Linear Regression and Linear Discriminant Analysis

Key Reference:

The Elements of Statistical Learning–Data Mining, Inference and Prediction by Hastie, Tibshirani and Friedman

Course Instructor: Shuen-Lin Jeng

Department of Statistics
National Cheng Kung University

July, 31, 2018

$+\epsilon$ 一小步也是一大步

- A relationship model of Y and X

$$Y = f(X),$$

where Y is a real number or a class label, f is a function and $X = (X_1, X_2, ..., X_p)$ is a p dimensional vector.

- A statistical model

$$Y = f(X) + \epsilon \qquad (E1)$$

where Y, X, and ϵ are random variables. A typical assumption of ϵ is $E(\epsilon) = 0$.

Linear Regression Model

- A linear regression model assumes that the regression function $f(X) = E(Y|X)$ is linear in the inputs $X_1, \ldots, X_p$.

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{E2}$$

- In statistical literature, Y is called the dependent or response variable, and the $X_1, \ldots, X_p$ are called the independent variables, explanatory variables, regressors, or predictors.
- In machine learning and pattern recognition, Y is called the output, and the $X_1, \ldots, X_p$ are called the inputs or features.

Linear Regression Model

- $$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \qquad (E2)$$
- For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.
- An understanding of linear methods is essential for understanding nonlinear ones.

Linear Regression Function

- $$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \qquad (E2)$$

  the variables X_j can come from different sources:

Estimation by Least Squares

- The most popular estimation method is *least squares* of residuals sum of squares (RSS),

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N}(y_i - f(X_i))^2 \\ &= \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \qquad (E3) \end{aligned}$$

- The criterion measures the average lack of fit.

Note what assumptions (E3) makes about:

- the validity of model (E2)?
- the distribution of ϵ?
- the correlation between $y_i s$?

Estimation by Least Squares

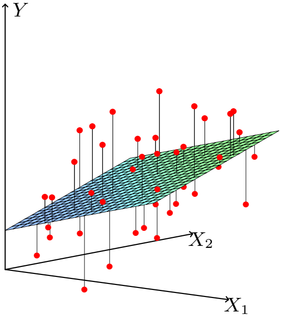

FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Estimation by Least Squares

- $$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (E5)$$

- $$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (E6)$$

  The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ appearing in equation (E6) is sometimes called the "projection" matrix because it projects $\mathbf{y}$ in the the space spanned by X.

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

Estimation by Least Squares

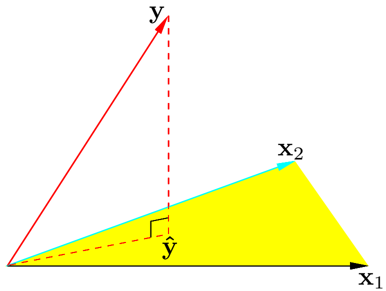

FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Estimation by Least Squares

- Figure 3.2 , We denote the column vectors of $\mathbf{X}$ by $x_0, x_1, \ldots, x_p$
  These vectors span a subspace of $\Re^N$, also referred to as the column space of $\mathbf{X}$. We minimize $RSS(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$ by choosing $\hat{\beta}$ so that the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to this subspace.
  $\hat{\mathbf{y}}$ is hence the *orthogonal projection* of $\mathbf{y}$ onto this subspace.
- When $\mathbf{X}$ is not of full rank, $\mathbf{X}^T\mathbf{X}$ is singular
  The fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ are still the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$;
  One may resolve the non-unique representation by .....

Inference of β

- In order to pin down the sampling properties of β, we now assume that the observations y_i are uncorrelated and have constant variance σ^2, and that the x_i are fixed (non random).

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The $N-p-1$ rather than N in the denominator makes $\hat{\sigma}^2$ an unbiased estimator.

Inference of β

- We also assume that the deviations of Y around its expectation are Gaussian, i.e. $\epsilon \sim N(0, \sigma^2)$. Then

$$\begin{aligned} Y &= \mathrm{E}(Y|X_1, \ldots, X_P) + \varepsilon \\ &= \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \end{aligned}$$

- $$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

- $$(N - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$$

In addition $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

- To test the nonzero of β_j by the standardized coefficient or Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

where v_j is the jth diagonal element of $(X^T X)^{-1}$.

- To test for the significance of groups of coefficients simultaneously.

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)}$$

where RSS_1 is the residual sum-of-squares for the least squares fit of the bigger model with $p_1 + 1$ parameters, and RSS_0 the same for the nested smaller model with $p_0 + 1$ parameters, having $p_1 - p_0$ parameters constrained to be zero.

R^2

- $$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$
  $$SSR = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$
  $$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$$
- R^2 - coefficient of determination
  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSR}{SST}$$
  Proportion of variation explained by the regressors.

R^2

R^2 can be misleading!

- Simply adding more terms to the model will increase R^2.
- As the range of the regressor variable increases (decreases), R^2 generally increases (decreases).
- R^2 does not indicate the appropriateness of a linear model.
- Adjusted R^2. Penalizes us for added terms to the model that are not significant

$$R^2_{adj} = 1 - \frac{SSR/(n-p)}{SST/(n-1)}$$

- To obtain a $1-2\alpha$ confidence interval for β_j

$$(\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}, \quad \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma})$$

- An approximate confidence set for entire parameter β

$$C_\beta = \left\{ \beta | (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \beta) \leq \hat{\sigma}^2 {\chi^2_{p+1}}^{(1-\alpha)} \right\}$$

Example: Prostate Cancer

TABLE 3.1. *Correlations of predictors in the prostate cancer data.*

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

3.2.1 Example: Prostate Cancer

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the* $p = 0.05$ *level.*

Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

Example: Prostate Cancer

- We randomly split the dataset into a training set of size 67 and a test set of size 30.
- Consider dropping all the non-significant terms in Table 3.2. The F test for the significance of group {age, lcp, gleason, and pgg45} is

$$F = \frac{(32.81 - 29.43)/(9-5)}{29.43/(67-9)} = 1.67$$

  P-value is 0.17.

- The mean prediction error on the test data is 0.521. In contrast, prediction using the mean training value of lpsa has a test error of 1.057, it is called the“base error rate.”

The Gauss-Markov Theorem

- The linear model
$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon_j$$
where ϵ_j's have zero mean and equal variance and are uncorrelated.
- We focus on estimation of any linear combination of the parameters $\theta = a^T \beta$;
$$\hat{\theta} = \alpha^T \hat{\beta} = \alpha^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$
- If we assume that the linear model is correct, $a^T \hat{\beta}$ is unbiased since
$$\begin{aligned} \mathrm{E}(\alpha^T \hat{\beta}) &= \mathrm{E}(\alpha^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) \\ &= \alpha^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}_\beta \\ &= \alpha^T \beta \end{aligned}$$

The Gauss-Markov Theorem

- The Gauss-Markov theorem states that if we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ that is unbiased for $a^T \hat{\beta}$, that is, $E(\mathbf{c}^T \mathbf{y}) = a^T \hat{\beta}$, then

$$\mathrm{Var}(\alpha^T \hat{\beta}) \leq \mathrm{Var}(\mathbf{c}^T \mathbf{y})$$

- This is one of the most famous results in statistics asserts that the least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates.

Error of Estimation and Prediction

- The mean squared error of an estimator $\hat{\theta}$ in estimating θ:

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}(\hat{\theta}-\theta)^2 \\ &= \mathrm{Var}(\hat{\theta}) + [\mathrm{E}(\hat{\theta})-\theta]^2\end{aligned}$$

- Consider the prediction of the new response at input x_0,

$$Y_0 = f(x_0) + \varepsilon_0$$

$$\begin{aligned}\mathrm{E}(Y_0 - \hat{f}(x_0))^2 &= \sigma^2 + \mathrm{E}(x_0^T\hat{\beta} - f(x_0))^2 \\ &= \sigma^2 + \mathrm{MSE}(\hat{f}(x_0))\end{aligned}$$

Subset Selection

- There are two reasons that we are often not satisfied with the least squares estimates
  - The first is *prediction accuracy*: the least squares estimates often have low bias but large variance.
  - The second reason is *interpretation*.

AIC, BIC

- $$AIC = -\frac{2}{n} loglike + 2\frac{d}{n}$$

  $$BIC = -2 * loglike + \log n * d$$

  where d is the value for model complexity, typically is the number of unmown parameters.
- Smaller values are better in relative sense.

Best-Subset Selection

- An efficient algorithm - the leaps and bounds procedure (Furnival and Wilson, 1974) - makes this feasible for p as large as 30 or 40.
  typically we choose the smallest model that minimizes an estimate of the expected prediction error.
  cross-validation to estimate prediction error and select k; the AIC criterion is a popular alternative.

Best-Subset Selection

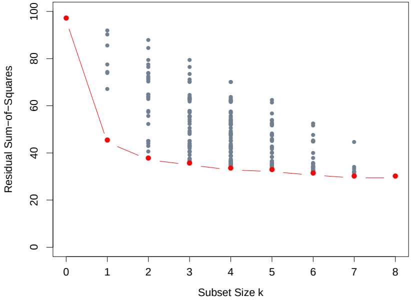

FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Forward- and Backward-Stepwise Selection

- Subset selection is infeasible for p much larger than 40
- Forward-stepwise selection is a *greedy algorithm*, producing a nested sequence of models.

  It might seem sub-optimal compared to best-subset selection.
  - *Computational*; for large p we cannot compute the best subset sequence,
  - *Statistical*; a price is paid in variance for selecting the best subset of each size; forward stepwise is a more constrained search, and will have lower variance, but perhaps more bias.
- *Backward-stepwise selection* starts with the full model, and sequentially deletes the predictor that has the least impact on the fit.
- Backward selection can only be used when $N > p$, while forward stepwise can always be used.

Forward- and Backward-Stepwise Selection

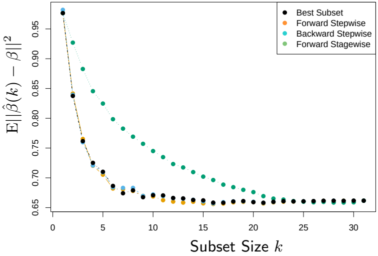

FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem* $Y = X^T\beta + \varepsilon$. *There are* $N = 300$ *observations on* $p = 31$ *standard Gaussian variables, with pairwise correlations all equal to* 0.85. *For* 10 *of the variables, the coefficients are drawn at random from a* $N(0, 0.4)$ *distribution; the rest are zero. The noise* $\varepsilon \sim N(0, 6.25)$, *resulting in a signal-to-noise ratio of* 0.64. *Results are averaged over* 50 *simulations. Shown is the mean-squared error of the estimated coefficient* $\hat\beta(k)$ *at each step from the true* β.

Forward- and Backward-Stepwise Selection

- In the R package the step function uses the AIC criterion for weighing the choices, which takes proper account of the number of parameters fit; at each step an add or drop will be performed that minimizes the AIC score.
- Other more traditional packages base the selection on F-statistics, adding “significant” terms, and dropping“non-significant” terms. These are out of fashion, since they do not take proper account of the multiple testing issues.

Prostate Cancer Data Example (Continued)

- Table 3.3 shows the coefficients from a number of different selection and shrinkage methods. They are *best-subset* selection using an all-subsets search, *ridge regression*, the *lasso, principal components regression* and *partial least squares*. Each method has a complexity parameter, and this was chosen to minimize an estimate of prediction error based on tenfold cross-validation;
- Note that we have already divided these data into a training set of size 67 and a test set of size 30. Cross-validation is applied to the training set, since selecting the shrinkage parameter is part of the training process.

Prostate Cancer Data Example (Continued)

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Prostate Cancer Data Example (Continued)

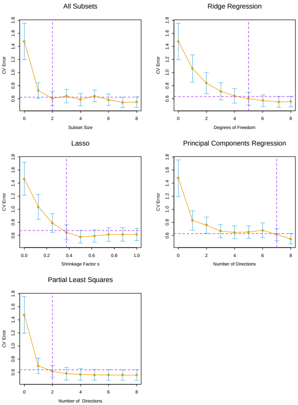

FIGURE 3.7. *Estimated prediction error curves and their standard errors for the various selection and shrinkage methods. Each curve is plotted as a function of the corresponding complexity parameter for that method. The horizontal axis has been chosen so that the model complexity increases as we move from left to right. The estimates of prediction error and their standard errors were obtained by tenfold cross-validation; full details are given in Section 7.10. The least complex model within one standard error of the best is chosen, indicated by the purple vertical broken lines.*

Shrinkage Methods: Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size.

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero (and each other).

Shrinkage Methods: LASSO

- $$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

  Notice the similarity to the ridge regression problem (3.42) or (3.41): the L_2 ridge penalty $\sum_1^p \beta_j^2$ is replaced by the L_1 lasso penalty $\sum_1^p |\beta_j|$.

The Lasso

- This latter constraint makes the solutions nonlinear in the y_i, and there is no closed form expression as in ridge regression.
- Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero. Thus the lasso does a kind of continuous subset selection.
- If t is chosen larger than $t_0 = \sum_1^p |\beta_j|$ (where $\hat{\beta}_j = \hat{\beta}_j^{ls}$, the least squares estimates), then the lasso estimates are the $\hat{\beta}_j' s$.
- On the other hand, for $t = t_0/2$ say, then the least squares coefficients are shrunk by about 50% on average.

the Lasso

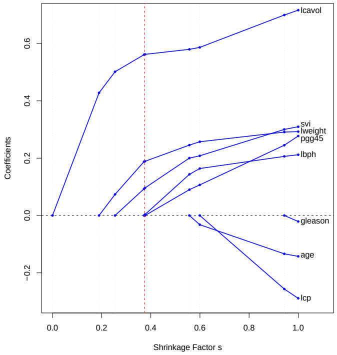

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Methods Using Derived Input Directions: Principal Components Regression

- X_{v_m}, and then regresses y on $z_1, z_2, \ldots, z_M$ for some $M \leq p$. $z_1, z_2, \ldots, z_M$ are the first M principle components of $\mathbf{X}$. Since the z_m are orthogonal, this regression is just a sum of univariate regressions:

$$\hat{\mathbf{y}}^{\mathsf{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

where $\hat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$.

$$\hat{\beta}^{\mathsf{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

Linear Methods for Classification

- For an important class of procedures, these decision boundaries are linear; this is what we will mean by linear methods for classification.
- $$\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$$

  The decision boundary between class k and ℓ is that set of points for which $\hat{f}_k(x) = \hat{f}_\ell(x)$, that is, the set $\{x : (\hat{\beta}_{k0} - \hat{\beta}_{\ell 0}) + (\hat{\beta}_k - \hat{\beta}_\ell)^T x = 0\}$, an affine set or hyperplane.

Multiple Outputs of Linear Regression

- $$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

  Here $\mathbf{Y}$ is the $N \times K$ response matrix, with ik entry y_{ik}, $\mathbf{X}$ is the $N \times (p+1)$ input matrix, $\mathbf{B}$ is the $(p+1) \times K$ matrix of parameters and $\mathbf{E}$ is the $N \times K$ matrix of errors.

  $$\begin{aligned} \text{RSS}(\mathbf{B}) &= \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \\ &= \text{tr}\Big[(\mathbf{Y} - \mathbf{X}\mathbf{B})^T (\mathbf{Y} - \mathbf{X}\mathbf{B})\Big] \end{aligned}$$

Multiple Outputs of Linear Regression

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

- $$\mathrm{RSS}(\mathbf{B};\boldsymbol{\Sigma}) = \sum_{i=1}^{N}(y_i - f(x_i))^T\boldsymbol{\Sigma}^{-1}(y_i - f(x_i))$$

Linear Regression of an Indicator Matrix

- $Y = (Y_1, ..., Y_K)$, and the N training instances of these form an NxK indicator response matrix Y. Y is a matrix of 0's and 1's, with each row having a single 1.
- Note that we have a coefficient vector for each response column y_k, and hence a $(p+1) \times K$ coefficient matrix $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
-
    - compute the fitted output $\hat{f}(x)^T = (1, x^T)\hat{\mathbf{B}}$, a K vector;
    - identify the largest component and classify accordingly:

$$\hat{G}(x) = \text{argmax}_{k \in \mathcal{G}} \hat{f}_k(x)$$

Linear Regression of an Indicator Matrix

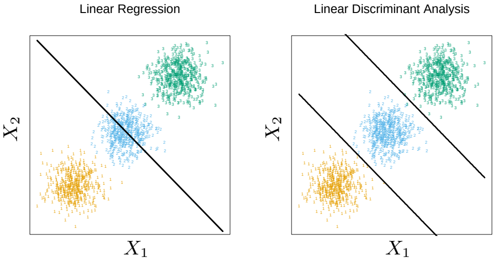

FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).*

Linear Regression of an Indicator Matrix

- There is a serious problem with the regression approach when the number of classes $K \geq 3$, especially prevalent when K is large.
- A loose but general rule is that if $K \geq 3$ classes are lined up, polynomial terms up to degree $K - 1$ might be needed to resolve them.

Linear Regression of an Indicator Matrix

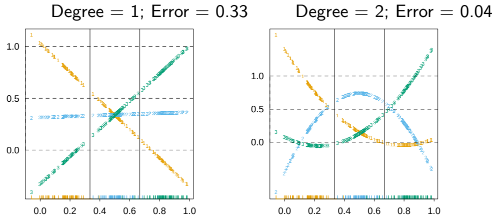

FIGURE 4.3. *The effects of masking on linear regression in* $\mathbb{R}$ *for a three-class problem. The* rug plot *at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, for the blue class,* y_{blue} *is* 1 *for the blue observations, and* 0 *for the green and orange. The fits are linear and quadratic polynomials. Above each plot is the training error rate. The Bayes error rate is* 0.025 *for this problem, as is the LDA error rate.*

LDA

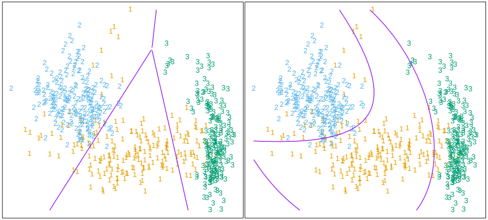

FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$*. Linear inequalities in this space are quadratic inequalities in the original space.*

Linear Discriminant Analysis

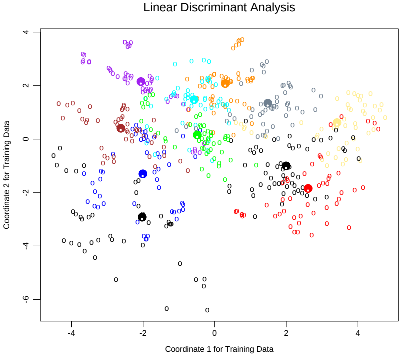

FIGURE 4.4. *A two-dimensional plot of the vowel training data. There are eleven classes with $X \in \mathbb{R}^{10}$, and this is the best view in terms of a LDA model (Section 4.3.3). The heavy circles are the projected mean vectors for each class. The class overlap is considerable.*

Linear Discriminant Analysis

TABLE 4.1. *Training and test error rates using a variety of linear techniques on the vowel data. There are eleven classes in ten dimensions, of which three account for* 90% *of the variance (via a principal components analysis). We see that linear regression is hurt by masking, increasing the test and training error by over* 10%*.*

Technique	Error Rates	
	Training	Test
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

LDA

- Assume that the classes have a multivariate normal distribution ($f_k(x)$) common covariance matrix $\Sigma_k = \Sigma, \forall k$.
- This linear log-odds function ($\log f_k(x)/f_l(x)$)implies that the decision boundary between classes k and ℓ - the set where $Pr(G = k|X = x) = Pr(G = \ell|X = x)$.is linear in x; in p dimensions a hyperplane.
- *linear discriminant functions*

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

are an equivalent description of the decision rule, with $G(x) = \arg\max_k \delta_k(x)$.

Linear Discriminant Analysis

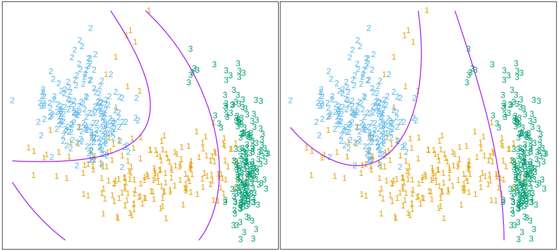

FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Linear Discriminant Analysis

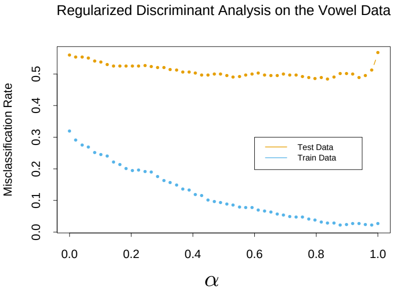

FIGURE 4.7. *Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of $\alpha \in [0, 1]$. The optimum for the test data occurs around $\alpha = 0.9$, close to quadratic discriminant analysis.*

LDA

- Fisher arrived at this decomposition via a different route, without referring to Gaussian distributions at all. He posed the problem:

  > Find the linear combination $Z = a^T X$ such that the between-class variance is maximized relative to the within-class variance.

  Again, the between class variance is the variance of the class means of Z, and the within class variance is the pooled variance about the means.

Linear Discriminant Analysis

Here W is the pooled within-class covariance matrix $\sum_1^K \pi_k W_k$, $W_k = \sum_1^{n_k} \pi_i(\bar{x}_{ki} - \bar{x}_k)(\bar{x}_{ki} - \bar{x}_k)^T$ and B is the between class covariance matrix $\sum_1^K \pi_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T$

$$\max_a \frac{a^T\mathbf{B}a}{a^T\mathbf{W}a}$$

or equivalently

$$\max_a a^T\mathbf{B}a \quad \text{subject to} \quad a^T\mathbf{W}a = 1$$

One can find the next direction a_2, orthogonal in W to a_1, such that $a_2^T Ba_2/a_2^T Wa_2$ is maximized; and so on. The $a_1, a_2, ...$ are referred to as discriminant coordinates, not to be confused with discriminant functions. They are also referred to as canonical or discriminant variates,

Linear Discriminant Analysis

- The a_ℓ are referred to as discriminant coordinates, not to be confused with discriminant functions.
-
  - Gaussian classification with common covariances leads to linear decision boundaries. Classification can be achieved by sphering the data with respect to W, and classifying to the closest centroid (modulo $\log \pi_k$) in the sphered space.
  - Since only the relative distances to the centroids count, one can confine the data to the subspace spanned by the centroids in the sphered space.
  - This subspace can be further decomposed into successively optimal subspaces in term of centroid separation. This decomposition is identical to the decomposition due to Fisher.

Linear Discriminant Analysis

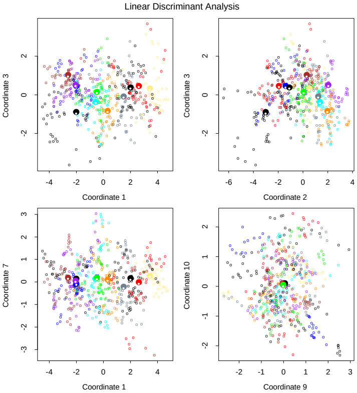

FIGURE 4.8. *Four projections onto pairs of canonical variates. Notice that as the rank of the canonical variates increases, the centroids become less spread out. In the lower right panel they appear to be superimposed, and the classes most confused.*

Linear Discriminant Analysis

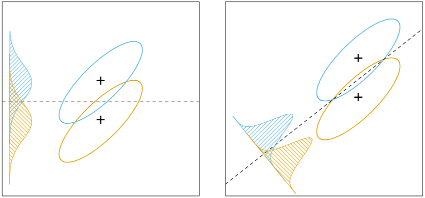

FIGURE 4.9. *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

Linear Discriminant Analysis

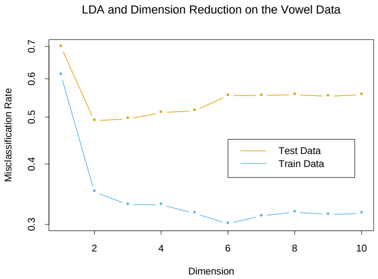

FIGURE 4.10. *Training and test error rates for the vowel data, as a function of the dimension of the discriminant subspace. In this case the best error rate is for dimension 2. Figure 4.11 shows the decision boundaries in this space.*

Linear Discriminant Analysis

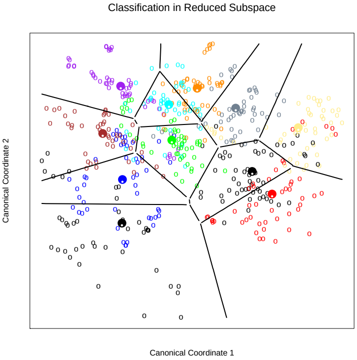

FIGURE 4.11. *Decision boundaries for the vowel training data, in the two-dimensional subspace spanned by the first two canonical variates. Note that in any higher-dimensional subspace, the decision boundaries are higher-dimensional affine planes, and could not be represented as lines.*