Regularized Linear Models in Stacked Generalization

Sam Reid and Greg Grudic

Department of Computer Science
University of Colorado at Boulder
USA

June 11, 2009

How to combine classifiers?

- Which classifiers?
- How to combine?
- Adaboost, Random Forest prescribe classifiers and combiner
- We want $L \geq 1000$ heterogeneous classifiers
- Vote/Average/Forward Stepwise Selection/Linear/Nonlinear?
- Our combiner: Regularized Linear Model

Outline

1. Introduction
   - How to combine classifiers?
2. Model
   - Stacked Generalization
   - StackingC
   - Linear Regression and Regularization
3. Experiments
   - Setup
   - Results
   - Discussion

Outline

1. Introduction
   - How to combine classifiers?
2. Model
   - Stacked Generalization
   - StackingC
   - Linear Regression and Regularization
3. Experiments
   - Setup
   - Results
   - Discussion

Stacked Generalization

- Combiner is produced by a classification algorithm
- Training set = base classifier predictions on unseen data + labels
- Learn to compensate for classifier biases
- Linear and nonlinear combiners
- What classification algorithm should be used?

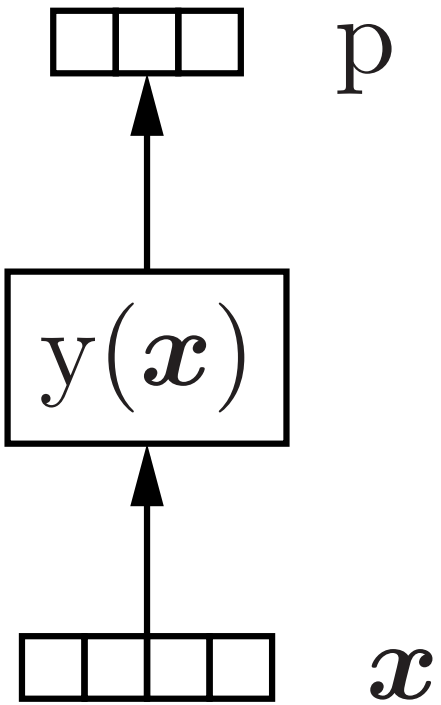
Stacked Generalization - Combiners

- Wolpert, 1992: relatively global, smooth combiners
- Ting & Witten, 1999: linear regression combiners
- Seewald, 2002: low-dimensional combiner inputs

Problems

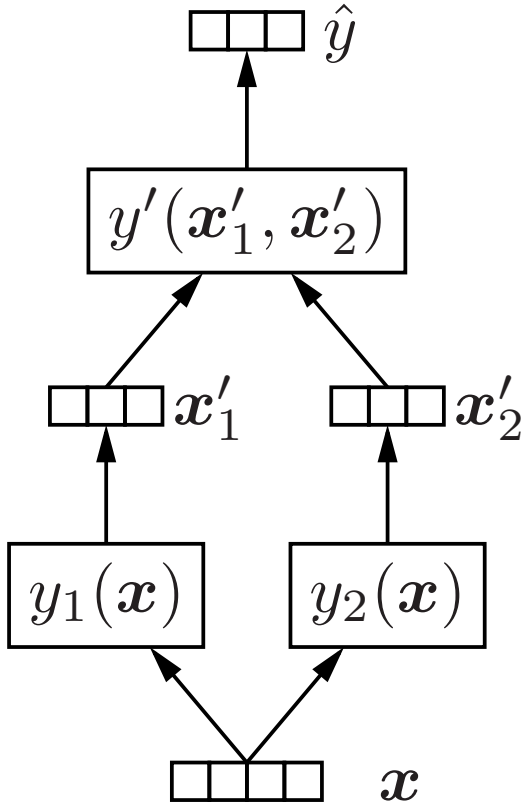
- Caruana et al., 2004: Stacking performs poorly because regression overfits dramatically when there are 2000 highly correlated input models and only 1k points in the validation set.
- How can we scale up stacking to a large number of classifiers?
- Our hypothesis: regularized linear combiner will
  - reduce variance
  - prevent overfitting
  - increase accuracy

Posterior Predictions in Multiclass Classification



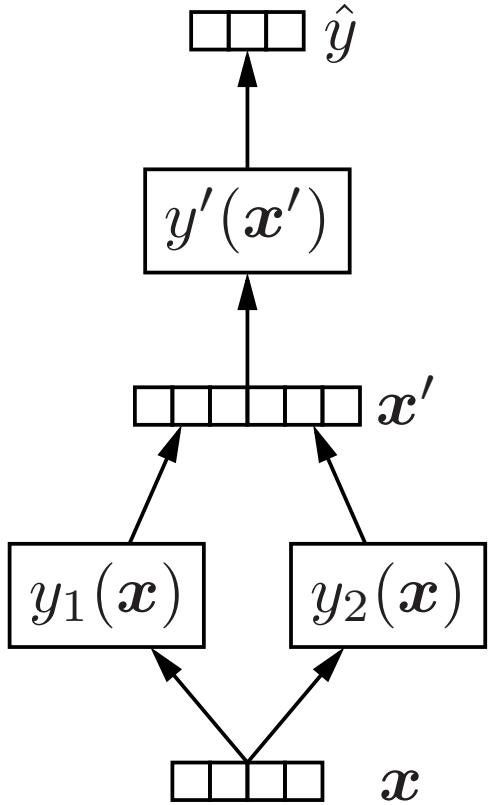
Classification with $d = 4$, $k = 3$

Ensemble Methods for Multiclass Classification



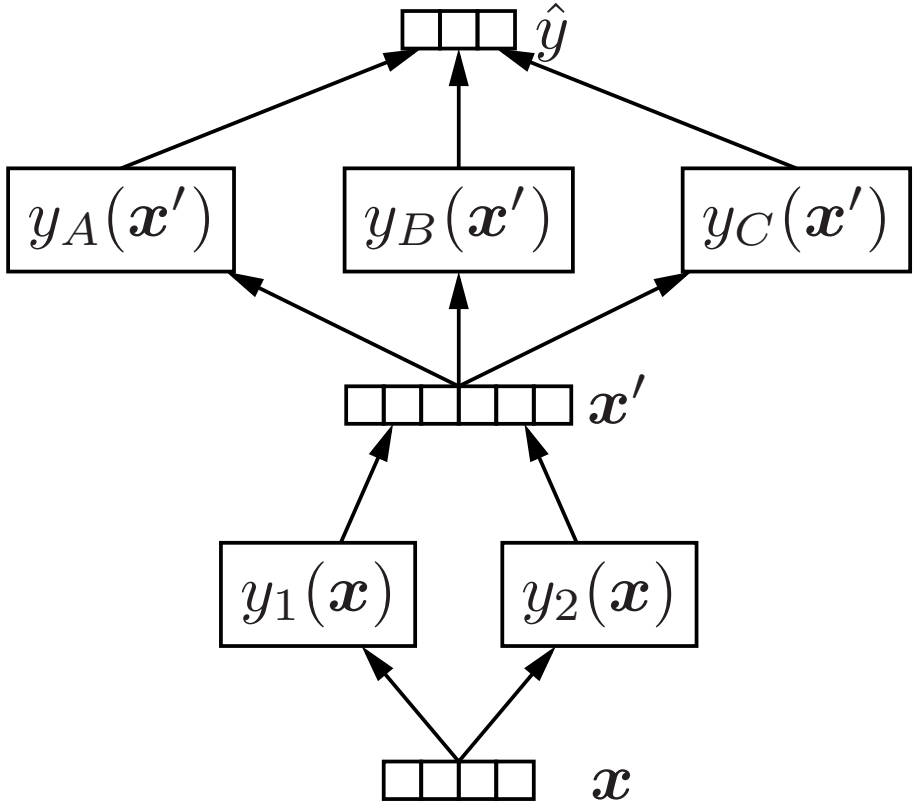
Multiple classifier system with 2 classifiers

Stacked Generalization



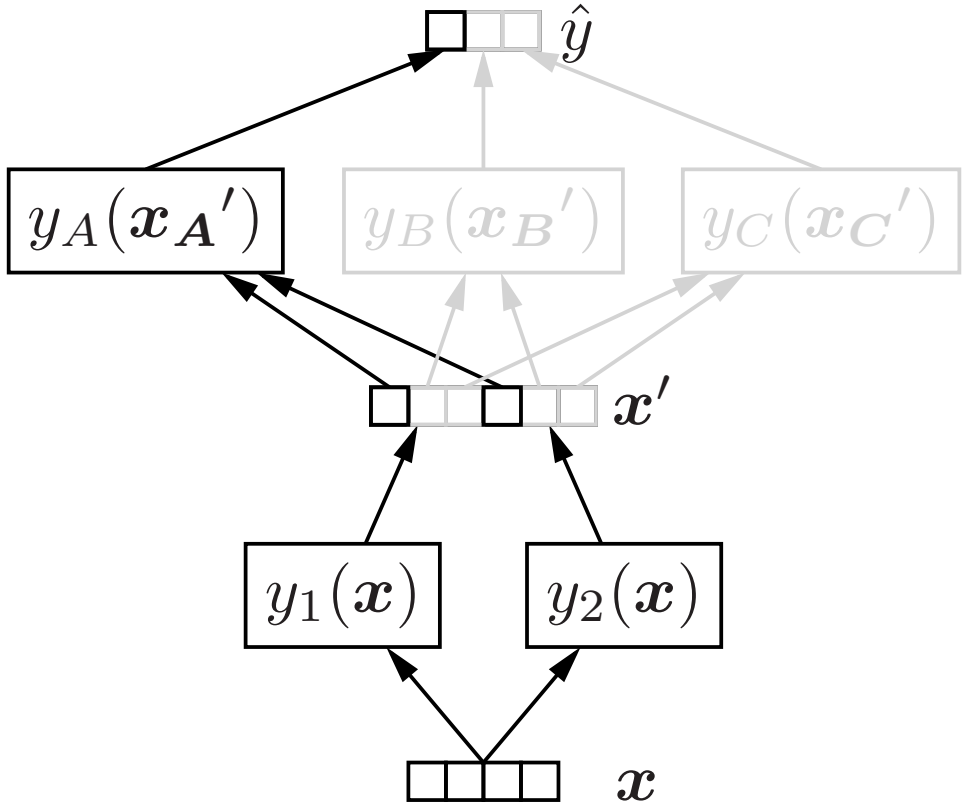
Stacked generalization with 2 classifiers

Classification via Regression



Stacking using Classification via Regression

StackingC



StackingC, class-conscious stacked generalization

Linear Models

- Linear model for use in Stacking or StackingC
- $\hat{y} = \sum_{i=1}^{d} \beta_i x_i + \beta_0$
- Least Squares: $L = |y - X\beta|^2$
- Problems:
  - High variance
  - Overfitting
  - Ill-posed problem
  - Poor accuracy

Regularization

- Increase bias a little, decrease variance a lot
- Constrain weights $\Rightarrow$ reduce flexibility $\Rightarrow$ prevent overfitting
- Penalty terms in our studies:
  - Ridge Regression: $L = |y - X\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|^2$
  - Lasso Regression: $L = |y - X\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|_1$
  - Elastic Net Regression: $L = |y - X\beta|^2 + \lambda|\boldsymbol{\beta}|^2 + (1-\lambda)|\boldsymbol{\beta}|_1$
- Lasso/Elastic Net produce sparse models

Model

- About 1000 base classifiers making probabilistic predictions
- Stacked Generalization to create combiner
- StackingC to reduce dimensionality
- Convert multiclass to regression
- Use linear regression
- Regularization on the weights

Model

- About 1000 base classifiers making probabilistic predictions
- Stacked Generalization to create combiner
- StackingC to reduce dimensionality
- Convert multiclass to regression
- Use linear regression
- **Regularization on the weights**

Outline

1. Introduction
   - How to combine classifiers?
2. Model
   - Stacked Generalization
   - StackingC
   - Linear Regression and Regularization
3. Experiments
   - Setup
   - Results
   - Discussion

Datasets

Table: Datasets and their properties

Dataset	*Attributes*	*Instances*	*Classes*
balance-scale	4	625	3
glass	9	214	6
letter	16	4000	26
mfeat-morphological	6	2000	10
optdigits	64	5620	10
sat-image	36	6435	6
segment	19	2310	7
vehicle	18	846	4
waveform-5000	40	5000	3
yeast	8	1484	10

Base Classifiers

- About 1000 base classifiers for each problem
  1. Neural Network
  2. Support Vector Machine (C-SVM from LibSVM)
  3. K-Nearest Neighbor
  4. Decision Stump
  5. Decision Tree
  6. AdaBoost.M1
  7. Bagging classifier
  8. Random Forest (Weka)
  9. Random Forest (R)

Results

	select–best	*vote*	*average*	*sg – linear*	*sg – lasso*	*sg – ridge*
balance	**0.9872**	0.9234	0.9265	0.9399	0.9610	0.9796
glass	0.6689	0.5887	0.6167	0.5275	0.6429	**0.7271**
letter	0.8747	0.8400	0.8565	0.5787	0.6410	**0.9002**
mfeat	0.7426	0.7390	0.7320	0.4534	0.4712	**0.7670**
optdigits	0.9893	0.9847	0.9858	0.9851	0.9660	**0.9899**
sat-image	0.9140	0.8906	0.9024	0.8597	0.8940	**0.9257**
segment	0.9768	0.9567	0.9654	0.9176	0.6147	**0.9799**
vehicle	0.7905	0.7991	0.8133	0.6312	0.7716	**0.8142**
waveform	0.8534	0.8584	**0.8624**	0.7230	0.6263	0.8599
yeast	**0.6205**	0.6024	0.6105	0.2892	0.4218	0.5970

Statistical Analysis

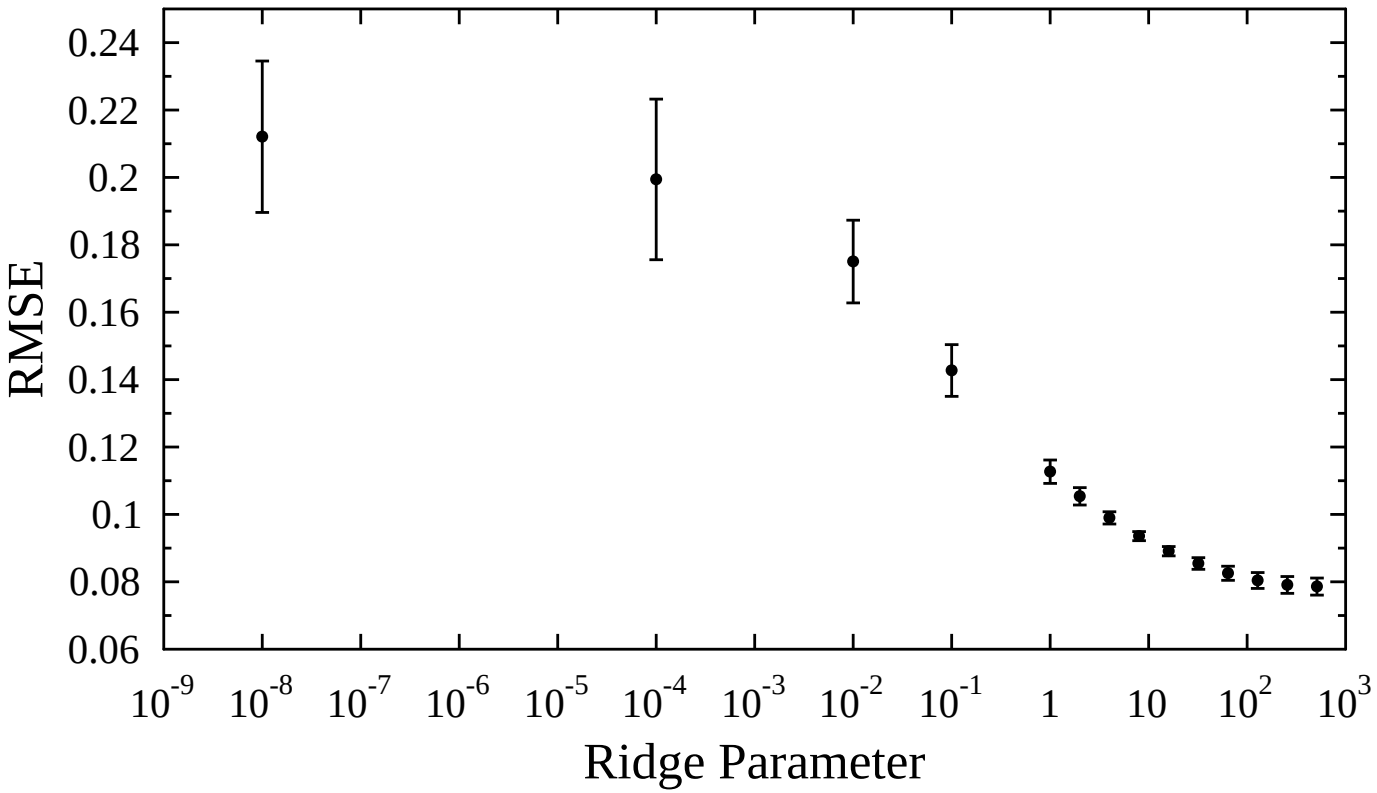
- Pairwise Wilcoxon Signed-Rank Tests

- Ridge outperforms unregularized at $p \leq 0.002$
- Lasso outperforms unregularized at $p \leq 0.375$
  - Validates hypothesis: regularization improves accuracy
- Ridge outperforms lasso at $p \leq 0.0019$
  - Dense techniques outperform sparse techniques
- Ridge outperforms Select-Best at $p \leq 0.084$
  - Properly trained model better than single best

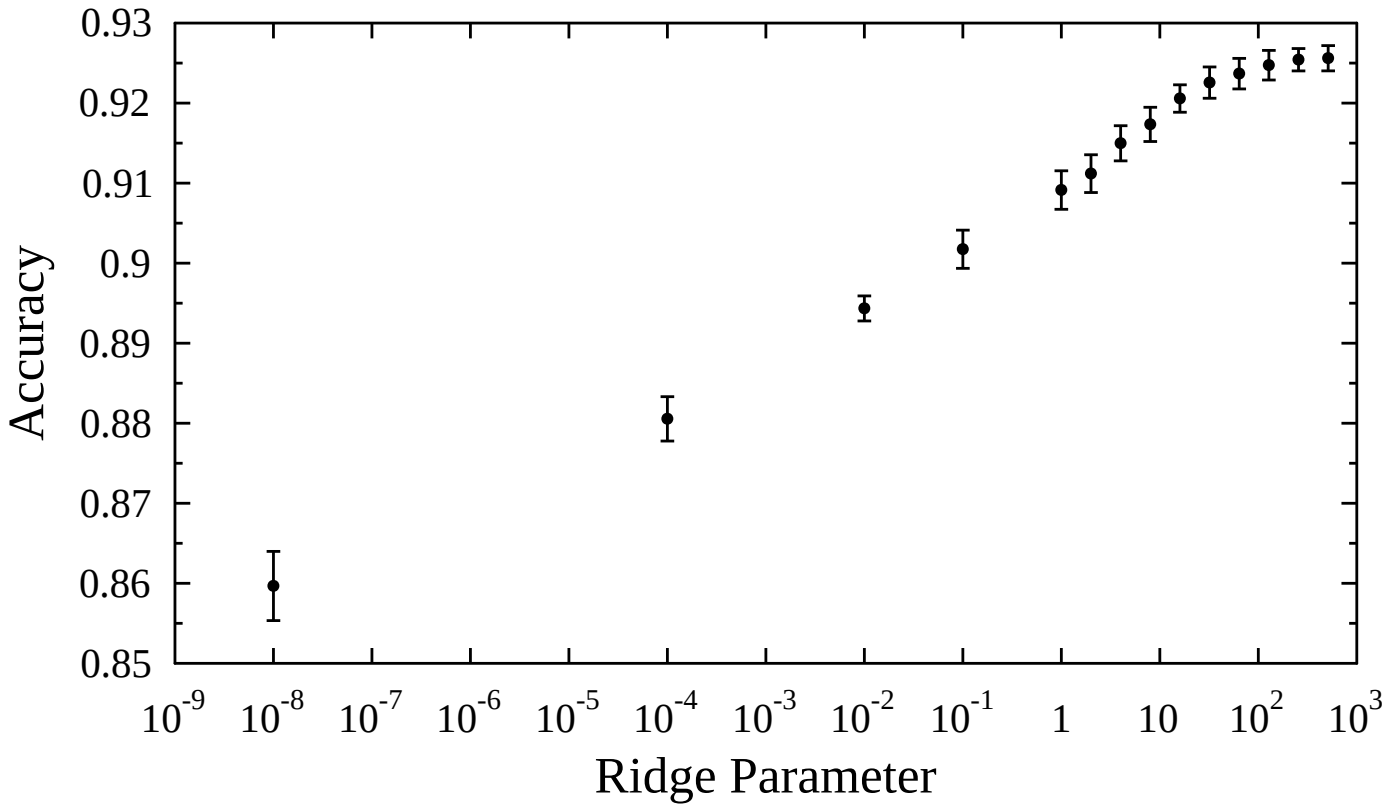
Baseline Algorithms

- Average outperforms Vote at $p \leq 0.014$
  - Probabilistic predictions are valuable
- Select-Best outperforms Average at $p \leq 0.084$
  - Validation/training is valuable

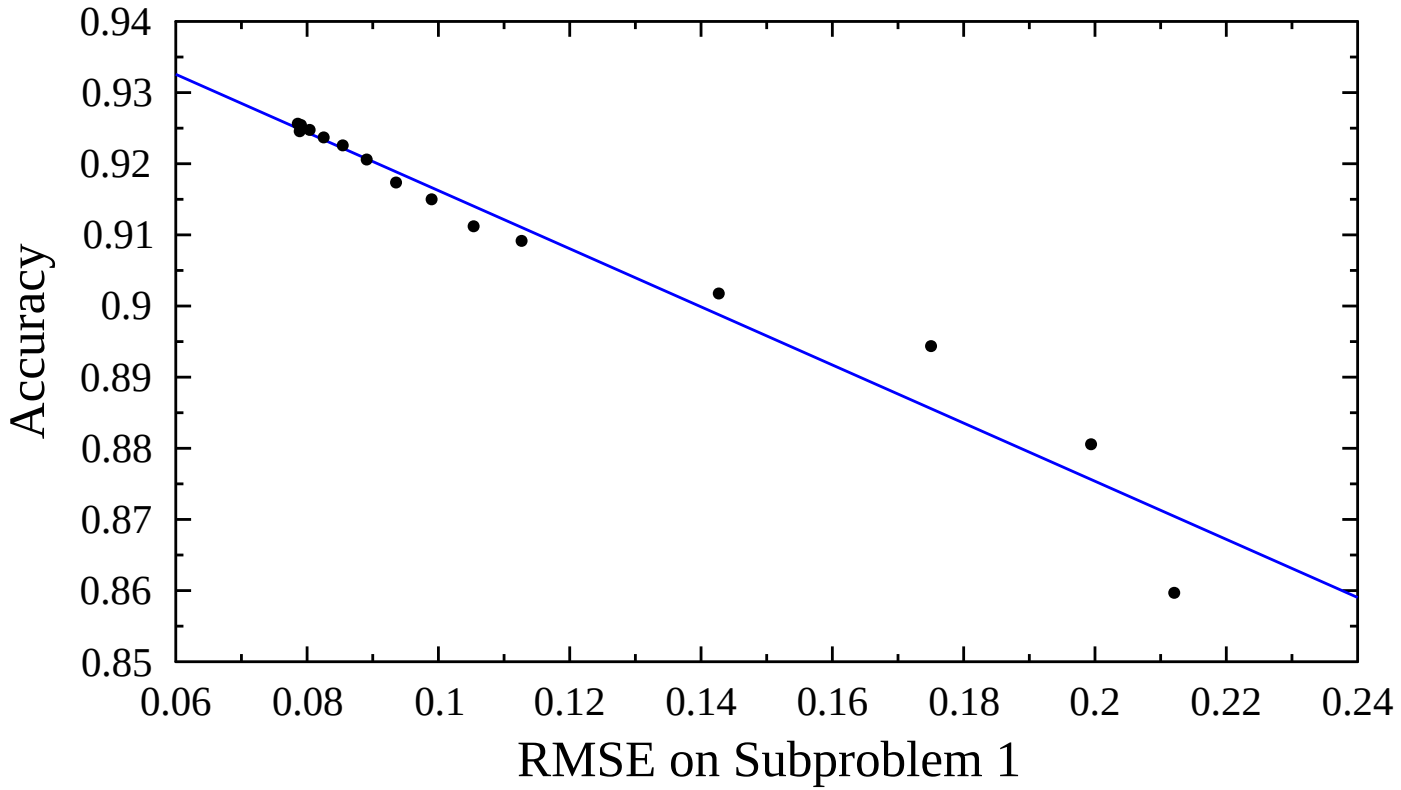
Subproblem/Overall Accuracy - I



Subproblem/Overall Accuracy - II



Subproblem/Overall Accuracy - III



Accuracy for Elastic Nets

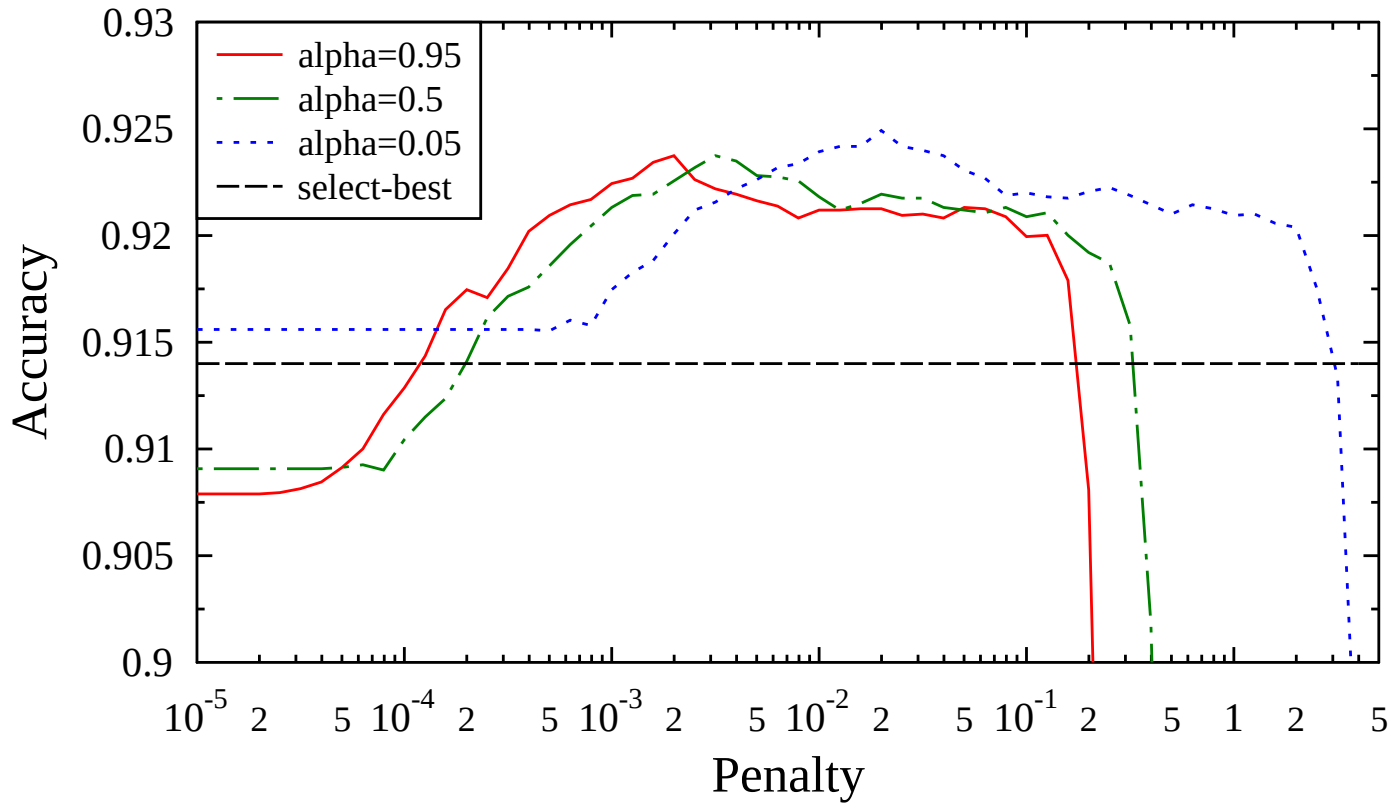


Figure: Overall accuracy on sat-image with various parameters for elastic-net.

Partial Ensemble Selection

- Sparse techniques perform Partial Ensemble Selection
- Choose from classifiers and predictions
- Allow classifiers to focus on subproblems
- Example: Benefit from a classifier good at separating A from B but poor at A/C, B/C

Partial Selection

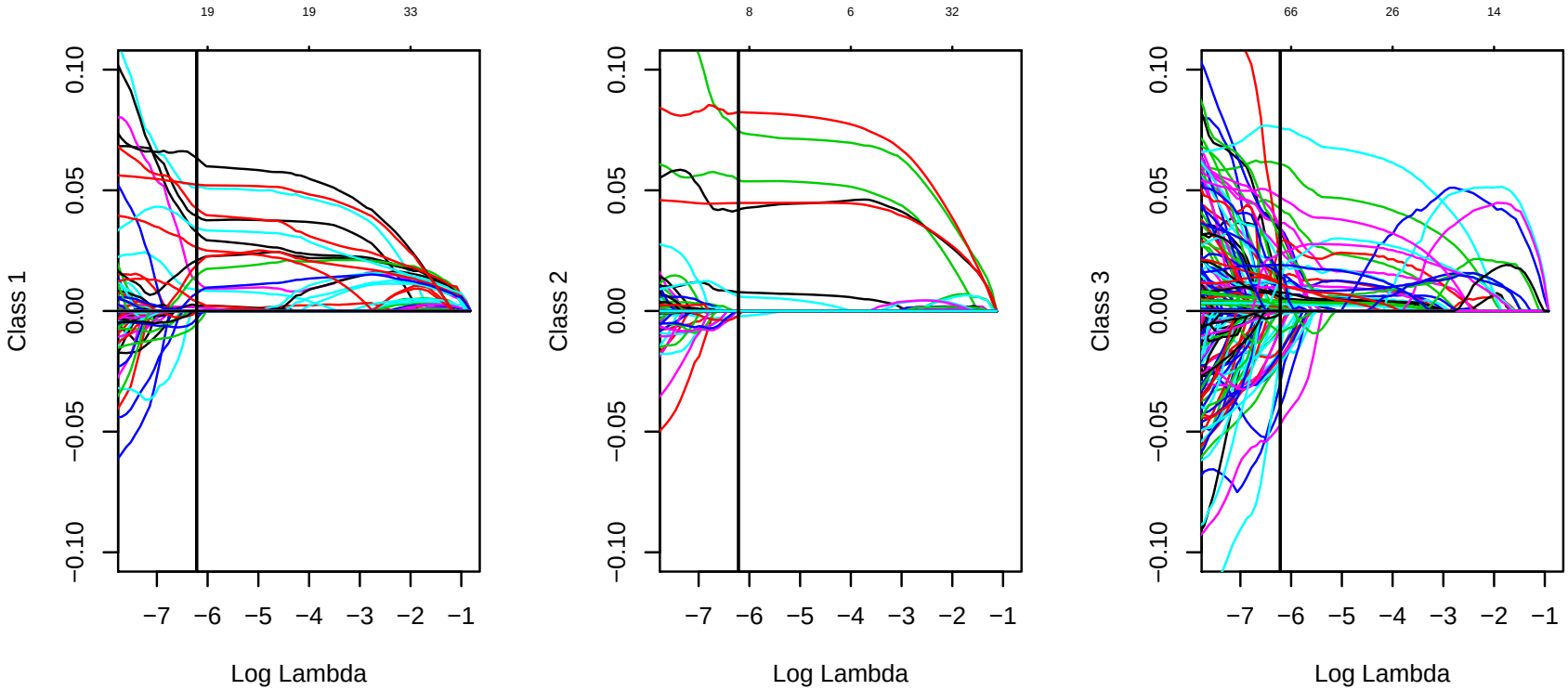


Figure: Coefficient profiles for the first three subproblems in StackingC for the *sat-image* dataset with elastic net regression at $\alpha = 0.95$.

Selected Classifiers

Classifier	*red*	*cotton*	*grey*	*damp*	*veg*	*v.damp*	*total*
adaboost-500	0.063	0	0.014	0.000	0.0226	0	0.100
ann-0.5-32-1000	0	0	0.061	0.035	0	0.004	0.100
ann-0.5-16-500	0.039	0	0	0.018	0.009	0.034	0.101
ann-0.9-16-500	0.002	0.082	0	0	0.007	0.016	0.108
ann-0.5-32-500	0.000	0.075	0	0.100	0.027	0	0.111
knn-1	0	0	0.076	0.065	0.008	0.097	0.246

Table: Selected posterior probabilities and corresponding weights for the *sat-image* problem for elastic net StackingC with $\alpha = 0.95$ for the 6 models with highest total weights.

Conclusions

- Regularization is essential in Linear StackingC
- Trained linear combination outperforms Select-Best
- Dense combiners outperform sparse combiners
- Sparse models allow classifiers to specialize in subproblems

Future Work

- Examine full Bayesian solutions
- Constrain coefficients to be positive
- Choose a single regularizer for all subproblems

Acknowledgments

- PhET Interactive Simulations
- Turing Institute
- UCI Repository
- University of Colorado at Boulder

Questions?

- Questions?