Lecture 6

Subspaces, Bases, and Projections in Machine Learning

SWCON253, Machine Learning

Won Hee Lee, PhD

Learning Goals

- Understand the fundamental concepts of **subspaces, bases, and projections** in machine learning

Recall geometric view of least squares

Given $(\underline{x}_i, y_i)$ for i=1, ...,n

Labels $\underline{y} \in \mathbb{R}^p$ for n training samples

Features $X \in \mathbb{R}^{n\mathrm{x}p}$ (p features)

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} \cdots \underline{x}_1^T \cdots \\ \cdots \underline{x}_2^T \cdots \\ \\ \cdots \underline{x}_n^T \cdots \end{bmatrix} \in \mathbb{R}^{n\mathrm{x}p}$$

We want to find $\underline{\hat{y}} = X\underline{w}$ such that $\left\|\underline{\hat{y}} - \underline{y}\right\|_2^2$ is as small as possible

Let X_1, X_2, …, X_p = p columns of X.

Then, $\hat{y} = w_1 X_1 + w_2 X_2 + \cdots + w_p X_p$

Basic least squares framework

This hyperplane corresponds to all the different vectors that could possibly be $\hat{y}$.

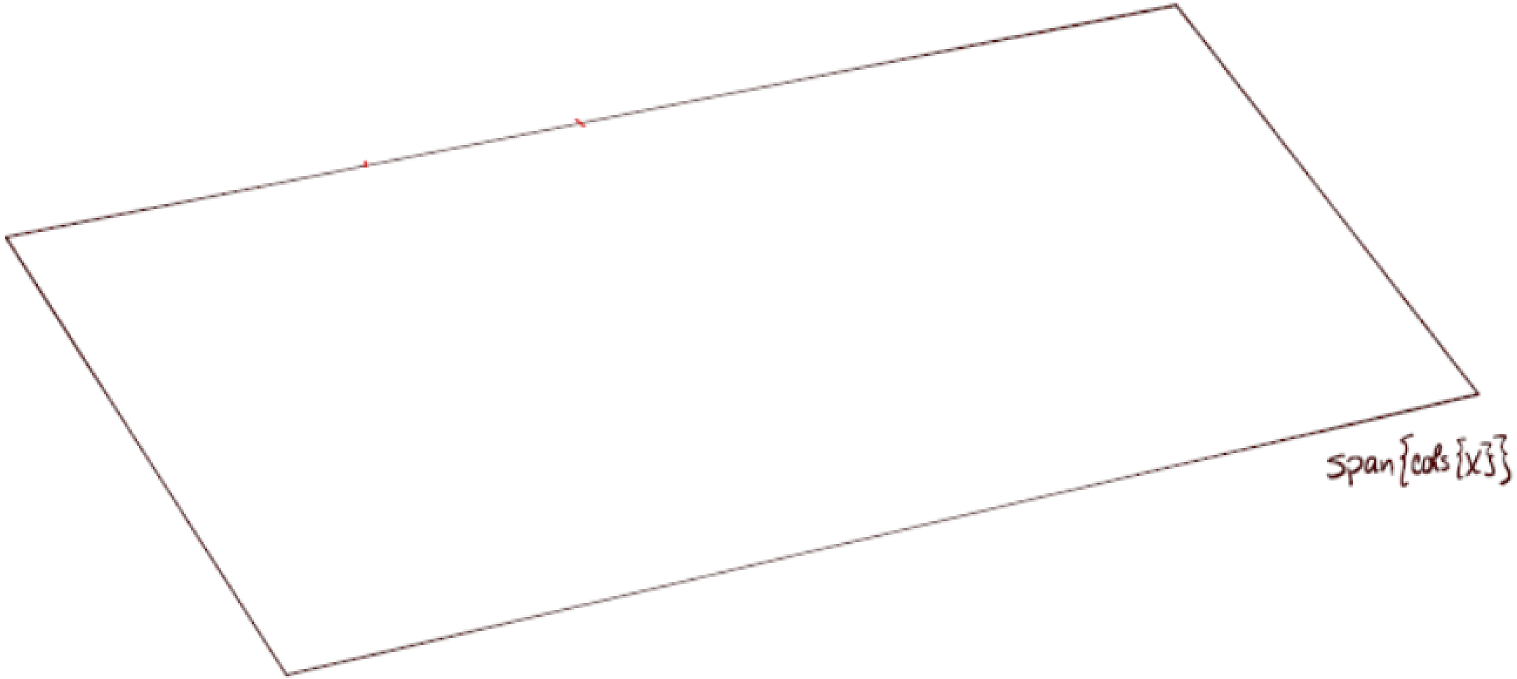
We will choose a specific $\hat{y}$ as close as possible to y.

But, these are choices that we are choosing $\hat{y}$ because $\hat{y}$ is a weighted sum of columns of X.

This means it has to be possible to write $\hat{y}$ in this form.

$\mathrm{Span}\big(\mathrm{cols}(\mathrm{X})\big) = \mathcal{X}$

$= \{\underline{\mathrm{v}} \in \mathbb{R}^{\mathrm{n}}: \underline{\mathrm{v}} = w_1 X_1 + w_2 X_2 + \cdots + w_p X_p$ for some $w_1, w_2, \ldots, w_p\}$

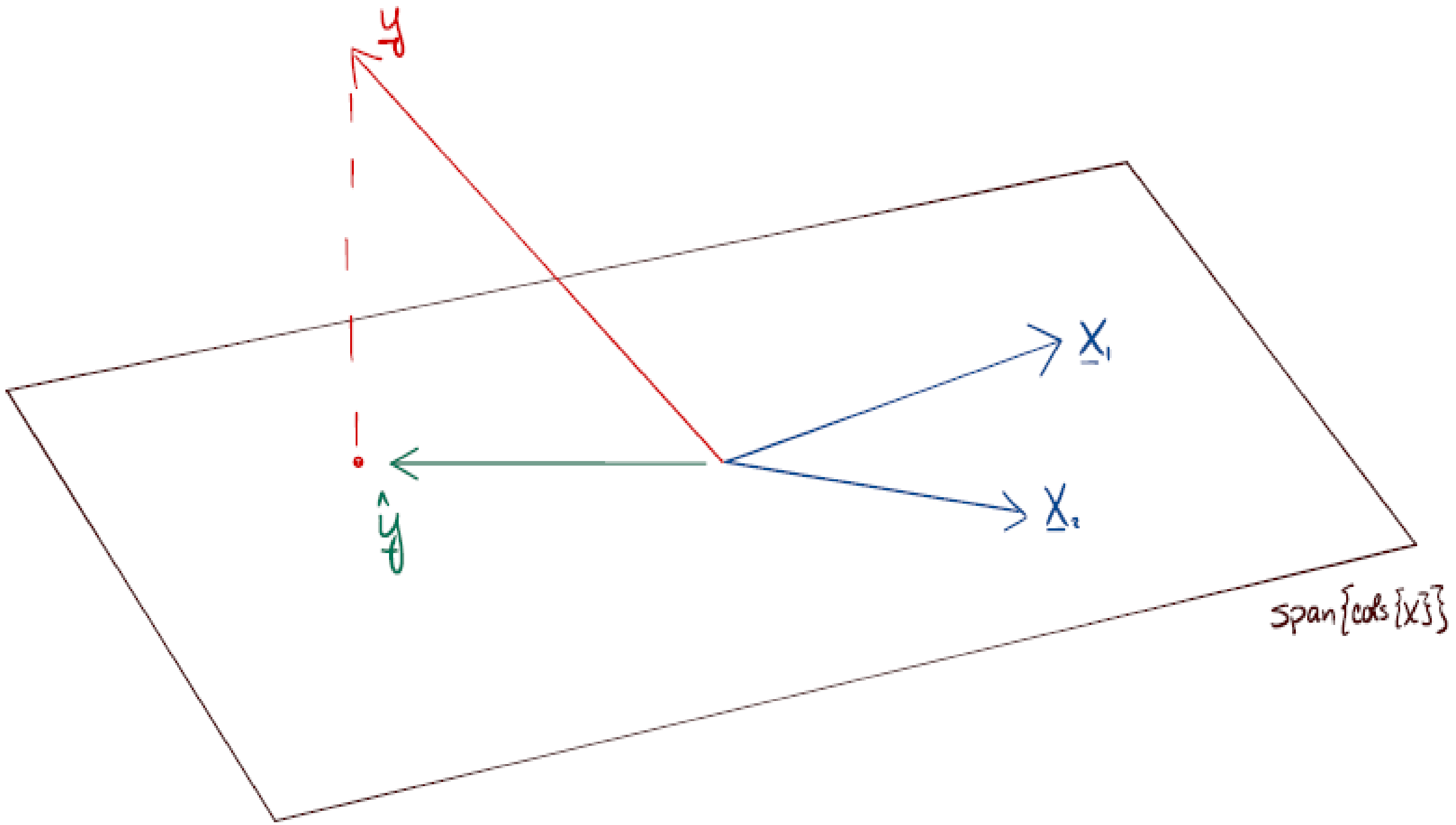


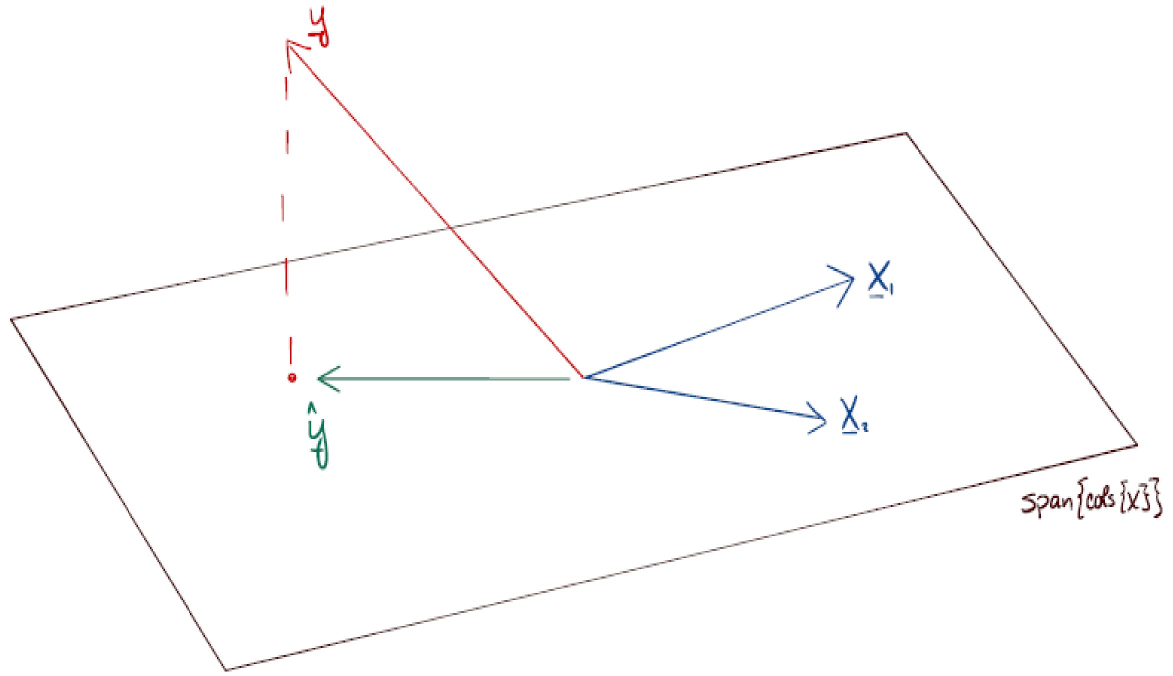
$\text{Span}(\text{cols}(X)) = \mathcal{X}$
$= \{\underline{v} \in \mathbb{R}^n : \underline{v} = w_1 X_1 + w_2 X_2 + \cdots + w_p X_p$ for some $w_1, w_2, \ldots, w_p\}$

y does not lie in this hyperplane.

We want the point inside the hyperplane that is as close as possible to y.

We will call $\hat{y}$.





The hyperplane span(cols(X)) $\mathcal{X}$ is called a **subspace** .

If the columns of X are linearly independent, then they form a **basis** for $\mathcal{X}$.

$\hat{y}$ is the **projection** of y onto the subspace $\mathcal{X}$.

We will use this notion of least squares with a motivating example.

Subspaces

Consider all points $\mathrm{X} \in \mathbb{R}^n$.

A subspace $\mathcal{S}$ is a subset of these points that satisfies a few key properties:

If $\underline{x}, \underline{y} \in \mathcal{S}$, then $\alpha\underline{x} + \beta\underline{y} \in \mathcal{S}$ for any α, β

Specifically, let $\mathcal{S}$ be a subspace and let $\underline{x}$ and $\underline{y}$ be any two points in the subspace.

Then for any scalars α and β, the weighted sum $\alpha\underline{x} + \beta\underline{y}$ must also be in the subspace.

Ex 1. n=3,

$\mathcal{S}=\{\underline{\mathrm{x}} \in \mathbb{R}^3 \colon x_1 = x_2 = -x_3\}$

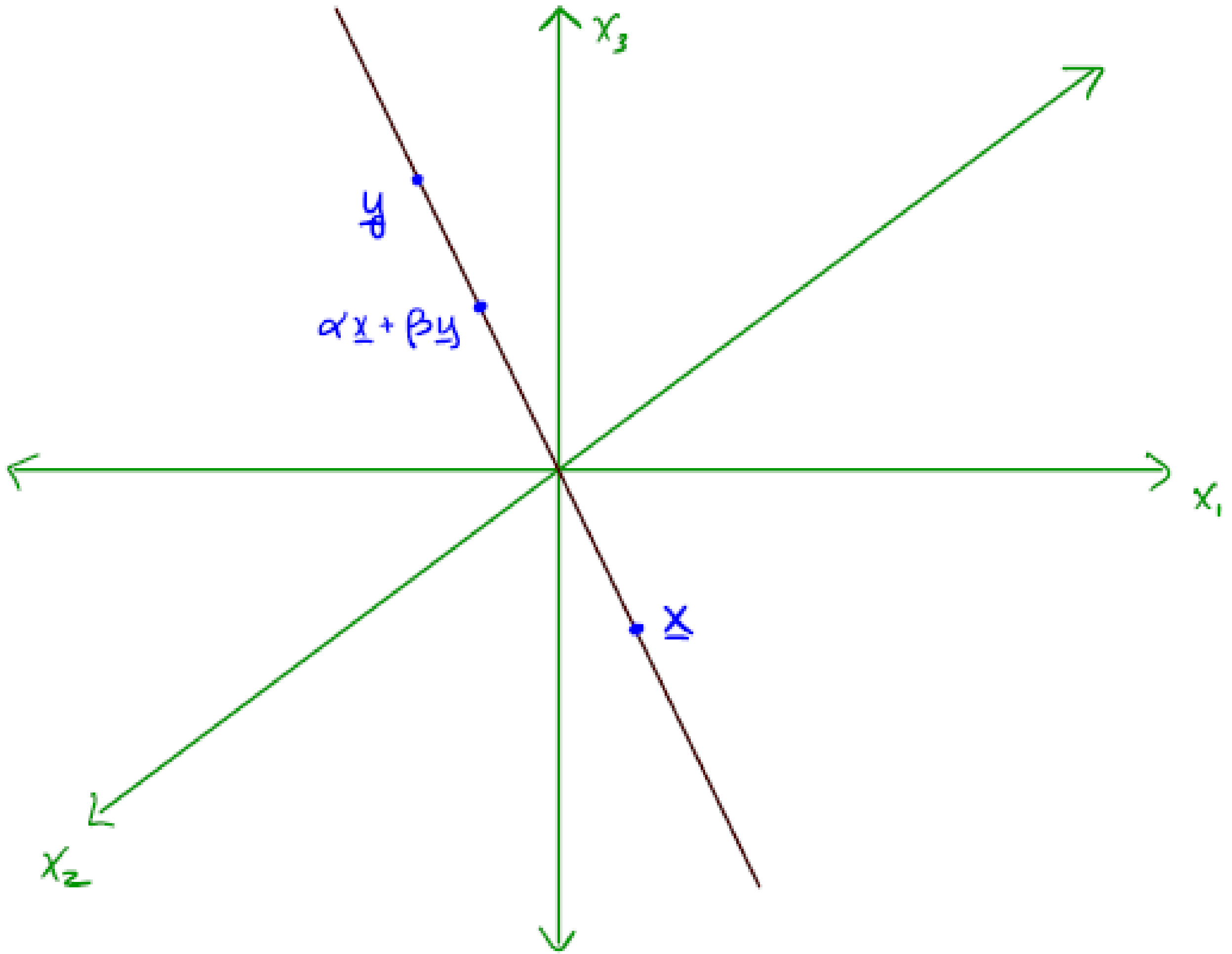
$\underline{\mathrm{x}} \in \mathcal{S}$

$\underline{\mathrm{x}} = a\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$ for some a

$x_1=a$, $x_2=a$, $x_3=-a$

$\underline{\mathrm{x}} \in \mathcal{S}$, $\underline{y} \in \mathcal{S}$

for any scalars α and β, the weighted sum $\alpha\underline{x} + \beta\underline{y}$ must also be in the subspace.

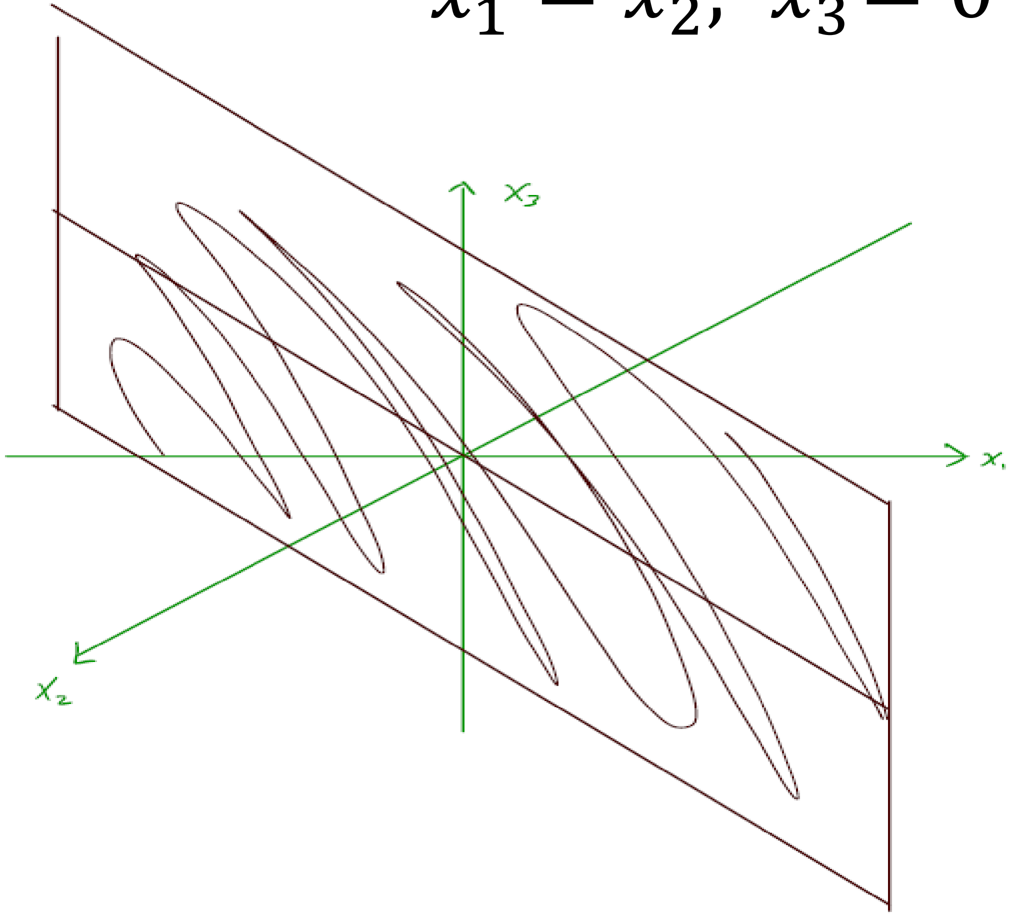


Ex 2. n=3,

$\mathcal{S}$={$\underline{\mathrm{x}} \in \mathbb{R}^3$: $x_1 = x_2$}

Vertical plane along diagonal

$x_1 = x_2,\ x_3 = 0$

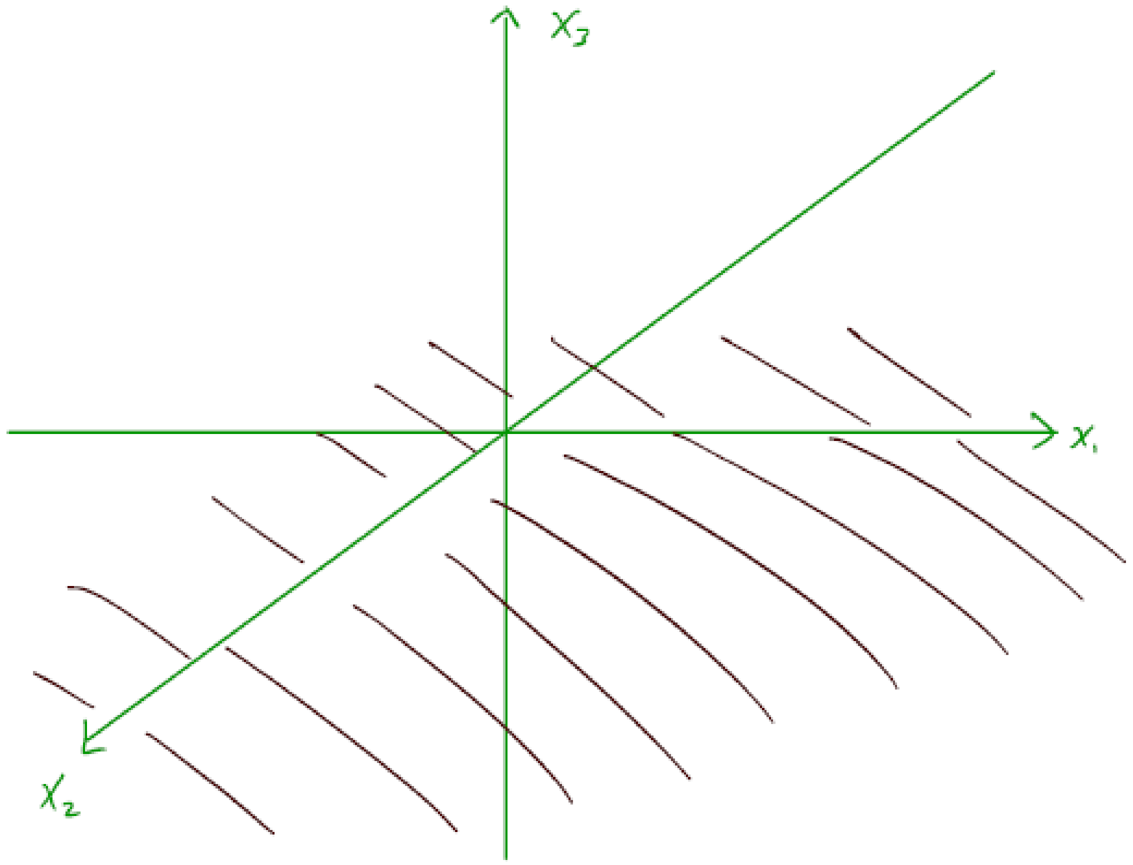


Ex 3. n=3,

$\mathcal{S}$={$\underline{\mathrm{x}} \in \mathbb{R}^3$: $x_3 = 0$}

Horizontal plane

$x_1 x_2\ plane$



Ex 4, recommender system

$$\underset{n\times p}{X} = \underbrace{\left.\begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix}\right\}\text{movie}}_{\text{customer}} = \begin{bmatrix} U \\ n\times r \end{bmatrix} \underbrace{\begin{bmatrix} V \\ r\times p \end{bmatrix}}_{r \text{ weights for each user}}$$

= r representative taste profiles

X_{ij} = rating of i^{th} movie by j^{th} customer (user)

The span of columns of U is a subspace.
This means that all columns of X lie in that subspace.

Ex 4, recommender system

$$\underset{n \times p}{X} = \underbrace{\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}}_{\text{customer}} \left.\vphantom{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}\right\} \text{movie} = \underbrace{\begin{bmatrix} U \\ n \times r \end{bmatrix}}_{= r \text{ representative taste profiles}} \underbrace{\begin{bmatrix} V \\ r \times p \end{bmatrix}}_{r \text{ weights for each user}}$$

X_{ij} = rating of i^{th} movie by j^{th} customer (user)

For example, for one column of X, we can think about this column as **a weighted sum** of the columns of U and the j^{th} column of V that tells us what those **weights** are.

So, every column of X is a **weighted sum of the columns of U** for some sort of **weights (V)** here and this again coincides with our notion of **subspaces** because the subspace corresponds to the span of the columns of U.

How to represent a subspace?

a. Represent $\mathcal{S}$ as the span of a set of vectors
b. Represent $\mathcal{S}$ as the span of a set of linearly independent vectors (called **basis**)
c. Represent $\mathcal{S}$ as the span of a set of orthonormal vectors (called **orthonormal basis**)

Recall

n=3, $\mathcal{S} = \{\underline{\mathrm{x}} \in \mathbb{R}^3 \colon x_3 = 0\}$ → horizontal plane

$a.$ $\mathcal{S}$ = span$\left\{\begin{pmatrix}1/2\\1/2\\0\end{pmatrix}, \begin{pmatrix}1/2\\-1/2\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}\right\}$

$b.$ $\mathcal{S}$ = span$\left\{\begin{pmatrix}1/2\\1/2\\0\end{pmatrix}, \begin{pmatrix}1/2\\-1/2\\0\end{pmatrix}\right\}$

$c.$ $\mathcal{S}$ = span$\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}\right\}$

How to represent a subspace?

a. Represent $\mathcal{S}$ as the span of a set of vectors

Recall

n=3, $\mathcal{S}$ =$\{\underline{\mathrm{x}} \in \mathbb{R}^3 \colon x_3 = 0\} \rightarrow$ horizontal plane

$a.$ $\mathcal{S}$ = span$\left\{\begin{pmatrix}1/2\\1/2\\0\end{pmatrix}, \begin{pmatrix}1/2\\-1/2\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}\right\}$

How to represent a subspace?

b. Represent $\mathcal{S}$ as the span of a set of linearly independent vectors (called **basis**)

Recall

n=3, $\mathcal{S}$ =$\{\underline{\mathrm{x}} \in \mathbb{R}^3\colon\ x_3 = 0\} \rightarrow$ horizontal plane

$$b.\ \ \mathcal{S} = \text{span}\left\{\begin{pmatrix} 1/2 \\ 1/2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/2 \\ -1/2 \\ 0 \end{pmatrix}\right\}$$

A collection of vectors $\underline{v}_1, \underline{v}_2, \dots, \underline{v}_p \in \mathbb{R}^n$ is linearly independent when $\sum_{i=1}^{p} a_i \underline{v}_i$=0 if and only if $a_i = 0$ for all i.

How to represent a subspace?

c. Represent $\mathcal{S}$ as the span of a set of orthonormal vectors (called **orthonormal basis**)

Orthogonal norm(length) = 1

Recall

n=3, $\mathcal{S}$ $=\{\underline{\mathrm{x}} \in \mathbb{R}^3\colon x_3 = 0\} \rightarrow$ horizontal plane

$c.\quad \mathcal{S} = \text{span}\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}\right\}$

How to represent a subspace?

c. Represent $\mathcal{S}$ as the span of a set of orthonormal vectors (called **orthonormal basis**)

Two vectors $\underline{u}_1$ and $\underline{u}_2$ are orthogonal if

$\langle \underline{u}_1, \underline{u}_2 \rangle = \underline{u}_1^T \underline{u}_2 = \underline{u}_2^T \underline{u}_1 = 0$

A vector u is normal if $\|u\|_2 = \|u\|_2^2 = \langle \underline{u}, \underline{u} \rangle = \underline{u}^T \underline{u} = 1$

A set of vectors $\underline{u}_1, \underline{u}_2, \dots, \underline{u}_p$ is orthonormal if

$\langle \underline{u}_i, \underline{u}_j \rangle = 1$ if i=j because they are normal

$\langle \underline{u}_i, \underline{u}_j \rangle = 0$ if i≠j because they are orthogonal

orthonormal basis
orthogonal basis
orthobasis

Properties of the orthonormal basis matrix

If $\mathcal{S}$ = span$\{\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_p\}$ where the vectors are orthonormal, then

U = $\begin{bmatrix} \vdots & \vdots & & \vdots \\ u_1 & u_2 & \ldots & u_p \\ \vdots & \vdots & & \vdots \end{bmatrix}$ is a orthogonal basis matrix; U is an orthogonal matrix

$U^T U = C \rightarrow C_{ij} = \langle \underline{u}_i, \underline{u}_j \rangle$

$\langle \underline{u}_i, \underline{u}_j \rangle = 1$ if i=j

$\langle \underline{u}_i, \underline{u}_j \rangle = 0$ if i≠j

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = I$$

U is (squared) length preserving

Let $\underline{\mathrm{v}} \in \mathbb{R}^p$ Let's consider Uv

$$\|Uv\|_2^2 = (Uv)^T(Uv) = v^T U^T U v = v^T v = \|v\|_2^2$$

We take any vector and multiply by an orthogonal matrix by it, then the squared length of that product is equal to the squared length of the original vector.

Dimension of subspace

$\dim(\mathcal{S})$ = number of vectors in subspace basis
e.g., dim(line)=1; dim(plane)=2

If $\mathcal{S}$=span(cols(X)), then $\dim(\mathcal{S})$=rank(X)

Rank corresponds to the number of linearly independent columns in a matrix X.
The dimension of the subspace corresponds exactly to the rank of the matrix X.

Projection

The projection of a point $\underline{y}$ onto a set is the point in the set closest to $\underline{y}$.

$$\underline{\hat{y}} = \text{projection of } \underline{y} \text{ onto set } \mathcal{X} = \mathrm{P}_{\mathcal{X}}\underline{y} \; = \underset{\underline{x}\in\mathcal{X}}{\operatorname{argmin}} \left\|\underline{y} - \underline{x}\right\|_2^2$$

We want to find the argument the point x in the space $\mathcal{X}$ that minimizes the distance between $\underline{y}$ and $\underline{x}$ which is the same as minimizing the squared distances.

We want to understand this notion in the context of subspaces and bases.

If $\mathcal{X}$ is a subspace spanned by columns of $X \in \mathbb{R}^{nxp}$ with LI columns, any point in $\mathcal{X}$ has form $\underline{\hat{y}} = w_1\underline{x}_1 + w_2\underline{x}_2 + \cdots + w_p\underline{x}_p$.

We want to find a point $\underline{\hat{y}}$ as close as possible to $\underline{y}$ in the subspace and we know that $\underline{\hat{y}}$ has this form. So all we have to do is to find the weights $\underline{\hat{\boldsymbol{w}}}$.

Let $\underline{\widehat{w}} = \underset{\underline{w}}{\operatorname{argmin}} \left\|\underline{y} - X\underline{w}\right\|_2^2$ and $\underline{\hat{y}} = X\underline{\widehat{w}}$

Least squares

$\underline{\widehat{\boldsymbol{w}}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underline{\boldsymbol{y}}$

Projection matrix

$\underline{\hat{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underline{y}$

$= \mathrm{P}_X\underline{y}$

Orthogonal Subspace Bases and Least Squares

Let $X \in \mathbb{R}^{nxp}$, $\underline{y} \in \mathbb{R}^{n}$

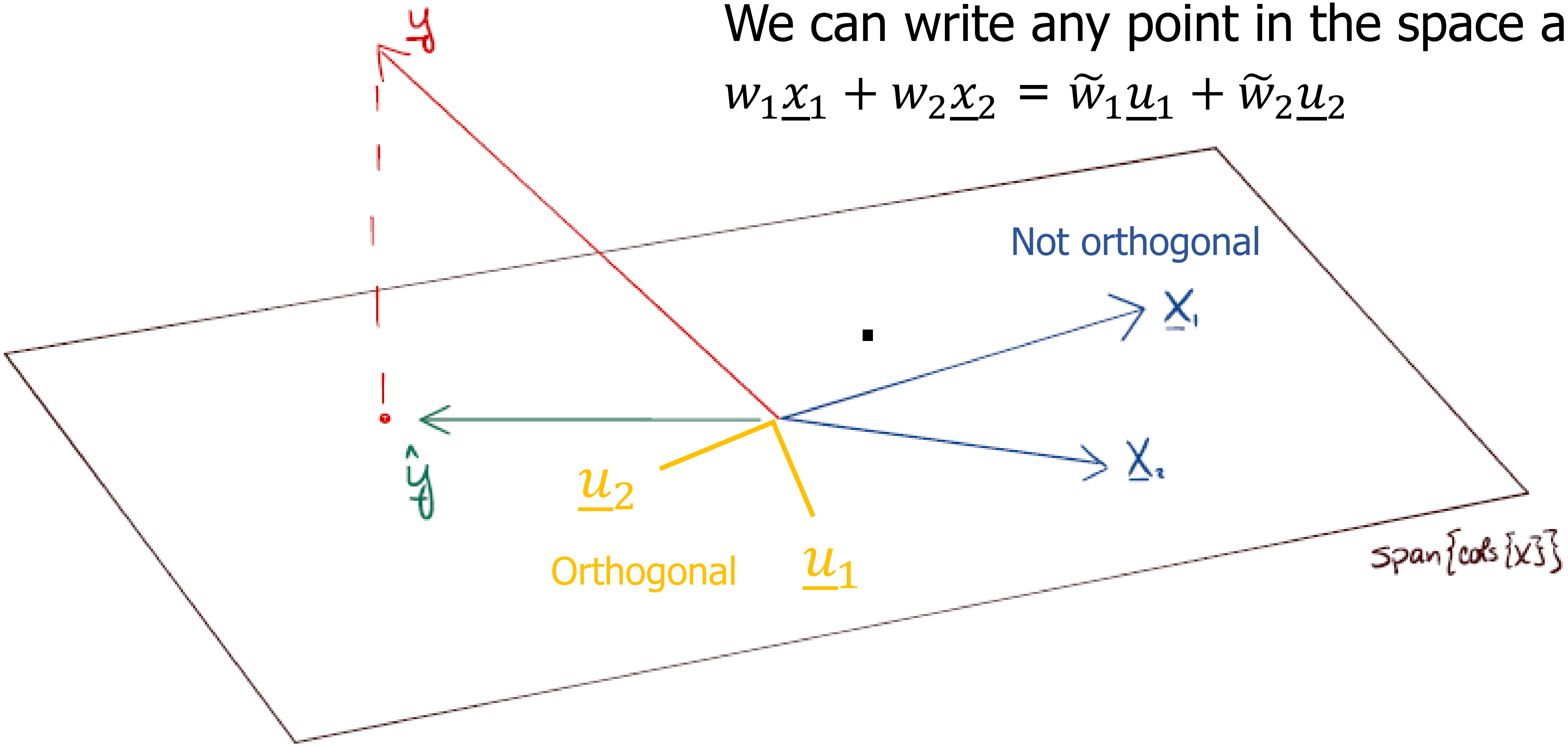
Let U be orthonormal basis matrix for subspace spanned by columns of X

i.e., span(cols(U)) = span(cols(X))

span(cols(U)) = span(cols(X))

We can write any point in the space as

$w_1\underline{x}_1 + w_2\underline{x}_2 = \widetilde{w}_1\underline{u}_1 + \widetilde{w}_2\underline{u}_2$



Representing the same space, the difference is that the columns of X might not be orthogonal and normalized vs. the columns of u are orthogonal and normalized.

Orthogonal Subspace Bases and Least Squares

Let $X \in \mathbb{R}^{nxp}$, $\underline{y} \in \mathbb{R}^{n}$

Let U be orthonormal basis matrix for subspace spanned by columns of X

i.e., span(cols(U)) = span(cols(X))

$\underline{\hat{y}} = X\underline{\widehat{w}} = U\underline{\widetilde{w}}$ for any $\underline{\hat{y}} \in \mathcal{X}$, there are both $\underline{\widehat{w}}, \underline{\widetilde{w}}$ so that $\underline{\hat{y}} = X\underline{\widehat{w}} = U\underline{\widetilde{w}}$

A weighted ($\underline{\widehat{w}}$) sum of the columns of X = a weighted ($\underline{\widetilde{w}}$) sum of the columns of U

Use least squares to find $\underline{\tilde{w}}$

Least squares

$$\underline{\hat{w}} = (X^T X)^{-1} X^T \underline{y}$$

$$\underline{\tilde{w}} = \underset{\underline{w}}{\operatorname{argmin}} \left\| \underline{y} - U\underline{w} \right\|_2^2$$

$$= (U^T U)^{-1} U^T \underline{y}$$

$$\underline{\hat{\mathbf{y}}} = \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\underline{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underline{\boldsymbol{y}}$$

Projection onto span(cols(U)) Projection onto span(cols(X))

$$U(\boldsymbol{U}^T\boldsymbol{U})^{-1}U^T = U\boldsymbol{I}U^T = UU^T$$

Least squares

$$\underline{\widehat{w}} = (X^TX)^{-1}X^T\underline{y}$$

It's all about the computation!

When this U is an orthobasis, U^TU is equal to an identity matrix I.

The inverse of an identity matrix is also the identity matrix.

Thus, UU^T, we **no longer have to invert a matrix.**

If this matrix X^TX is big, inverting it can be really difficult to do (more memory required for large-scale ML problems).

If we can find an orthobasis U for the same space, we can get the exact same solution (predicted labels $\underline{\hat{y}}$), but **without computing any matrix inverse**.

$$\underline{\hat{\mathbf{y}}} = \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\underline{\boldsymbol{y}} \quad = \quad \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underline{\boldsymbol{y}}$$

The notion of projection is very helpful!

because it helps to see that even though we derived the formula in terms of our original features X, we don't necessarily have to do exactly this computation in the computer.

Mathematically it is equivalent.

There could potentially be a huge advantage.

Announcements

- Homework #1 out
  - 스스로 복습한 내용을 A4용지에 손글씨로 작성/스캔하여 제출
  - 1-page 이상 per a lecture (lectures 3-5)
  - Any forms (pdf, jpeg, etc.) should be fine!
  - **Due March 23th Tuesday at 11:59 pm**