Dimension Reduction

CS534

Why dimension reduction?

- High dimensionality – large number of features
  - E.g., documents represented by thousands of words, millions of bigrams
  - Images represented by thousands of pixels
- Redundant and irrelevant features (not all words are relevant for classifying/clustering documents
- Difficult to interpret and visualize
- Curse of dimensionality

Extract Latent Linear Features

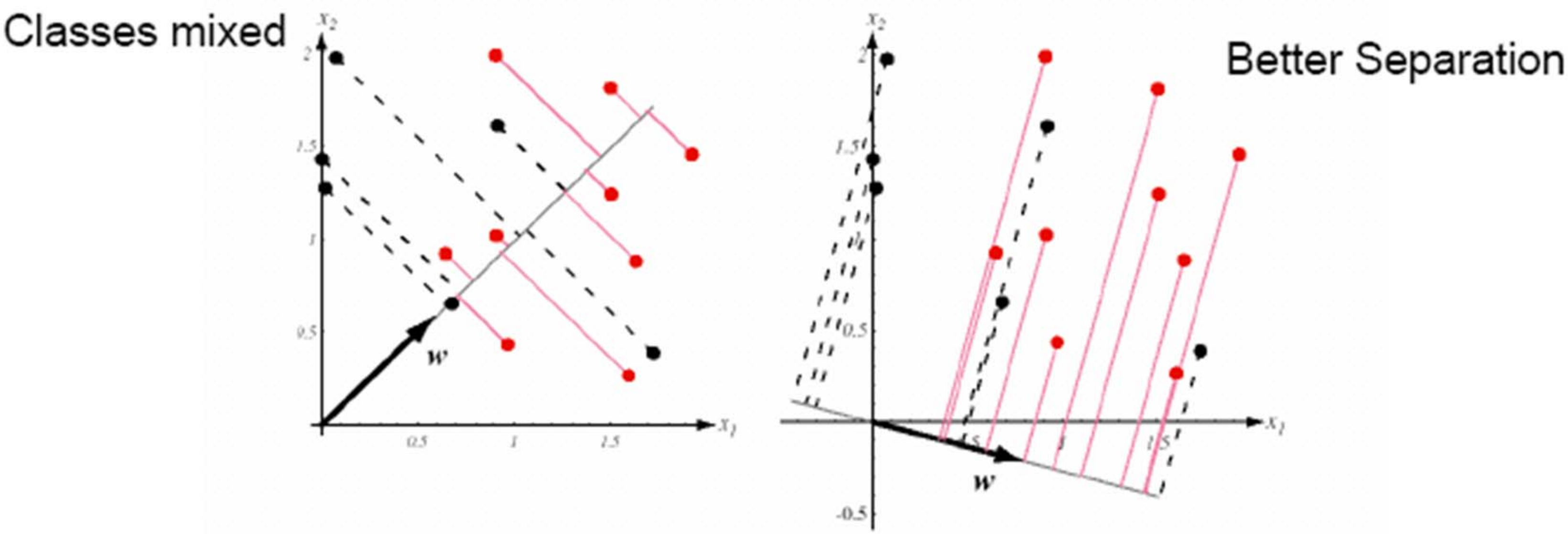
- Linearly project *n*-d data onto a *k*-d space
  - e.g., project space of 10^4 words into 3-dimensions
- There are infinitely many k-d subspaces that we can project the data into, which one should we choose
- This depends on the task at hand
  - If supervised learning, we would like to maximize the separation among classes: Linear discriminant analysis (LDA)
  - If unsupervised, we would like to retain as much data variance as possible: principal component analysis (PCA)

LDA: linear discriminant analysis

- Also named Fisher Discriminant Analysis
- It can be viewed as
  - *a dimension reduction* method
  - a generative classifier (p(x|y): Gaussian with distinct $\boldsymbol{\mu}$ for each class but shared Σ
- We will now look at its dimension reduction interpretation

Intuition

- Find a project direction so that the separation between classes is maximized
- In other words, we are looking for a projection that best discriminates different classes



Objectives of LDA

- One way to measure separation is to look at the class means

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{\mathbf{x}\in c_1}\mathbf{x} \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{\mathbf{x}\in c_2}\mathbf{x}$$

Original means

$$m'_1 = \frac{1}{N_1}\sum_{\mathbf{x}\in c_1}\mathbf{w}^T\mathbf{x} \qquad m'_2 = \frac{1}{N_2}\sum_{\mathbf{x}\in c_2}\mathbf{w}^T\mathbf{x}$$

Projected means

$$\left|m'_1 - m'_2\right|^2 = \left|\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right|^2$$

We want the distance between the projected means to be as large as possible

Objectives of LDA

- We further want the data points from the same class to be as close as possible
- This can be measured by the class ***scatter*** *(variance within the class)*

$$s_i^{\ 2} = \sum_{x \in c_i} (\mathbf{w}^T \mathbf{x} - m'_i)^2$$

Total within class scatter for projected class i

$$s_1^2 + s_2^2$$

Total within class scatter

Combining the two sides

- There are a number of different ways to combine these two sides of the objective
- LDA seeks to optimize the following objective:

$$\arg\max_{\mathbf{w}} \frac{|m'_1 - m'_2|^2}{s_1^2 + s_2^2}$$

$$|m_1' - m_2'|^2 = (w^T m_1 - w^T m_2)^2$$

$$= w^T (m_1 - m_2)(m_1 - m_2)^T w$$

$$\boxed{= w^T S_B w}$$

$$s_1^2 = \sum_{x \in C_1} (w^T x - w^T m_1)^2 = \sum_x w^T (x - m_1)(x - m_1)^T w$$

$$= w^T \left(\sum_x (x - m_1)(x - m_1)^T \right) w = w^T S_1 w$$

$$s_1^2 + s_2^2 = w^T (S_1 + S_2) w$$

$$\boxed{= w^T S_w w}$$

The LDA Objective

$$\boxed{J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}}$$

$$S_i = \sum_{x \in \mathbf{C}_i} (x - m_i)(x - m_i)^T$$

$S_B = (m_1 - m_2)(m_1 - m_2)^T$
the between class scatter matrix

$S_w = S_1 + S_2$
the total within class scatter matrix, where

$$S_i = \sum_{x \in C_i} (x - m_i)(x - m_i)^T$$

- The above objective is known as generalized Reyleigh quotient, and it's easy to show a w that maximizes $J(w)$ must satisfy $S_B w = \lambda S_w w$
- Noticing that $S_B w = (m_1 - m_2)(m_1 - m_2)^T w$ always take the direction of $m_1 - m_2$

Scalar

- Ignoring the scalars, this leads to:

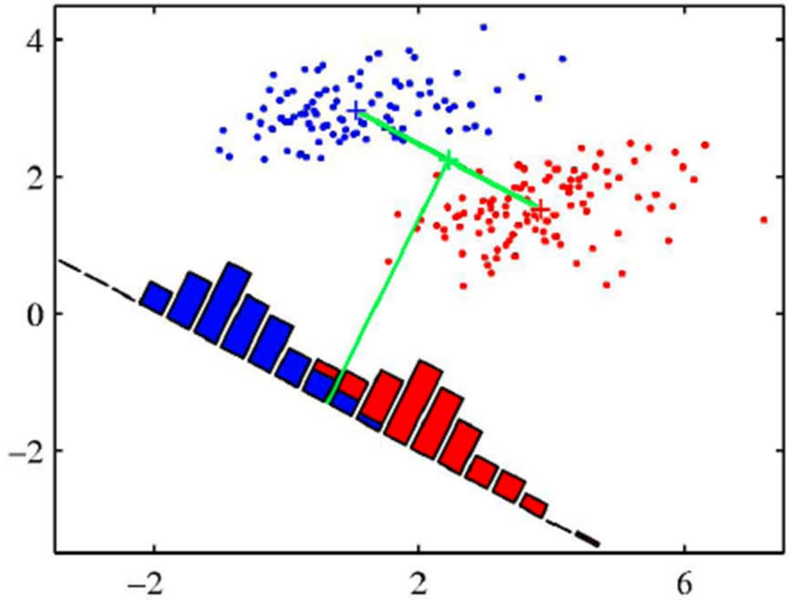
$$(m_1 - m_2) = S_w w$$

$$w = S_w^{-1}(m_1 - m_2)$$

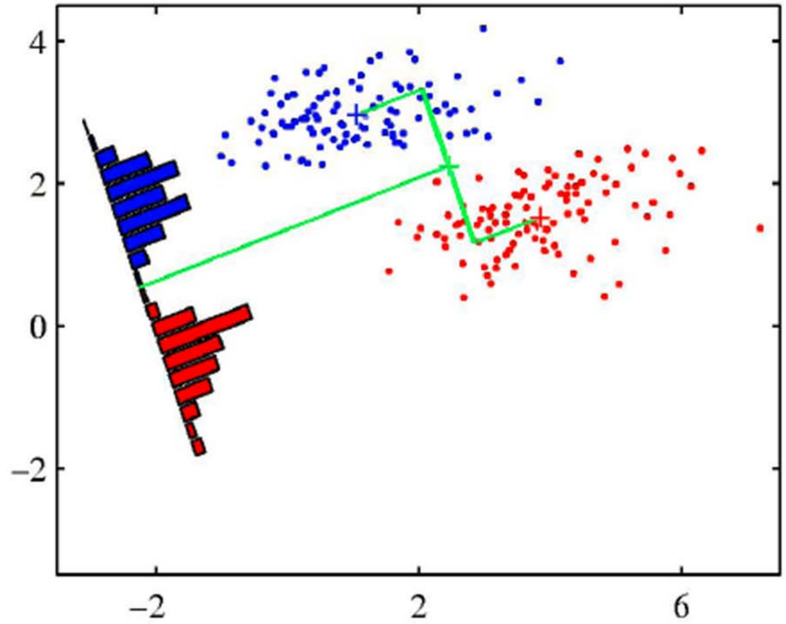
LDA for two classes

$$\mathbf{w} = S_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Projecting data onto one dimension that maximizes the ratio of between-class scatter and total within-class scatter



$w = m_1 - m_2$



$w = S_w^{-1}(m_1 - m_2)$

LDA for Multi-Classes

$$\boxed{J(w) = \frac{w^T S_B w}{w^T S_w w}}$$

- Objective remains the same, with slightly different definition for between-class scatter:

$$S_B = \frac{1}{k}\sum_{i=1}^{k}(m_i - m)(m_i - m)^T$$

m is the overall mean

- Solution: k-1 eigenvectors of $S_w^{-1} S_B$