

Lecture 10: Linear discriminant analysis

CSI 5v93: Introduction to machine learning

Baylor University
Computer Science Department

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

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 1/18

Announcements

- Homework 3 assigned by tomorrow – start early!

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 2/18

Questions?

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 3/18

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

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 4/18

Linear decision boundaries for classification (4.1)

Classifier on real-valued data

Decision boundaries

Linear decision boundaries

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 5/18

Linear regression of an indicator matrix

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

Now we assume the model:

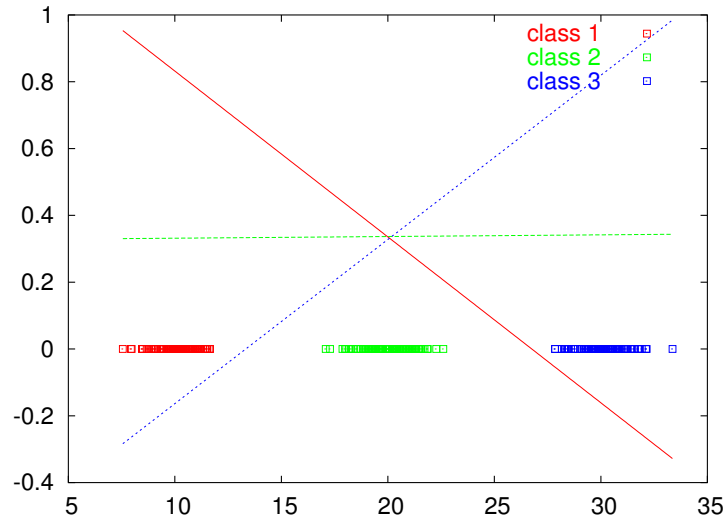
$$\mathbf{Y} = \mathbf{X}\mathbf{B}$$

And estimate:

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

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 6/18

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?

Discriminant functions and linear decision boundaries

Recall our problem framework for classification: k discriminant functions $\delta_k(x)$

The decision boundary between classes a and b :

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

We are interested in decision boundaries which are linear.

The discriminant function could be linear to achieve this, but it is not necessary (example).

Posterior probabilities

We want to classify to the class that gives the largest posterior probability:

For class l :

- Prior probability: π_l
- Model probability: $f_l(x)$

Using Bayes' rule, the class posterior probability is:

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

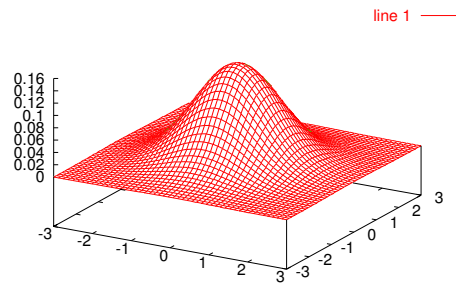
Linear discriminant analysis

Log-odds:

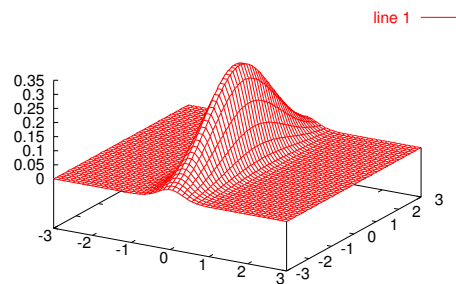
$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = l|X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

The boundary is where the log-odds equals 0.

Multivariate Gaussian probabilities



Spherical 2-d Gaussian



Non-spherical 2-d Gaussian

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 11/18

LDA and Gaussian probabilities

Gaussian probability distribution in one dimension $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Spherical Gaussian in d dimensions $N(\mu, \sigma^2 \mathbf{I})$:

$$f(x) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp \left\{ -\frac{\|x - \mu\|^2}{2\sigma^2} \right\}$$

General Gaussian in d dimensions $N(\mu, \Sigma)$:

$$f(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 12/18

LDA and Gaussian probabilities

Assume that $\Sigma_k = \Sigma$ for all classes, which simplifies the log-odds:

$$\begin{aligned}\log \frac{\Pr(G = k|X = x)}{\Pr(G = l|X = x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\ &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l)\end{aligned}$$

The boundary is (still) where the log-odds equals 0.

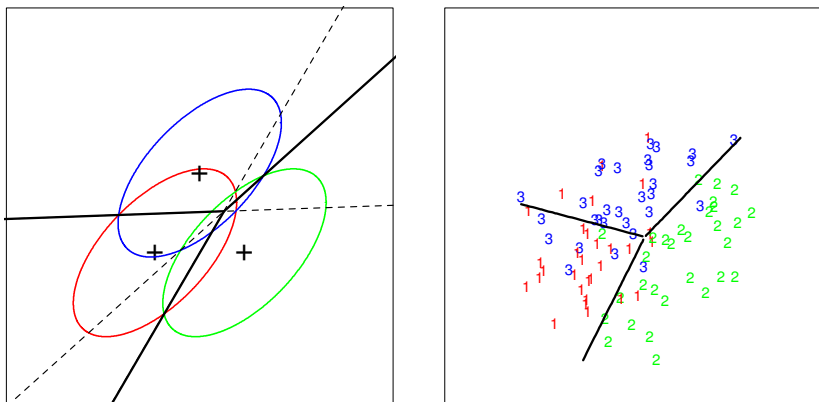
The boundary has a linear dependence on x .

From this we can see that the *linear discriminant function* is:

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

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 13/18

LDA examples



Left: idealized model (three Gaussians with common covariance)
Note that decision boundaries are not perpendicular to lines between Gaussian centers (why?).

Right: data and learned model's decision boundaries

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 14/18

Learning (estimating) the parameters

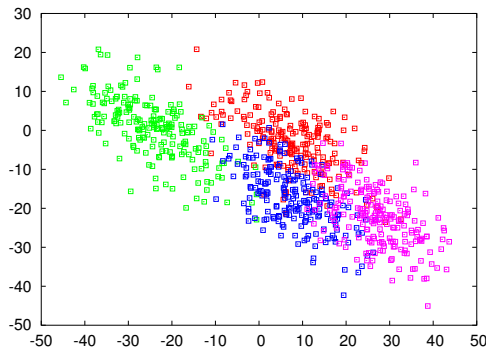
We do not know π_k , μ_k , or Σ .

We must learn them from the data:

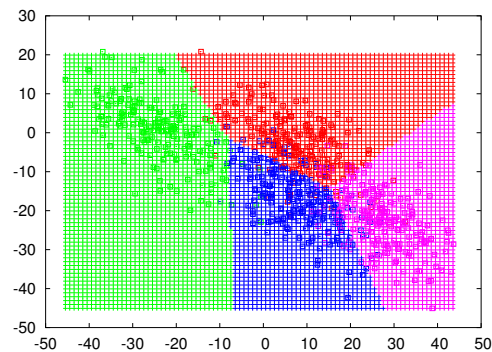
$$\begin{aligned}\hat{\pi}_k &= \frac{N_k}{N} \\ \hat{\mu}_k &= \frac{1}{N_k} \sum_{g_i=k} x_i \\ \hat{\Sigma} &= \frac{1}{N - k} \sum_{k=1}^K \sum_{g_i=k} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)\end{aligned}$$

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 15/18

Examples of LDA (in Matlab)



Data from 4 classes (colored red, green, blue, purple), each distributed as 2-d Gaussian with same covariance



After learning LDA parameters – the classification regions

Note that the decision boundaries are linear.

CSI 5v93: Introduction to machine learning, Lecture 10 – p. 16/18

Related to LDA

QDA (quadratic discriminant analysis): covariances (Σ_k) are not restricted

Regularized discriminant analysis: cross between LDA and QDA

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

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