CSCI3230 (ESTR3108)
Fundamentals of Artificial Intelligence

Tutorial 5

Yuan Zhong

Email: yzhong22@cse.cuhk.edu.hk
Office: Room 1024, 10/F, SHB

Dept. of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

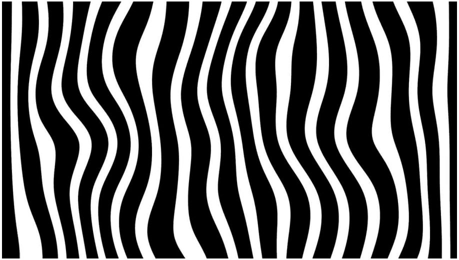
Part 1. Change of basis

Part 2. Spectral theory

Part 3. Statisical view of PCA

Principal components

Principal components



Principal components

Principal components

Curse of dimensionality in binary classification

Consider a statistical binary classification model which contains no assumptions and constraints.

- When we have an infinite dataset

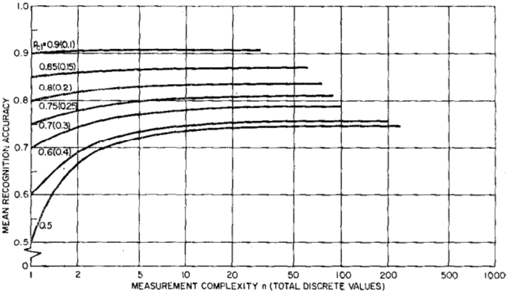


Fig. 2. Infinite data set accuracy.

Consider a statistical binary classification model which contains no assumptions and constraints.

- When we have a finite dataset: peaking phenomenon

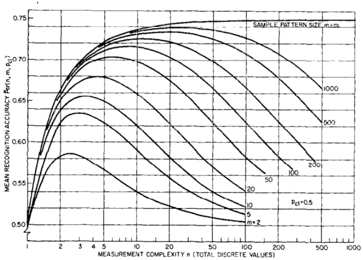


Fig. 3. Finite data set accuracy ($p_{c1} = \frac{1}{2}$).

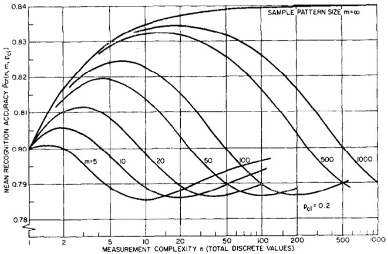
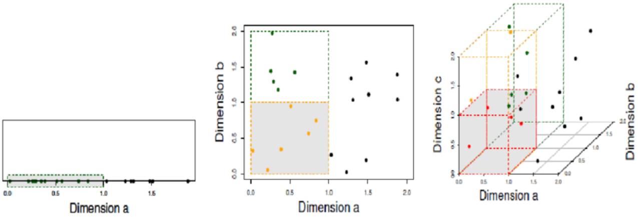


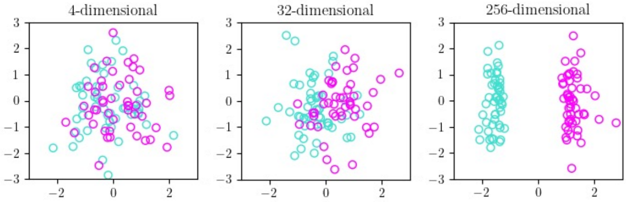
Fig. 4. Finite data set accuracy ($p_{c1} = \frac{1}{5}$).

An intuitive explanation of curse of dimensionality in binary classification

- The space is getting sparser with increasing dimensions



- Projecting onto 2-dimension:



Part 1. Change of basis

Subspace

A subset $\mathcal{S}$ of $\mathbb{R}^m$ is said to be a subspace if

$$\mathbf{x}, \mathbf{y} \in \mathcal{S}, \quad \alpha, \beta \in \mathbb{R} \qquad \Rightarrow \qquad \alpha\mathbf{x} + \beta\mathbf{y} \in \mathcal{S} \tag{1}$$

- If $\mathcal{S}$ is a subspace and $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathcal{S}$, any linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_n$, i.e., $\sum_{i=1}^n \alpha_i \mathbf{a}_i$ for any $\boldsymbol{\alpha} \in \mathbb{R}^n$ lies in $\mathcal{S}$.
- Some quick facts: let $\mathcal{S}_1, \mathcal{S}_2$ be subspaces of $\mathbb{R}^m$
  - $\mathcal{S}_1 + \mathcal{S}_2$ is a subspace.
  - $\mathcal{S}_1 \cap \mathcal{S}_2$ is a subspace.

Span

The span of a collection of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ is defined as

$$\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \left\{ \mathbf{y} \in \mathbb{R} \;\middle|\; \mathbf{y} = \sum_{i=1}^{n} \alpha_i \mathbf{a}_i,\ \boldsymbol{\alpha} \in \mathbb{R}^n \right\} \tag{2}$$

- the set of all linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_n$.
- a subspace
- Question: any span is a subspace. But can any subspace be written as a span?

Span

The span of a collection of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ is defined as

$$\text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \left\{ \mathbf{y} \in \mathbb{R} \;\middle|\; \mathbf{y} = \sum_{i=1}^{n} \alpha_i \mathbf{a}_i, \;\; \boldsymbol{\alpha} \in \mathbb{R}^n \right\} \tag{2}$$

- the set of all linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_n$.
- a subspace
- Question: any span is a subspace. But can any subspace be written as a span?
  - **Theorem**: let $\mathcal{S}$ be a subspace of $\mathbb{R}^m$. There exists a integer n and a collection of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathcal{S}$ such that $\mathcal{S} = \text{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$.
  - **Implication**: we can always represent a subspace by a span.

Linear independence

- Suppose we have vectors $u_1 \dots u_n \in \mathbb{R}^n$
- If $u_1 \dots u_n$ are independent then:

$$c_1 u_1 + \cdots + c_n u_n = 0, \quad c_i \in \mathbb{R}, i = 1 \dots n$$

iff $c_1 = \cdots = c_n = 0$

- Alternatively, we can write this statement in matrix-vector multiplication: Let $A = [u_1 \dots u_n]$, $c \in \mathbb{R}^n$:

$$Ac = 0$$

iff $c = 0$, which also implies A is full-rank.

Quiz

Given $u_1 \dots u_n \in \mathbb{R}^n$, for any pair (u_i, u_j), $i \neq j$, u_i and u_j are independent. Can we say $u_1 \dots u_n$ are independent to each other?

Basis

- Basis is a set of independent vectors, which can span the vector space by linear combination.
- Denote basis as $B = \{u_1, \dots, u_n\}$, where $u_1 \dots u_n$ are independent.
- Let $U = [u_1 \dots u_n]$. The subspace spanned by B is exactly the range space (column space) of U.
- If S is spanned by B, then any vector $v \in S$, we can find scalars $c_1, \dots, c_n$ such that:

$$v = c_1 u_1 + \cdots + c_n u_n$$

- The number of vectors in the basis = dimension: $dim(S) = n$.

Basis

Typical basis:

- Canonical basis: $\{e_1, \ldots, e_n\}$, a very useful and common basis of $\mathbb{R}^n$. For any i, e_i is a vector where only the i-th element equal to 1 and others are zeros.
    1. For example, in R^2 space, $e_1 = [1, 0]$ and $e_2 = [0, 1]$ are the two canoical basis vectors.
    2. Coordinates like $(x_1, \ldots, x_n)$ without other context is actually the shorthand of $x_1 e_1 + \cdots + x_n e_n$.
- Orthonormal basis: a basis $[u_1, \ldots, u_n]$ is orthonormal iff $u_j^T u_i = 0$ when $i \neq j$ and $u_i^T u_i = 1$ (orthogonal and normalized).
    1. For example, canonical basis is orthonormal.
    2. Basis $\{[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}], [\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]\}$ is orthonormal.

Basis

- Note that basis of a subspace is not unique.
  1. Besides of the canonical basis e_1, e_2, e_3, $u_1 = (1,1,1)^T, u_2 = (0,1,0)^T, u_3 = (1,2,2)^T$ can also be a basis for 3-D Euclidean space.
  2. Basis for a plane in $\mathbb{R}^3$ is not unique as well. Spanning of $(0,1,0)^T, (1,0,0)^T$ and $(1,1,0)^T, (1,0,0)^T$ are the same plane.
- Different basis may have different use.
  1. In computer graphics, we often need to transform a point from the world coordinate system to the camera coordinate system.
  2. Discrete signals are generally recorded in time domain. But we usually do Fourier transform to get the frequency domain information. Fourier transform: canonical basis $\Rightarrow$ fourier basis.
- This is the reason why we need change of basis.

Change of basis

Change of basis transformation

Given a basis $B = \{u_1, \ldots, u_n\}$, let $A = [u_1, \ldots, u_n]$, the change of basis transformation from canonical basis to B is A^{-1}.

- Convsersely, A is the change of basis transformation from B to canonical basis.
- For example: Suppose we have basis $[1, 0, 0]$ and $[0, 0, 1]$. A coordinate $p = [2, 3]$ represented by this basis is actually the coordinate $[2, 0, 3]$ under the canoical basis. By applying the theorem, $Ap = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}$.

Part 2. Spectral theory

Eigenvalues and eigenvectors

$\forall A \in \mathbb{R}^{n\times n}$, if we have the equation below for some λ and v,

$$Av = \lambda v$$

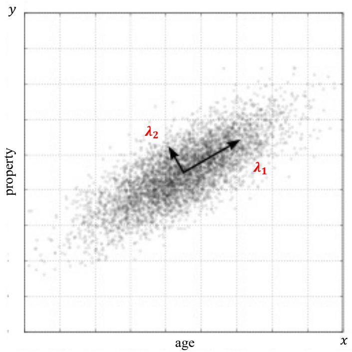
- λ is called eigenvalue and the corresponding non-zero v is called eigenvector asscoiated with the λ.
- Eigenvalues are solutions to $det(A - \lambda I) = 0$, which is a n-order polynomial.
- When there are n independent eigenvectors, A can be diagonalizable.

$$A = U\Lambda U^{-1}$$

where Λ is a diagonal matrix with eigenvalues $\lambda_1 \dots \lambda_n$ on its diagonal, U's i-th column is the eigenvector associated with λ_i

Eigenvalues and eigenvectors: a toy example

Consider a dataset of people with 2 attributes: age and property.



- Two eigenvectors with corresponding eigenvalues λ_1, λ_2.
- Projection onto the direction of λ_1 leads to minimal information loss.
- Statistical intuition: the direction of λ_1 shows a larger variance.

Spectral theory for symmetric matrix

- For symmetric A (i.e., $A^T = A$), A must have n real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and the associated orthonormal eigenvectors $v_1, v_2, \ldots, v_n$:

$$Av_i = \lambda_i v_i$$

  - A can be diagonalized as

$$A = U \Lambda U^T$$

    where $U = [v_1 \ldots v_n]$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$
  - Spectral resolution: $A = \sum_{i=1}^{n} \lambda_i v_i v_i^T$
  - Courant-Fischer theorem:

$$\lambda_i = \max_{dim V = i} \min_{x \in V, \|x\|_2 = 1} x^T A x$$

Quiz

Let this quiz help with recapping spectral theory. Which statement below about eigenvalues and eigenvectors is correct?

- A Eigenvectors of $A \in \mathbb{R}^{n \times n}$ need not to be independent.
- B Any matrix $A \in \mathbb{R}^{n \times n}$ must have n eigenvalues.
- C Symmetric matrix $A \in \mathbb{R}^{n \times n}$ must have n non-repeated eigenvalues.
- D Any matrix $A \in \mathbb{R}^{n \times n}$ is NOT invertible iff 0 is an eigenvalue of A.

Quiz

Let this quiz help with recapping spectral theory. Which statement below about eigenvalues and eigenvectors is correct?

A Eigenvectors of $A \in \mathbb{R}^{n\times n}$ need not to be independent.

B Any matrix $A \in \mathbb{R}^{n\times n}$ must have n eigenvalues.

C Symmetric matrix $A \in \mathbb{R}^{n\times n}$ must have n non-repeated eigenvalues.

D Any matrix $A \in \mathbb{R}^{n\times n}$ is NOT invertible iff 0 is an eigenvalue of A.

Correct Answer: D

Part 3. Statisical view of PCA

Statistical view

Recall the statistical view of PCA:

Statistical view of PCA

For a multi-variate random variable $X \in \mathbb{R}^D$ (zero-mean), we need to find $d \ll D$ independent unit vectors $u_1, \ldots, u_d \in \mathbb{R}^D$ such that $y_i = u_i^T X$, $u_i^T u_i = 1$ and we have: $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \cdots \geq \mathrm{Var}(y_d)$. The optimal solution to $u_1, \ldots, u_d$ is the first d eigenvectors of $\mathrm{E}[XX^T]$ asscoiated with its d largest eigenvalues.

- $\mathrm{E}[XX^T]$ is the covariance matrix of X.
- Each sample (column) of data matrix $\mathbf{X}$ can be seen as an observation of random feature vector X.
- Given a zero-mean data matrix $\mathbf{X} \in \mathbb{R}^{D \times m}$, the sample covariance matrix can be computed as $\frac{1}{m}\mathbf{X}\mathbf{X}^T$.

Proof of PCA in statistical view

Firstly, note that random variable $X \in \mathbb{R}^D$ is zero-mean, then $y_i = u_i^T X$ is also zero-mean due to the linearity of $\mathrm{E}(\cdot)$.

$$\mathrm{Var}(y_i) = \mathrm{E}(y_i y_i^T) = \mathrm{E}(u_i^T X X^T u_i) = u_i^T \mathrm{E}(X X^T) u_i$$

Denote $[v_1, v_2, \ldots, v_D]$ as eigenvectors of $\mathrm{E}(XX^T)$ associated with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$. Eigenvectors $[v_1, \ldots, v_D]$ is an orthonormal basis of $\mathbb{R}^D$ since $\mathrm{E}(XX^T)$ is symmetric.

PCA in statistical view is equivalent to:

1. $u_1 := \arg\max_{u \in \mathbb{R}^D} u^T \mathrm{E}(XX^T) u = v_1$
2. $u_2 := \arg\max_{u \in \mathrm{span}(v_2, \ldots, v_D)} u^T \mathrm{E}(XX^T) u = v_2$

   $\cdots\cdots\cdots$
3. $u_d := \arg\max_{u \in \mathrm{span}(v_d, \ldots, v_D)} u^T \mathrm{E}(XX^T) u = v_d$

Proof of PCA in statistical view (cont.)

We will prove the proposition below and then statistical view of PCA will be proved by the proposition.

Proposition 1

Any symmetric $A \in \mathbb{R}^{n \times n}$ with eigenvectors $[v_1 \dots v_n]$ associated with $\lambda_1 \geq \dots \lambda_n$. For every $i \in \{1, \dots, n\}$, we have:

$$\lambda_i = \max_{u \in \mathrm{span}(v_i, \dots, v_n)} u^T A u$$

The optimal solution to achieving the equality is $u = v_i$.

By inserting $\mathrm{E}(XX^T)$ as A to this proposition, we can directly see the optimal solution of PCA in statistical view.

Proof of Proposition 1:
$\forall \xi \in \mathrm{span}(v_i, \dots, v_n), \|\xi\|_2 = 1$, $\exists\ c = [c_i, \dots, c_n]$.
We have $\xi = \sum_{j=i}^{n} c_j v_j$ since $[v_1 \dots v_n]$ is a basis of $\mathbb{R}^n$.

Firstly, we will show $\sum_{i=1}^{n} c_i^2 = 1$:

$$\xi^T \xi = 1 = \sum_{j=i}^{n} c_j v_j^T \sum_{j=i}^{n} c_j v_j = \sum_{j=i}^{n} \sum_{k=i}^{n} c_j c_k v_j^T v_k$$

Since $v_i \dots v_j^T v_k = \begin{cases} 1, j = k \\ 0, j \neq k \end{cases}$ $\quad \xi^T \xi = \sum_{j=i}^{n} c_j c_j v_j^T v_j = \sum_{j=i}^{n} c_j^2 = 1$

Proof of PCA in statistical view (cont.)

Also, by spectral resolution, we have

$$A = \sum_{k=1}^{n} \lambda_k v_k v_k^T$$

Then, we can write $\xi^T A \xi$ as:

$$\begin{aligned}
\xi^T A \xi &= \sum_{j=i}^{n} c_j v_j^T \sum_{k=1}^{n} \lambda_k u_k u_k^T \sum_{l=i}^{n} c_l v_l \\
&= \sum_{j=i}^{n} \sum_{k=1}^{n} \sum_{l=i}^{n} c_j c_l \lambda_k v_j^T v_k v_k^T v_l = \sum_{j=i}^{n} c_j^2 \lambda_j \\
&\leq (\sum_{j=i}^{n} c_j^2) \lambda_i = \lambda_i
\end{aligned}$$

Furthermore, by taking $\xi = v_i$, we can see the equality will be reached.