (Q-1)\text{max}_{i,j}(E[f_if_j] - E[f_i]E[f_j])$.

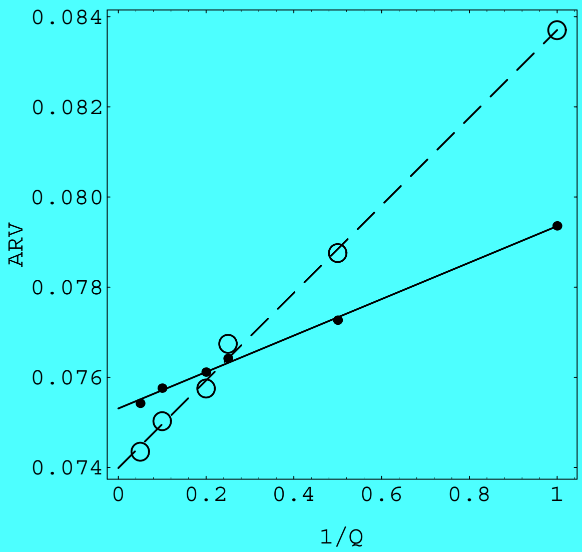
It follows $[ab \leq \frac{a^2+b^2}{2} \Rightarrow E[f_if_j] - E[f_i]E[f_j] \leq \text{max}_i \text{Var}(f_i)]$ that

$$\text{Var}(\bar{f}) \leq \frac{1}{Q}\gamma \leq \max_i \text{Var}(f_i). \qquad (2)$$

Error as a function of ensemble size and training time (Horn et al., 98)





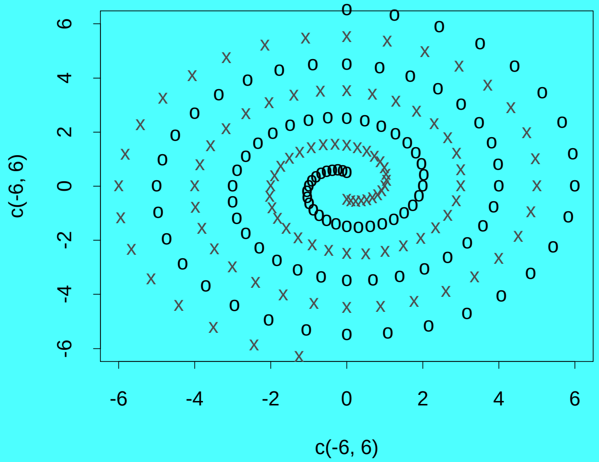


Different ensembles of two predictors as a function of training time. The variance goes down as 1/Q.

Regularization revisited

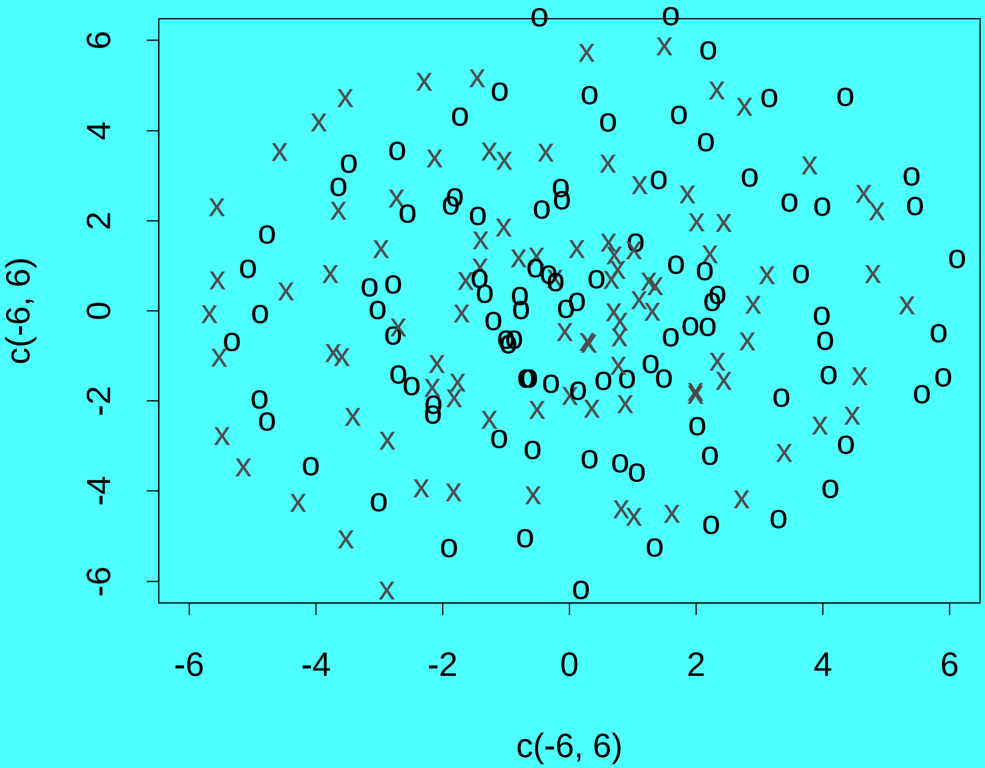
- Consider a highly non-natural problem for NN
- Low dimensional (highly nonlinear)
- Study the ability to control model properties *Capacity, Variance, Bias/Smoothness*
- Easy visualization of Generalization Properties

The Two-Spiral Problem



- 194 X,Y values. Half produce a 1 output, and half produce 0
- Lang and Witbrock (1988) proposed a 2-5-5-5-1 net (138 weights)
- Fahlman Lebiere (1990) Cascade Correlation architecture
- Baum and Lang (1991) Net of $2-50-1$ could be consistent with training set, but could not be found from random initial weights
- Deffuant (1995) suggested the "Perceptron Membrane": piecewise linear discriminating surfaces using 29 perceptrons. Non smooth solution

The noisy spirals

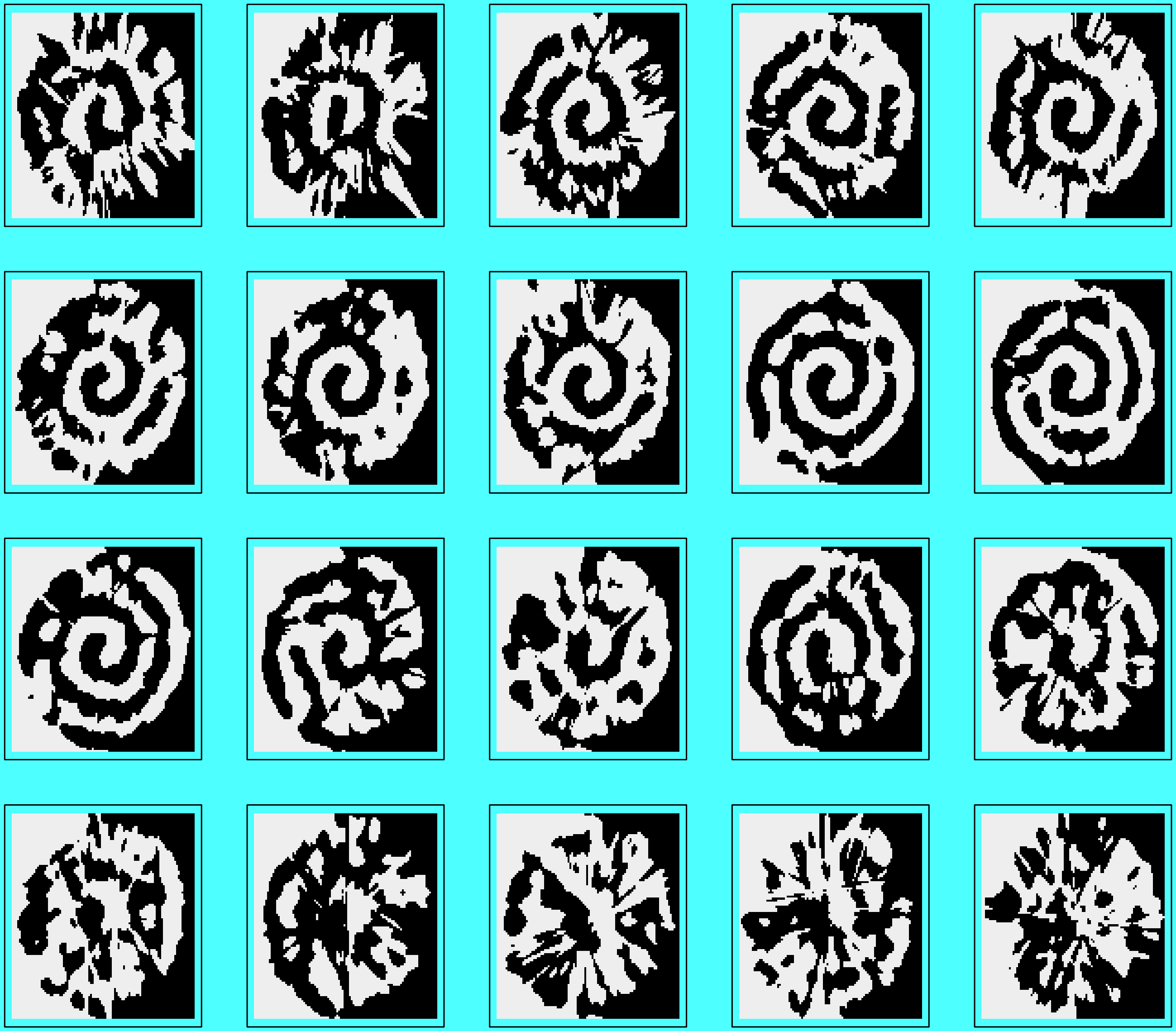


Additional Gaussian noise (SD=0.3)

Different noise levels

Results of training with different noise levels. $\epsilon = 0, \ldots, 0.8$

Different noise levels and optimal weight decay



Summary: 40 Net Ensemble

Top left: No constrains. *Top right:* Optimal WD, no noise. *Bottom left:* Optimal noise, no WD. *Bottom right:* Optimal noise & WD.

Local GAM

- Local generalized additive model (Hastie Tibshirani, 1986)
- Uses a polynomial fit of degree 1 (optimal)
- The span parameter determines the degree of locality of the estimation
- Ideal model for the problem
  - Local with control on locality
  - No ridge constraints
  - Provides a unique model (less variability)
  - Smoothness controlled by locality and degree of the polynomial

Noisy GAM

Average of 20 GAMs with varying degrees of noise

Take home from the Spirals

- NN are easy to regularize
  - Weight decay (smoothing)
  - Ensemble average
- Bootstrap with noise is useful for other models' regularization and is **not** equivalent to smoothing

Challenge:

show similar performance using Stacking, Bagging, Boosting, Arcing, Randomization, etc.

Problems in Interpretability of NN

- The model is not identifiable since there is no unique solution to a fixed ANN architecture and learning rule.

- Estimation with gradient descent increases variability of the model due to local minima

- There is no clear Optimal network architecture (number of hidden layers, number of hidden units, recurrent, second order, etc.)

- Nonlinear model: all effects should only be calculated locally (per input observation)

- How to devise summary statistics for ranking between variables?

Summary

- While most activity is geared towards same architecture ensembles, Different architecture ensemble is promising

- Model assessment was presented for same or different architecture ensembles

- Variance control is possible with simple averaging

- Large ensemble performance can be predicted from small set