Lecture 9: Linear methods for classification

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

Dr. Greg Hamerly
http://cs.baylor.edu/~hamerly/

Announcements

- Homework 2 due today
- Homework 3 assigned Thursday
- Getting work done early

Questions?

Chapter 4: Linear methods for classification

- 4.1 – Introduction
- 4.2 – Linear regression of an indicator matrix
- 4.3 – Linear discriminant analysis
- 4.4 – Logistic regression
- 4.5 – Separating hyperplanes

Linear decision boundaries for classification (4.1)

Classifier on real-valued data

Decision boundaries

Linear decision boundaries

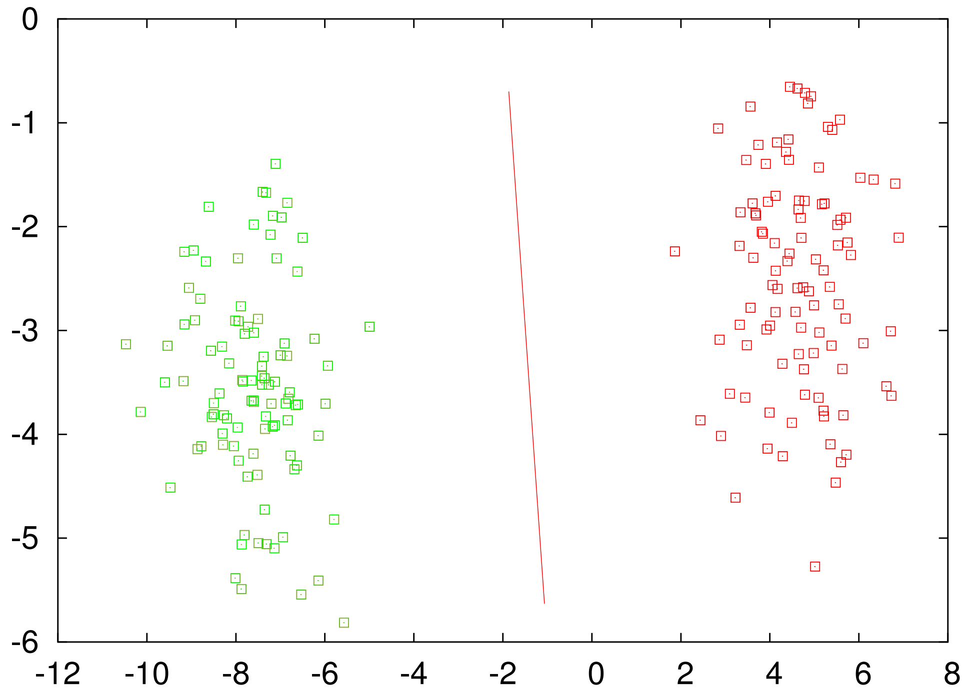
Linear regression for two classes

Two outputs: $y = 0$ or $y = 1$ (for example)

$$y = \beta_0 + \beta^T x$$

Learn β that gives least-squared error.

Example of 2-class classification w/linear regression



Discriminant functions

A discriminant function for class k is a function $\delta_k(x)$ of x that gives a score to the input.

For classification, do the following:

- compute the $\delta_k(x)$ for all classes ($1 \ldots k$)
- choose the class with the largest $\delta_k(x)$

We will be investigating discriminant functions which have linear boundaries between classes.

Finding the decision boundaries

The decision boundary between two classes is where the discriminant functions are equal.

Between class a and class b, the decision boundary is made up of the points:

$$\{x | \delta_a(x) = \delta_b(x)\}$$

Linear regression with multiple classes (4.1)

Two classes is easy with linear regression.

Structure for k classes: indicator matrix.

For $k = 3$:

$$Y = \overbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}}^{k \text{ classes}} \qquad X \in \mathbb{R}^{n \times d}$$

The sum of each row of Y is 1.

Learning the parameters

We assume the linear model for each of the k classes simultaneously:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}$$

We get the least-square estimates in the same way:

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

Where now

$$\hat{\mathbf{B}} \in \mathbb{R}^{d \times k}$$

So we have one k linear models; one for each class.

Classification using regression

We can view each linear model as a discriminant function:

$$\delta_k(x) = \beta_k x$$

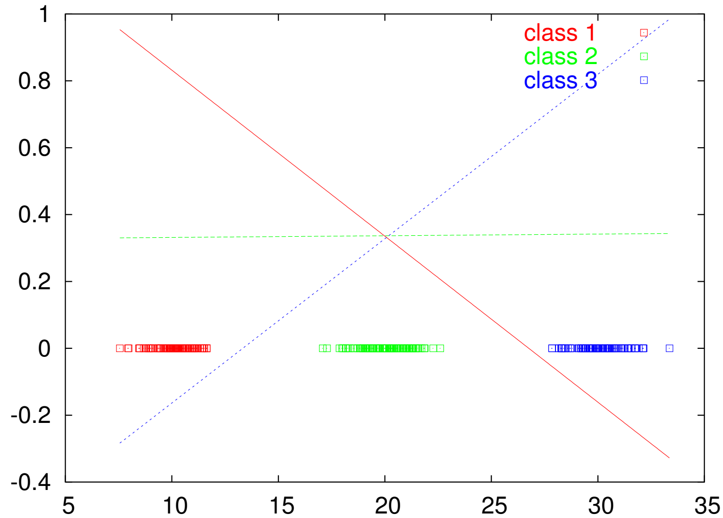
Then we can classify to the class with the largest discriminant.

Note that because of the linear constraints,

$$\sum_{i=1}^{k} \delta_k(x) = 1$$

So δ_k could be viewed as a probability (but δ_k could be negative!).

Problem: masking

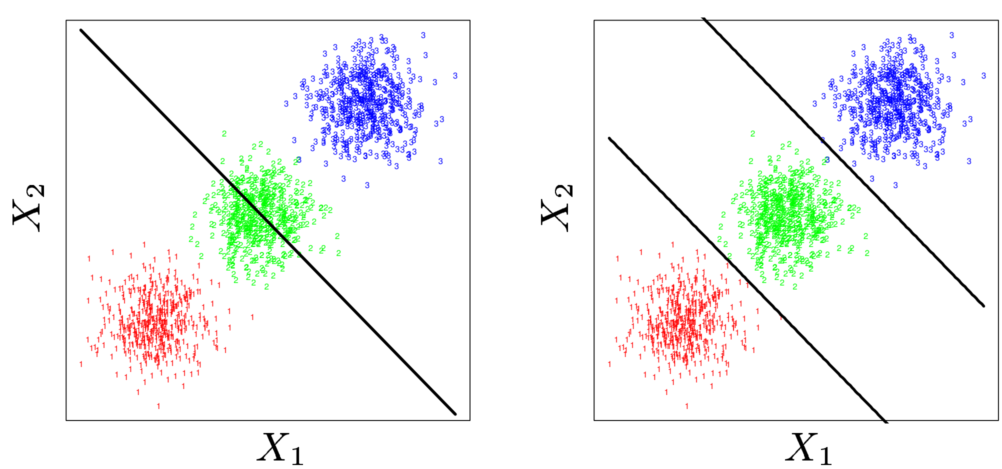


$k = 3$ classes along $y = 0$, and the three learned β functions.

Note that class 2 never dominates.

Can you move the β functions so it works?

Masking is hard to get rid of



Here is another example (linear regression on the left, LDA on the right). The lines are decision boundaries (not βs).

Next we will investigate other methods that do not have the masking problem.

2-minute journal

Please write a response to the following on a piece of paper and hand it in immediately. Please make it anonymous (no names). Write about:

- major points you learned today
- areas not understood or requiring clarification