Introduction to Machine Learning
CS195-5: Supplementary Lecture A

Thomas Hofmann

September 30, 2003

Linear Discriminant Analysis

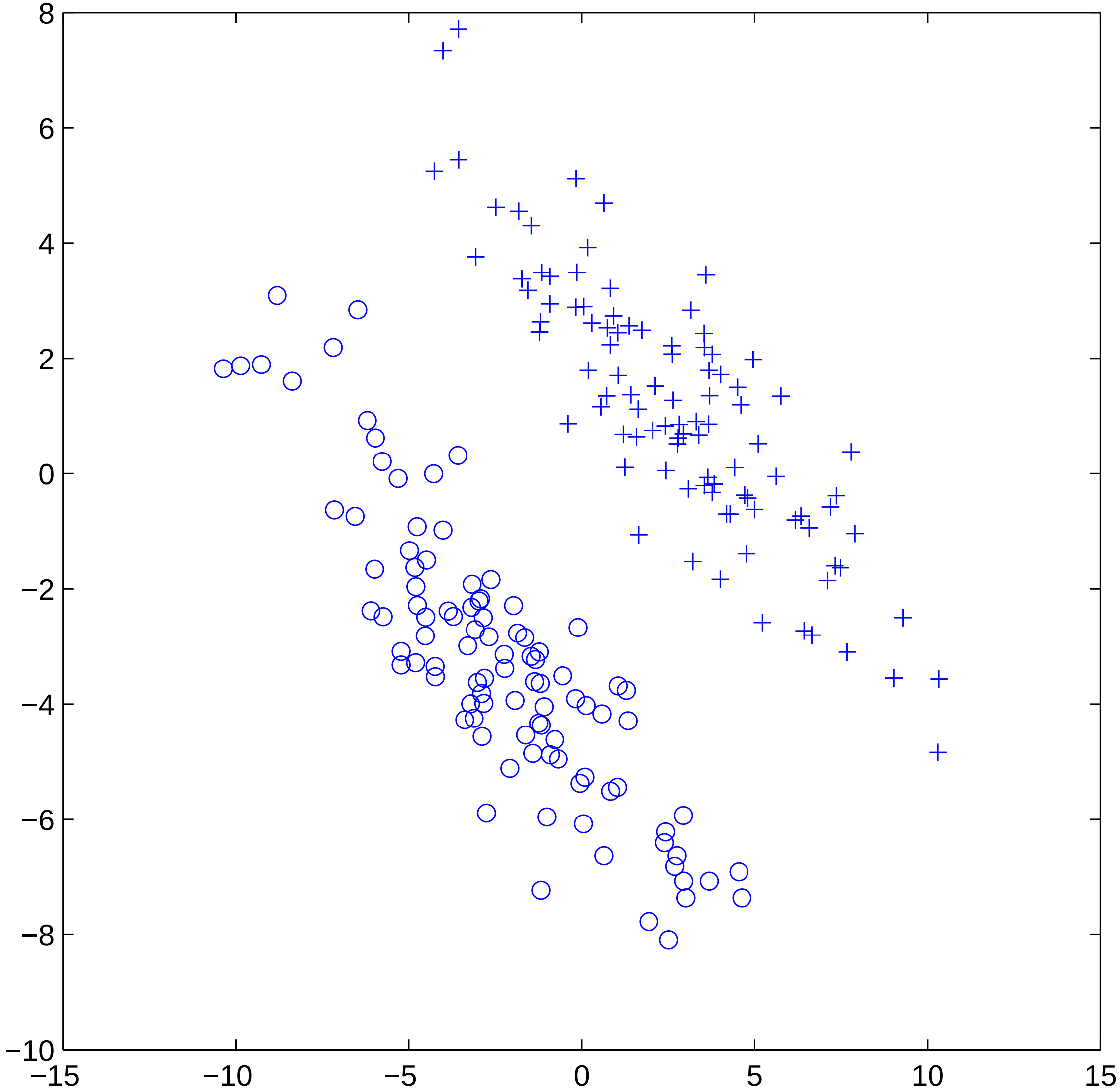
- Linear classification: projection to one-dimensional subspace (direction parametrized by $\mathbf{w}$) plus thresholding (parametrized by bias b).
- Ideal Discrimination: Project data onto a line such that patterns become "**well separated**".
- For given $\mathbf{w}$, each pattern will be represented by

$$\pi(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle,$$

where π defines a projection onto the line defined by $\mathbf{w}$, if $\|\mathbf{w}\| = 1$.

Looking for Directions

Which direction is a good one to pick?



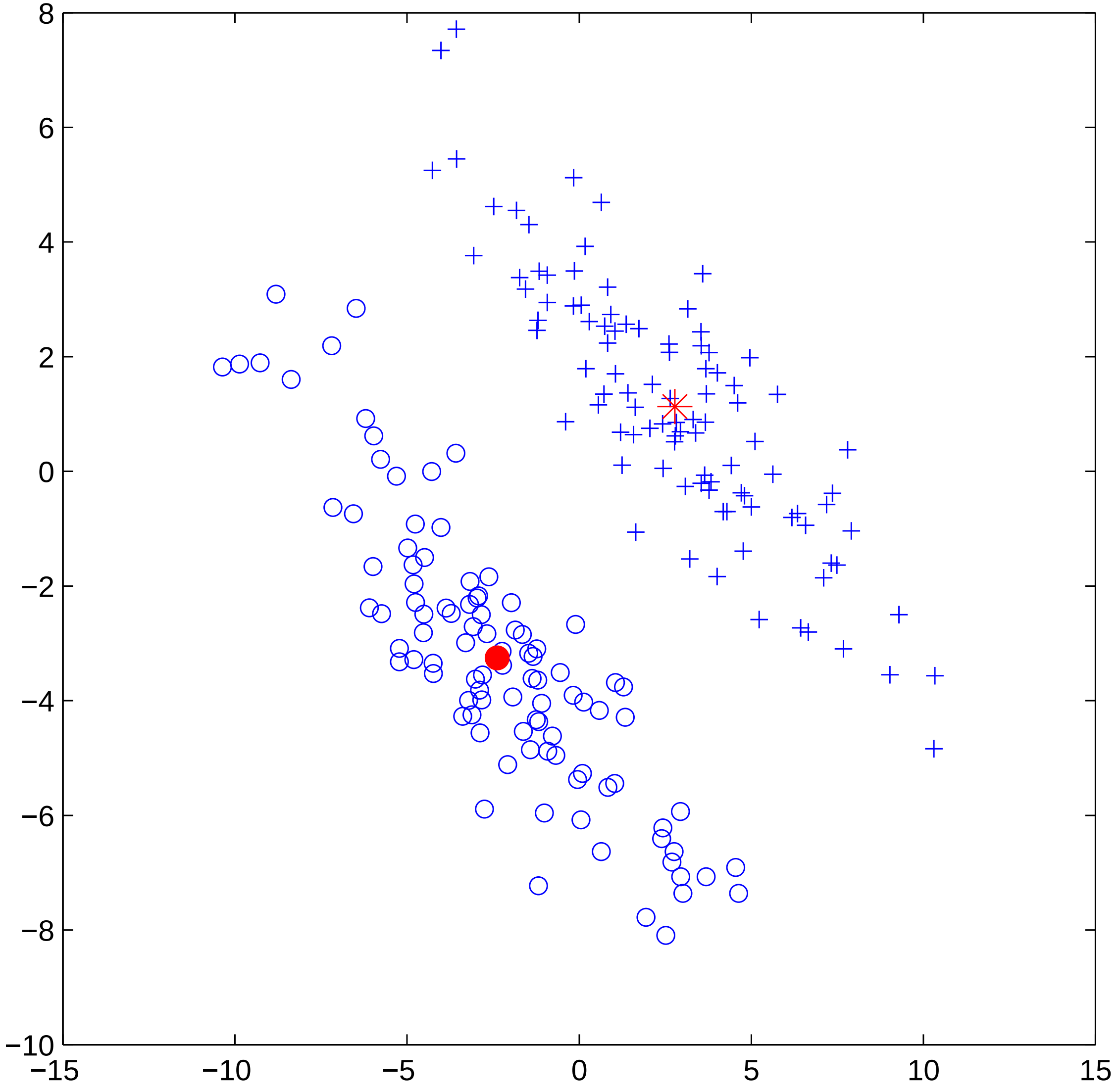
Linear Discriminant Analysis

- What are the requirements for a "good" direction / line to project onto? Focus on two-class case.
- We want the projections of the class means to be maximally separated.
- Make sure the projection of the positive centroid is as far away as possible from the projection of the negative centroid.
- Positive and negative centroid:

$$\mathbf{m}_+ \equiv \frac{\sum_{i:y_i=1} \mathbf{x}_i}{\sum_{i:y_i=1} 1}, \qquad \mathbf{m}_- \equiv \frac{\sum_{i:y_i=-1} \mathbf{x}_i}{\sum_{i:y_i=-1} 1}$$

Looking for Directions: The Centroids

Here are the centroids, m_-, m_+:



Linear Discriminant Analysis

- Absolute difference of their projections under $\mathbf{w}$:

$$|\pi(\mathbf{m}_+) - \pi(\mathbf{m}_-)| = |\langle \mathbf{w}, \mathbf{m}_+ - \mathbf{m}_- \rangle|$$

  What matters is the projection of the difference vector between the two centroids. [Which direction $\mathbf{w}$ achieves this?]

- This alone is clearly not sufficient [Think about it!]. The variance of the one-dimensional projections within each class should be as small as possible.

- Formally: scatter of the projected points in positive/negative class is defined as

$$s_+^2 \equiv \sum_{\mathbf{x}_i : y_i = 1} (\pi(\mathbf{x}_i) - \pi(\mathbf{m}_+))^2, \qquad s_-^2 \equiv \sum_{\mathbf{x}_i : y_i = -1} (\pi(\mathbf{x}_i) - \pi(\mathbf{m}_-))^2$$

Linear Discriminant Analysis: Fisher's Criterion

- Fisher's criterion:

$$J(\mathbf{w}) = \frac{|\pi(\mathbf{m}_+) - \pi(\mathbf{m}_-)|^2}{s_+^2 + s_-^2}$$

  The denominator $s_+^2 + s_-^2$ is called the **total within class scatter**.

- This criterion is invariant w.r.t. scaling of $\mathbf{w}$.

- Ronald Fisher (1890-1962): The 'father' of statistics.
  Natural selection is a mechanism for generating an exceedingly high degree of improbability.

Fisher's Linear Discriminant Analysis

- How can we maximize J? Use matrix notation first!
- Define scatter matrices

$$\mathbf{S}_{\pm} \equiv \sum_{\mathbf{x}_i : y_i = \pm 1} (\mathbf{x}_i - \mathbf{m}_{\pm})(\mathbf{x}_i - \mathbf{m}_{\pm})'$$

$$\mathbf{S}_W = \mathbf{S}_+ + \mathbf{S}_-$$

$\mathbf{S}_W$ is called the within class scatter matrix.

- Now one can write

$$\begin{aligned} s_{\pm}^2 &= \sum_{\mathbf{x}_i : y_i = \pm 1} \left(\langle \mathbf{w}, \mathbf{x}_i \rangle - \langle \mathbf{w}, \mathbf{m}_{\pm} \rangle \right)^2 \\ &= \sum_{\mathbf{x} : y_i = \pm 1} \mathbf{w}'(\mathbf{x}_i - \mathbf{m}_{\pm})(\mathbf{x}_i - \mathbf{m}_{\pm})'\mathbf{w} = \mathbf{w}'\mathbf{S}_{\pm}\mathbf{w} \end{aligned}$$

Fisher's Linear Discriminant Analysis

- Hence

$$\tilde{s}_+^2 + \tilde{s}_-^2 = \mathbf{w}'\mathbf{S}_W\mathbf{w}\,.$$

- Similarly

$$(\pi(\mathbf{m}_+) - \pi(\mathbf{m}_-))^2 = \mathbf{w}'\mathbf{S}_B\mathbf{w},$$

with the between class scatter matrix

$$\mathbf{S}_B \equiv (\mathbf{m}_+ - \mathbf{m}_-)(\mathbf{m}_+ - \mathbf{m}_-)'$$

- This results in an equivalent expression for Fishers discriminant criterion as a ratio between two quadratic forms:

$$J(\mathbf{w}) = \frac{\mathbf{w}'\mathbf{S}_B\mathbf{w}}{\mathbf{w}'\mathbf{S}_W\mathbf{w}}$$

Fisher's Linear Discriminant Analysis

- How can we maximize J - now that we have converted the criterion to matrix notation?!
- The above ratio is also known as the generalized Rayleigh quotient in physics. [*so what?*]
- Let's try and solve it...

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \frac{2\mathbf{S}_B \mathbf{w}}{\mathbf{w}' \mathbf{S}_W \mathbf{w}} - \frac{\mathbf{w}' \mathbf{S}_B \mathbf{w}}{\mathbf{w}' \mathbf{S}_W \mathbf{w}} \frac{2\mathbf{S}_W \mathbf{w}}{\mathbf{w}' \mathbf{S}_W \mathbf{w}} = 0$$

- Hence one gets for the optimal $\mathbf{w}$

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}, \quad \lambda = \frac{\mathbf{w}' \mathbf{S}_B \mathbf{w}}{\mathbf{w}' \mathbf{S}_W \mathbf{w}} \ .$$

Fisher's Linear Discriminant Analysis

- Rewriting one gets... An eigenequation!

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}$$

- Since

$$\mathbf{S}_B\mathbf{w} = \langle\mathbf{w}, \mathbf{m}_+ - \mathbf{m}_-\rangle(\mathbf{m}_+ - \mathbf{m}_-)$$

  it is not necessary to actually determine the eigenvalues.

- One simply gets

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_+ - \mathbf{m}_-)$$

Matlab's Linear Discriminant Analysis

```
% number of positive and negative examples
num_plus  = size(Xplus,1);
num_minus = size(Xminus,1);

% class  means and difference of class means
m_plus    = sum(X_plus, 1) / num_plus;
m_minus   = sum(X_minus,1) / num_minus
m_diff    = m_plus - m_minus;

% subtract class mean from data
mX_plus   = X_plus  - repmat(m_plus, num_plus, 1);
mX_minus  = X_minus - repmat(m_minus,num_minus,1);
```

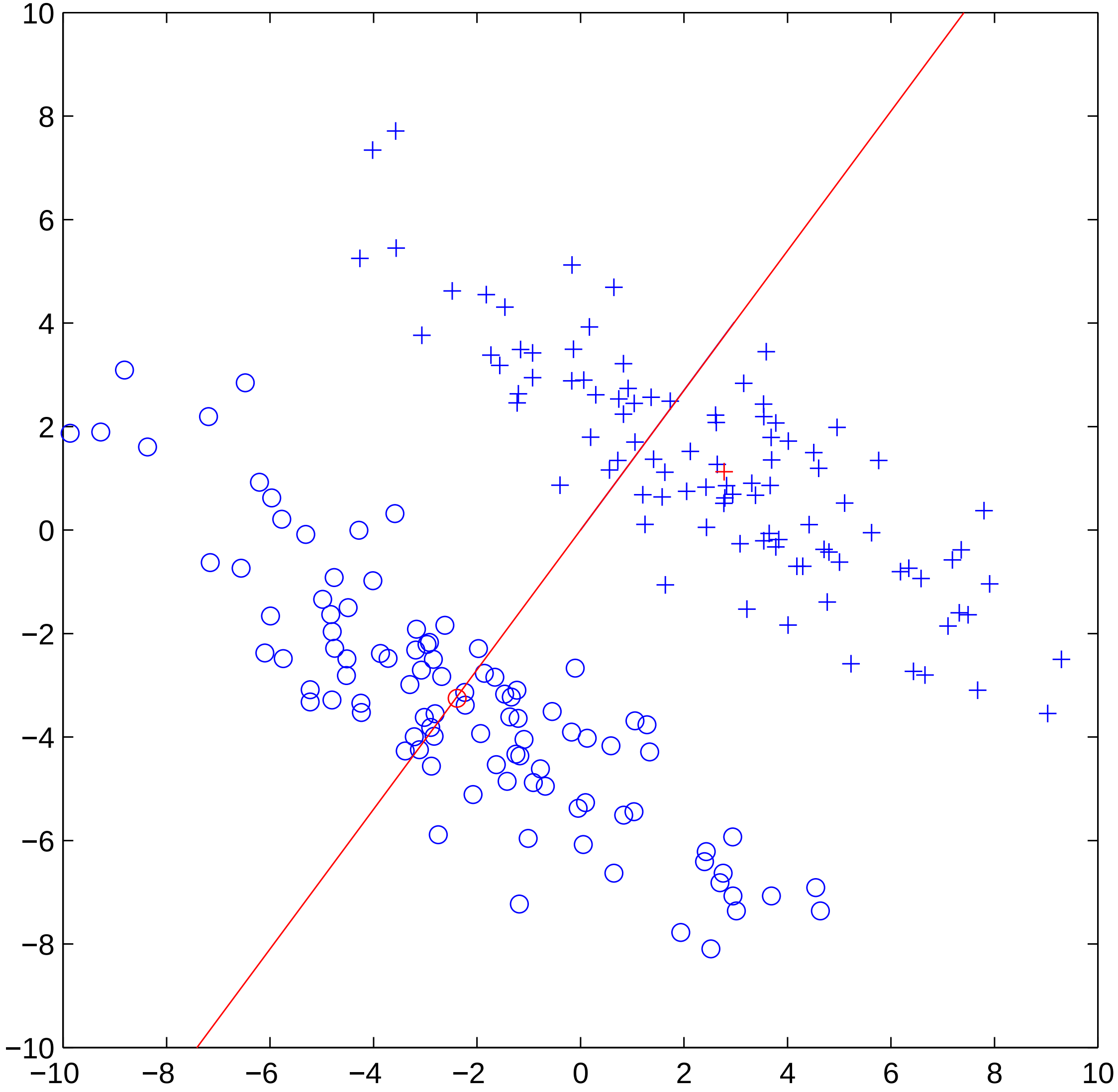
Matlab's Linear Discriminant Analysis

```
% compute within class scatter
S_plus     = mX_plus'  * mX_plus ;
S_minus    = mX_minus' * mX_minus ;
S          = S_plus + S_minus ;

% optimal w
w_opt = inv(S) * m_diff';

% normalize (arbitrary)
w_opt = w_opt / norm(w_opt);
```

Matlab's Linear Discriminant Analysis



Fisher's Linear Discriminant Analysis

- What have we achieved?
- General solution to project data onto a one-dimensional subspace (a line) with optimal class separation.
- Computational: Inversion of within class scatter matrix.
- One needs to fix a threshold in order to define a classifier.