INF8953CE - Fall 2020

Machine Learning

- Sarath Chandar

6. Dimensionality Reduction.

6. Dimensionality Reduction

Let $x \in \mathbb{R}^d$ be a random vector. We wish to find $k < d$ directions that captures as much as possible of the variance of X (It contains the most important info in the dataset)

Applications: ① feature extraction
② data visualization
③ lossy data compression.

1-D Case: Find the single direction that captures as much as possible of the variance of X.

Formally: We want $p \in \mathbb{R}^d$ (the direction) such that $\|p\| = 1$, so as to maximize $\text{var}(p^T x)$.

why $\|p\| = 1$?
Else you maximize the variance by letting $\|p\| \to \infty$.

Let $\mu = E[x]$

$$S = \text{Cov}(x) = E\left[(x-\mu)(x-\mu)^T\right]$$

For any $p \in \mathbb{R}^d$, the projection $p^T x$ has

mean: $E[p^T x] = p^T \mu$

variance: $\text{Var}(p^T x) = E\left[(p^T x - p^T \mu)^2\right]$

$$= E\left[p^T (x-\mu)(x-\mu)^T p\right]$$

$$= p^T S p$$

Goal: $\max_{\|p\|=1} p^T S p = \max_{p \neq 0} \frac{p^T S p}{p^T p}$

$$= \max_{p \neq 0} \frac{p^T Q \Lambda Q^T p}{p^T Q Q^T p}$$

[$Q \Lambda Q^T$ is the spectral decomposition of S]

[$\because QQ^T = I$] since it is orthonormal

matrix with orthonormal row vectors → eigen vectors.

diagonal matrix with eigen values

$$= \max_{y \neq 0} \frac{y^T \Lambda y}{y^T y} \quad [\text{writing } y = Q^T p]$$

$$= \max_{y \neq 0} \frac{\lambda_1 y_1^2 + \ldots + \lambda_d y_d^2}{y_1^2 + \ldots + y_d^2}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ are the eigen values and corresponding eigen vectors $u_1 \ldots u_d$.

$$= \max_{y \neq 0} \frac{\lambda_1 y_1^2}{\sum_i y_i^2} + \ldots + \lambda_d \frac{y_d^2}{\sum_i y_i^2}$$

$$\leq \lambda_1$$

where equality is attained in the last step when

$$y = e_1 \quad \text{i.e.} \quad Z = Qe_1 = u_1$$

where u_1 is the first eigen vector of S.

We call this the principal component.

First principle component → first eigen vector of Cov(x).

Similarly, k-dimensional subspace that captures as much as possible of the variance of X is simply the subspace spanned by the top-k eigen vectors of Cov(x) : $u_1, \ldots u_k$.

This is known as Principal Component Analysis (PCA).

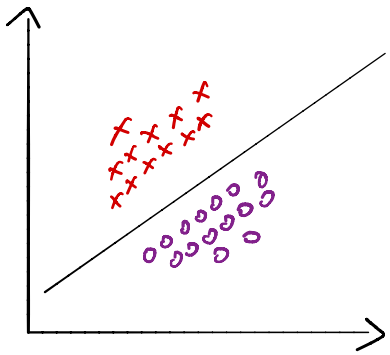
Procedure:-

(1) Compute the mean μ and the Covariance matrix S of the data X.

(2) Compute the top-k eigen vectors $u_1 \ldots u_k$ of S.

(3) Project $x \rightarrow P^T x$, where P^T is the $k \times d$ matrix whose rows are $u_1 \ldots u_k$.

Can we use PCA dimensions for classification ?



Projecting data points to the first principal component collapses the classes!

Better solution:- Find the projection that maximizes the class separation.

Consider a 2-class problem in which there are N_1 points of class C_1 and N_2 points of class C_2.

Mean vectors of C_1, C_2 are:

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x^{(n)}$$

$$m_2 = \frac{1}{N_2} \sum_{n \in C_2} x^{(n)}$$

Choose w so as to maximize

$$w^T m_2 - w^T m_1$$

However this expression can be made arbitrarily large by increasing the magnitude of w.

↳ Constrain $\|w\| = 1$

maximize $w^T m_2 - w^T m_1$
Subject to $w^T w = 1$

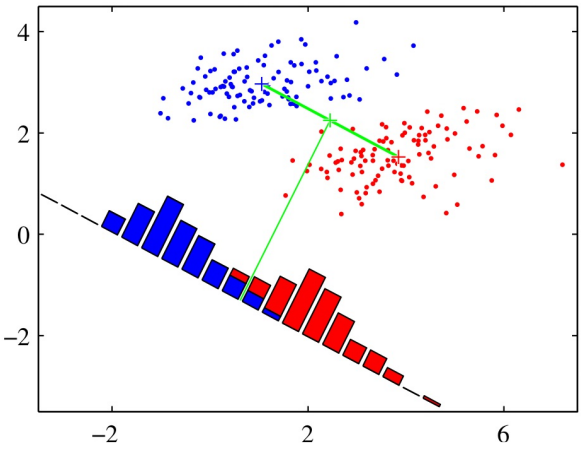
Lagrangian formulation:

$$L = w^T(m_2 - m_1) + \lambda (w^T w - 1)$$

$$\frac{\partial L}{\partial w} = (m_2 - m_1) + 2\lambda w = 0$$

$$w = -\frac{1}{2\lambda}(m_2 - m_1)$$

$$w \propto (m_2 - m_1)$$



Two classes are well separated in the original space.

But have considerable overlap in the projected space.

→ because of the strongly nondiagonal covariances of class distributions.

Solution proposed by Fisher:-

Maximize a function that will give a large separation between the projected class means while also give a small variance within each class, thereby minimizing class overlap.

within-class variance of the transformed data for class C_k is

$$s_k^2 = \sum_{n \in C_k} \left(w^T x^{(n)} - w^T m_k \right)^2$$

Total within-class variance $= s_1^2 + s_2^2$

Fisher criterion: $J(w) = \dfrac{\left(w^T m_2 - w^T m_1 \right)^2}{s_1^2 + s_2^2}$

$$\left(w^T m_2 - w^T m_1 \right)^2 = \left(w^T (m_2 - m_1) \right)^2$$

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

$$= w^T S_B w \quad \text{——} \quad (1)$$

$\hookrightarrow$ between class covariance matrix.

$$s_1^2 + s_2^2 = \sum_{n \in C_1} \left(w^T (x^{(n)} - m_1) \right)^2 + \sum_{k \in C_2} \left(w^T (x^{(k)} - m_2) \right)^2$$

$$= \sum_{n \in C_1} w^T (x^{(n)} - m_1)(x^{(n)} - m_1)^T w + \sum_{k \in C_2} w^T (x^{(k)} - m_2)(x^{(k)} - m_2)^T w$$

$$= w^T S_w w \quad \text{——} \quad (2)$$

where $S_w = \sum_{n \in C_1} (x^{(n)} - m_1)(x^{(n)} - m_1)^T + \sum_{n \in C_2} (x^{(k)} - m_2)(x^{(k)} - m_2)^T$

Using ① and ②,

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

diff. w.r.t. w and set it to 0.

$$(w^T S_B w) S_w w - (w^T S_w w) S_B w = 0$$

$$(w^T S_B w) S_w w = (w^T S_w w) S_B w \quad \text{—— ④}$$

We do not care about the magnitude of w, only its direction. So drop the scalars $(w^T S_B w)$, $(w^T S_w w)$ in ④.

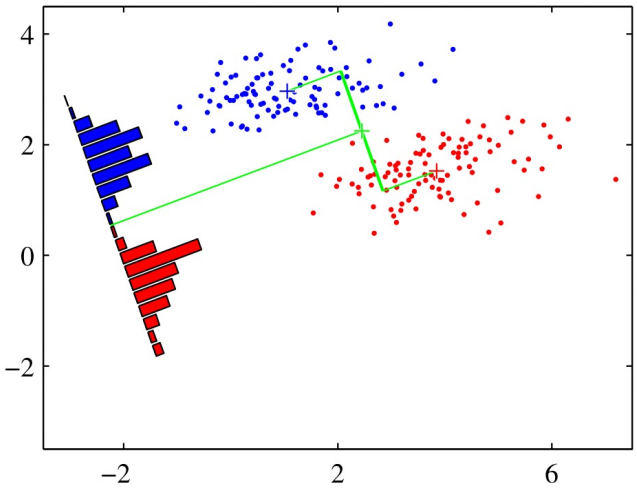
$$S_w w = S_B w.$$

We know that $S_B = (m_2 - m_1)(m_2 - m_1)^T$.

So $S_B w$ is always in the direction of $(m_2 - m_1)$.

$$S_w w \propto (m_2 - m_1)$$

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



Note: If the within-class covariance is isotropic, so that S_W is proportional to the unit matrix, w is proportional to the difference of the class means.

This model is known as Fisher's linear discriminant.

However it is not a discriminant. It is a specific choice of direction for projection of data.

The projected data can be used to construct a discriminant - by choosing a threshod θ.

$$w^T x \geq \theta \Rightarrow C_1$$

$$w^T x < \theta \Rightarrow C_2$$

θ can be found by minimizing the training error.

You should know!

1. Dimensionality reduction.
2. Applications of dimensionality reduction.
3. Principal Component Analysis (PCA).
4. Fisher's linear discriminant analysis.