Chapter 4. Linear Models for Classification

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455
Email: weip@biostat.umn.edu

PubH 7475/8475

Linear Model and Least Squares

- Data: (Y_i, X_i), $X_i = (X_{i1}, ..., X_{ip})'$, $i = 1, ..., n$.
  Y_i: categorical with K classes; often $K = 2$.
- LM: defined $y_k = I(\text{class } k)$,

$$f_k(x) = E(y_k|x) = Pr(G = k|x) = \beta_{0k} + \beta_k' x$$

- Dedcision boundary: $f_k(x) = f_l(x)$, linear
- Use LSE
- Is it ok to use LM?
  Yes, but ...masking
- can be formulated as multivariat/multiple response LM.

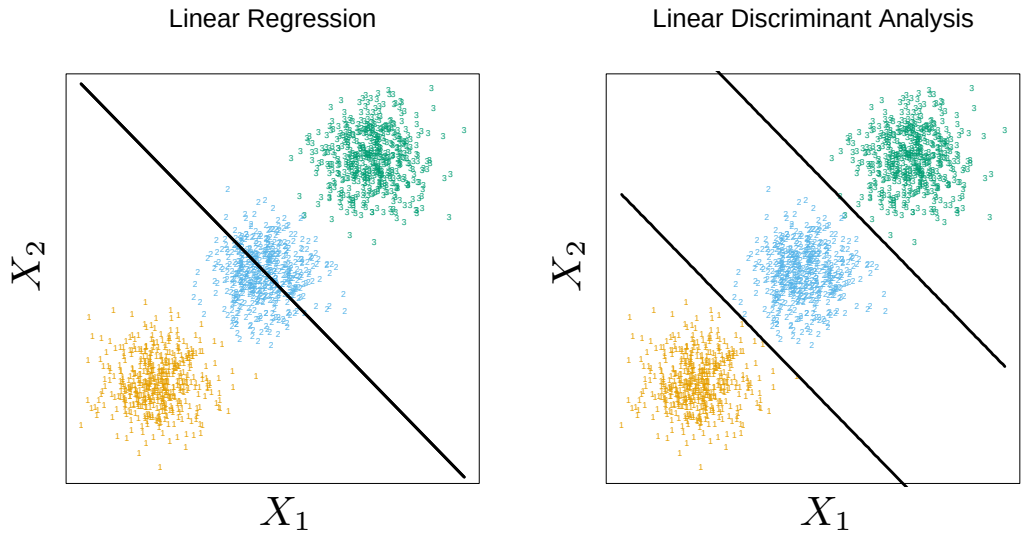


FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle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

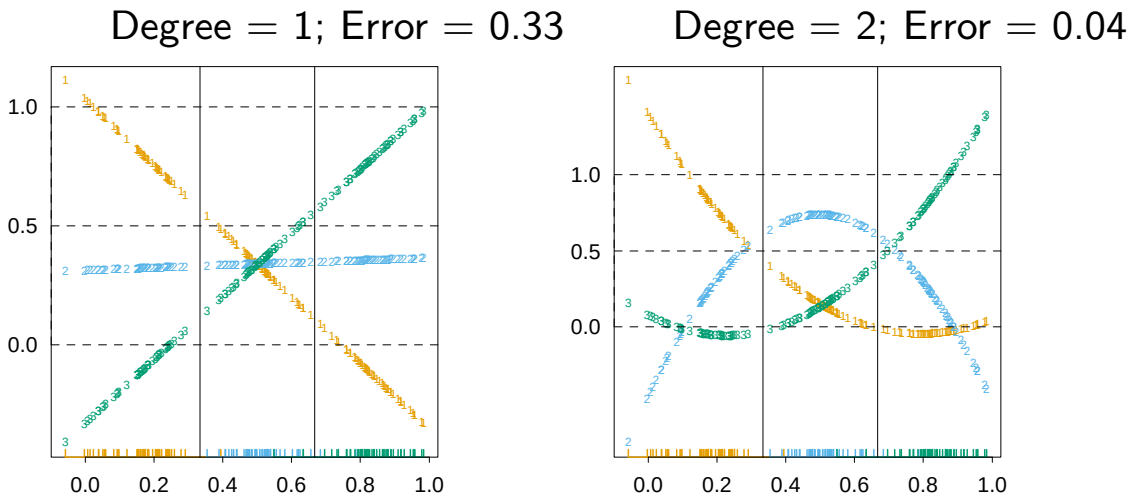


FIGURE 4.3. *The effects of masking on linear regression in* $\mathbb{R}$ *for a three-class problem. The* rug plot *at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, 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.*

Discriminant Analysis

- optimal Bayes rule: $\hat{G}(x) = \arg\max_k Pr(k|x)$.
  $Pr(k|x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$,
  π_k: prior prob $Pr(k)$,
  f_k: density of x in class k.
- Assume $f_k = N(\mu_k, \Sigma) \Longrightarrow$ LDA.
- Assume $f_k = N(\mu_k, \Sigma_k) \Longrightarrow$ QDA.
- Assume $f_k = \sum_j p_j N(\mu_j, \Sigma_j) \Longrightarrow$ MDA. §12.7
- Estimat f_k nonparametrically, e.g. by kernel density estimation $\Longrightarrow$ KDA.
  General, but not working well for large p – "curse of dim."
- Naive Bayes: assuming independence among the predictors,
  $f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)$.
  often work quite well!

LDA

- Assume: $x|k \sim N(\mu_k, \Sigma)$.
- $\log[\pi_k f_k(x)] \propto \log \pi_k + x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k = \delta_k(x)$.
  $\hat{G}(x) = \arg\max_k Pr(k|x) = \arg\max_k \delta_k(x)$,
  Linear.
- In practice, estimate
  $\hat{\pi}_k = n_k/n$,
  $\hat{\mu}_k = \sum_{G_i=k} x_i/n_k$,
  $\hat{\Sigma} = \sum_{k=1}^{K}\sum_{G_i=k}(x_i - \hat{\mu}_k)'(x_i - \hat{\mu}_k)/(n-K)$,

LDA

- LDA:

$$\begin{aligned} \log \frac{Pr(k|x)}{Pr(l|x)} &= \delta_k(x) - \delta_l(x) \\ = \quad & \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)'\Sigma^{-1}(\mu_k - \mu_l) + x'\Sigma^{-1}(\mu_k - \mu_l) \\ = \quad & \beta_{0,kl} + x'\beta_{kl}, \end{aligned}$$

  is a linear logistic reg model!

- $K = 2$, code $y = \pm 1$, is LM ok?
  $E(y) = b_0 + x'b$, then LSE
  $\hat{b} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$,
  but intercept differs.
  Explains LM $\approx$ LDA $\approx$ Logistic reg for $K = 2$?

QDA

- Assume: $x|k \sim N(\mu_k, \Sigma_k)$.

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

quadratic, or

$$\log \frac{Pr(k|x)}{Pr(l|x)} = \beta_{0,kl} + x' \beta_{1,kl} + x' B_{2,kl} x.$$

quadratic logit model.

- Estimation: similar to LDA ...
- QDA: more general and thus better than LDA?
- Figs 4.1 and 4.6.
- Example code: ex4.1.r

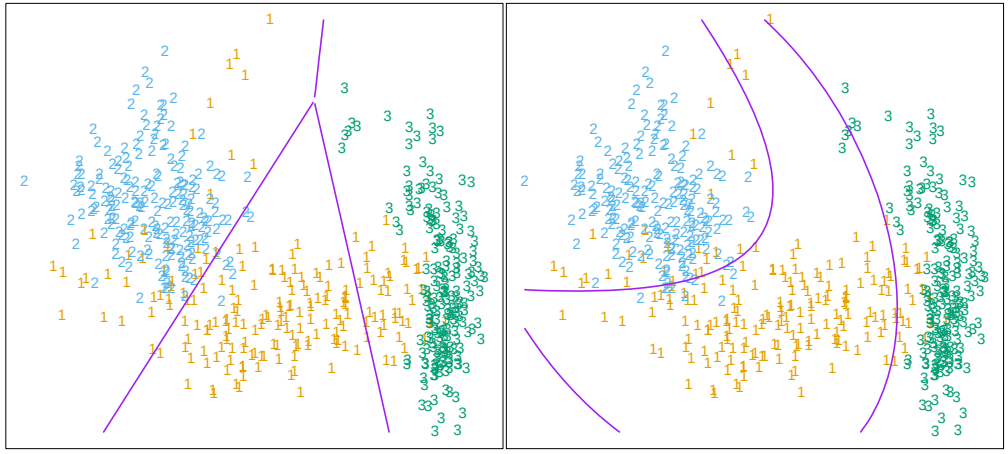


FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$. *Linear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

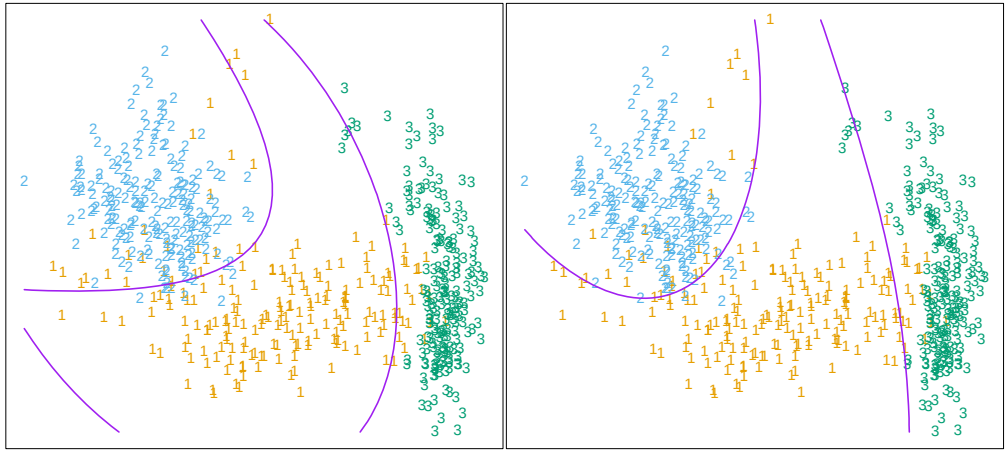


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$*). The right plot shows the quadratic decision boundaries found by QDA. The differences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

- LDA and QDA: too strong assumptions and thus do not work well?
  e.g. with categorical predictors?
- LDA: can be applied to x with high-order terms; Fig 4.6.
- STATLOG: evaluations based on 22 real datasets,
  LDA: in the top 3 for 7 out of 22;
  QDA: in the top 3 for 4 out of 22;
  L/QDA: in the top 3 for 11 out of 22.
- Dudoit & Speed (JASA, 2001): for high-dim gene expression data, LDA/QDA performed well.
  In fact, a diagonalized LDA or QDA could perform even better!
  DLDA or DQDA: only use the diagonal elements of $\hat{\Sigma}$ or $\hat{\Sigma}_k$.
  (8000) Theory: Bickel and Levina (2004).
- Why?

DLDA (§18.2)

- DLDA: only use $\text{diag}(\Sigma)$,

$$\delta_k(x^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \bar{x}_{kj})^2}{s_j^2} + 2\log \pi_k,$$

  where $x^* = (x_1^*, ..., x_p^*)'$ is a test obs, $\bar{x}_{kj}$ is the within-class sample mean for predictor j in class k, and s_j^2 is the pooled sample variance for predictor j, both based on a training set.
- Classification rule: $C(x^*) = \arg\max_k \delta_k(x^*)$.
- It is a naive Bayes rule and a nearest centroid rule.
- Problem: if p too large ... need to do VS!
  (8000) Theory: Fan & Fan (2008); DLDA becomes a random guessing as $p \to \infty$ unless the signal levels are extremely high ...

Nearest Shrunken Centroids (§18.2)

- Still use the DLDA rule, but first do VS.
- Define

$$d_{kj} = \frac{\bar{x}_{kj} - \bar{x}_j}{m_k(s_j + s_0)},$$

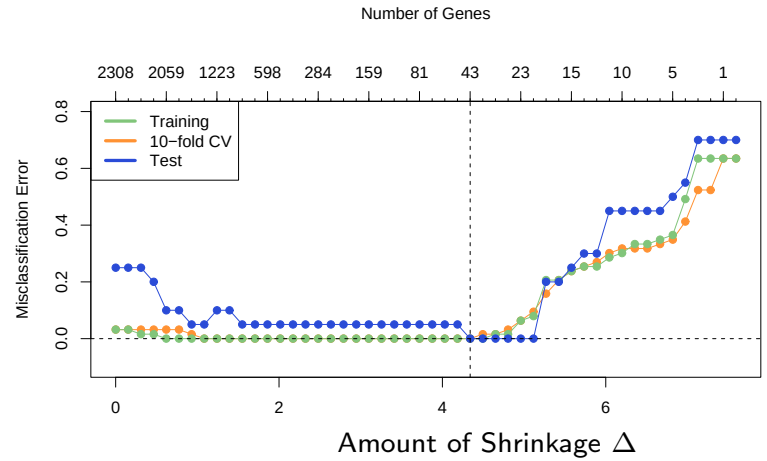
  where $\bar{x}_j$ is the overall mean for predictor j, $s0$ is a small constant, e.g. the median of s_j's, $m_k^2 = 1/N_k - 1/N$, and $\text{Var}(\bar{x}_{kj} - \bar{x}_j) = m_k^2\sigma^2$.
  Like a (regularized) t-statistic!
- ST: $d'_{kj} = \text{sign}(d_{kj})(|d_{kj}| - \Delta)_+$, or HT: $d'_{kj} = d_{kj} I(|d_{kj}| \geq \Delta)$, where the tuning parameter Δ is chosen by CV.
- new centroids: $\bar{x}'_{kj} = \bar{x}_j + m_k(s_j + s_0)d'_{kj}$, and $\bar{x}'_{kj} = \bar{x}_j$ if $d'_{kj} = 0$; if so for all k, no use of predictor j–why?
- Still use the DLDA rule (with NEW centroids):

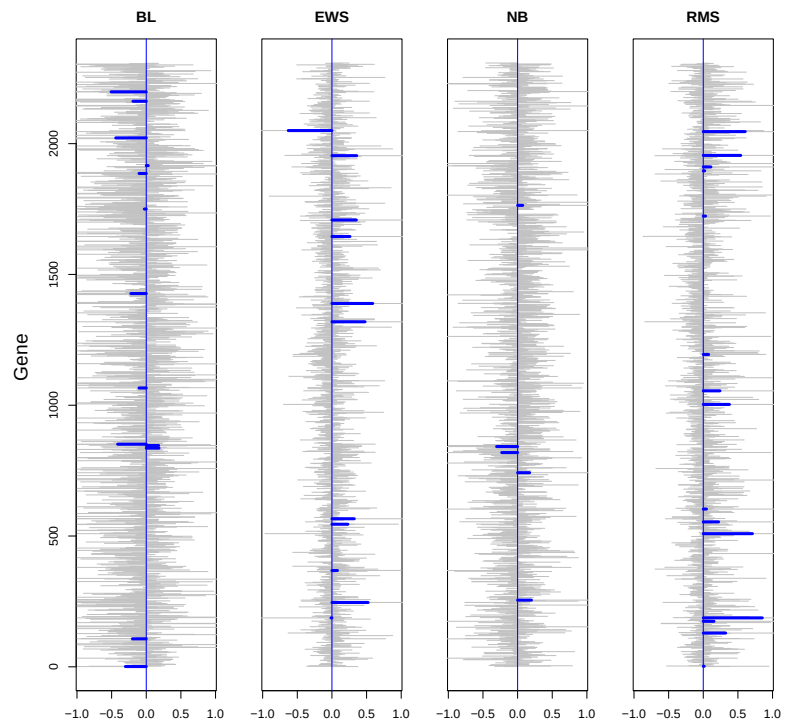
$$\delta_k(x^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \bar{x}'_{kj})^2}{s_j^2} + 2\log \pi_k, \qquad C(x^*) = \arg\max_k \delta_k(x^*).$$

- Estimate the class probability:

$$\hat{p}_k(x^*) = \frac{\exp[\frac{1}{2}\delta_k(x^*)]}{\sum_{l=1}^{K} \exp[\frac{1}{2}\delta_l(x^*)]}.$$

- A penalized approach; simple and yet often effective, but ...
- Example code: ex4.2.r





Centroids: Average Expression Centered at Overall Centroid

Regularized Discriminant Analysis

- A compromise b/w LDA and QDA: use $\tilde{\Sigma}_k(\alpha)$ in QDA,

$$\tilde{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma},$$

  where $\alpha \in [0,1]$ to be determined by CV.
- Fig 4.7.
- Similarly (§18.3.1),
  $\tilde{\Sigma}_k(\alpha) = \alpha\hat{\Sigma} + (1-\alpha)\hat{\sigma}^2 I$,
  $\tilde{\Sigma}_k(\alpha) = \alpha\hat{\Sigma} + (1-\alpha)\text{diag}(\hat{\Sigma})$,
  or, ..., covering DLDA, DQDA, ...
- (8000) A direct approach (Mai et al 2012): no need to estimate Σ! how?
  A connection b/w LDA and LSE!
  So, use penalized LR!

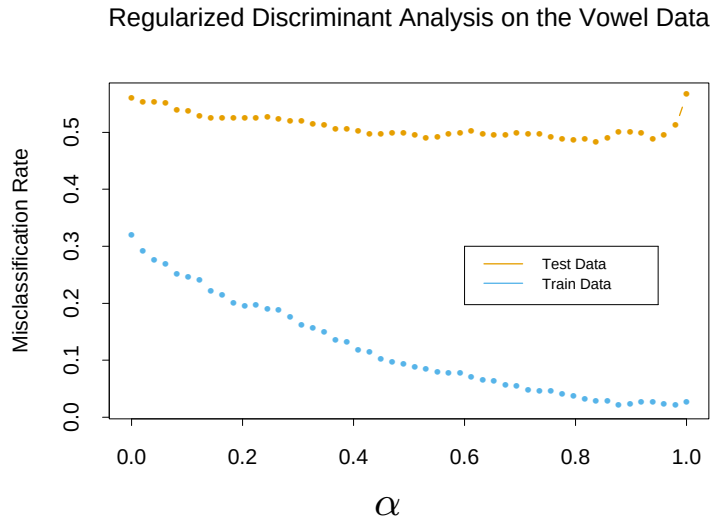


FIGURE 4.7. *Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of* $\alpha \in [0, 1]$. *The optimum for the test data occurs around* $\alpha = 0.9$, *close to quadratic discriminant analysis.*

Logistic regression

- Binary or multinomial logit model: for $k = 1, ..., K-1$,

$$\log \frac{Pr(k|x)}{Pr(K|x)} = \beta_{0,k} + x'\beta_{1,k},$$

or equivalently,

$$Pr(k|x) = \frac{\exp(\beta_{0,k} + x'\beta_{1,k})}{1 + \sum_{l=1}^{K-1} \exp(\beta_{0,k} + x'\beta_{1,k})}.$$

Then $\hat{G}(x) = \arg\max_k Pr(k|x)$.

- x can be expanded to include high-order terms.
- Parameter estimation: MLE
  Note: approx equivalent to fitting multiple binary logit models separetely (Begg & Gray, 1984, Biometrika).
- Logistic reg vs L/QDA: the former is more general; the latter has a stronger assumption and thus possibly more efficient if ...; Logistic reg is quite good.
- Example code: ex4.1.r

Penalized logistic regression (§18.3.2, 18.4)

- Need VS or regularization for a large p.
- Add a penalty term $J(\beta)$ to $-\log L$
  $J(\beta)$ can be Lasso, ..., as before.
- Computing algorithms: a Taylor expansion (i.e. quadratic approx) of $\log L$, then the same as penalized LR.
- R package `glmnet`: an elastic net penalty.
  hence do either Lasso or Ridge (or both).