Efficient Model Selection for Regularized Discriminant Analysis

John A. Ramey

Department of Statistical Science
Baylor University

Wednesday, August 4, 2010

Classification

- Supervised learning area where the *i*th response y_i is one of K discrete **classes**.
  - $y_i \in \{0, 1, \ldots, K-1\}$ for $i = 1, 2, \ldots, N$.
  - Also called **groups** or **labels**.
- We know the K classes *a priori*.
- $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$
- Covariates (or features) $\mathbf{x}_i \in \mathbb{R}_p$
- $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)'$
- The classical approach is to partition $\mathbb{R}_p$ into K mutually exclusive sets.
- For a future observation without a recorded response, we wish to predict the response.
  - The goal is to minimize the prediction error.

Classification

- From our training data, we build a **classifier**.
  - This is a function of the training covariates $\mathbf{X}$.
- For an unlabeled observation, we **classify** (or **predict**) it to one of the K classes with the classifier.
- The **error rate** of a classifier is the true probabibility that an unlabeled observation will be incorrectly classified.
- Often we partition the available data into training data and **validation** (or **test**) data to estimate the error rate.
  - This reduces the number of available training observations and can be counterproductive if $p >> N$.

High-Dimensional Data

- High-dimensional data are becoming increasingly common.
  - Automatic collection of large quantities of data.
  - Large storage space (e.g. hard drives, cloud storage).
- Design shift in statistical data analysis in many disciplines.
  - Less focus on a few well-selected variables
  - More focus on identifying the most relevant variables among a large number of variables.
- Increases the difficulty of classification and other machine learning methods.
  - Difficult to visualize beyond $p = 3$.
  - "Curse of dimensionality" (Bellman, 1961).

Curse of Dimensionality

- N/p is preferred to be large.
- If not, we have data sparsity
- Sometimes it can be difficult to obtain sufficiently more observations than the feature space dimension, p.
    - Often $p >> n$
    - Example: Microarray data
- Classical estimators are usually unstable when $p >> N$
- Asymptotic results become problematic

Possible Solutions

- Feature/variable selection.
  - Determine which, if any, variables are relevant.
  - Omit variables from the model that appear relatively unimportant through feature selection.
- Emphasize variables algorithmically through methods such as regularization.
- Dimension reduction.
- Ignore interactions.

Quadratic Discriminant Analysis (QDA)

- Quadratic classification boundaries

$$d_k(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)^T + \ln |\boldsymbol{\Sigma}_k| - 2 \ln \pi_k \tag{1}$$

- Often MLEs are substituted for unknown parameters
  - $\mathbf{S}_k$ for $\boldsymbol{\Sigma}_k$
  - $\bar{\mathbf{x}}_k$ for $\boldsymbol{\mu}_k$
  - $\hat{\pi}_k$ for π_k
- Classify $\mathbf{x}$ to class $\hat{k}$ where

$$d_{\hat{k}}(\mathbf{x}) = \min_{1 \le k \le K} d_k(\mathbf{x})$$

Linear Discriminant Analysis (LDA)

- Linear classification boundaries
- Special case of QDA
  - $\boldsymbol{\Sigma}_k \equiv \boldsymbol{\Sigma}$
- $d_k(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)^T - 2 \ln \pi_k$
- MLE for $\boldsymbol{\Sigma}$ is the pooled sample covariance matrix $\mathbf{S}_p$

QDA with High-Dimensional Data

- Spectral Decomposition of $\mathbf{\Sigma}_k^{-1}$

$$\mathbf{\Sigma}_k^{-1} = \sum_{i=1}^{p} \frac{\mathbf{v}_i \mathbf{v}_i^T}{e_i} \tag{2}$$

- $\mathbf{v}_i$ is the ith eigenvector of $\mathbf{\Sigma}_k$ and corresponding eigenvalue e_i
- (2) is heavily weighted by the smallest eigenvalues and the directions associated with their eigenvector
    - Hence (1) is also

Shrinking

- $\mathbf{S}_k$ is singular when $n_k < p$
  - Eigenvalues are near 0.
  - $\mathbf{S}_k^{-1}$ is numerically unstable
- Stabilize eigenvalues $\mathbf{S}_k^{-1}$ by computing the ridge estimator

$$\mathbf{S}_k(\gamma) = S_k + \gamma \mathbf{I}_p, \quad \gamma > 0$$

- An equivalent form is given by

$$\mathbf{S}_k(\gamma) = \gamma S_k + (1 - \gamma)\mathbf{I}_p, \quad \gamma \in [0, 1] \tag{3}$$

- γ is chosen by crossvalidation
- (1) is stabilized by substituting (3) for $\boldsymbol{\Sigma}_k$

Regularized Discriminant Analysis (RDA)

- Friedman (1989) proposed a classifier that is a convex combination of the class covariance matrices and the pooled sample covariance matrix

$$\hat{\Sigma}_k(\lambda) = (1-\lambda)\mathbf{S}_k + \lambda \mathbf{S}_p, \quad \lambda \in [0,1] \tag{4}$$

- Then we shrink (4) and scale by the average of the eigenvalues of $\hat{\Sigma}_k(\lambda)$ to obtain

$$\hat{\Sigma}_k(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \gamma \frac{\text{tr}\{\hat{\Sigma}_k(\lambda)\}}{p}\mathbf{I}_p, \quad \gamma \in [0,1] \tag{5}$$

- Substitute (5) into (1) for the RDA classifier
  - LDA: $\lambda = 1, \gamma = 0$
  - QDA: $\lambda = 0, \gamma = 0$
  - Nearest Means: $\lambda = 0, \gamma = 1$

RDA Model Selection

- Friedman recommends to construct a unit grid of (λ, γ) values
  - $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_B), \quad \lambda_j \in [0, 1]$
  - $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_B), \quad \gamma_j \in [0, 1]$
  - Grid: $\boldsymbol{\lambda} \times \boldsymbol{\gamma}$
- Compute the conditional error rate (CER) at each grid point
- The grid point with the minimum CER is the selected model
- Other model selection methods in the literature (highly variable)
  - Particle Swarm Optimization
  - Nelder-Mead

Problems with the Grid Method

- The minimum CER is often not unique, resulting in ties
  - For this poster we choose the (λ, γ) point closest (in squared distance) to LDA $(\lambda = 1, \gamma = 0)$
- Computationally expensive
  - Eased by Friedman's "down-dating" algorithm for the leave-one-out (LOO) CER
  - The grid is "embarrassingly parallel" and can take advantage of parallel processing, such as the R package `foreach`
  - GreedyGrid

GreedyGrid Algorithm

- Heuristic algorithm that explores the grid to find the (λ, γ) pairs to find those that are minimum
- Leads to tremendous savings while allowing for very precise grids

GreedyGrid Algorithm

1. Initial: Compute CER_{ij} at $\lambda_i = \lambda_{\lfloor B/2 \rfloor}, \gamma_j = \gamma_{\lfloor B/2 \rfloor}$
2. Compute **CER** at $(\lambda_{i-1}, \gamma_j)$, $(\lambda_{i+1}, \gamma_j)$, $(\lambda_i, \gamma_{j-1})$, $(\lambda_i, \gamma_{j+1})$
3. If $CER_{ij} \geq \min \mathbf{CER}_{i'j'}$
    1. Set $i = i'$ and $j = j'$
    2. Go to Step #2
4. Return $\hat{\lambda} = \lambda_i$ and $\hat{\gamma} = \gamma_j$

Simulation Experiments

- We study two of the simulation experiments considered by Friedman
- $K = 3$ populations (classes)
- $N = 45$ labels are randomly drawn
- Consider feature space dimension $p = 10, 30, 60, 90$
- Expected Error Rate (EER) estimated by generating 1000 observations from each class and classifying with training classifiers

Simulation Experiment #1

- Orthogonal Means
  - $\mu_1 = (0, 0, 0, 0, \ldots, 0)^T$
  - $\mu_2 = (0, 3, 0, 0, \ldots, 0)^T$
  - $\mu_3 = (0, 0, 4, 0, \ldots, 0)^T$
- Unequal Spherical Covariance Matrices
  - $\mathbf{\Sigma}_1 = \mathbf{I}_p$
  - $\mathbf{\Sigma}_2 = 2\mathbf{I}_p$
  - $\mathbf{\Sigma}_3 = 3\mathbf{I}_p$

Simulation Experiment #1 - Heatmaps

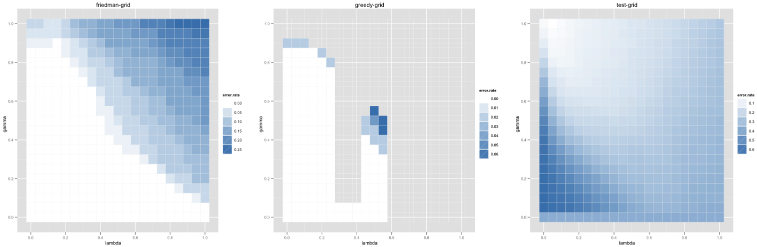

Figure: Heatmaps for Grid CER, GreedyGrid CER, and EER with $p = 90$. The minimum EER of 0.078 (0.008) is attained at $(\lambda, \gamma) = (0.05, 0.85)$. Grid size $= 441$. Greedy Grid Size $= 150$.

Simulation Experiment #2

- Class means:
  - $\mu_{i1} = 0$
  - $\mu_{i2} = 14/\sqrt{p}$
  - $\mu_{i3} = (-1)^i \mu_{2i}$
- Unequal Highly Ellipsoidal Covariance Matrices
  - High and low variance subspaces of classes 1 and 2 are complementary to each other
  - Third class has low variance and high variance in the intermediate subspace and complementary high/low variance subspaces 1 and 2, respectively
  - e_{ik} is the ith eigenvalue of Σ_k for $1 \leq i \leq p$
  - $e_{i1} = [9(i-1)/(p-1)+1]^2$
  - $e_{i2} = [9(p-i)/(p-1)+1]^2$
  - $e_{i3} = \{9[i-(p-1)/2]/(p-1)\}^2$

Simulation Experiment #2 - Heatmaps

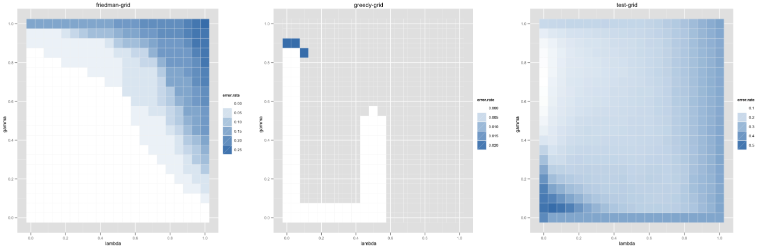

Figure: Heatmaps for Grid CER, GreedyGrid CER, and EER with $p = 90$. The minimum EER of 0.097 (0.009) is attained at $(\lambda, \gamma) = (0.00, 0.70)$. Grid size = 441. Greedy Grid Size = 87.

Simulation Experiment #2 - GreedyGrid Sizes

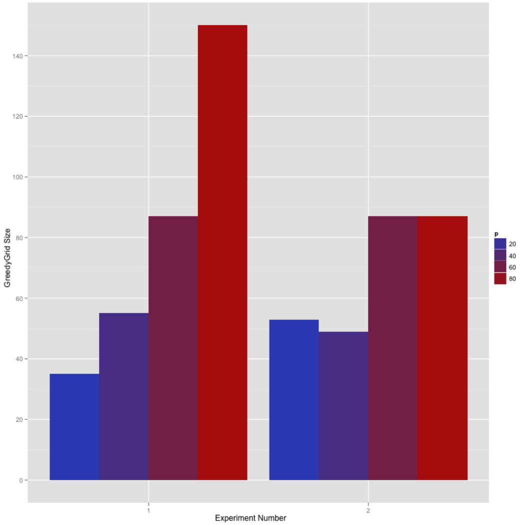

Figure: GreedyGrid Size when Grid Size is 441

Conclusion

- Classification of high-dimensional is a difficult but common problem.
- RDA is able to classify with low EER when $p >> N$.
- This will require more model selection methods.
- Future Work:
  - Apply to real high-dimensional data sets.
  - Use other error rate estimators such as .632.
  - Mathematically-based model selection algorithm for RDA.
  - Break ties.

References

- Bellman, R.E. (1961). *Adaptive Control Processes*. Princeton University Press: Princeton, NJ.
- Friedman, J. H. (1989). "Regularized Discriminant Analysis." *Journal of American Statistical Association*, 84, 165-175.