Announcements

HW1

- Due Mon 9/20, 11:59 pm
- Online + Written components
- Reach out for help (probability, LaTeX, anything!)

Quiz

- First quiz on Wed 9/22
- Details posted on Piazza soon

Mathematical Foundations for Machine Learning

Calculus

Instructor: Pat Virtue

Today

Linear regression

- Linear algebra formulation
  - Linear algebra properties
- Calculus
  - Multivariate calculus
  - Calculus with linear algebra
- Optimization
  - Convex functions (briefly)
  - Closed-form solutions

Linear Regression

Last time

Linear algebra needed to:

- Find m that minimizes MSE with model $y = mx$, $x \in \mathbb{R}$

Today

Optimization

- Linear, convex, closed-form solution

Calculus needed to:

- Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$
- Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{x} \in \mathbb{R}^M$

Calculus need for neural networks

Reminder: Derive the following expansion

Figure out each step, given the provided justification

1 input feature
No bias term
$\hat{y} = mx \qquad m, x \in \mathbb{R}$

All data
$\mathbf{y}, \mathbf{x} \in \mathbb{R}^N$

	Justification
$J(m; \mathbf{x}, \mathbf{y}) = \frac{1}{N} \lVert \underbrace{\mathbf{y} - m\mathbf{x}}_{z} \rVert_2^2$	
$= \frac{1}{N} (y - mx)^T (y - mx)$	$\lVert \mathbf{z} \rVert_2^2 = \mathbf{z}^T \mathbf{z}$
$= \frac{1}{N} (y^T - m x^T)(y - mx)$	$(A + B)^T = A^T + B^T$
$= \frac{1}{N} [y^T y - m y^T x - m x^T y + m^2 x^T x]$	$(A + B)C = AC + BC$ and $C(A + B) = CA + CB$
$= \frac{1}{N} [\mathbf{y}^T \mathbf{y} - 2m\mathbf{y}^T \mathbf{x} + m^2 \mathbf{x}^T \mathbf{x}]$	$\mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u}$

Reminder: Linear Regression

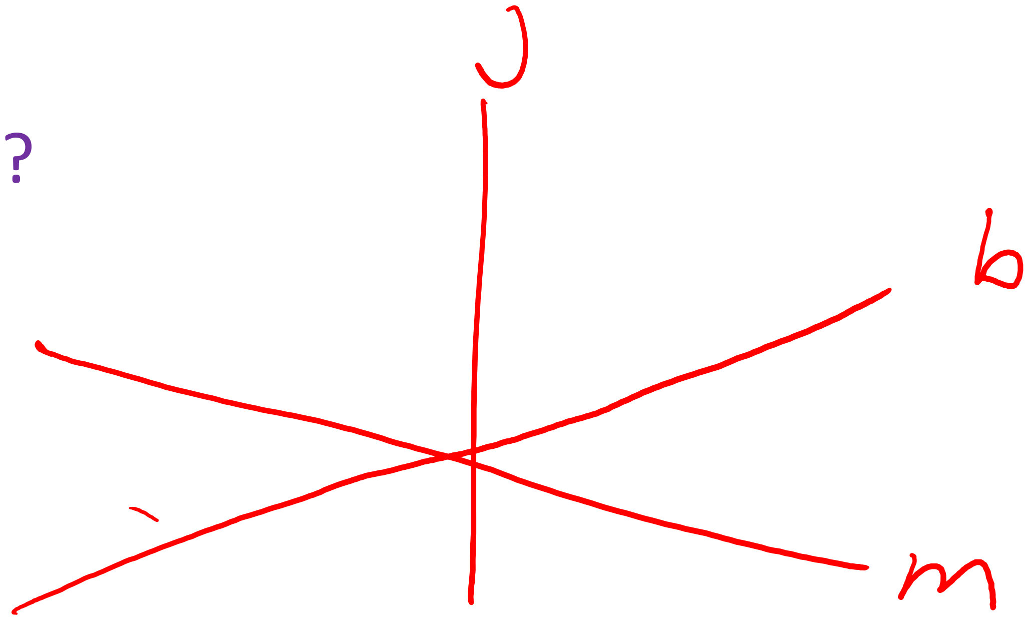
Linear algebra formation

Find m that minimizes MSE with model $\hat{y} = mx, \quad x \in \mathbb{R}$

$$J(m; \mathbf{x}, \mathbf{y}) = \frac{1}{N} \|\mathbf{y} - m\mathbf{x}\|_2^2$$

$$= \frac{1}{N} [\mathbf{y}^T\mathbf{y} - 2m\mathbf{y}^T\mathbf{x} + m^2\mathbf{x}^T\mathbf{x}]$$

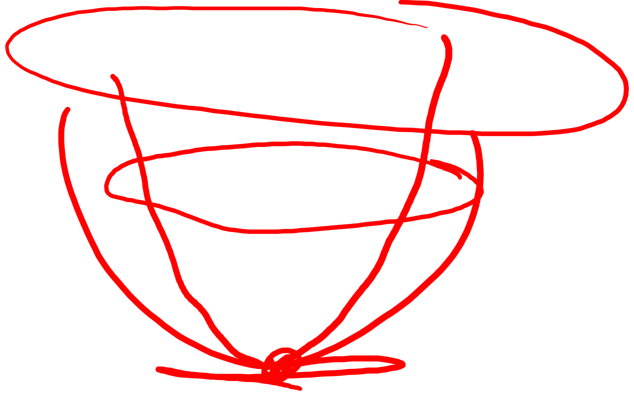
What shape is $J(m)$?



$\mathcal{D} = \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{4}$

$x^{(i)}$	$y^{(i)}$
2	0
0	-1
-3	0
1	2

Reminder: Linear Regression



$$J(m;\mathbf{x},\mathbf{y}) = \frac{1}{N}[\mathbf{y}^T\mathbf{y} - 2m\mathbf{y}^T\mathbf{x} + m^2\mathbf{x}^T\mathbf{x}]$$

$$\frac{dJ}{\partial m} = \frac{1}{N}[0 - 2y^Tx + 2mx^Tx]$$

$$0 = -y^Tx + mx^Tx$$

$$\hat{m} = \frac{y^Tx}{x^Tx} \leftarrow 0$$

Linear and Convex Functions

Optimization

Linear function

If $f(\boldsymbol{x})$ is linear, then for any $\boldsymbol{u}, \boldsymbol{v}$:

- $f(\boldsymbol{u} + \boldsymbol{z}) = f(\boldsymbol{u}) + f(\boldsymbol{z})$
- $f(\alpha \mathbf{z}) = \alpha f(\mathbf{z}) \quad \forall \alpha$
- $f(\alpha \boldsymbol{u} + (1 - \alpha)\boldsymbol{z}) = \alpha f(\boldsymbol{u}) + (1 - \alpha) f(\boldsymbol{z}) \quad \forall \alpha$

Optimization

Convex function

If $f(\boldsymbol{x})$ is convex, then for any $\boldsymbol{u}, \boldsymbol{v}$:

- $f(\alpha\boldsymbol{u} + (1-\alpha)\boldsymbol{z}) \leq \alpha f(\boldsymbol{u}) + (1-\alpha) f(\boldsymbol{z}) \quad \forall\ 0 \leq \alpha \leq 1$

Convex optimization

If $f(\boldsymbol{x})$ is convex, then:

- Every local minimum is also a global minimum ☺

Optimization

Linear function

If $f(\boldsymbol{x})$ is linear, then for any $\boldsymbol{u}, \boldsymbol{v}$:

- $f(\boldsymbol{u} + \boldsymbol{z}) = f(\boldsymbol{u}) + f(\boldsymbol{z})$
- $f(\alpha \boldsymbol{z}) = \alpha f(\boldsymbol{z}) \quad \forall \alpha$
- $f(\alpha \boldsymbol{u} + (1 - \alpha)\boldsymbol{z}) = \alpha f(\boldsymbol{u}) + (1 - \alpha) f(\boldsymbol{z}) \quad \forall \alpha$

Proof: Is $f(x) = mx + b$ linear?

Optimization

Linear function

If $f(\boldsymbol{x})$ is linear, then for any $\boldsymbol{u}, \boldsymbol{v}$:

- $f(\boldsymbol{u} + \boldsymbol{z}) = f(\boldsymbol{u}) + f(\boldsymbol{z})$
- $f(\alpha \boldsymbol{z}) = \alpha f(\boldsymbol{z}) \quad \forall \alpha$
- $f(\alpha \boldsymbol{u} + (1 - \alpha)\boldsymbol{z}) = \alpha f(\boldsymbol{u}) + (1 - \alpha) f(\boldsymbol{z}) \quad \forall \alpha$

Proof: Is $f(x) = mx + b$ linear?

Linear Regression

Last time

Linear algebra needed to:

- Find m that minimizes MSE with model $y = mx$, $x \in \mathbb{R}$

Today

Optimization

- Linear, convex, closed-form solution

Calculus needed to:

- Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$
- Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{x} \in \mathbb{R}^M$

Calculus need for neural networks

Optimization: Notation

$$J(m) = \frac{1}{N}\left[\mathbf{y}^T\mathbf{y} \ - 2m\mathbf{y}^T\mathbf{x} + m^2\mathbf{x}^T\mathbf{x}\right]$$

Optimization: Closed-form Solution

We have a closed-form solution when we can write the optimal parameter equal to an expression that can evaluate to a value(s).

$$J(m) = \frac{1}{N}[\mathbf{y}^T\mathbf{y} - 2m\mathbf{y}^T\mathbf{x} + m^2\mathbf{x}^T\mathbf{x}]$$

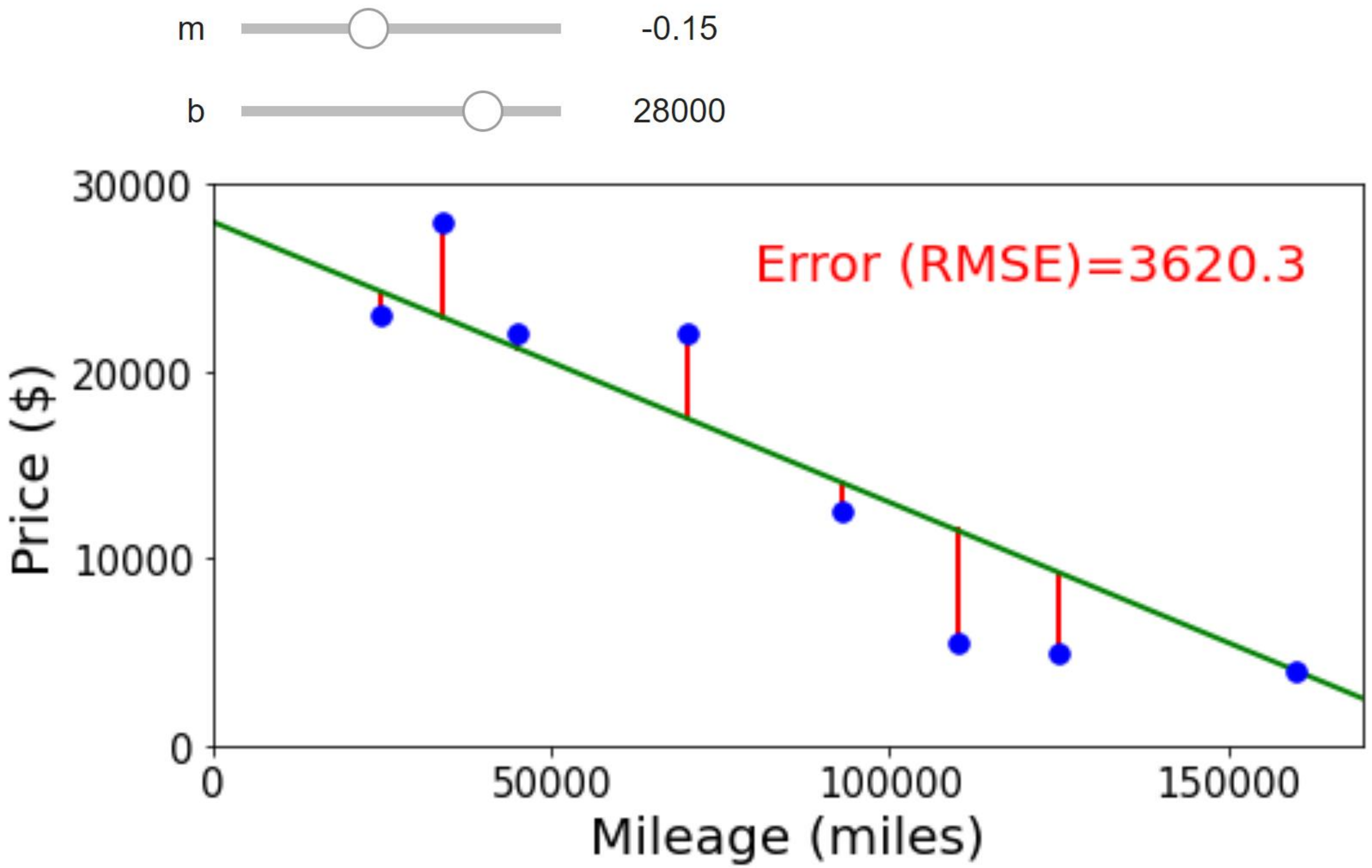
$$\frac{\partial J}{\partial m} = \frac{1}{N}[0 - 2\mathbf{y}^T\mathbf{x} + 2m\mathbf{x}^T\mathbf{x}]$$

$$0 = -\mathbf{y}^T\mathbf{x} + m\mathbf{x}^T\mathbf{x}$$

$$\widehat{m} = \frac{\mathbf{y}^T\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

Linear Regression

Adding Input Features



Linear Regression

Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$

$$\mathcal{D} = \left\{\left(x_1^{(i)}, x_2^{(i)}, y^{(i)}\right)\right\}_{i=1}^{4}$$

$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
2	-1	0
0	2	-1
-3	-2	0
1	3	2

Linear Regression

Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$

$$\mathcal{D} = \left\{\left(x_1^{(i)}, x_2^{(i)}, y^{(i)}\right)\right\}_{i=1}^{4}$$

$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
2	-1	0
0	2	-1
-3	-2	0
1	3	2

Calculus

Calculus with Linear Algebra

Vector in, scalar out

Gradient

$y = f(\mathbf{x}) \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^M$

Calculus with Linear Algebra

Vector in, scalar out

Calculus with Linear Algebra

Vector in, scalar out

Exercise

Prove $\frac{\partial}{\partial \mathbf{v}} \mathbf{v}^T A \mathbf{v} = \left(A^T + A\right)\mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2\times 2}$

$f(\mathbf{v}) = \mathbf{v}^T A \mathbf{v} \quad \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$

Prove $\frac{\partial f}{\partial \mathbf{v}} = \left(A^T + A\right)\mathbf{v}$

Hint: Start by expanding $A\mathbf{v}$ and then expanding $\mathbf{v}^T A \mathbf{v}$, i.e., write $f(v_1, v2)$ in terms of scalars $v_1, v_2, A_{1,1}, A_{1,2}, A_{2,1}, A_{2,2}$

Hint: Write out scalar partial derivatives, $\frac{\partial f}{\partial v_1}$ and $\frac{\partial f}{\partial v_2}$

Hint: Work both top-down and bottom-up, i.e., also expand $\left(A^T + A\right)\mathbf{v}$, so you can see where you are going.

Linear Regression

Linear Regression

Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$

$$J(\boldsymbol{w}) = \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}\right)$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}}{\partial \boldsymbol{z}} = \left(\boldsymbol{A} + \boldsymbol{A}^T\right)\boldsymbol{z}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}}{\partial \boldsymbol{z}} = \boldsymbol{z}^T\left(\boldsymbol{A} + \boldsymbol{A}^T\right)$$

Linear Regression

Last time

Linear algebra needed to:

- Find m that minimizes MSE with model $y = mx$, $x \in \mathbb{R}$

Today

Optimization

- Linear, convex, closed-form solution

Calculus needed to:

- Find $\mathbf{w}$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x}$, $\mathbf{x} \in \mathbb{R}^2$
- Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{x} \in \mathbb{R}^M$

Calculus need for neural networks

Linear Regression

Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\mathbf{x} \in \mathbb{R}^M$

$$\mathcal{D} = \left\{\left(x_1^{(i)}, x_2^{(i)}, y^{(i)}\right)\right\}_{i=1}^{4}$$

$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
2	-1	0
0	2	-1
-3	-2	0
1	3	2

Linear Regression

Find $\mathbf{w}, b$ that minimizes MSE with model $y = \mathbf{w}^T\mathbf{x} + b$, $\quad \mathbf{x} \in \mathbb{R}^M$

$$\mathcal{D} = \left\{\left(x_1^{(i)}, x_2^{(i)}, y^{(i)}\right)\right\}_{i=1}^{4}$$

	$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
	2	-1	0
	0	2	-1
	-3	-2	0
	1	3	2

Linear Regression

$\|\vec{z}\|_2^2 = \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} z_i z_i = \vec{z}^T\vec{z}$

Expanding objective before computing gradient

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X\theta}\|_2^2$$

$$= \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta})$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T - \boldsymbol{\theta}^T\boldsymbol{X}^T\right)(\boldsymbol{y} - \boldsymbol{X\theta})$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X\theta} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta}\right)$$

these two are the same

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta}\right)$$

$\frac{\partial J}{\theta}$

$\frac{\partial J}{\partial \theta} = \begin{bmatrix} \vdots \\ 0 \\ \vdots \end{bmatrix}$

Linear Regression

Gradient of objective with respect to parameters

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X\theta}\|_2^2$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta}\right)$$

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{N}\left(0 - 2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\right)$$

Dimension mismatch

$$= \frac{1}{N}\left(0 - 2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X\theta}\right)$$

$$= \frac{2}{N}\left(-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X\theta}\right)$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}}{\partial \boldsymbol{z}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{z}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}}{\partial \boldsymbol{z}} = \boldsymbol{z}^T(\boldsymbol{A} + \boldsymbol{A}^T)$$

Linear Regression

Closed-form solution

$$\nabla J(\boldsymbol{\theta}) = \frac{2}{N}\left(-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}\right)$$

$$\nabla J(\theta) = 0$$

$$X^T X \theta = X^T y \quad \leftarrow \text{Normal equation}$$

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

$$\boldsymbol{\theta} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$$

$$\mathcal{D} = \left\{\left(1, x_1^{(i)}, x_2^{(i)}, y^{(i)}\right)\right\}_{i=1}^{4}$$

$x_0^{(i)}$	$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
1	2	-1	0
1	0	2	-1
1	-3	-2	0
1	1	3	2