

Linear Methods for Classification

Jia Li

Department of Statistics
The Pennsylvania State University

Email: jjali@stat.psu.edu
<http://www.stat.psu.edu/~jjali>

Classification

- ▶ Supervised learning
 - ▶ Training data: $\{(x_1, g_1), (x_2, g_2), \dots, (x_N, g_N)\}$
 - ▶ The feature vector $X = (X_1, X_2, \dots, X_p)$, where each variable X_j is quantitative.
 - ▶ The response variable G is categorical. $G \in \mathcal{G} = \{1, 2, \dots, K\}$
 - ▶ Form a predictor $G(x)$ to predict G based on X .
 - ▶ Email spam: G has only two values, say 1 denoting a useful email and 2 denoting a junk email. X is a 57-dimensional vector, each element being the relative frequency of a word or a punctuation mark.
- ▶ $G(x)$ divides the input space (feature vector space) into a collection of regions, each labeled by one class.

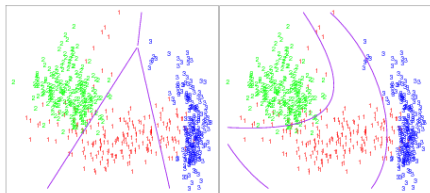


Figure 4.1: *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_{12}, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.*

Linear Methods

- ▶ *Decision boundaries* are linear: linear methods for classification.
- ▶ Two class problem:
 - ▶ The decision boundary between the two classes is a hyperplane in the feature vector space.
 - ▶ A hyperplane in the p dimensional input space is the set:

$$\{x : \alpha_0 + \sum_{j=1}^p \alpha_j x_j = 0\} .$$

The two regions separated by a hyperplane:

$$\{x : \alpha_0 + \sum_{j=1}^p \alpha_j x_j > 0\} \text{ and}$$

$$\{x : \alpha_0 + \sum_{j=1}^p \alpha_j x_j < 0\} .$$

- ▶ More than two classes:
 - ▶ The decision boundary between any pair of class k and l is a hyperplane (shown in previous figure).
- ▶ Question: which hyperplanes to use?
 - ▶ Different criteria lead to different algorithms.
 - ▶ Linear regression of an indicator matrix.
 - ▶ Linear discriminant analysis.
 - ▶ Logistic regression.
 - ▶ Rosenblatt's perceptron learning algorithm.
- ▶ Linear decision boundaries are not necessarily constrained.

The Bayes Classification Rule

- ▶ Suppose the marginal distribution of G is specified by the pmf $p_G(g)$, $g = 1, 2, \dots, K$.
- ▶ The conditional distribution of X given $G = g$ is $f_{X|G}(x | G = g)$.
- ▶ The training data (x_i, g_i) , $i = 1, 2, \dots, N$, are independent samples from the joint distribution of X and G ,

$$f_{X,G}(x, g) = p_G(g)f_{X|G}(x | G = g) .$$

- ▶ Assume the loss of predicting G as $G(X) = \hat{G}$ is $L(\hat{G}, G)$.
- ▶ Goal of classification: minimize the expected loss

$$E_{X,G}L(G(X), G) = E_X(E_{G|X}L(G(X), G)) .$$

- ▶ To minimize the left hand side, it suffices to minimize $E_{G|X}L(G(X), G)$ for each X . Hence the optimal predictor

$$G(x) = \arg \min_g E_{G|X=x}L(g, G) .$$

—*Bayes classification rule.*

- ▶ For 0-1 loss, i.e.,

$$L(g, g') = \begin{cases} 1 & g \neq g' \\ 0 & g = g' \end{cases}$$

$$E_{G|X=x}L(g, G) = 1 - Pr(G = g | X = x) .$$

- ▶ The Bayes rule becomes the rule of maximum a posteriori probability:

$$\begin{aligned} G(x) &= \arg \min_g E_{G|X=x}L(g, G) \\ &= \arg \max_g Pr(G = g | X = x) . \end{aligned}$$

- ▶ Many classification algorithms attempt to estimate $Pr(G = g | X = x)$, then apply the Bayes rule.

Linear Regression of an Indicator Matrix

- ▶ If \mathcal{G} has K classes, there will be K class indicators Y_k , $k = 1, \dots, K$.

| \mathbf{g} | \mathbf{y}_1 | \mathbf{y}_2 | \mathbf{y}_3 | \mathbf{y}_4 |
|--------------|----------------|----------------|----------------|----------------|
| 3 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 |

- ▶ Fit a linear regression model for each Y_k , $k = 1, 2, \dots, K$, using X :

$$\hat{\mathbf{y}}_k = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_k .$$

- ▶ Define $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k)$:

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

Classification Procedure

- ▶ Define $\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.
- ▶ For a new observation with input x , compute the fitted output

$$\begin{aligned} \hat{f}(x) &= [(1, x) \hat{\mathbf{B}}]^T \\ &= [(1, x_1, x_2, \dots, x_p) \hat{\mathbf{B}}]^T \\ &= \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \dots \\ \hat{f}_K(x) \end{pmatrix} \end{aligned}$$

- ▶ Identify the largest component of $\hat{f}(x)$ and classify accordingly:

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

Rationale

- ▶ The linear regression of Y_k on X is a linear approximation to $E(Y_k | X = x)$.



$$\begin{aligned} E(Y_k | X = x) &= Pr(Y_k = 1 | X = x) \cdot 1 + Pr(Y_k = 0 | X = x) \cdot 0 \\ &= Pr(Y_k = 1 | X = x) \\ &= Pr(G = k | X = x) \end{aligned}$$

- ▶ According to the Bayes rule, the optimal classifier:

$$G^*(x) = \arg \max_{k \in \mathcal{G}} Pr(G = k | X = x) .$$

- ▶ Linear regression of an indicator matrix:
 - ▶ Approximate $Pr(G = k | X = x)$ by a linear function of x using linear regression.
 - ▶ Apply the Bayes rule to the approximated probability.

Example: Diabetes Data

The diabetes data set is taken from the UCI machine learning database repository at:

<http://www.ics.uci.edu/~mlearn/Machine-Learning.html> . The original source of the data is the National Institute of Diabetes and Digestive and Kidney Diseases. There are 768 cases in the data set, of which 268 show signs of diabetes according to World Health Organization criteria. Each case contains 8 quantitative variables, including diastolic blood pressure, triceps skin fold thickness, a body mass index, etc.

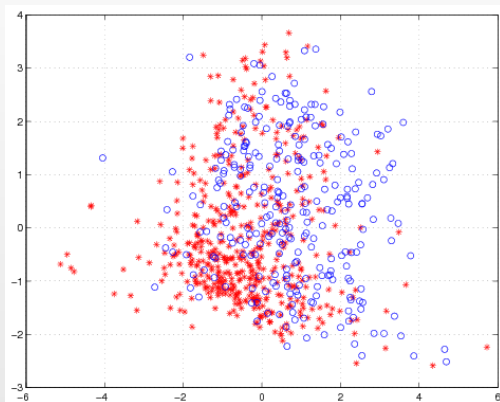
- ▶ Two classes: with or without signs of diabetes.
- ▶ Denote the 8 original variables by $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_8$.
- ▶ Remove the mean of \tilde{X}_j and normalize it to unit variance.

- ▶ The two principal components X_1 and X_2 are used in classification:

$$X_1 = 0.1284\tilde{X}_1 + 0.3931\tilde{X}_2 + 0.3600\tilde{X}_3 + 0.4398\tilde{X}_4 \\ + 0.4350\tilde{X}_5 + 0.4519\tilde{X}_6 + 0.2706\tilde{X}_7 + 0.1980\tilde{X}_8$$

$$X_2 = 0.5938\tilde{X}_1 + 0.1740\tilde{X}_2 + 0.1839\tilde{X}_3 - 0.3320\tilde{X}_4 \\ - 0.2508\tilde{X}_5 - 0.1010\tilde{X}_6 - 0.1221\tilde{X}_7 + 0.6206\tilde{X}_8$$

The scatter plot follows. Without diabetes: stars (class 1), with diabetes: circles (class 2).



$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ &= \begin{pmatrix} 0.6510 & 0.3490 \\ -0.1256 & 0.1256 \\ -0.0729 & 0.0729 \end{pmatrix}\end{aligned}$$

$$\hat{Y}_1 = 0.6510 - 0.1256X_1 - 0.0729X_2$$

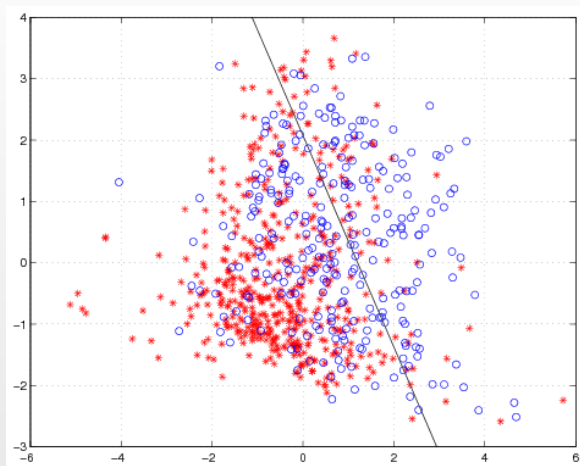
$$\hat{Y}_2 = 0.3490 + 0.1256X_1 + 0.0729X_2$$

Note $\hat{Y}_1 + \hat{Y}_2 = 1$.

Classification rule

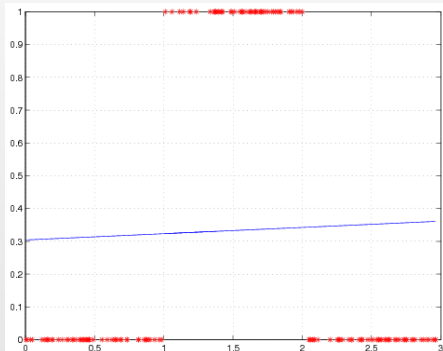
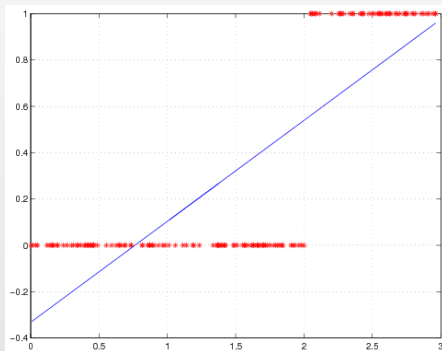
$$\begin{aligned}\hat{G}(x) &= \begin{cases} 1 & \hat{Y}_1 \geq \hat{Y}_2 \\ 2 & \hat{Y}_1 < \hat{Y}_2 \end{cases} \\ &= \begin{cases} 1 & 0.151 - 0.1256X_1 - 0.0729X_2 \geq 0 \\ 2 & \textit{otherwise} \end{cases}\end{aligned}$$

- ▶ Within training data classification error rate: 28.52%.
- ▶ *Sensitivity* (probability of claiming positive when the truth is positive): 44.03%.
- ▶ *Specificity* (probability of claiming negative when the truth is negative): 86.20%.



The Phenomenon of Masking

- ▶ When the number of classes $K \geq 3$, a class may be masked by others, that is, there is no region in the feature space that is labeled as this class.
- ▶ The linear regression model is too rigid.



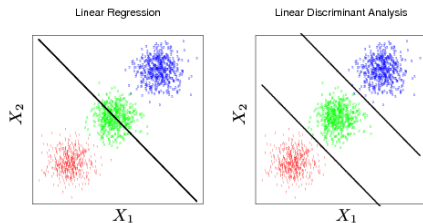


Figure 4.2: The data come from three classes in \mathbb{R}^2 and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle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

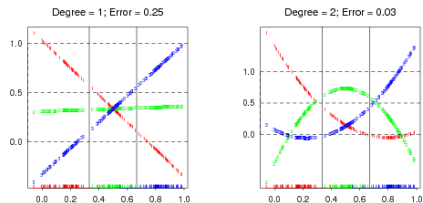


Figure 4.3: *The effects of masking on linear regression in \mathbb{R} for a three-class problem. The rug plot at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, for the red class, y_{red} is 1 for the red observations, and 0 for the green and blue. 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.*