Linear Classifiers

Dept. Computer Science & Engineering,
Shanghai Jiao Tong University

Classification Problems

- Digital Recognition/ Plate number Recognition
- Face ID Recognition
- Speech ID Recognition
- Finger Print Recognition
- Spam Mail Detection
- Disease Diagnosis
- ……

Outline

- Linear Regression
- Linear and Quadratic Discriminant Functions
- Reduced Rank Linear Discriminant Analysis
- Logistic Regression
- Separating Hyperplanes

ODE for Linear Classification

- How to formulate a classification problem
  - Posterior maximum
- How to regularize the classification problem
- How to find an appropriate dimension for classification?
- How to find a classification model with good generalization?

Linear Regression

Indicator response matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_1^T \\ 1 & x_2^T \\ 1 & x_3^T \\ & \vdots \\ 1 & x_N^T \end{pmatrix}, \quad g = \begin{pmatrix} 3 \\ 1 \\ 4 \\ \vdots \\ 2 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & & & \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\hat{f}(x) = \hat{\beta}^T x = \left(\hat{f}_1(x), \hat{f}_2(x), \cdots, \hat{f}_K(x)\right)^T$$

Linear Regression

- The response classes are coded by a indicator variable. A K classes depend on K indicator variables, as $\boldsymbol{y}(k)$, $k=1,2,\ldots,K$, each indicates a class. And N training instances of the indicator vector could form a indicator response matrix $\boldsymbol{y}$.

- To the Data set $\boldsymbol{x}(\mathrm{k})$, there is a mapping:

$$f: \quad \mathbf{x}(k) \rightarrow \mathbf{y}(k)$$

- According to the linear regression model:

$$f(X) = (1, X^T)\hat{\beta}$$

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Linear Regression

- Given X, the classification should be:

$$\hat{G}(x) = \arg\max_{k \in G} \hat{f}_k(x)$$

- In another form, a target t_k is constructed for each class, the t_k presents the k-th column of a K identity matrix, according to the a sum-of-squared-norm criterion :

$$\min_B \sum_{i=1}^{N} \left\| y_i - \left[\left(1, x_i^T\right) B \right]^T \right\|^2$$

and the classification is:

$$\hat{G}(x) = \arg\min_{k \in G} \left\| \hat{f}(x) - t_k \right\|^2$$

Problems of the linear regression

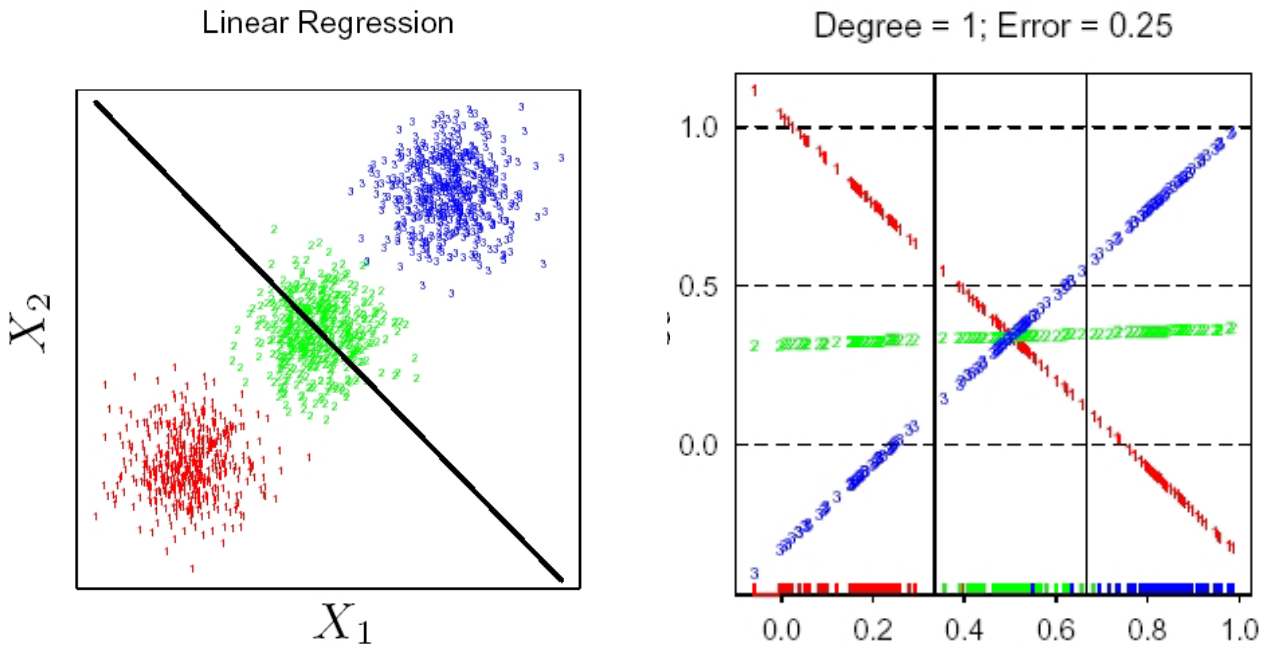


Figure 4.2: *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indica-*

- The data come from three classes in IR^2 and easily separated by linear decision boundaries.
- The left plot shows the boundaries found by linear regression of indicator response variables.
- The middle class is completely masked .

The rug plot at bottom indicates the positions and the class membership of each observations. The 3 curves are the fitted regressions to the 3-class indicator variables.

Problems of the linear regression

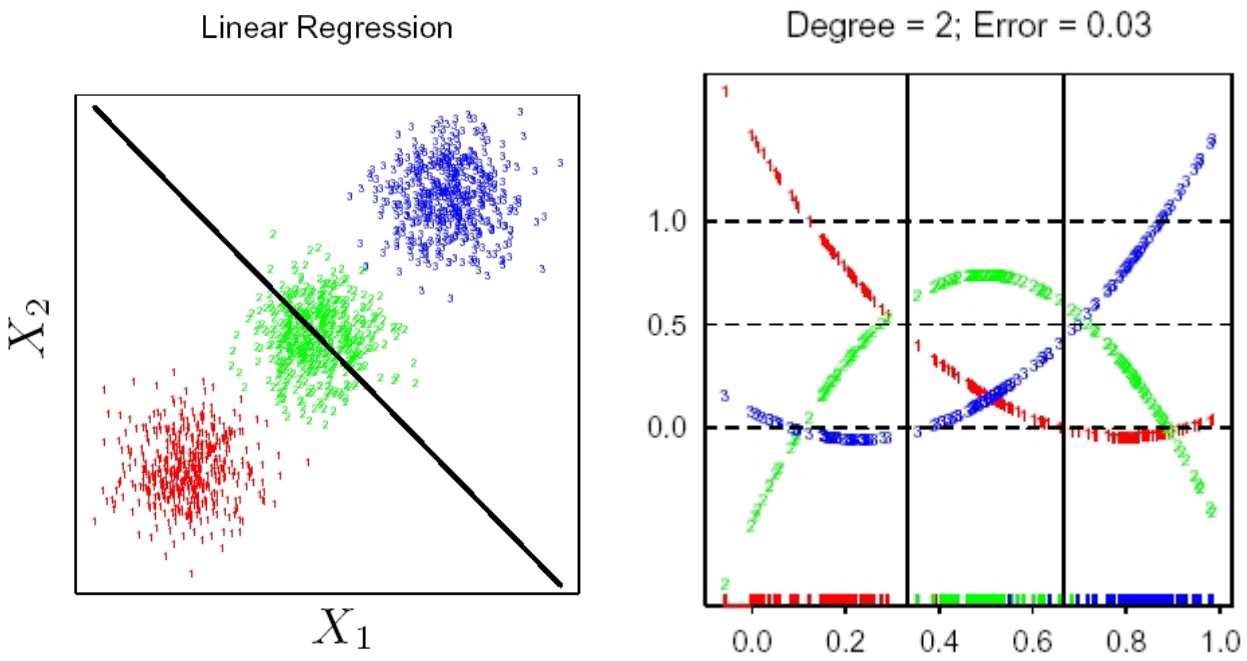


Figure 4.2: *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indica-*

- The left plot shows the boundaries found by linear discriminant analysis. And the right shows the fitted regressions to the 3-class indicator variables.

Outline

- Linear Regression
- Linear and Quadratic Discriminant Functions
- Reduced Rank Linear Discriminant Analysis
- Logistic Regression
- Separating Hyperplanes

Linear Discriminant Analysis

- According to the Bayes optimal classification mentioned in chapter 2, the posteriors is needed.

  post probability : $\Pr(G \mid X)$

  Assume:

  $f_k(x)$ —— condition-density of $\mathbf{X}$ in class $G=k$.

  π_k—— prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$

  Bayes theorem give us the discriminant:

$$\Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Linear Discriminant Analysis

- Multivariate Gaussian density:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

- Comparing *pdfs* of two classes k and l , assume $\Sigma_k = \Sigma, \forall k$

$$\log \frac{\Pr(G=k \mid X=x)}{\Pr(G=l \mid X=x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)\Sigma^{-1}(\mu_k - \mu_l)$$

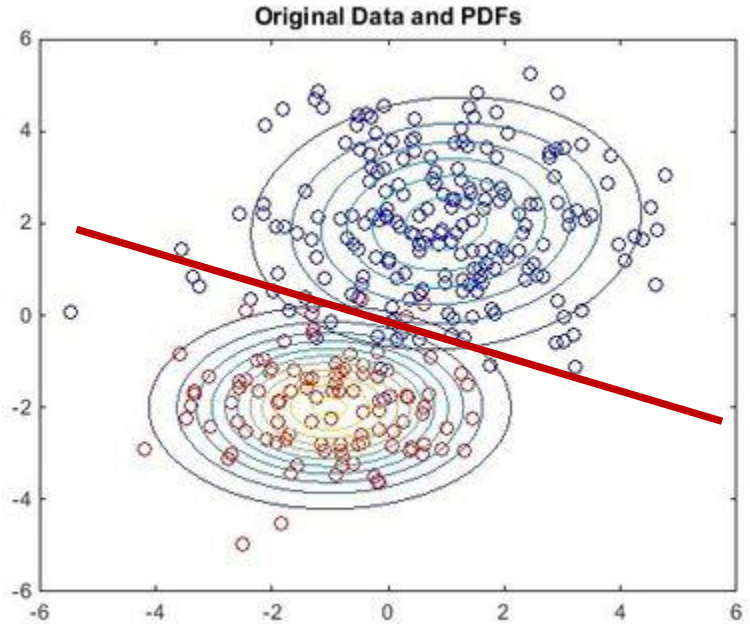
$$+ x^T \Sigma^{-1}(\mu_k - \mu_l)$$

Linear Discriminant Analysis

- The linear log-odds function above implies that the boundary of class k and l is linear in x in p dimension a hyperplane.
- Linear Discriminant Function:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k + \log \pi_k$$

So we estimate $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}$



Parameter Estimation

$$\hat{\pi}_k = \frac{N_k}{N}, N_k \text{ is the number of Class k data}$$

$$\hat{\mu}_k = \sum_{g_i = k} x_i / N_k;$$

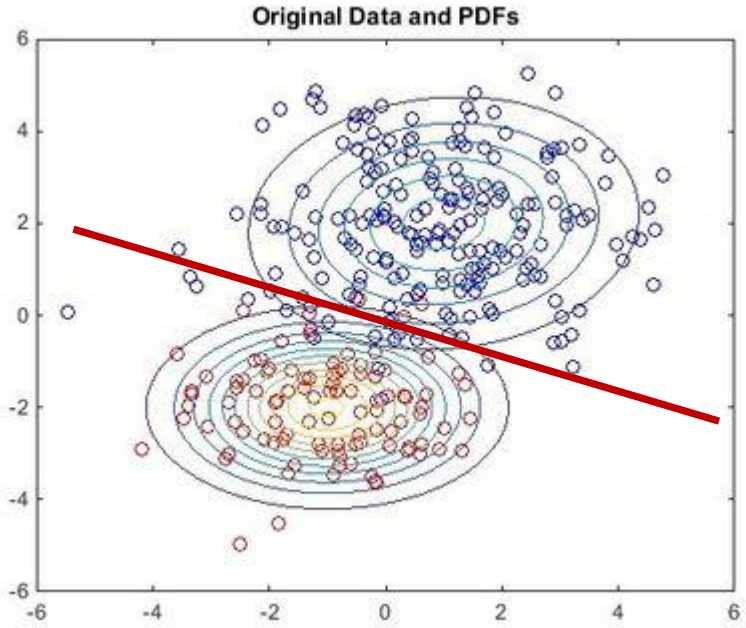
$$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T /(N - K).$$

LDA Rule

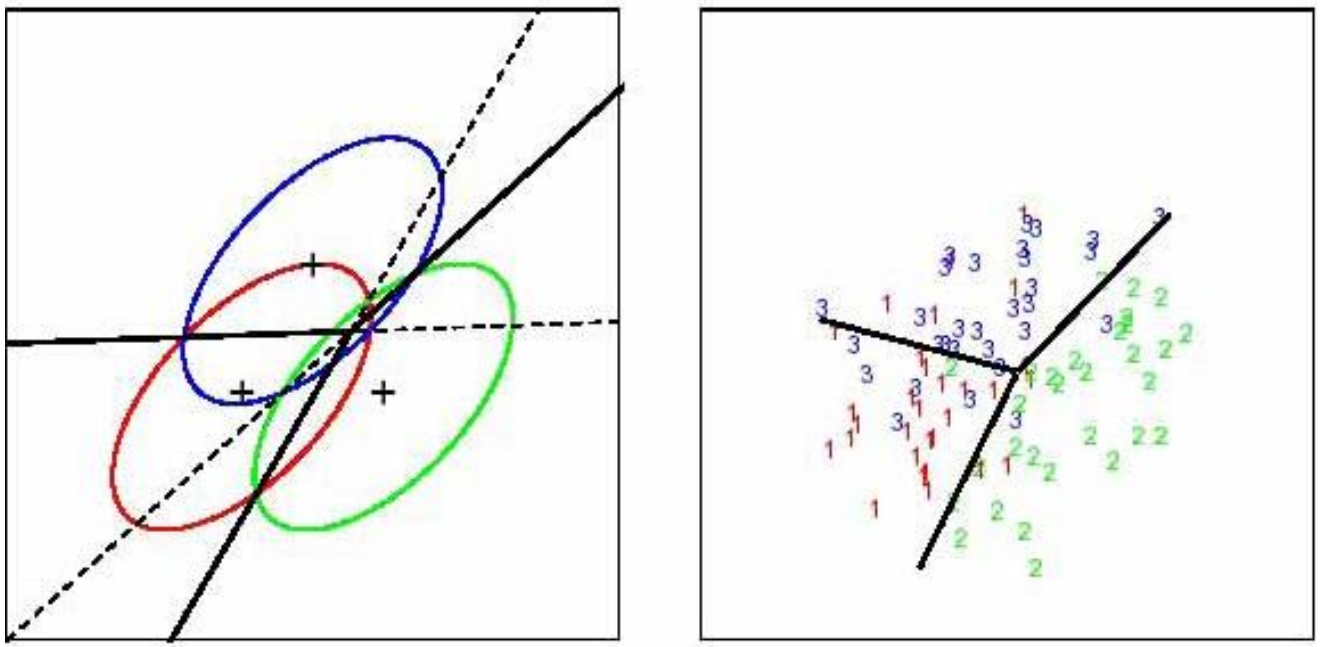
LDA rule: $g_k(x) = \text{var}\max_l \{\delta_l(x)\}$

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k + \log \pi_k$$

Decision boundary $\{ x \mid \delta_k(x) = \delta_l(x) \}$



Linear Discriminant Analysis



- Three Gaussian distribution with the same covariance and different means. The Bayes boundaries are shown on the left (solid lines). On the right is the fitted LDA boundaries on a sample of 30 drawn from each Gaussian distribution.

Quadratic Discriminant Analysis

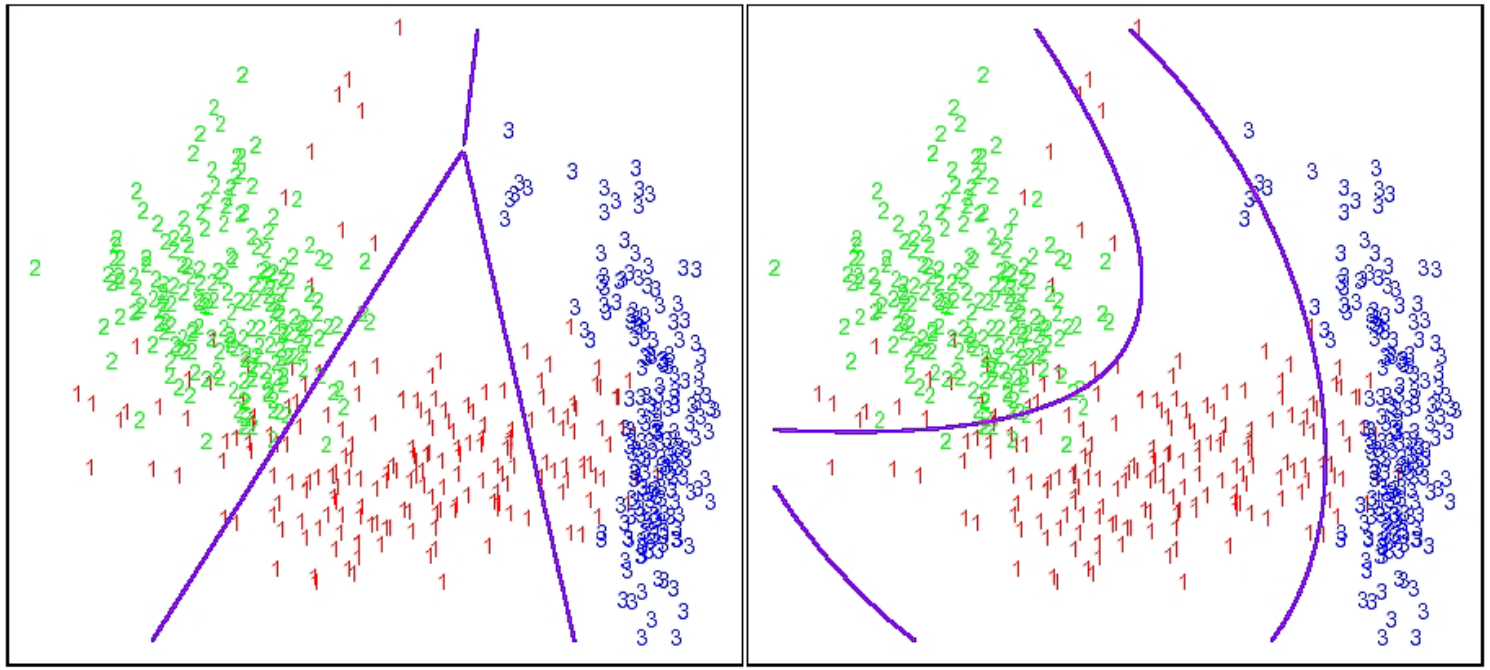
- When the covariances of Class k and l are different

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

- This is the Quadratic Discriminant Function
- The decision boundary is described by a quadratic equation

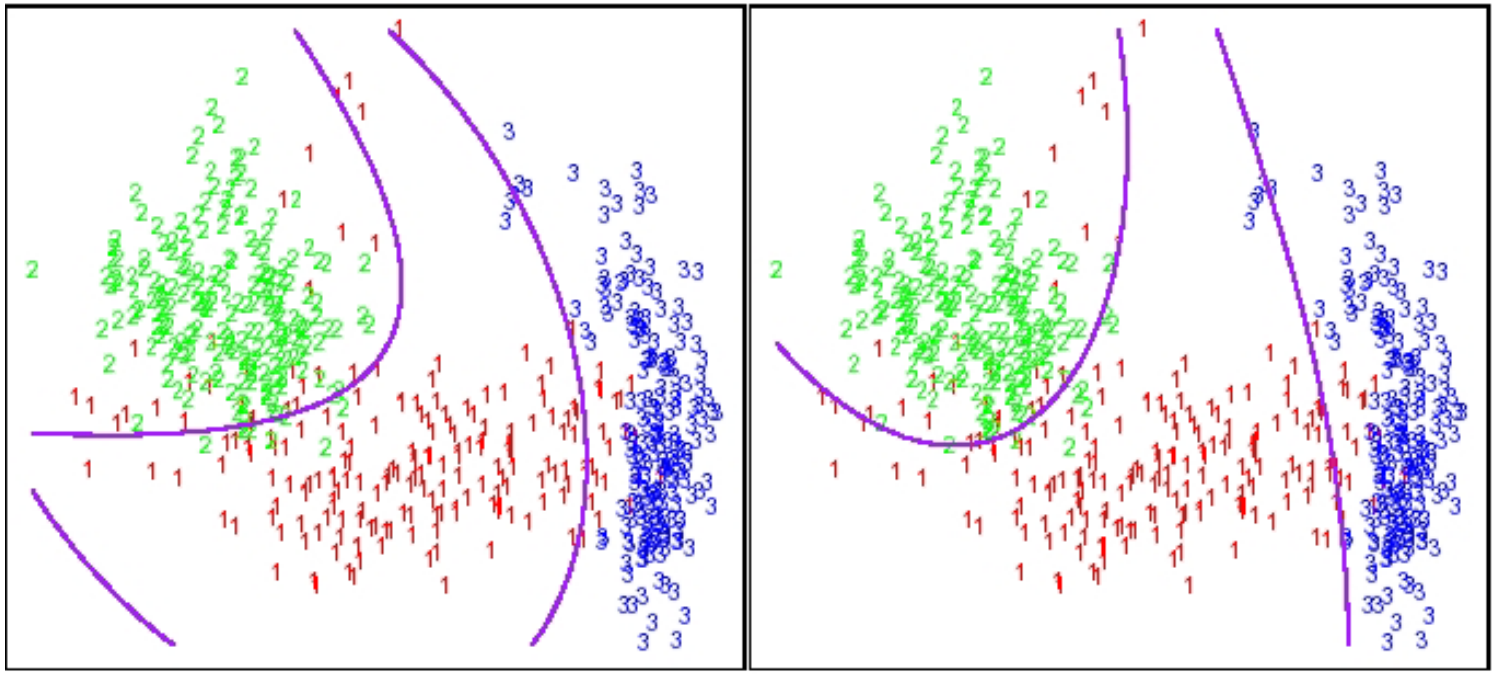
$$\{x : \delta_k(x) = \delta_l(x)\}$$

LDA & QDA



- Boundaries on a 3-classes problem found by both the linear discriminant analysis in the original 2-dimensional space $\mathbf{X_1}$, $\mathbf{X_2}$ (the left) and in a 5-dimensional space $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_{12}}$, $\mathbf{X_1^2}$, $\mathbf{X_2^2}$ (the right).

LDA & QDA



- Boundaries on the 3-classes problem found by LDA in the 5-dimensional space above (the left) and by Quadratic Discriminant Analysis (the right).

Regularized Discriminant Analysis

- Shrink the separate covariances of QDA toward a common covariance as in LDA.

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}, \quad \alpha \in [0,1]$$

Regularized QDA

- $\hat{\Sigma}$ was allowed to be shrunk toward the scalar covariance.

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2\mathrm{I}, \quad \gamma \in [0,1]$$

Regularized LDA

- Together :

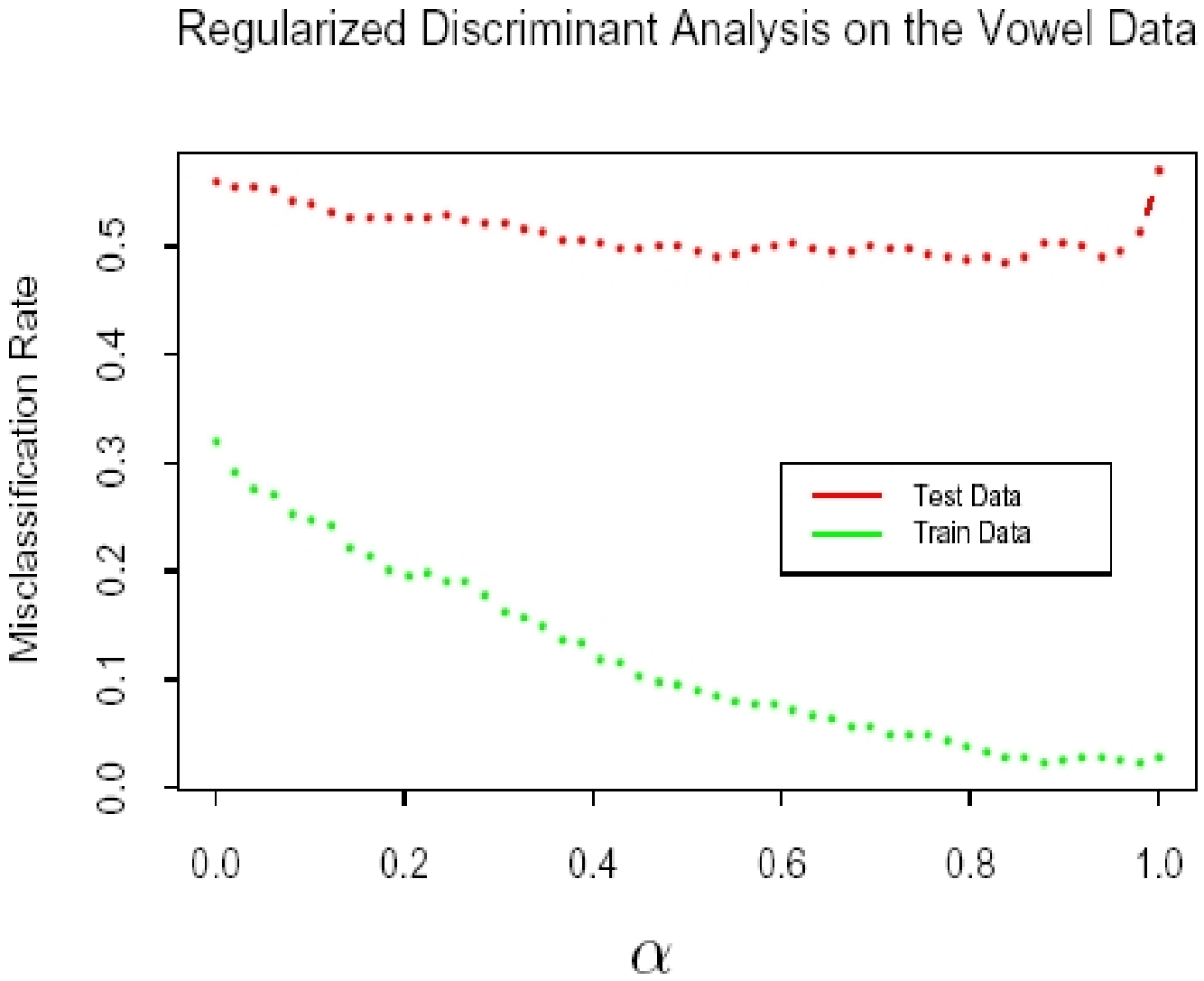
$$\hat{\Sigma}(\alpha, \gamma)$$

Regularized Discriminant Ana.

- Could use $$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\,\mathrm{diag}(\hat{\Sigma})$$
- In recent micro expression work, we can use

$$\delta_K(x) = \sum_{j=1}^{P} \frac{(x_j - \hat{u}'_{jk})^2}{S_j^2} - \frac{1}{2}\log \pi_K$$

where $\hat{u}'_{jk}$ in a “SHRUNKEN CENTROID”



- Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of $\alpha \in [0,1]$.
- The optimum for the test data occurs around $\alpha = 0.9$, close to quadratic discriminant analysis

Computations for LDA

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k) + \log\pi_k$$

- The eigen-decomposition for each $\hat{\Sigma}_k = U_k D_k U_k^T$ where U_k is $p \times p$ orthonormal, and D_k is a diagonal matrix of positive eigenvalues d_{kl}.
- So the ingredients for $\delta_k(x)$ are:

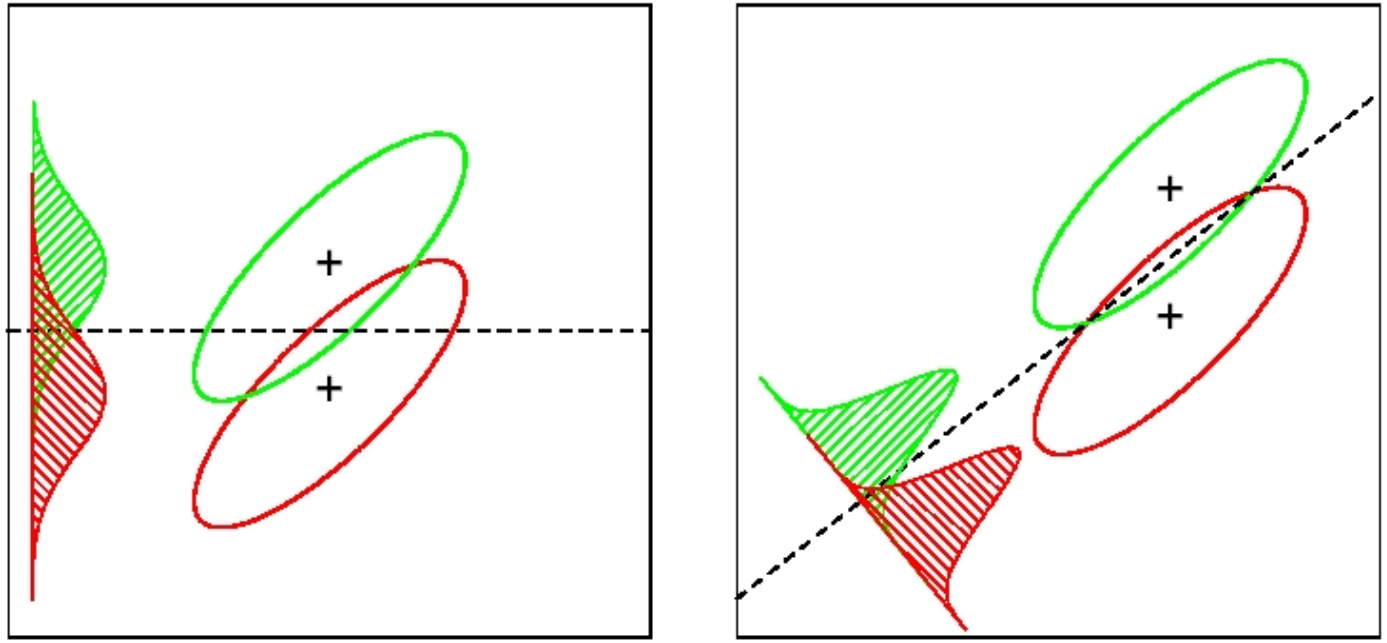
$$(x-\hat{\mu}_x)^T \hat{\Sigma}_k^{-1}(x-\hat{\mu}_x) = \left[U_k^T(x-\hat{\mu}_x)\right]^T D_k^{-1}\left[U_k^T(x-\hat{\mu}_x)\right]$$

$$\log\left|\hat{\Sigma}_k\right| = \sum_l \log d_{kl}$$

Outline

- Linear Regression
- Linear and Quadratic Discriminant Functions
- Reduced Rank Linear Discriminant Analysis
- Logistic Regression
- Separating Hyperplanes

Reduced Rank LDA



- Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel).
- The discriminant direction minimizes this overlap for Gaussian data (right panel).

Reduced Rank LDA

- Let $\hat{\Sigma} = UDU^T$
- Let $X^* = D^{-1/2}U^T X$ i.e. $\hat{\Sigma}^{-1/2} X$

  $\hat{U}_K^* = D^{-1/2}U^T\hat{U}_K$ $\hat{\Sigma}^{-1/2} U_K$
- LDA: $\delta_K(x) = \frac{1}{2}\left\|x^* - \hat{\mu}_K^*\right\|^2 - \log\hat{\pi}_K$
  - Closest centroid in sphered space(apart from $-\log\hat{\pi}_K$)
- Can project data onto K-1 dim subspace spanned by $\hat{U}_1^*, \cdots, \hat{U}_K^*$, and lose nothing!
- Can project even lower dim using principal components of $\hat{U}_K^*$, $k=1, \ldots, K$.

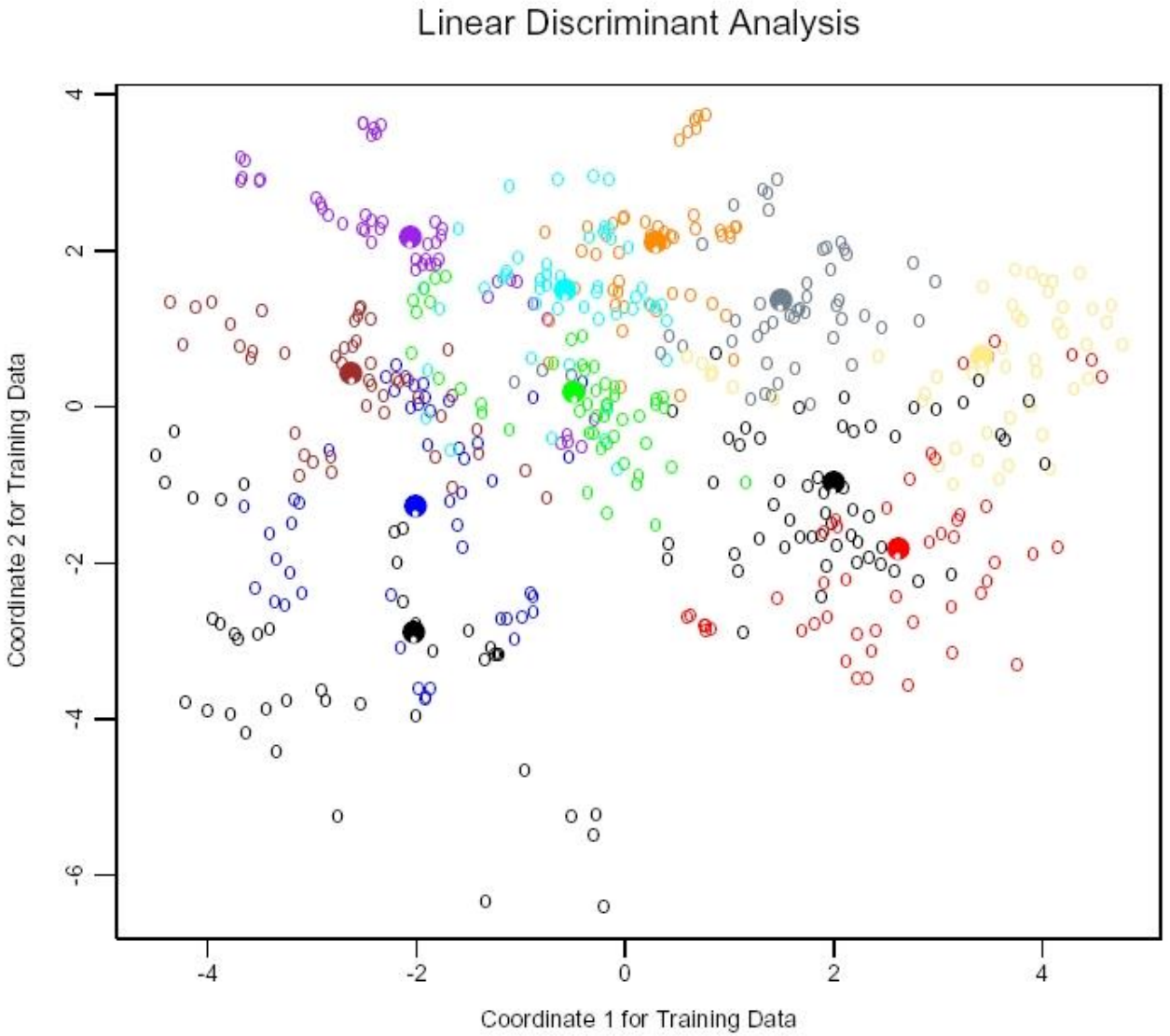
Reduced Rank LDA

- Compute $K \times p$ matrix M of centroids
- Compute $W = \hat{\Sigma}$, and $M^* = MW^{-1/2}$
- Compute B*, cov matrix of M* , and

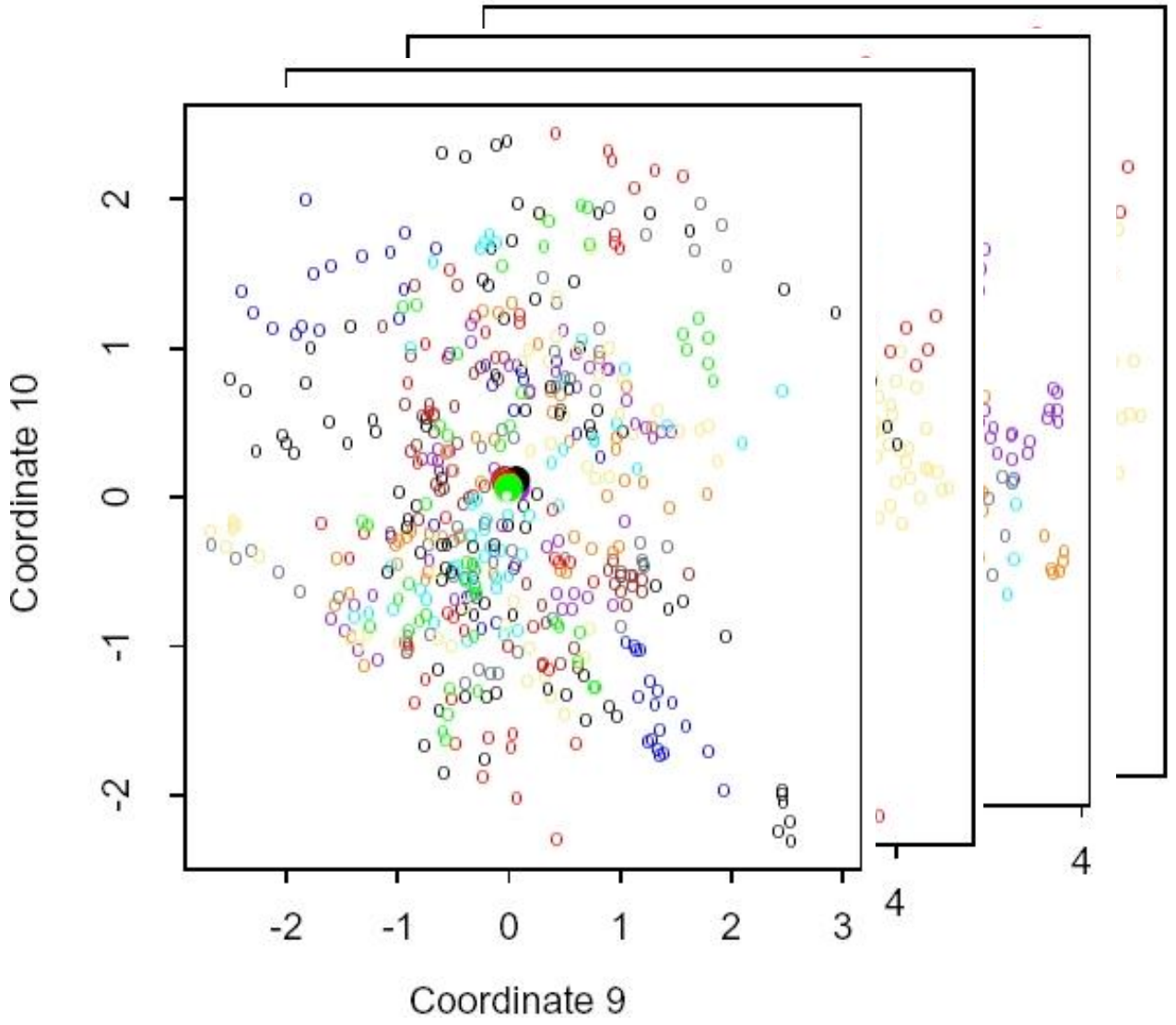
$$B^* = V^* D_B V^{*T}$$

- $Z_l = v_l^T X$ with $v_l = W^{-1/2} v_l^*$ is l-th discriminant variable
  (canonical variable)

- A two-dimensional plot of the vowel training data. There are eleven classes with $X \in \mathrm{IR}^{10}$,and this is the best view in terms of a LDA model. The heavy circles are the projected mean vectors for each class.



- Projections onto different pairs of canonical varieties



Fisher's problem

- Find $Z = a^T X \quad s.t.$ "between-class var" is maximized relative to "within-class var".

- Maximize "Rayleigh quotient" :

$$\max\left(\frac{a^T B a}{a^T W a}\right)$$

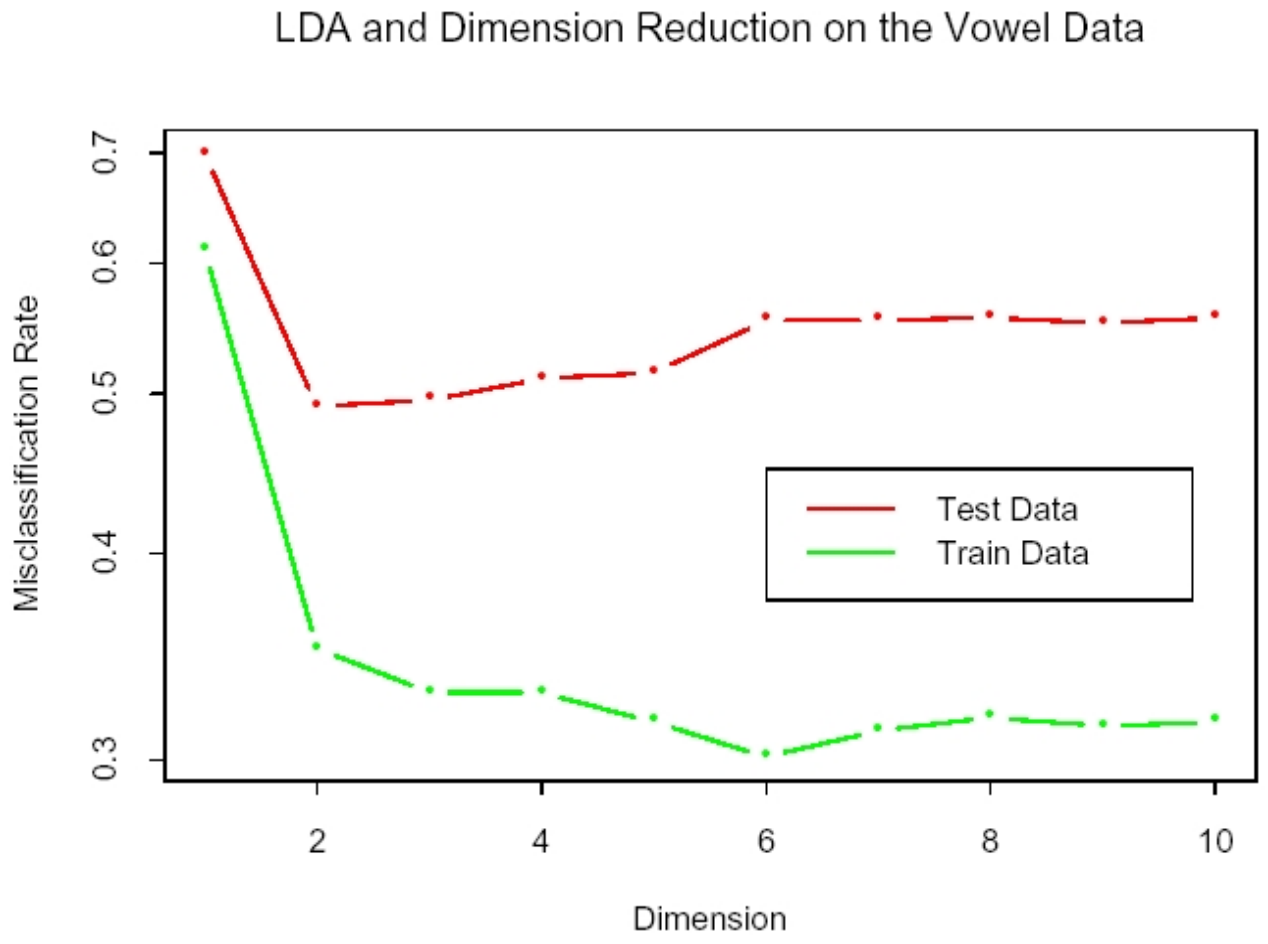
$$\max(a^T B a), \quad s.t.\ a^T W a = 1$$

$a = v_1$!

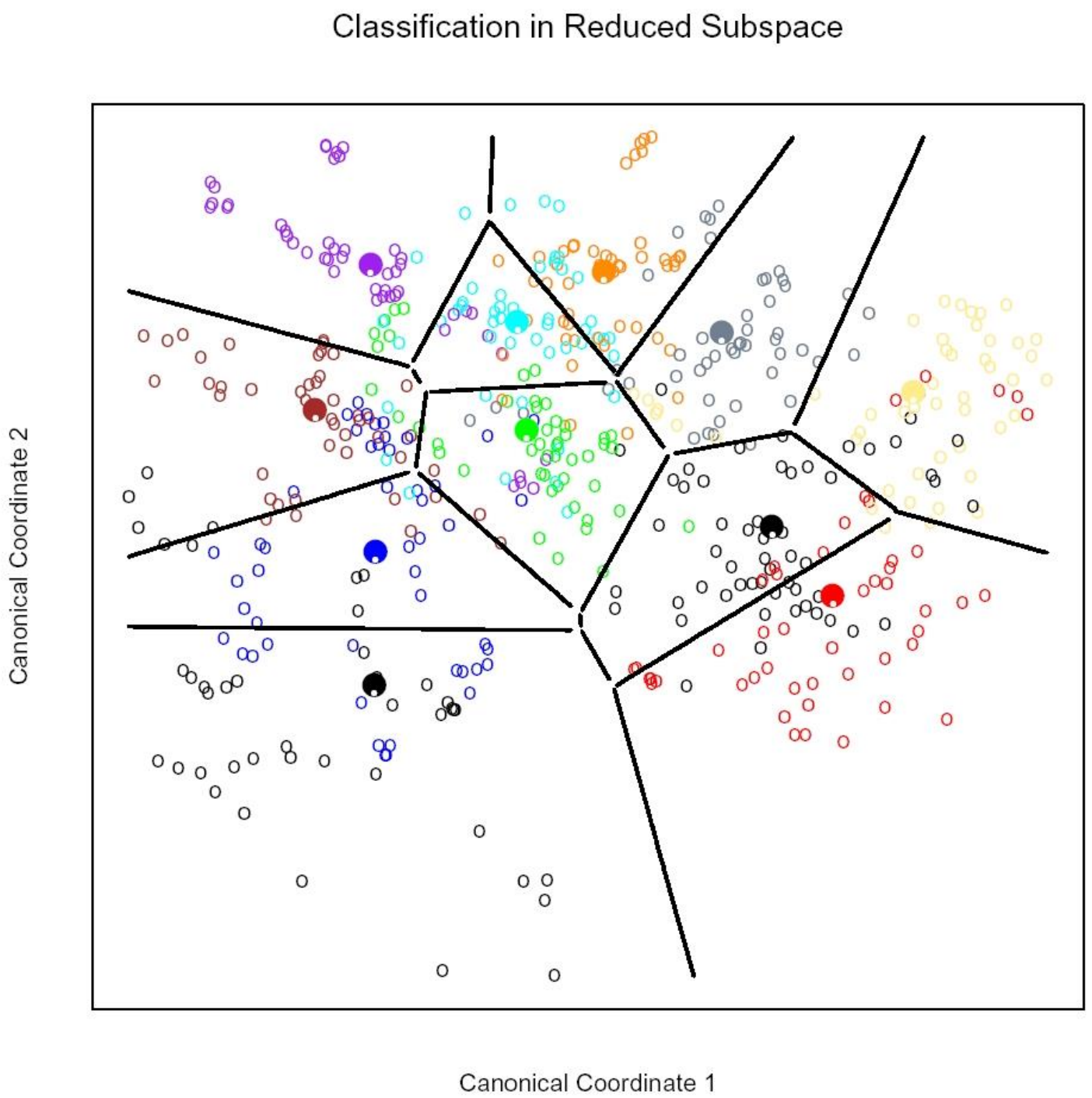
$$\rightarrow \max(a_2{}^T B a_2), \quad s.t.\ a_2{}^T W a_2 = 1$$

$$a_2{}^T W a_1 = 0$$

$a = v_2$ etc.



- Training and test error rates for the vowel data, as a function of the dimension of the discriminant subspace.
- In this case the best rate is for dimension 2.


Classification in Reduced Subspace
Canonical Coordinate 2
Canonical Coordinate 1

Outline

- Linear Regression
- Linear and Quadratic Discriminant Functions
- Reduced Rank Linear Discriminant Analysis
- Logistic Regression
- Separating Hyperplanes

Logistic Regression

- Model:

$$Log\frac{\Pr(g=1|X=x)}{\Pr(g=k|X=x)}=\beta_{1,0}+\beta_1^T x$$

$$\vdots$$

$$Log\frac{\Pr(g=K-1|X=x)}{\Pr(g=K|X=x)}=\beta_{K-1,0}+\beta_{K-1}^T x$$

Log Likelihood :

$$L(\theta)=\sum_{i=1}^{n}\log P_{gi}(x_i;\theta)$$

$$\Pr(g=k|X=x)=\frac{\exp(\beta_{k,0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l,0}+\beta_l^T x)},\ k=1,\cdots,K-1$$

$$\Pr(g=K|X=x)=\frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l,0}+\beta_l^T x)}$$

Logistic Regression 2

- Parameters estimation
  - Objective function

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{N} \log \Pr_{\beta}(y_i \mid x_i)$$

  - Parameters estimation

IRLS (iteratively reweighted least squares)

Particularly, for two-class case, using Newton-Raphson algorithm to solve the equation, the objective function:

$$p(x,\beta) = Pr_{\beta}(y = 1 \mid x); \quad Pr_{\beta}(y = 0 \mid x) = 1 - p(x,\beta)$$

$$p(x,\beta) = \Pr_{\beta}(y = 1 \mid x) = \frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)};$$

$$l(\beta) = \sum_{i=1}^{N} y_i \log p(x_i,\beta) + (1 - y_i)\log(1 - p(x_i,\beta))$$

Logistic Regression 3

$$p(x,\beta) = \Pr_\beta(y=1 \mid x) = \frac{\exp(\beta^T x)}{1+\exp(\beta^T x)};$$

$$l(\beta) = \sum_{i=1}^{N} y_i \log p(x_i,\beta) + (1-y_i)\log(1-p(x_i,\beta))$$

$$= \sum_{i=1}^{N} \{y_i \beta^T x_i - \log(1+e^{\beta^T x_i})\}$$

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = 0$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1-p(x_i;\beta))$$

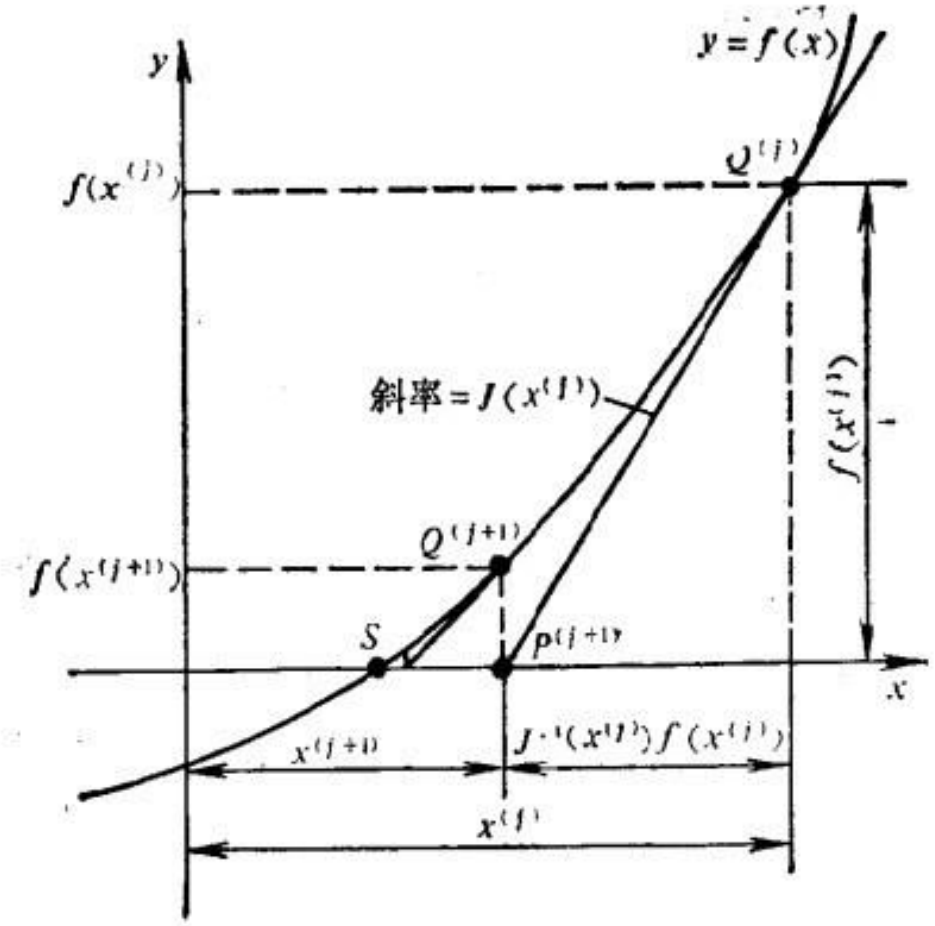
Logistic Regression 4

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = 0$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1 - p(x_i;\beta))$$

The Newton Iterative method

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta}$$



一维牛顿-拉夫森算法的几何解释

Logistic Regression 5

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = X^T(\mathbf{y} - \mathbf{p})$$

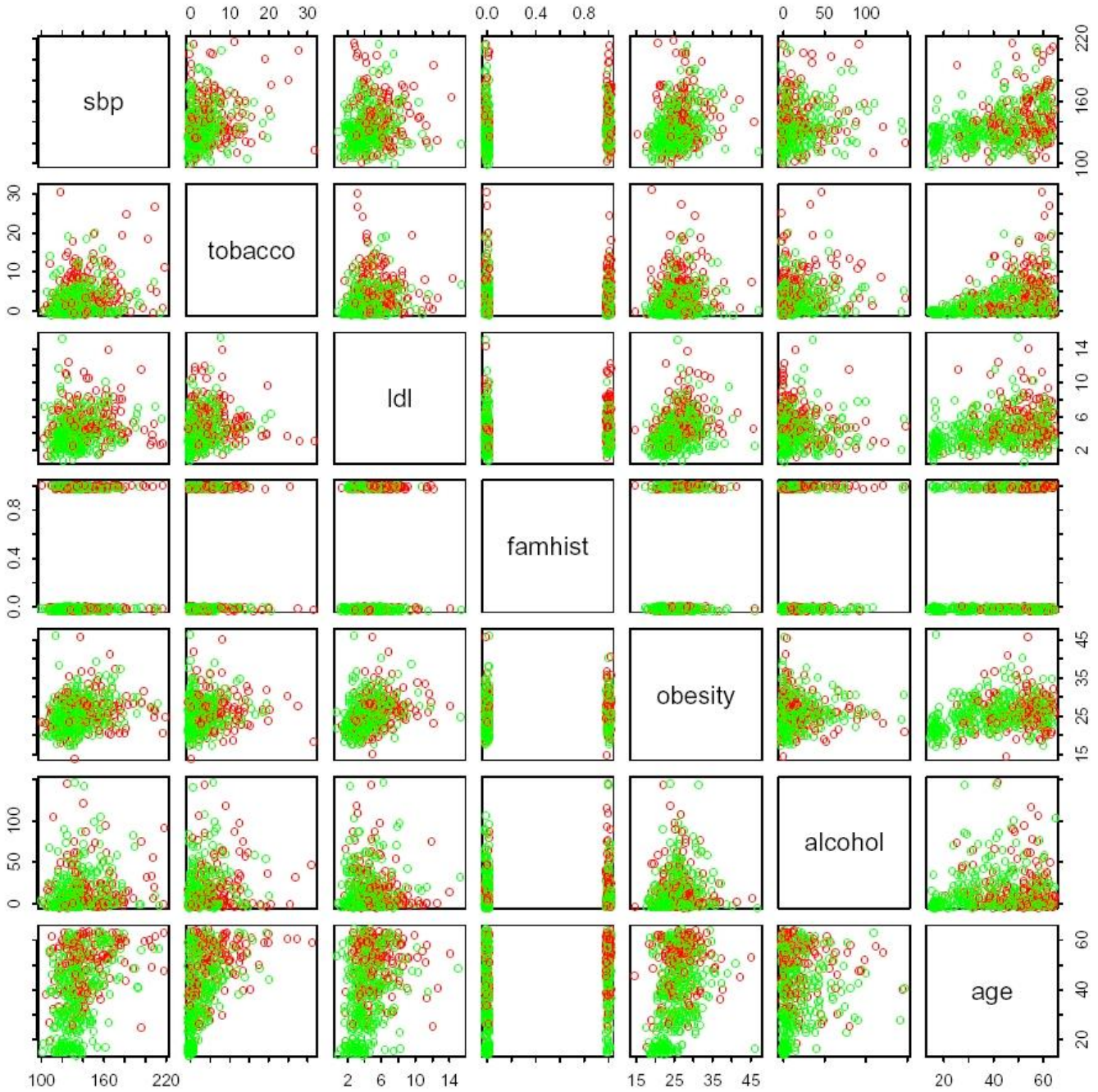
$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1 - p(x_i;\beta)) = -X^T W X$$

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta}$$

$$\beta^{new} = \beta^{old} + (X^T W X)^{-1} X^T (y - p) = (X^T W X)^{-1} X^T W z$$

$$z = X\beta^{old} + W^{-1}(y - p),$$

$$p_i = p(x_i;\beta^{old}), W_i = p_i(1 - p_i),$$



South African Heart Disease Data

ldl=low density lipoprotein

sbp=Systolic blood pressure

Results from a logistic regression fit to the South African heart disease data.

	Coefficient	Std. Error	*Z* Score
(Intercept)	-4.130	0.964	-4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	-1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

ldl=low density lipoprotein

sbp=Systolic blood pressure

Logistic Regression vs LDA

$$\log\frac{\Pr(g=k|X=x)}{\Pr(g=K|X=x)}=\log\frac{\pi_k}{\pi_K}-\frac{1}{2}(\mu_k+\mu_K)^T\Sigma^{-1}(\mu_k-\mu_K)$$
$$+x^T\Sigma^{-1}(\mu_k-\mu_K)$$
$$=\alpha_{k0}+\alpha_k^T x$$

Same form as Logistic Regression

$$LR:\Pr(X,g=k)=\Pr(X)\Pr(g=k|X=x)$$

Conditional Likelihood

Probability

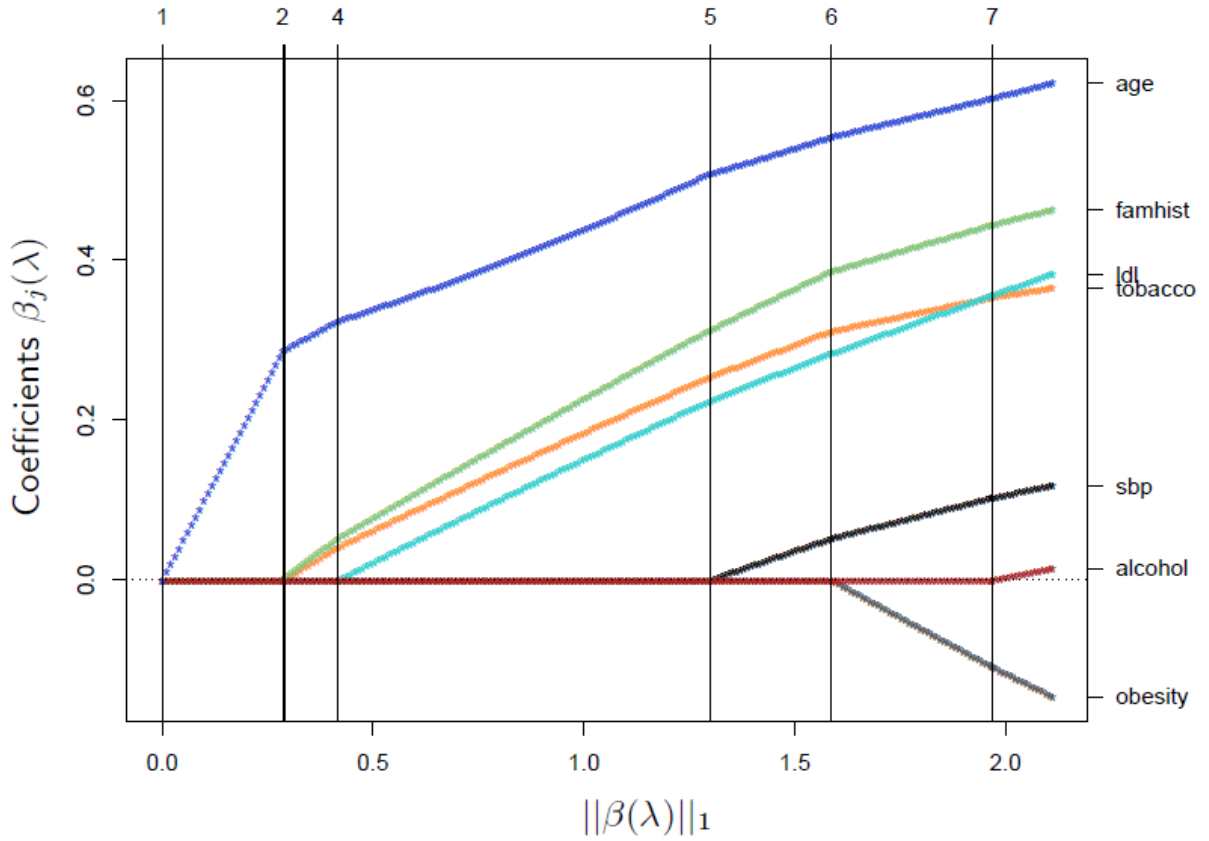
$$\frac{\exp(\beta_{k,0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l,0}+\beta_l^T x)}$$

$$LDA:\Pr(X,g=k)=\phi(X;\mu_k,\Sigma)\pi_k$$

$$\Pr(X)=\sum_{k=1}^{K}\pi_k\phi(X;\mu_k,\Sigma)$$

L_1 Regularized Logistic Regression

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0+\beta^T x_i})\right] - \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$



Outline

- Linear Regression
- Linear and Quadratic Discriminant Functions
- Reduced Rank Linear Discriminant Analysis
- Logistic Regression
- Separating Hyperplanes
  - Problem: Is there an optimal classifier?

Rosenblatt's Perceptron Learning Algorithm

$$D(\beta,\beta_0) = -\sum_{i\in M} y_i(x_i^T\beta+\beta_0)$$

$$y_i \in \{-1,1\} \quad M = Miscassified\ Observations$$

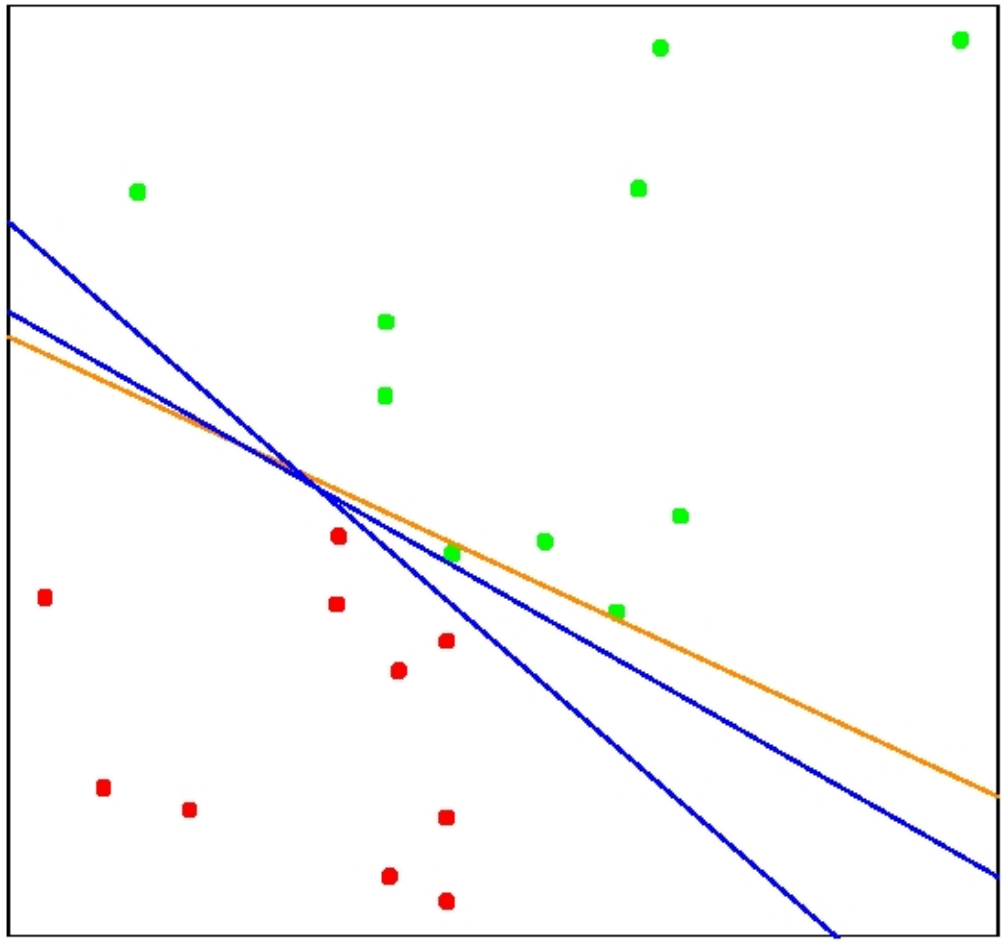
- $D(\beta,\beta_0)$--distance of points in M to boundary

  Stochastic Gradient Descent

$$\frac{\partial D}{\partial \beta} = -\sum_{i\in M} y_i x_i \qquad \frac{\partial D}{\partial \beta_0} = -\sum_{i\in M} y_i$$

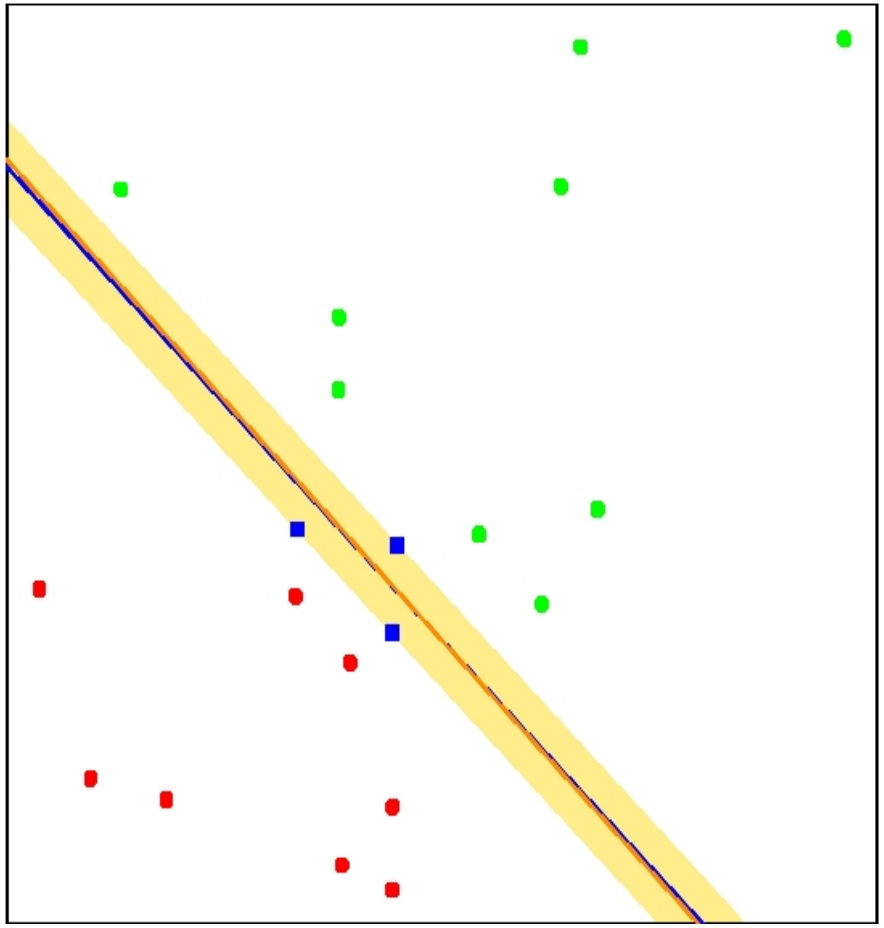
$$\begin{pmatrix}\beta\\ \beta_0\end{pmatrix} \leftarrow \begin{pmatrix}\beta\\ \beta_0\end{pmatrix} + \rho\begin{pmatrix}y_i x_i\\ y_i\end{pmatrix}$$

Separating Hyperplanes



- A toy example with two classes separable by hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the perceptron learning algorithm with different random starts.
- $$\{x : \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0\}$$

Optimal Separating Hyperplanes



$$\max_{\beta,\beta_0,\|\beta\|=1} C$$

subject to

$$y_i(x_i^T\beta+\beta_0)\geq C, i=1,\ldots,N$$

$$\hat{\beta} = \sum_{i=1}^{N}\hat{\alpha}_i y_i x_i$$

$\hat{\alpha}_i > 0$ if x_i on the boundary, else 0.

Such points are called *support points*

Review

- Linear Regression
- Linear and Quadratic Discriminant Functions
- Reduced Rank Linear Discriminant Analysis
- Logistic Regression
- Separating Hyperplanes

The End of the Talk

SVD for Matrix

- Assume $X \in R^{N \times p}$
- SVD ： $X = \mathrm{U} D V^T$, U is $\mathrm{N} \times \mathrm{N}$ orthognal matrix,

  V is $\mathrm{p} \times \mathrm{p}$ orthognal matrix

  $$D = \begin{pmatrix} \mathrm{diag}(d_1, \cdots, d_p) \\ \mathbf{0} \end{pmatrix}$$ is a $\mathrm{N} \times \mathrm{p}$ matrix
- To prove that the norm of regression solution $\hat{\beta} = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{y}$ will not increase as λ increases

SVD for Matrix

$$X = \mathrm{U} D V^T$$

$$\left\|\hat{\beta}\right\|^2 = \hat{\beta}^T \hat{\beta} = \mathbf{y}^T \mathbf{X} \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

$$\mathbf{X} = \mathrm{U} D V^T, \quad \mathbf{X}^T \mathbf{X} = V D^T D V^T; \quad \mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} = V \left(D^T D + \lambda \mathbf{I}\right) V^T$$

$$\left\|\hat{\beta}\right\|^2 = \mathbf{y}^T U \left[D \left(D^T D + \lambda \mathbf{I}\right)^{-1} D^T \right] U^T \mathbf{y}$$