Course: MATH 6397
Semester: Spring 2015
Instructor: Dr. Robert Azencott
Lecture Date: 02/17/2015
Prepared by: Zachery Hernandez

Reminder: Project 1 Due Date

- this Thursday (send by email)
- send 2 emails w/detailed recommendations
- provide 10 attributes
- perform linear discrimination between 2 classes using LDA
- visualize data points in 2 or 3 dimensions, treating each attribute as a dimension

---

1 Dimension Reduction by PCA and Mapping SVM into Hilbert Space

Recall objects described by vectors $v \in \mathbb{R}^k$
$v(1), v(2), \ldots, v(N)$: N objects

$$v(m) = \begin{cases} v_1(m) \\ \vdots & \text{each object described by k attributes} \\ v_k(m) \end{cases}$$

1.1 Elementary Linear Analysis:

PCA: Principal Component Analysis (Tibshirani)
For more information, please refer to Hastie, ..., Freedman.

1.1.1 Implementation of PCA:

Let $\bar{v}$ = center of cloud (in $\mathbb{R}^k$) of data vectors, where

$$\bar{v} = \frac{1}{N} \sum_{n=1}^{N} v(n)$$

Next, center the data by defining $W(m)$ such that $W(m) = V(m) - \bar{V}$ and $\bar{W} = 0$. Now we have centered cloud $W(1), \ldots, W(N)$.

Assume that the covariance matrix, Σ, of W is equal to the covariance matrix of V, where

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} W(x) W^T(x)$$

Note: Σ is always symmetric, semi positive definite, has real eigenvalues, $\lambda_1, \ldots, \lambda_k$ and eigenvectors $U_1, \ldots, U_k$.

Best case example: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$
Ill-conditioned example: $\epsilon \geq \lambda_{k-3} \geq \lambda_{k-2} \geq \lambda_{k-1} \geq \lambda_k \geq 0$, where ϵ is very small.

Can be found by graphing the eigenvalues (Fig. 1).
The bad gap shown means 3 attributes (6,7,8) can be written as a linear combination of other attributes. $1 \leq j \leq k$

1.1.2 History of PCA:

Developed and used widely between 1945 and 1975, before kernel PCA became more widespread. Main contributors to the method include Tucker, Ben Zekri, and others.

Goal of PCA: Find a subspace of $\mathbb{R}^k$ called $\mathcal{E}$ then project $\mathbb{R}^k$ to $\mathcal{E}$ ($Proj\mathbb{R}^k \longrightarrow \mathcal{E}$) and be able to visualize the data.

What is a good projection?

Best projection can be found by taking the top eigenvector of the covariance of the data (Fig. 2).

What is the best space $\mathcal{E}$ of dim r?

The best $\mathcal{E}$ can be generated by r top (largest) eigenvectors of Σ.

How do we choose r?

Mining and Automatic Learning/sortedEigsPlot.png

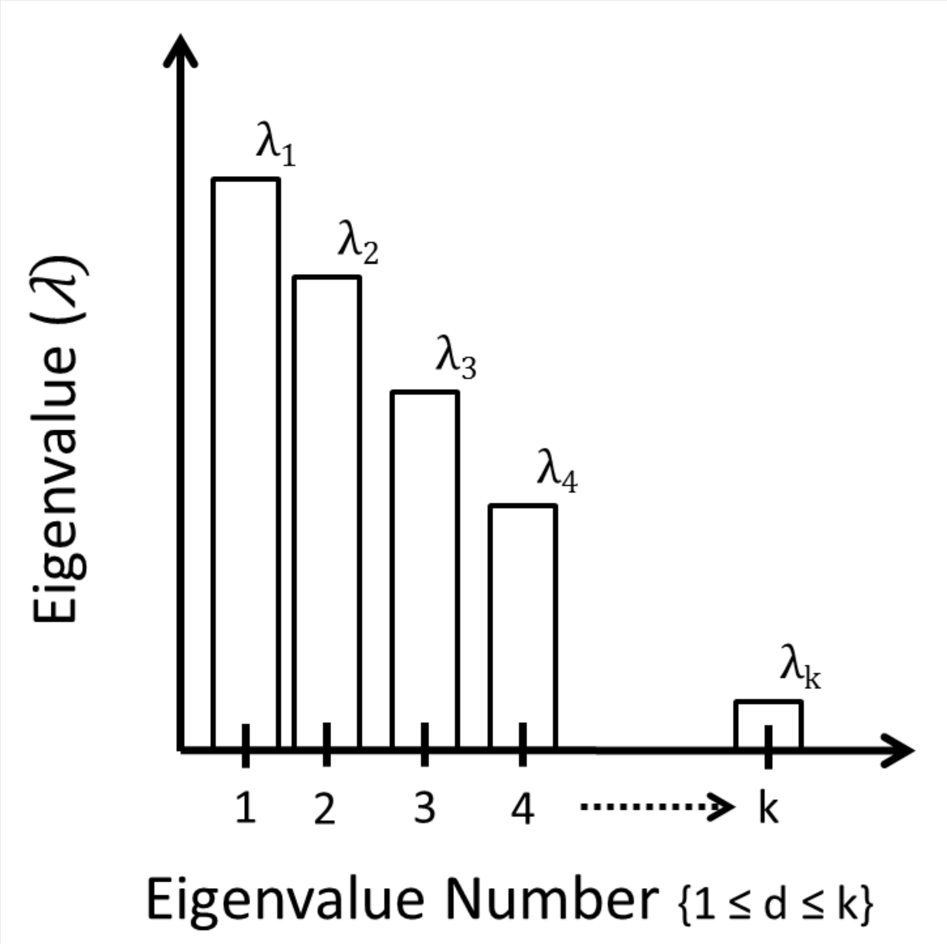

Figure 1: Bar graph depicting the decreasing eigenvalues d of the covariance matrix Σ.

Assume we want to project $v \to U_1, U_2, \ldots, U_r$ (Eigenvectors of Σ).

Let's call them $x_i(v) = \langle v, U_i \rangle$

$$Proj_{\mathcal{E}}(v) = \sum_{j=1}^{r} x_j(v) U_j = P(v)$$

Energy Ratio

$$ERAT(r) = \frac{\lambda_1 + \cdots + \lambda_r}{\lambda_1 + \cdots + \lambda_k}$$

Can look for energy gap (Fig. 3) or look for r such that $ERAT(r) \cong 90\%$

Classical PCA

$$Proj_{\mathcal{E}}(v) = \begin{bmatrix} x_1(v) \\ \vdots \\ x_r(v) \end{bmatrix} = \begin{bmatrix} \langle v, U_1 \rangle \\ \vdots \\ \langle v, U_r \rangle \end{bmatrix}$$

1.2 Extending the Maximum Margin Linear Discrimination to Hilbert Spaces:

- Crucial step to define SVM in infinite dimensions
- SVM = Support Vector Machines

Recall in finite dimensions:

- using maximum margin linear discrimination
- in $\mathbb{R}^k$
- and a dataset with $x(1), \ldots, x(N)$ vectors in $\mathbb{R}^k$
- compute a separator $sep(x) = \langle w, x \rangle + b \begin{cases} \text{+1 class} & \text{if sign} < 0 \\ \text{-1 class} & \text{if sign} < 0 \end{cases}$
- Goal: find $w \in \mathbb{R}^k$ and $b \in \mathbb{R}$
- use a training set: $x(1) \cdots x(N)$ and corresponding target vector $y(1) \cdots y(N)$, where $x(i) \rightsquigarrow y(i), 1 \leq i \leq N$
- find the maximum margin though the minimization problem

  $$\min_{w,b,\xi_j} \frac{1}{2}\|w\|^2 + C\sum_{j=1}^{N} \xi_j$$

  parameter: C such that $C > 0$

  constraints: $\xi_i - [1 - y(i)[\langle w, x(i) \rangle + b)] \geq 0$, $\xi_i \geq 0$

Previously used the Lagrangian (or Karaush-Kuhn-Tucker) method, where

$$w = \sum_{i=1}^{N} \beta_i x(i) \begin{cases} \text{where} & 0 \leq \beta_i \leq C \\ \text{where} & \alpha_i + \beta_i = C \end{cases}$$

and ended up using an alternative minimization problem, defined as

$$\min_{\alpha_i} \alpha_i y(i) - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j \langle x(i), x(j) \rangle_{\mathbb{R}^k} y(i)y(j), \text{ where } 0 \leq \alpha_i \leq C$$

Now: we introduce a kernel $k(v, w)$, where $v, w \in \mathbb{R}^k$ (i.e. the kernel k operates in Hilbert space H).

Can also be defined as $\langle \varphi_x, \varphi_x' \rangle_H = k(x, x')$, where $x(i) \mapsto \varphi_{x(i)}$ and $\mathbb{R}^k \mapsto \mathbb{R}$

Consider a vector $u \in \mathbb{R}^k$. It's projection into Hilbert space can be defined as

$$\varphi_{x(i)}(u) = k(x(i), u)$$

The operator $\varphi_{x(i)}$ can thus be represented as $X_i \in H$.

And now a new separator is defined as $sep(X)$, in Hilbert space, such that

$$sep(X) = \langle W, X \rangle_H + b, \text{ where } W \in H,\ b \in \mathbb{R}$$

and the rest is the same as above with products W and X in the Hilbert space H.

So the problem is to find $W \in H$, $b \in \mathbb{R}$, and ξ_i

The weighting variable W can be optimized as, $W = \sum\limits_{i=1}^{N} \beta_i x_i$.

Differences between original space and Hilbert space:

1. all vectors x become X
2. all products in $\mathbb{R}^k$ become products in H
3. $\langle x_i, x_j \rangle \equiv \langle \varphi_{x(i)}, \varphi_{x(i)} \rangle = k(x(i), x(j)) = M_{i,j}$

So we get α_i from

$$\min_{\alpha_i} \alpha_i y(i) - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j \langle x_i, x_j \rangle_H y_i, y_j, \text{ where } 0 \leq \alpha_i \leq C$$

And by $\alpha_i + \beta_i = C$, we get W from $W = \sum\limits_{i=1}^{N} \beta_i x_i$.

This yields $\langle W, X \rangle_H = \sum\limits_i \alpha \langle x_i, x \rangle_H = \sum\limits_i \alpha_i k(x_i, x)$

Thus $sep(\sum\limits_i \alpha_i k(x_i, x) + b)$ is considered as the new separator function $sep(X)$ for linear discrimination in Hilbert space.

Mining and Automatic Learning/scatterPlot2DforPCA.png

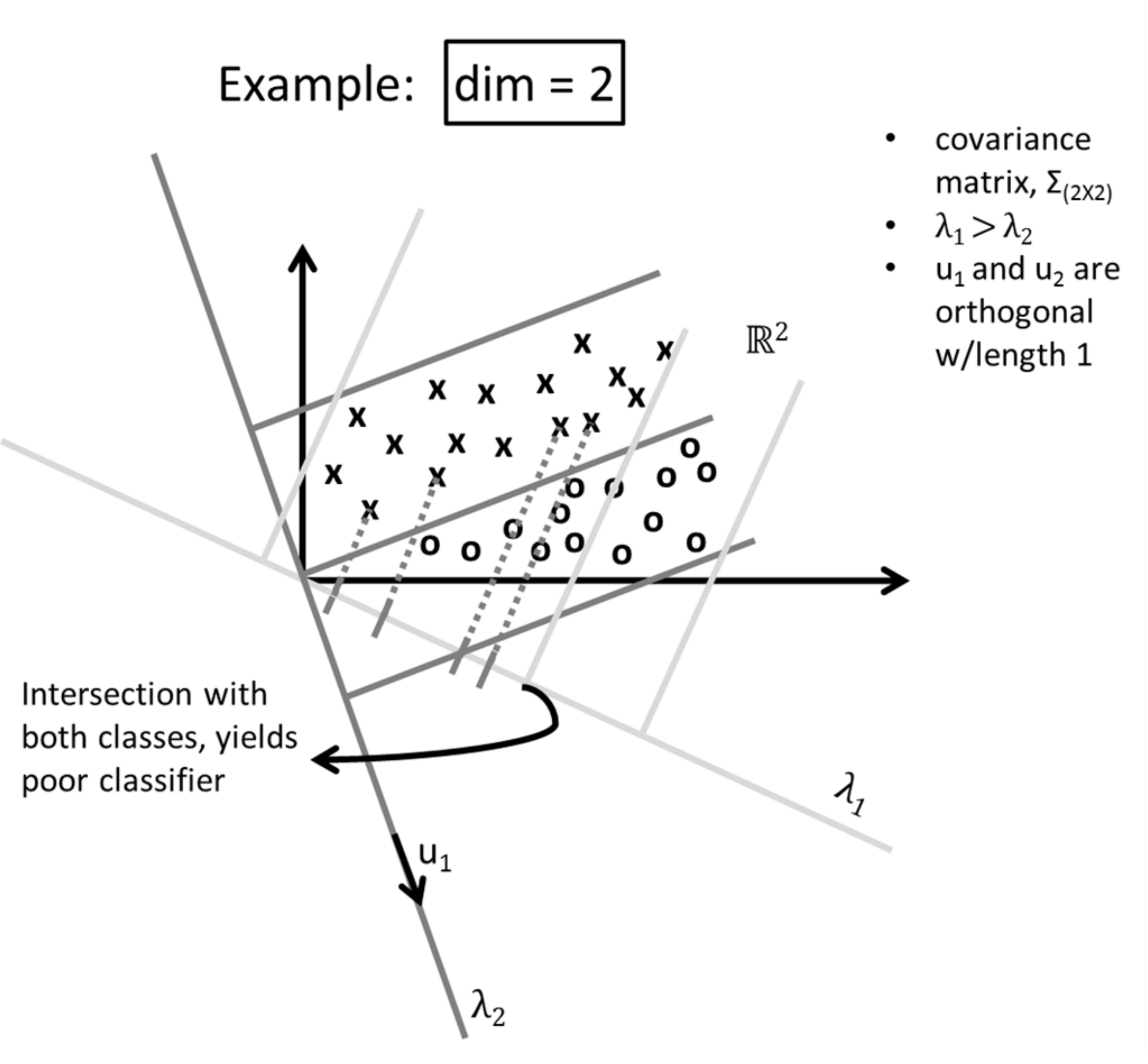
Example: dim = 2
• covariance matrix, $\Sigma_{(2X2)}$
• $\lambda_1 > \lambda_2$
• u_1 and u_2 are orthogonal w/length 1
$\mathbb{R}^2$
Intersection with both classes, yields poor classifier
λ_1
u_1
λ_2

Figure 2: Two-dimensional scatter plot of two classes of data point, represented by 'x's and 'o's.

Mining and Automatic Learning/ERATplot.png

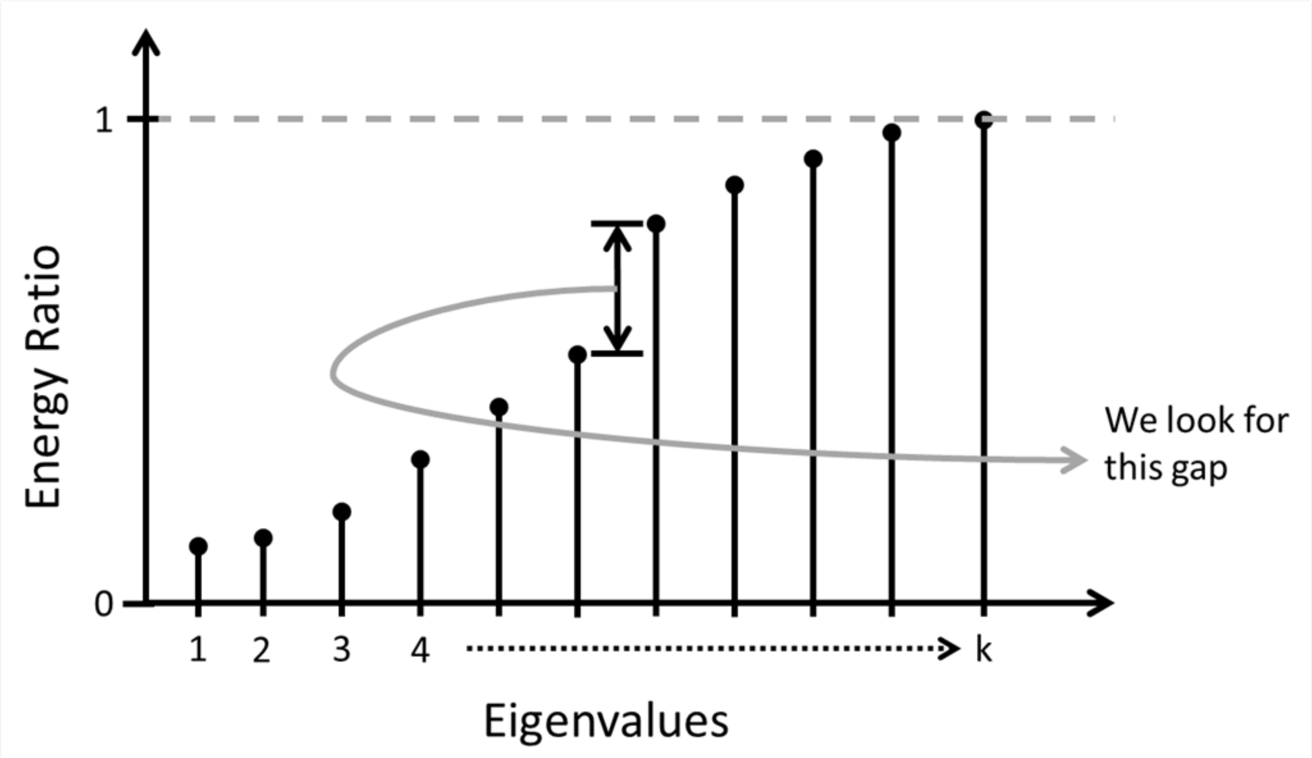

Figure 3: Example plot showing the energy ratio $ERAT(r)$ of each eigenvalue r.