
Support Vector Machines and Flexible Discriminants

Chapter 12: Hastie et al. (2001)

Madhusudana Shashanka

Department of Cognitive and Neural Systems
Boston University

Overview

● Part I – SVMs

- Support Vector Classifier
- Support Vector Machines

● Part II – Flexible Discriminants

- Flexible Discriminant Analysis
- Penalized Discriminant Analysis
- Mixture Discriminant Analysis

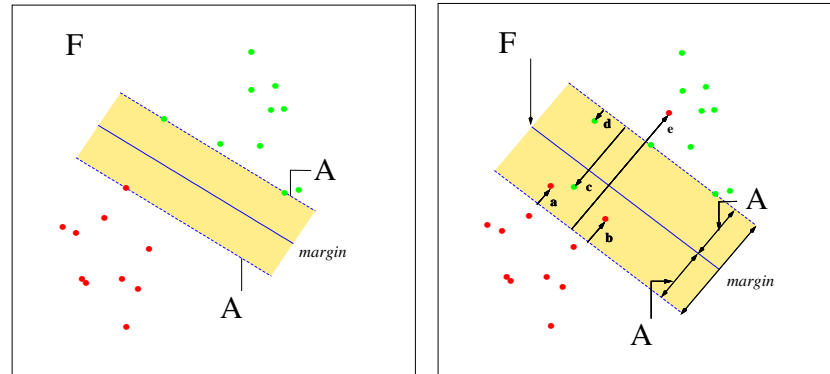
Introduction

- Generalizations of linear decision boundaries for classification.
- Optimal separating hyperplanes for linearly separable classes.
- Extensions to the non-separable case generalize to *support vector machines*.
- SVM - produce nonlinear decision boundaries by constructing a linear boundary in a large, transformed space.

Separating Hyperplanes - Recap

- Consider classes separable by a linear boundary.
- Data of the form (x_i, y_i) with $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$
 $\forall i = 1, 2, \dots, N$.
- Define a hyperplane $\{x : f(x) = x^T \beta + \beta_0 = 0\}$.
- Find the hyperplane that creates the biggest margin between the training points for classes -1 and 1.
 - $\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2$ subject to $y_i(x_i^T \beta + \beta_0) \geq 1, i = 1, 2, \dots, N$.
Margin is $C = 1/\|\beta\|$.
- Classification rule $G(x) = \text{sign}[x^T \beta + \beta_0]$.
- Convex optimization problem: quadratic criterion and linear inequality constraints.

Non-separable classes



- Allow some points on the wrong side.
- Define slack variables $\xi = (\xi_1, \xi_2, \dots, \xi_n)$.
- Modify the constraint as: $y_i(x_i^T \beta + \beta_0) \geq C(1 - \xi_i)$,
 $\forall i, \xi_i \geq 0, \sum_{i=1}^N \xi_i \leq \text{constant}$.
- Misclassifications occur when $\xi_i > 1$.

Support Vector Classifier

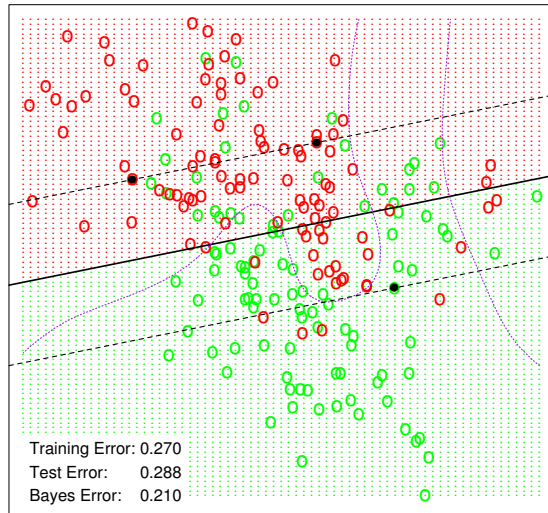
- $\min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i$
subject to $\xi_i \geq 0, y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i \forall i$.

$$L_P = \frac{1}{2} \|\beta\|^2 + \gamma \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i(x_i^T \beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^N \mu_i \xi_i$$

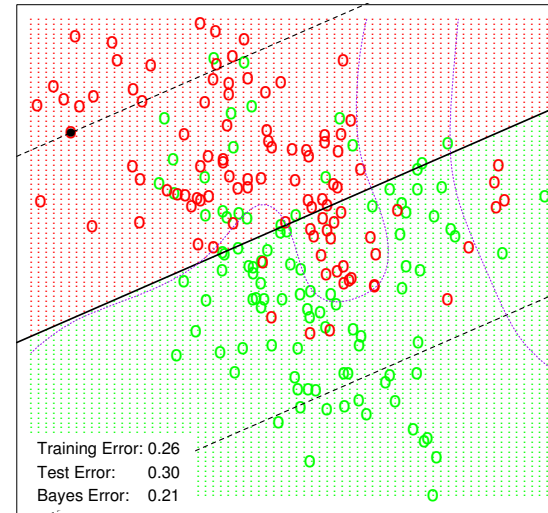
$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'}. \quad (1)$$

- Maximize L_D subject to $0 \leq \alpha_i \leq \gamma$ and $\sum_{i=1}^N \alpha_i y_i = 0$.
 - Solution β has the form $\hat{\beta} = \sum_{i=1}^N \hat{\alpha}_i y_i x_i$.
 - Decision function $\hat{G}(x) = \text{sign}[x^T \hat{\beta} + \hat{\beta}_0]$.
-

Mixture Example



Gamma = 10000



Gamma = 0.01

- Tuning parameter γ . Margin larger for smaller γ .
- $\gamma = 10000$: 62% support points and $\gamma = 0.01$: 85% support points.
- γ estimated by cross-validation.

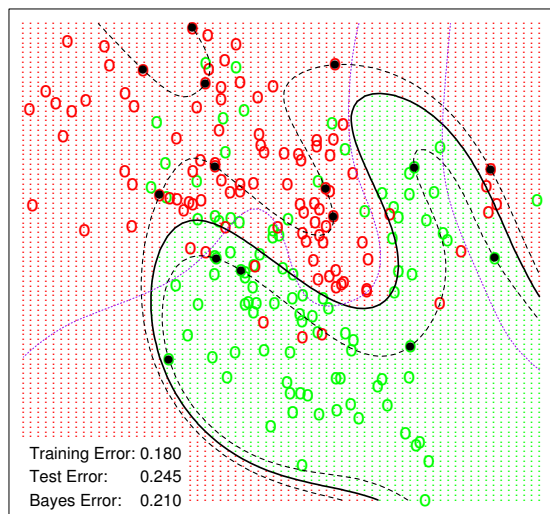
Support Vector Machines

- Idea: Enlarge the feature space using basis expansions.
- Perfect separation typically achievable in the new space.
- SVM - dimension of the enlarged space allowed to get very large, infinite in some cases.
- Represent the problem (1) such that
 - It only involves input (transformed) features via inner products.
 - $$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N \alpha_i \alpha_{i'} y_i y_{i'} \langle h(x_i) h(x_{i'}) \rangle.$$
 - Solution can be written as
$$f(x) = h(x)^T \beta + \beta_0 = \sum_{i=1}^N \alpha_i y_i \langle h(x) h(x_i) \rangle + \beta_0.$$
 - Need to specify only the kernel $K(x, x') = \langle h(x), h(x') \rangle.$
 - K symmetric positive (semi-) definite.

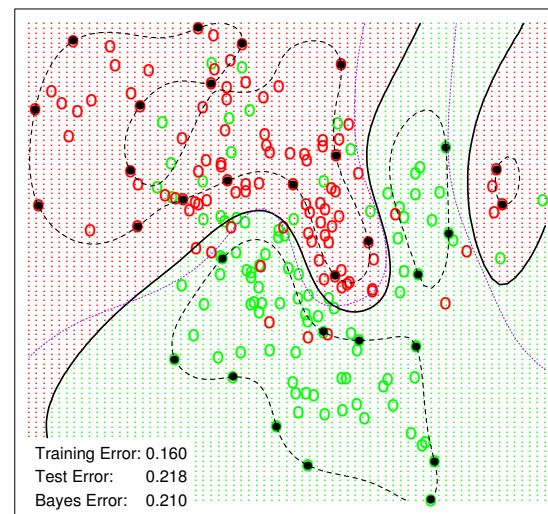
SVM kernel

- Popular choices for the kernel K
 - Polynomial: $K(x, x') = (1 + \langle x, x' \rangle)^d$.
 - Radial basis: $K(x, x') = \exp(-\|x - x'\|^2 / c)$.
 - Neural Network: $K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$.

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space

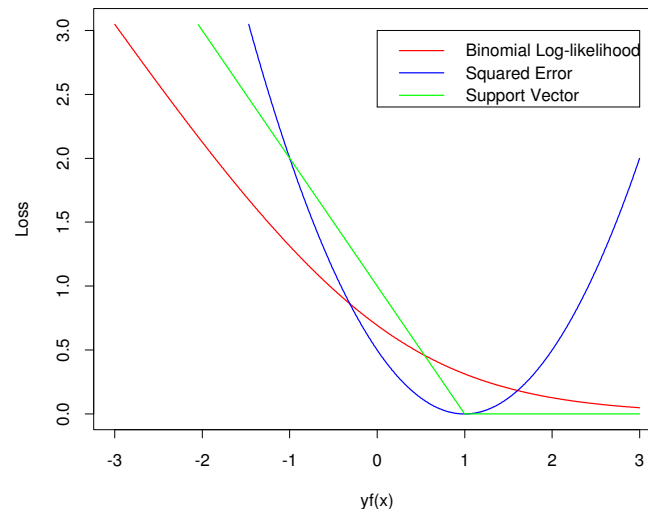


SVM as penalization

- SVM problem can be expressed in the *loss + penalty* form.

$$\min_{\beta_0, \beta} \sum_{i=1}^N [1 - y_i f(x_i)]_+ + \lambda \|\beta\|^2. \quad (2)$$

- Solution will be the same as obtained by previous methods if $\lambda = 1/(2\gamma)$.



Function Estimation

- Suppose basis h arises from the eigen expansion of a positive definite kernel K .
 - $K(x, x') = \sum_{m=1}^{\infty} \phi_m(x)\phi_m(x')\delta_m$ and $h_m(x) = \sqrt{\delta_m}\phi_m(x)$.
- With $\theta_m = \sqrt{\delta_m}\beta_m$, (2) can be written as

$$\min_{\beta_0, \theta} \sum_{i=1}^N [1 - y_i(\beta_0 + \sum_{m=1}^{\infty} \theta_m \phi_m(x_i))]_+ + \lambda \sum_{m=1}^{\infty} \frac{\theta_m^2}{\delta_m}.$$

- Theory of reproducing kernel Hilbert spaces guarantees a finite-dimensional solution of the form
$$f(x) = \beta_0 + \sum_{i=1}^N \alpha_i K(x, x_i).$$
- Equivalent version of (1):
$$\min_{\alpha_0, \alpha} \sum_{i=1}^N (1 - y_i f(x_i))_+ + \lambda \alpha^T \mathbf{K} \alpha.$$

Generality

- Models quite general – include smoothing splines, additive and interaction spline models.

$$\min_{f \in \mathcal{H}} \sum_{i=1}^N [1 - y_i f(x_i)]_+ + \lambda J(f), \quad (3)$$

where \mathcal{H} is the structured space of functions and $J(f)$ an appropriate regularizer on that space.

- If \mathcal{H} is the space of additive functions $f(x) = \sum_{j=1}^p f_j(x_j)$, and $J(f) = \sum_j \int f_j''(x_j) dx_j$, solution to (3) is
 - An additive cubic spline,
 - Kernel representation $K(x, x') = \sum_{j=1}^p K_j(x_j, x'_j)$.

SVMs for regression

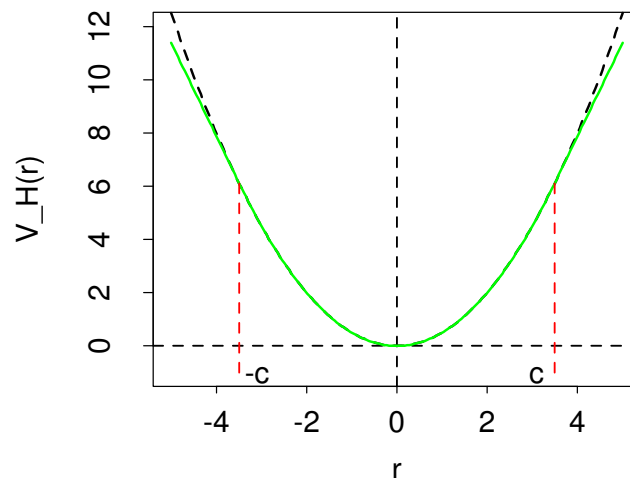
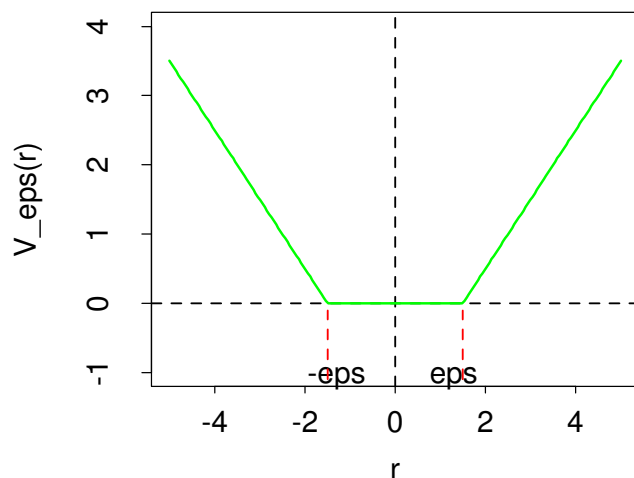
- Consider the linear regression model $f(x) = x^T \beta + \beta_0$.

- To estimate β , minimize

$$H(\beta, \beta_0) = \sum_{i=1}^N V_{\epsilon}(y_i - f(x_i)) + \frac{\lambda}{2} \|\beta\|^2, \text{ where}$$

$$V_{\epsilon}(t) = \begin{cases} 0 & \text{if } |t| < \epsilon, \\ |t| - \epsilon & \text{otherwise.} \end{cases} \quad V_H(r) = \begin{cases} r^2/2 & \text{if } |r| < c, \\ c|r| - c^2/2 & |r| > c. \end{cases}$$

- V_{ϵ} is an ϵ -insensitive measure, V_H is the Huber measure.



SVMs for regression

If $\hat{\beta}$, $\hat{\beta}_0$ are the minimizers of H , solutions of the form

$$\hat{\beta} = \sum_{i=1}^N (\hat{\alpha}_i^* - \hat{\alpha}_i) x_i \quad \hat{f}(x) = \sum_{i=1}^N (\hat{\alpha}_i^* - \hat{\alpha}_i) \langle x, x_i \rangle + \beta_0$$

where $\alpha_i, \alpha_i^* > 0$ and solve the quadratic programming problem

$$\min_{\alpha_i, \alpha_i^*} \epsilon \sum_{i=1}^N (\alpha_i^* + \alpha_i) - \sum_{i=1}^N y_i (\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i, i'=1}^N (\alpha_i^* - \alpha_i) (\alpha_{i'}^* - \alpha_{i'}) \langle x_i, x_{i'} \rangle$$

subject to constraints

$$0 \leq \alpha_i, \alpha_i^* \leq 1/\lambda, \quad \sum_{i=1}^N (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i \alpha_i^* = 0$$

Regression and Kernels

- Approximation of the regression function in terms of basis functions: $f(x) = \sum_{m=1}^M \beta_m h_m(x) + \beta_0$.
- To estimate β and β_0 , minimize
$$H(\beta, \beta_0) = \sum_{i=1}^N V(y_i - f(x_i)) + \frac{\lambda}{2} \sum \beta_m^2.$$
- Soln. has the form $\hat{f}(x) = \sum_{i=1}^N \hat{\alpha}_i K(x, x_i)$ with
$$K(x, y) = \sum \hat{\beta}_m h_m(x) + \hat{\beta}_0.$$
- For eg., if $V(r) = r^2$ and \mathbf{H} the $N \times M$ basis matrix with i th element $h_m(x_i)$,
 - $\{\mathbf{H}\mathbf{H}^T\}_{i,i'} = K(x_i, x_{i'})$ and $\hat{\alpha} = (\mathbf{H}\mathbf{H}^T + \lambda\mathbf{I})^{-1}\mathbf{y}$.
 - Need to evaluate only the inner product kernel $K(x_i, x_{i'})$.

SVM – Discussion

- Multi-class problems:
 - A classifier for each pair of classes. Final classifier is the one that dominates the most.
 - Multinomial loss function with a suitable kernel.
- Structural Risk Minimization
 - Suppose training points in a sphere of radius R .
 - Let $G(x) = \text{sign}[f(x)] = \text{sign}[\beta^T x + \beta_0]$.
 - $\{G(x), \|\beta\| \leq A\}$ has VC-dimension h satisfying $h \leq R^2 A^2$.
 - If $f(x)$ separates the training data optimally for $\|\beta\| \leq A$, then with probability at least $1 - \eta$ over training sets:
$$\text{Error}_{\text{Test}} \leq \frac{h(\log(2N/h)+1) - \log(\eta/4)}{N}.$$
 - Regularization parameter γ could be chosen following the SRM paradigm.

You are here

- Part I – SVMs
 - Support Vector Classifier
 - Support Vector Machines
- **Part II – Flexible Discriminants**
 - Flexible Discriminant Analysis
 - Penalized Discriminant Analysis
 - Mixture Discriminant Analysis

Generalizing LDA

- Flexible Discriminant Analysis (FDA):
 - Recast LDA as linear regression.
 - Enlarged set of predictors via basis expansions.
 - FDA \rightarrow LDA in this enlarged space.
- Penalized Discriminant Analysis (PDA):
 - Used in case of too many (correlated) predictors.
 - Fit an LDA model but penalize its coefficients.
- Mixture Discriminant Analysis (MDA):
 - Model each class by a mixture of Gaussians with different centroids.
 - Every component Gaussian shares the same covariance matrix, both within and between classes.
 - Allows for subspace reduction as in LDA.

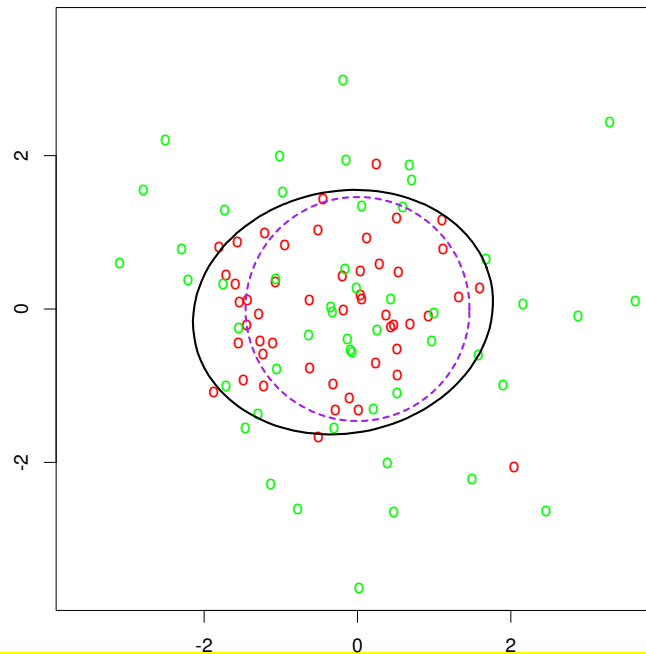
FDA – background

- LDA using linear regression on derived responses
 - Observations with responses $\mathcal{G} = \{1, \dots, K\}$.
 - $\theta : \mathcal{G} \rightarrow \mathbb{R}^1$ assigns scores to the classes.
 - Training sample $(g_i, x_i), i = 1, \dots, N$.
 - Solve $\min_{\beta, \theta} \sum_{i=1}^N (\theta(g_i) - x_i^T \beta)^2$.
- Discriminant vectors ν_ℓ of reduced-rank LDA
 - Independent scorings $\theta_1, \dots, \theta_L, L \leq K - 1$, and
 - L linear maps $\eta_\ell(X) = X^T \beta_\ell$, optimal for multiple regression in \mathbb{R}^p .
 - $\theta_\ell(g), \beta_\ell$ to minimize $\frac{1}{N} \sum_{\ell=1}^L \left[\sum_{i=1}^N (\theta_\ell(g_i) - x_i^T \beta_\ell)^2 \right]$.
 - Sequence of $\nu_\ell \iff$ sequence β_ℓ up to a constant.

FDA

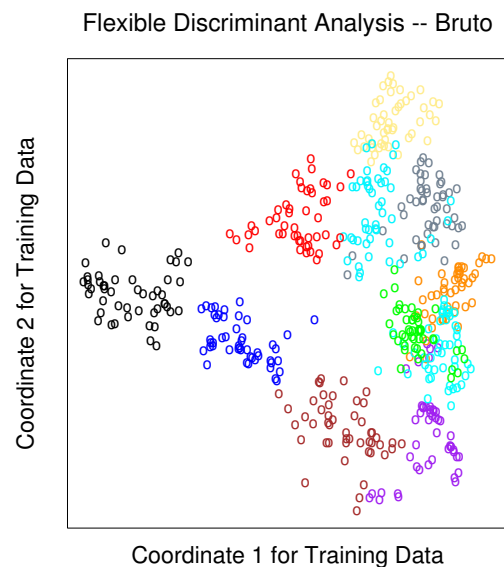
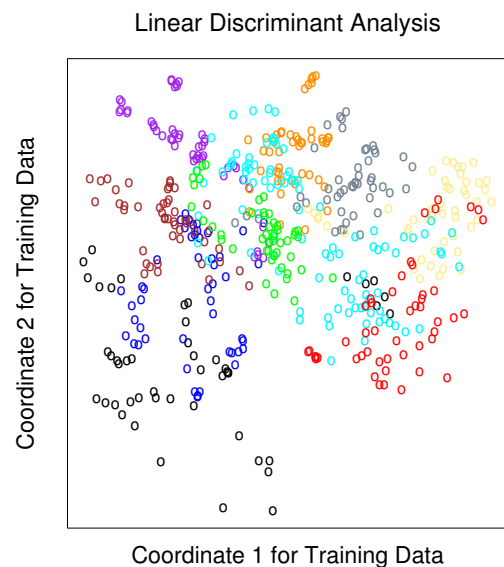
- Replace linear regression fits $\eta_\ell(x) = x^T \beta_\ell$ by more flexible nonparametric fits. The new criterion:

$$ASR(\{\theta_\ell, \eta_\ell\}_{\ell=1}^L) = \frac{1}{N} \sum_{\ell=1}^L \left[\sum_{i=1}^N (\theta_\ell(g_i) - \eta_\ell(x_i))^2 + \lambda J(\eta_\ell) \right].$$



Computing FDA

- Computations can be simplified in certain cases.
- Suppose the nonparametric regression procedure can be represented as a linear operator S_λ .
- Create an $N \times K$ indicator response matrix \mathbf{Y} such that $y_{ik} = 1$ if $g_i = k$ and $y_{ik} = 0$ otherwise.



Computing FDA – Steps

- *Multivariate nonparametric regression*
 - Let S_λ be the linear operator that fits the model and $\eta^*(x)$ be the vector of fitted regression functions.
- *Optimal scores*
 - Compute eigen-decomposition of $Y^T \hat{Y} = Y^T S_\lambda Y$, where the eigen vectors Θ are normalized: $\Theta^T D_\pi \Theta = I$.
 - $D_\pi = Y^T Y / N$ is a diagonal matrix of the estimated class priors.
- *Update* the model using optimal scores $\eta(x) = \Theta^T \eta^*(x)$.
- The first function in $\eta(x)$ is constant – rest are discriminant functions.

Penalized Discriminant Analysis

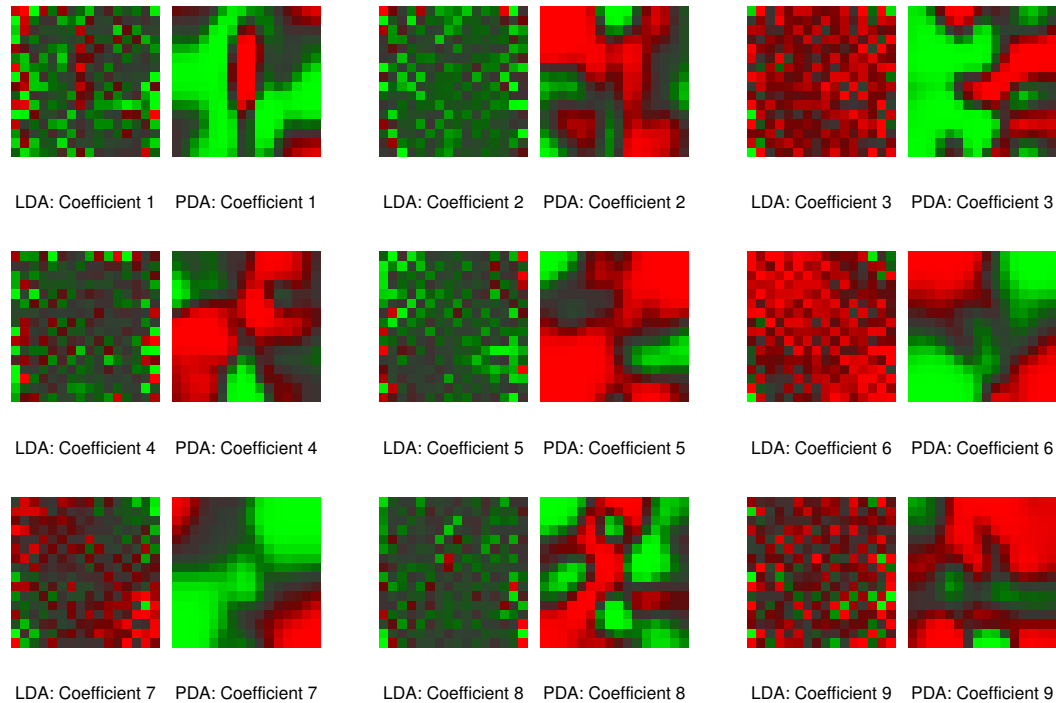
Linear regression onto a basis expansion $h(X)$ with a quadratic penalty on the coefficients:

$$ASR(\{\theta_\ell, \beta_\ell\}_{\ell=1}^L) = \frac{1}{N} \sum_{\ell=1}^L \left[\sum_{i=1}^N (\theta_\ell(g_i) - h^T(x_i)\beta_\ell)^2 + \lambda \beta_\ell^T \Omega \beta_\ell \right].$$

- Enlarge the set of predictors X via a basis expansion $h(X)$.
- Use (penalized) LDA in the enlarged space.
 - The penalized Mahalanobis distance
$$D(x, \mu) = (h(x) - h(\mu))^T (\Sigma_W + \Omega)^{-1} (h(x) - h(\mu)).$$
 - Σ_W is the within-class covariance matrix of $h(x_i)$.
- Decompose the classification subspace using a penalized metric: $\max u^T \Sigma_{\text{Bet}} u$ subject to $u^T (\Sigma_W + \Omega) u = 1$.

PDA – example

- For some problems, first step is not required.
- Example: digitized analog signals.
- LDA - *salt and pepper* images; PDA - smooth images.



Mixture Discriminant Analysis

- Generalization of Gaussian mixture models via FDA and PDA.

- Gaussian mixture model for k th class has density

$$P(X|G = k) = \sum_{r=1}^{R_k} \pi_{kr} \phi(X; \mu_{kr}, \Sigma).$$

- Class posterior probabilities

$$P(G = k|X = x) = \frac{\sum_{r=1}^{R_k} \pi_{kr} \phi(X; \mu_{kr}, \Sigma) \Pi_k}{\sum_{\ell=1}^K \sum_{r=1}^{R_\ell} \pi_{\ell r} \phi(X; \mu_{\ell r}, \Sigma) \Pi_\ell}.$$

- Estimate parameters by maximum likelihood, using the joint log-likelihood based on $P(G, X)$:

$$\sum_{k=1}^K \sum_{g_i=k} \log \left[\sum_{r=1}^{R_k} \pi_{kr} \phi(x_i; \mu_{kr}, \Sigma) \Pi_k \right]$$

- Use EM algorithm for MLEs.

Mixture Discriminant Analysis

- E-step: Several strategies. As an example,
 - Fit a k -means model with multiple random starts within class k .
 - Create an initial weight matrix from the R_k disjoint partitions.
- M-step: weighted LDA with $R = \sum_{k=1}^K R_k$ classes.
 - We can use optimal scoring to solve weighted LDA.
 - Enlarged indicator \mathbf{Y} matrix collapses to a *blurred* response matrix \mathbf{Z} .
 - Remaining steps:
 - $\hat{\mathbf{Z}} = \mathbf{S}\mathbf{Z}$
 - $\mathbf{Z}^T \hat{\mathbf{Z}} = \mathbf{\Theta}\mathbf{D}\mathbf{\Theta}^T$
 - Update π s and Π s.

Computational Considerations

N training cases, p predictors, m support vectors.

- SVM

- $m^3 + mN + mpN$, assuming $m \equiv N$.

- Does not scale well with N .

- LDA and PDA – $Np^2 + p^3$.

- FDA – depends on the regression method.

- Additive models and MARS: linear in N .

- Splines and kernel methods: typically N^3 .

References

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer-Verlag.