Linear Regression

Announcements

Recap on Logistic Regression / Optimization

Binary classification with $\mathscr{D} = \{x_i, y_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w) = \dfrac{1}{1 + \exp(-y(w^\top x))}$

Recap on Logistic Regression / Optimization

Binary classification with $\mathcal{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w) = \dfrac{1}{1 + \exp(-y(w^\top x))}$

Using MLE, we get our optimization objective:

$$\hat{w} := \arg\min_w \sum_{i=1}^n \ln(1 + \exp(-y_i(w^\top x_i)))$$

Recap on Logistic Regression / Optimization

Binary classification with $\mathcal{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \{-1,1\}$

Logistic Regression asumes $P(y \mid x; w) = \dfrac{1}{1 + \exp(-y(w^\top x))}$

Using MLE, we get our optimization objective:

$$\hat{w} := \arg\min_w \sum_{i=1}^n \ln(1 + \exp(-y_i(w^\top x_i)))$$

Given a test example x_{test}, we can make prediction:

$$\hat{y} = \begin{cases} +1 & P(+1 \mid x_{test}; \hat{w}) > 0.5 \\ -1 & \text{else} \end{cases}$$

Recap on Logistic Regression / Optimization

$$\arg\min_{w} \sum_{i=1}^{n} \ln(1 + \exp(-y_i(w^\top x_i)))$$

Logistic Regression with SGD as the optimizer:

Initialize $w^0 = \mathbf{0}$

While not converged:

Recap on Logistic Regression / Optimization

$$\arg\min_{w} \sum_{i=1}^{n} \ln(1 + \exp(-y_i(w^\top x_i)))$$

Logistic Regression with SGD as the optimizer:

Initialize $w^0 = \mathbf{0}$

While not converged:

Randomly sample one training pair $(x, y) \sim \mathscr{D}$

Recap on Logistic Regression / Optimization

$$\arg\min_{w} \sum_{i=1}^{n} \ln(1 + \exp(-y_i(w^\top x_i)))$$

Logistic Regression with SGD as the optimizer:

Initialize $w^0 = \mathbf{0}$

While not converged:

- Randomly sample one training pair $(x, y) \sim \mathscr{D}$
- Compute gradient: $g = (1 - P(y \mid x; w^t)) \cdot (-yx)$

Recap on Logistic Regression / Optimization

$$\arg\min_{w} \sum_{i=1}^{n} \ln(1 + \exp(-y_i(w^\top x_i)))$$

Logistic Regression with SGD as the optimizer:

Initialize $w^0 = \mathbf{0}$

While not converged:

Randomly sample one training pair $(x, y) \sim \mathscr{D}$

Compute gradient: $g = (1 - P(y \mid x; w^t)) \cdot (-yx)$

SGD update: $w^{t+1} = w^t + \alpha(1 - P(y \mid x; w^t)) \cdot (yx)$

Recap on Logistic Regression / Optimization

$$\arg\min_{w} \sum_{i=1}^{n} \ln(1 + \exp(-y_i(w^\top x_i)))$$

Logistic Regression with SGD as the optimizer:

Initialize $w^0 = \mathbf{0}$

While not converged:

- Randomly sample one training pair $(x, y) \sim \mathcal{D}$
- Compute gradient: $g = (1 - P(y \mid x; w^t)) \cdot (-yx)$
- SGD update: $w^{t+1} = w^t + \alpha(1 - P(y \mid x; w^t)) \cdot (yx)$

Compare this to Perceptron!

Outline for Today

1. Intro on Linear Regression

2. Normal equation for linear Regression

3. Interpretation of Linear Regression using MLE / MAP

Ex: Predicting the house price

Dataset:

Living area ($feet^2$)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮
x	y

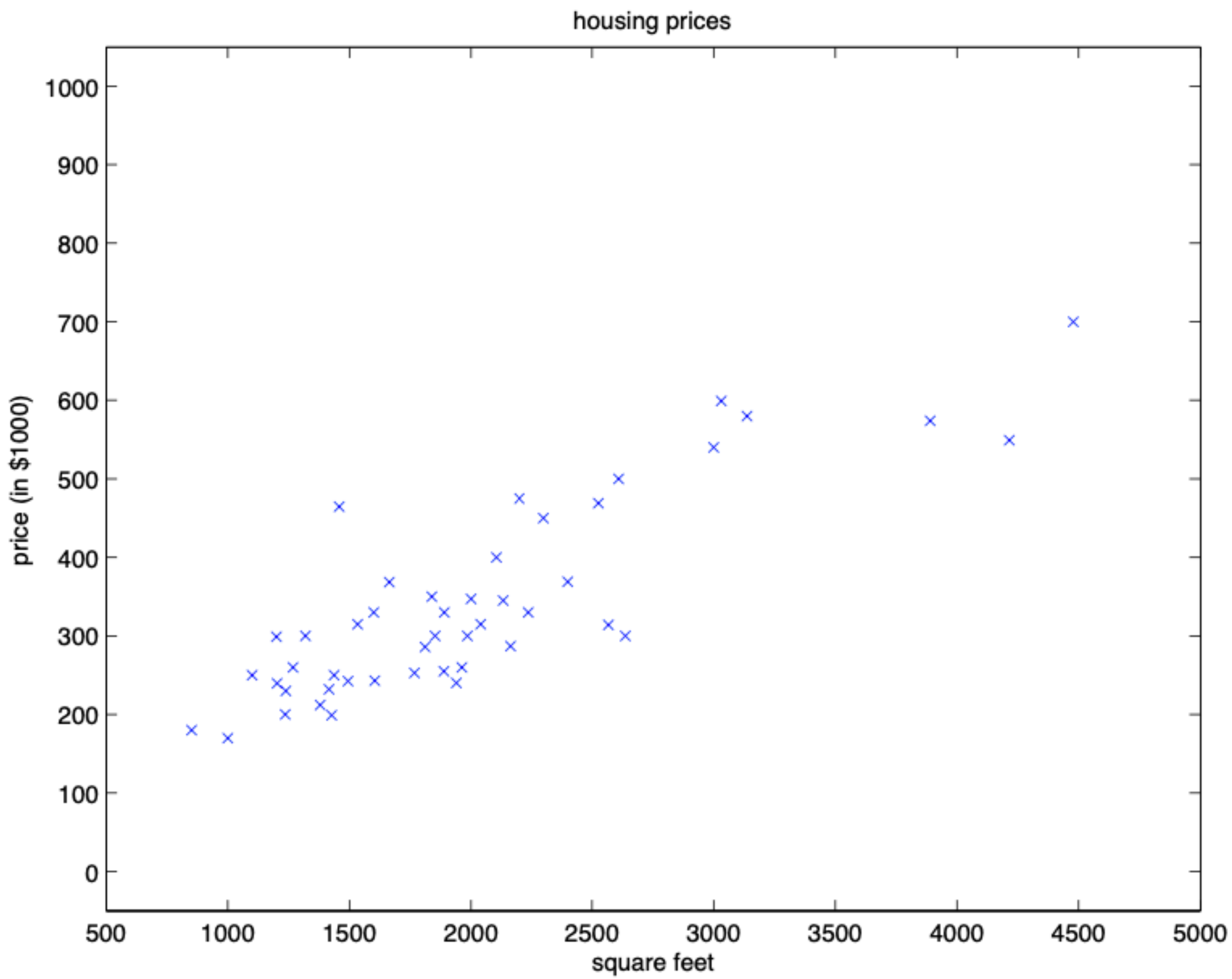
(Example from Stanford CS229)

Ex: Predicting the house price

Dataset:

Living area (feet^2)	Price (1000\$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮
x	y

Plot:



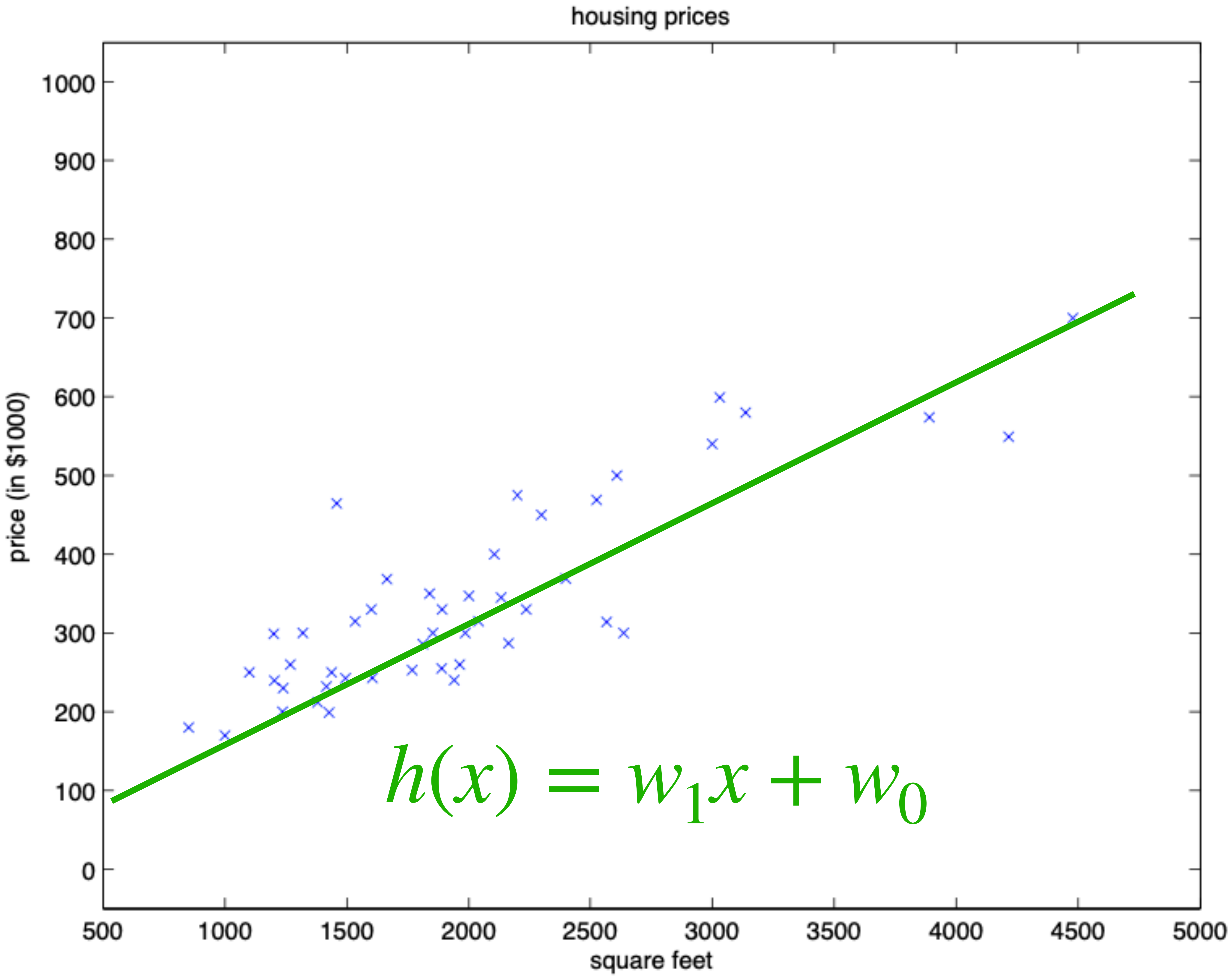
(Example from Stanford CS229)

Ex: Predicting the house price

Dataset:

Living area (feet^2)	Price (1000\$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮
x	y

Plot:



(Example from Stanford CS229)

Ex: Predicting the house price (2d case)

Dataset:

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮
$x[1]$	$x[2]$	y

(Example from Stanford CS229)

Ex: Predicting the house price (2d case)

Dataset:

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮
$x[1]$	$x[2]$	y

Goal: finding the linear function

$$h(x) = w_1 x[1] + w_2 x[2] + w_0$$

that fits the data well

(Example from Stanford CS229)

Ex: Predicting the house price (2d case)

Dataset:

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮
$x[1]$	$x[2]$	y

As usual, we append 1 to the feature, i.e.,

$$x = \begin{bmatrix} x[1] \\ x[2] \\ 1 \end{bmatrix}$$

So the linear function can be written as:

$$h(x) = w^\top x$$

Outline for Today

1. Intro on Linear Regression

2. Normal equation for linear Regression

3. Interpretation of Linear Regression using MLE / MAP

Mathematical formulation of linear regression

Input: dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Mathematical formulation of linear regression

Input: dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Hypothesis: linear function $h(x) = w^\top x$

Mathematical formulation of linear regression

Input: dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Hypothesis: linear function $h(x) = w^\top x$

Hypothesis class: all possible linear functions $\{w^\top x, \forall w \in \mathbb{R}^d\}$

Mathematical formulation of linear regression

Input: dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Hypothesis: linear function $h(x) = w^\top x$

Hypothesis class: all possible linear functions $\{w^\top x, \forall w \in \mathbb{R}^d\}$

Loss function: squared loss $\ell(w^\top x, y) = (w^\top x - y)^2$

Mathematical formulation of linear regression

Input: dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Hypothesis: linear function $h(x) = w^\top x$

Hypothesis class: all possible linear functions $\{w^\top x, \forall w \in \mathbb{R}^d\}$

Loss function: squared loss $\ell(w^\top x, y) = (w^\top x - y)^2$

Q: can we use absolute loss, i.e., $|w^\top x - y|$?

Mathematical formulation of linear regression

Input: dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Hypothesis: linear function $h(x) = w^\top x$

Hypothesis class: all possible linear functions $\{w^\top x, \forall w \in \mathbb{R}^d\}$

Loss function: squared loss $\ell(w^\top x, y) = (w^\top x - y)^2$

Q: can we use absolute loss, i.e., $|w^\top x - y|$?

Q: can we use $(w^\top x - y)^3$?

Mathematical formulation of linear regression

Formulating the optimization problem:

$$\arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Linear regression solution

$$\arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Let's compute the closed-form solution:

Linear regression solution

$$\arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Let's compute the closed-form solution:

Define $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$

Linear regression solution

$$\arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Let's compute the closed-form solution:

Define $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$

$$\sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \|X^\top w - Y\|_2^2$$

Linear regression solution

$$\arg\min_w \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Let's compute the closed-form solution:

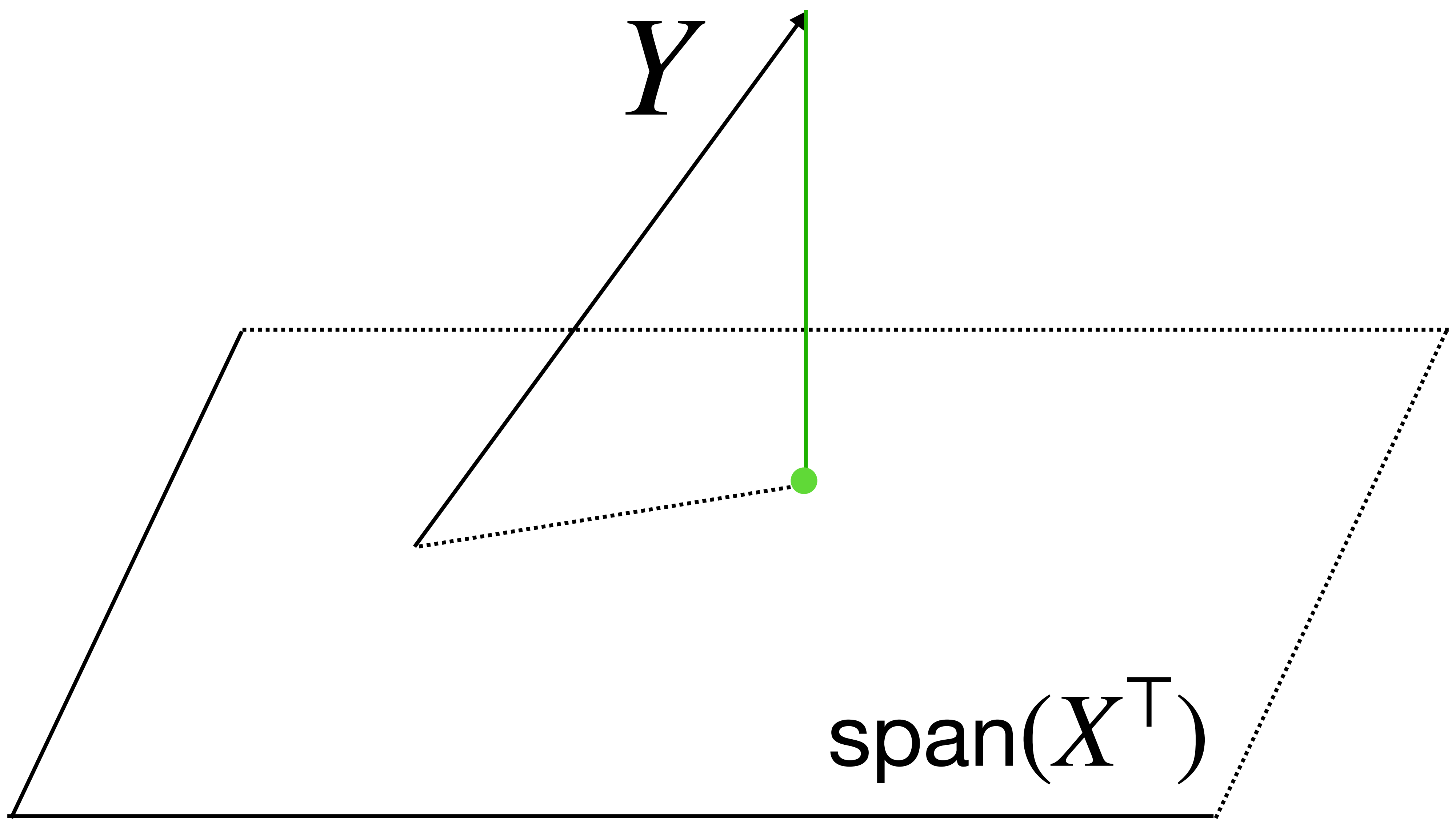
Define $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$

$$\sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \|X^\top w - Y\|_2^2$$

$$\Rightarrow \arg\min_w \|X^\top w - Y\|_2^2$$

Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2 \qquad X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$$



Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2$$

Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2$$

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY$$

Linear regression solution

$$\arg\min_w \|X^\top w - Y\|_2^2$$

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY$$

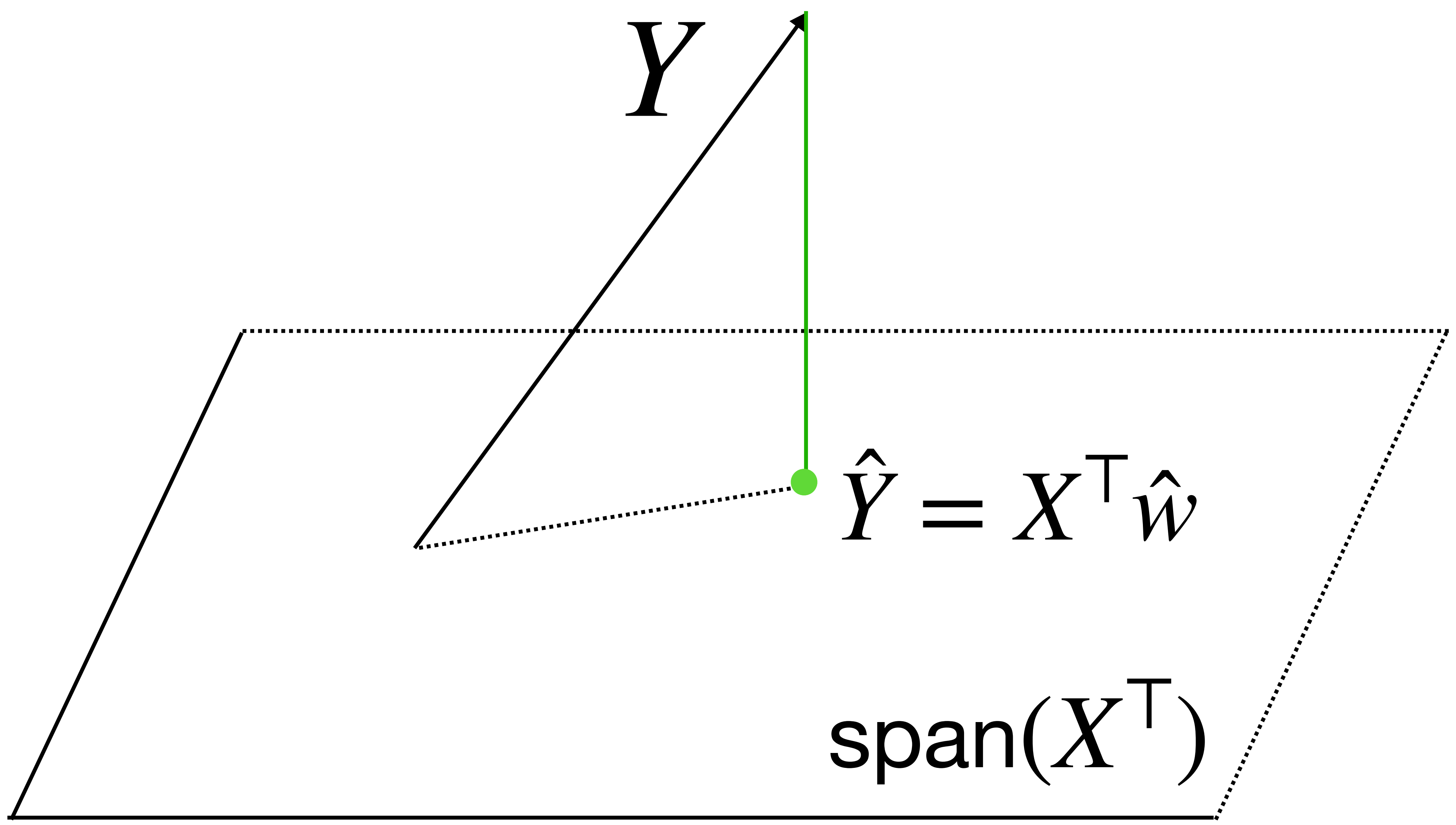
if $XX^\top$ is full rank, then $\hat{w} = (XX^\top)^{-1}XY$

Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2$$

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY$$

if $XX^\top$ is full rank, then $\hat{w} = (XX^\top)^{-1}XY$

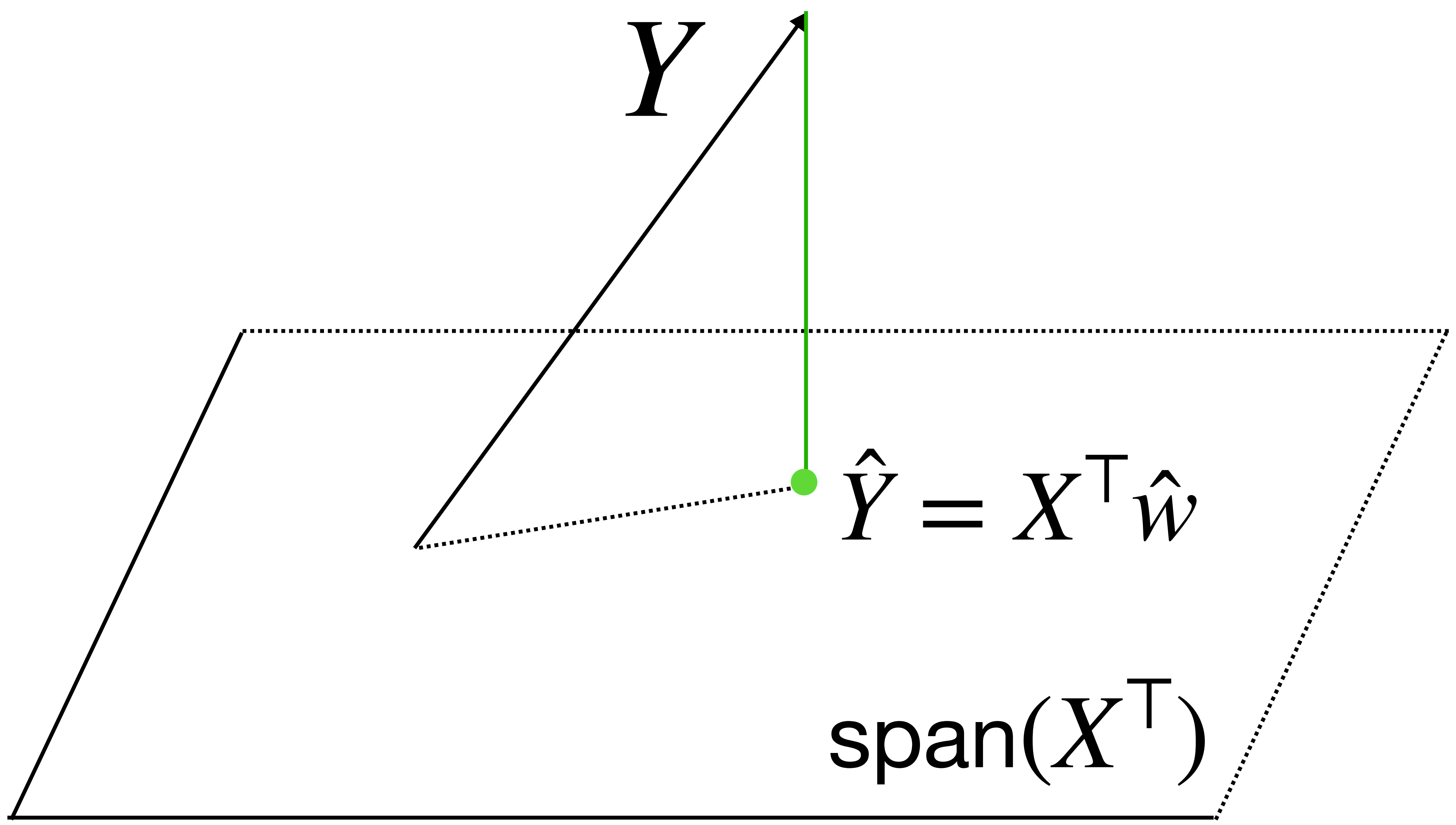


Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2$$

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY$$

if $XX^\top$ is full rank, then $\hat{w} = (XX^\top)^{-1}XY$



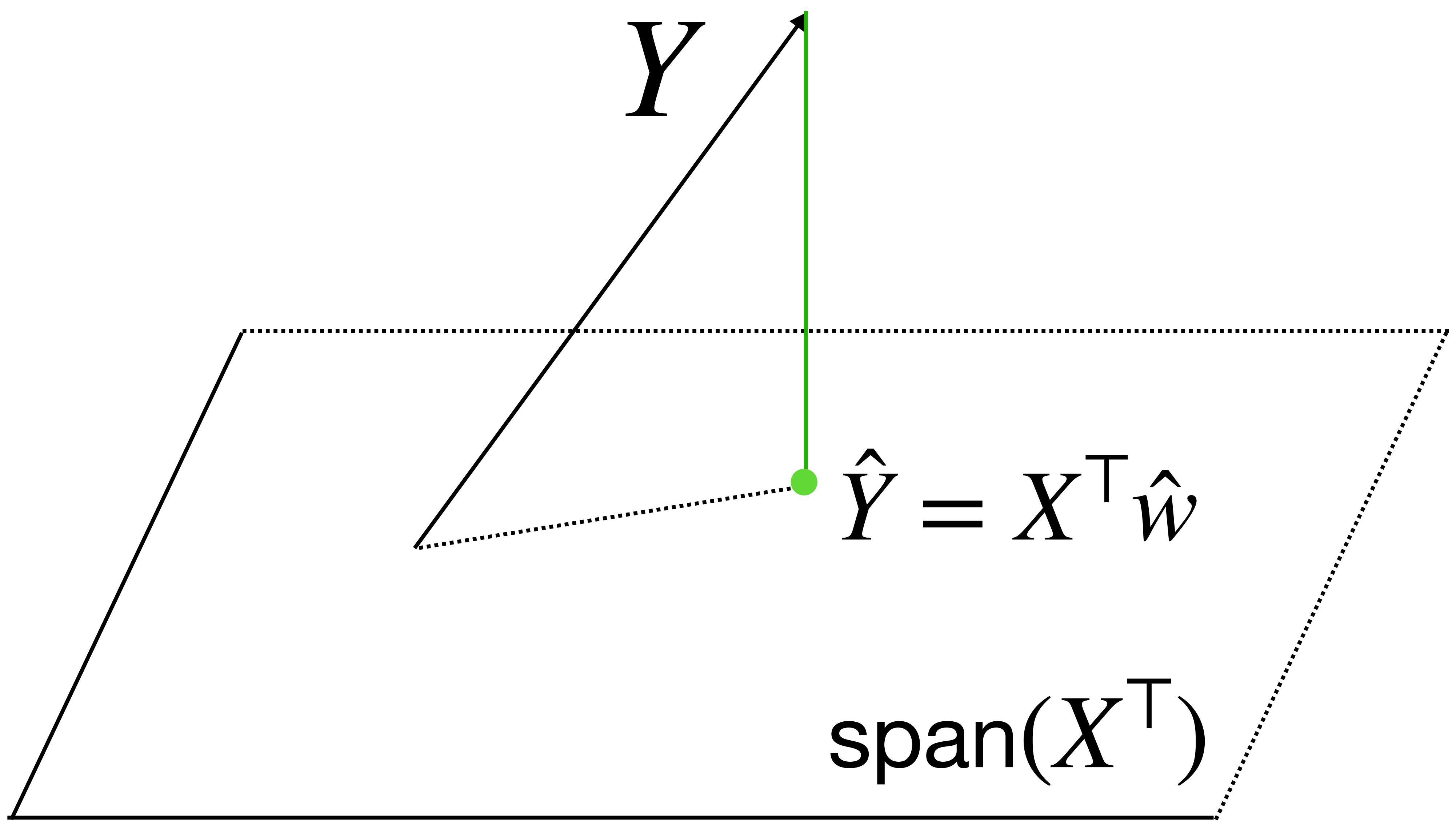
What if $XX^\top$ is not full rank?

Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2$$

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY$$

if $XX^\top$ is full rank, then $\hat{w} = (XX^\top)^{-1}XY$

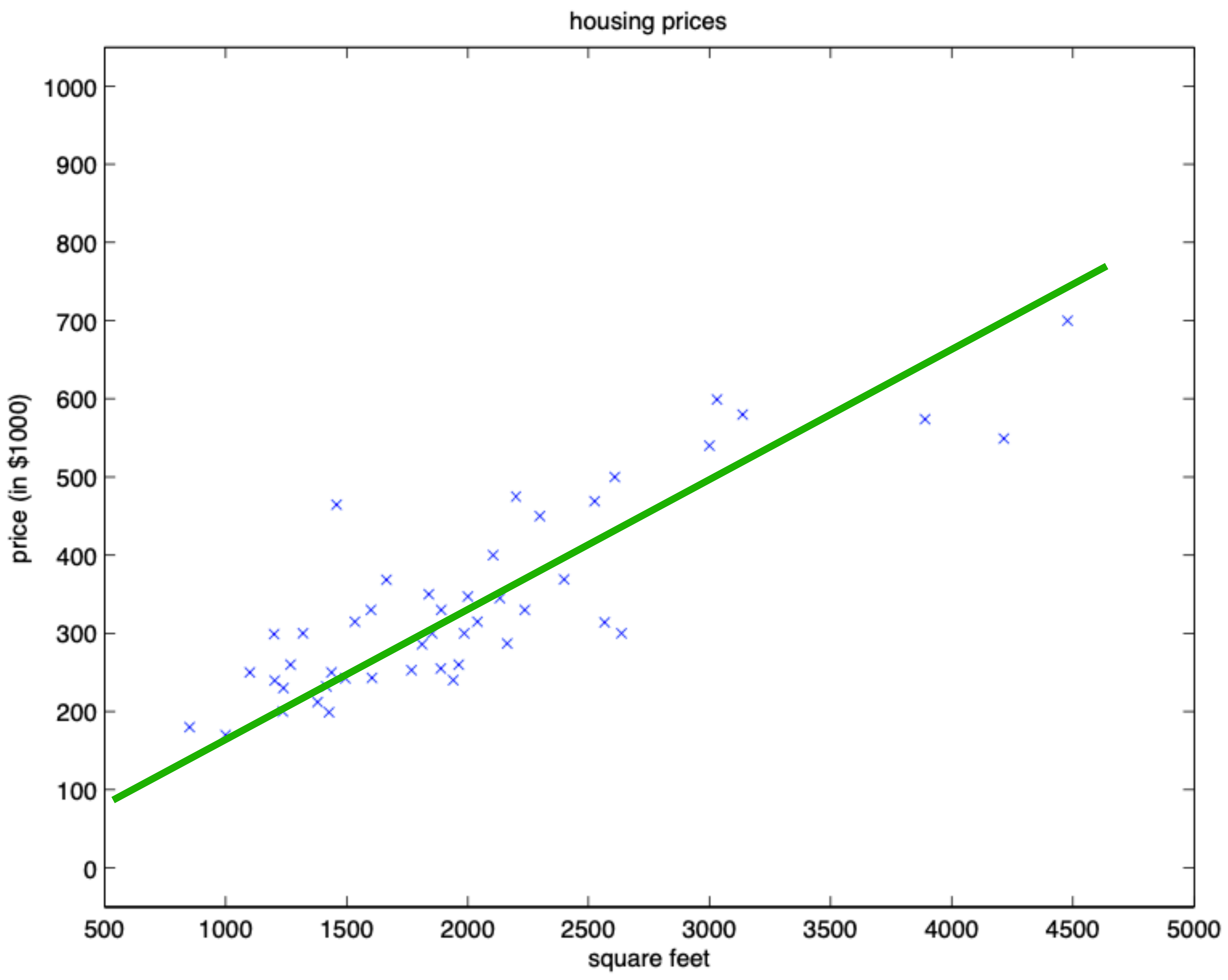


What if $XX^\top$ is not full rank?

(We will talk about regularization soon)

Prediction using linear regression

Once we learned $\hat{w}$, we can use it to make prediction on any new feature x

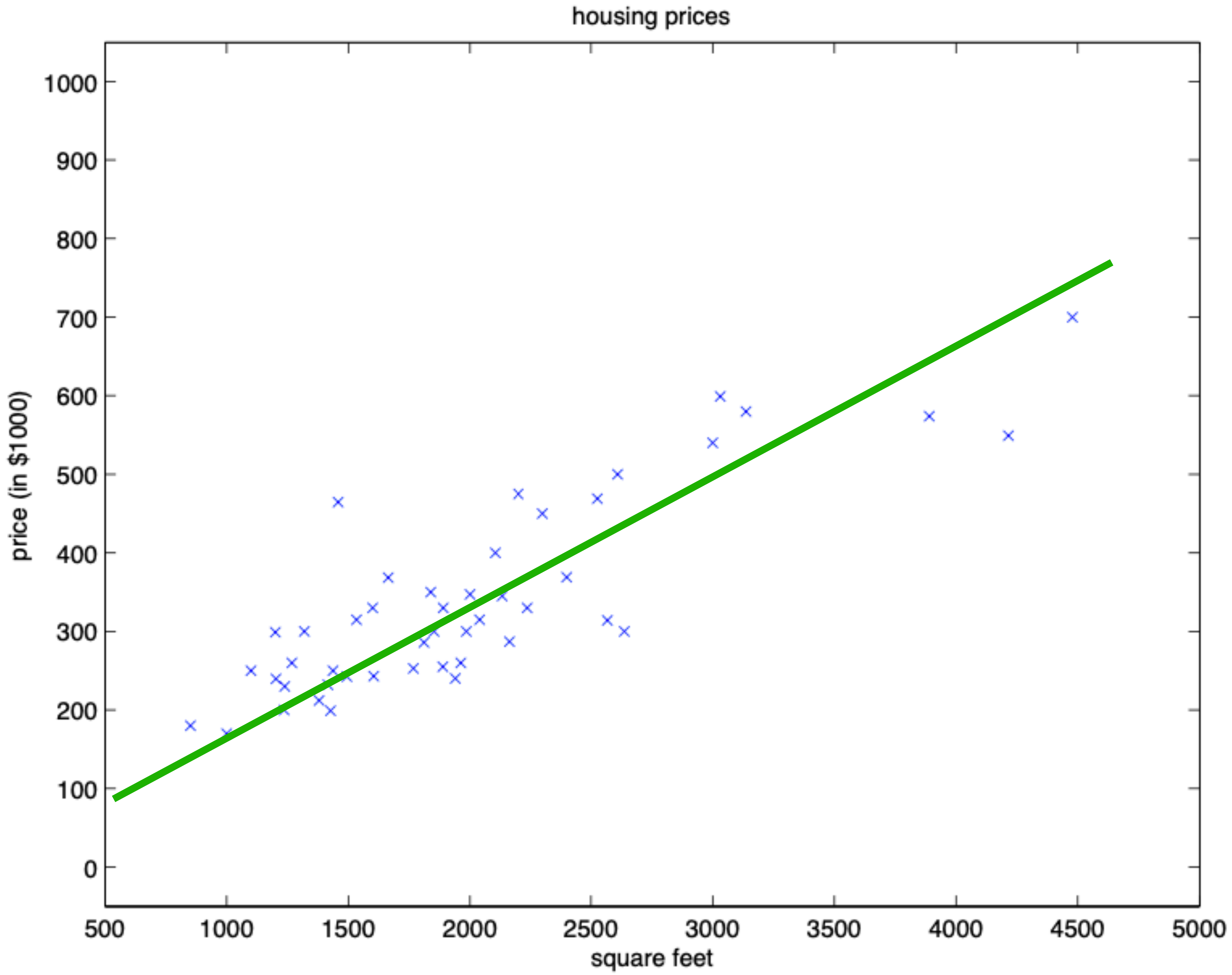


Given x_{test}, our prediction is:

$$\hat{y} = x_{test}^{\top}\hat{w}$$

Prediction using linear regression

Once we learned $\hat{w}$, we can use it to make prediction on any new feature x

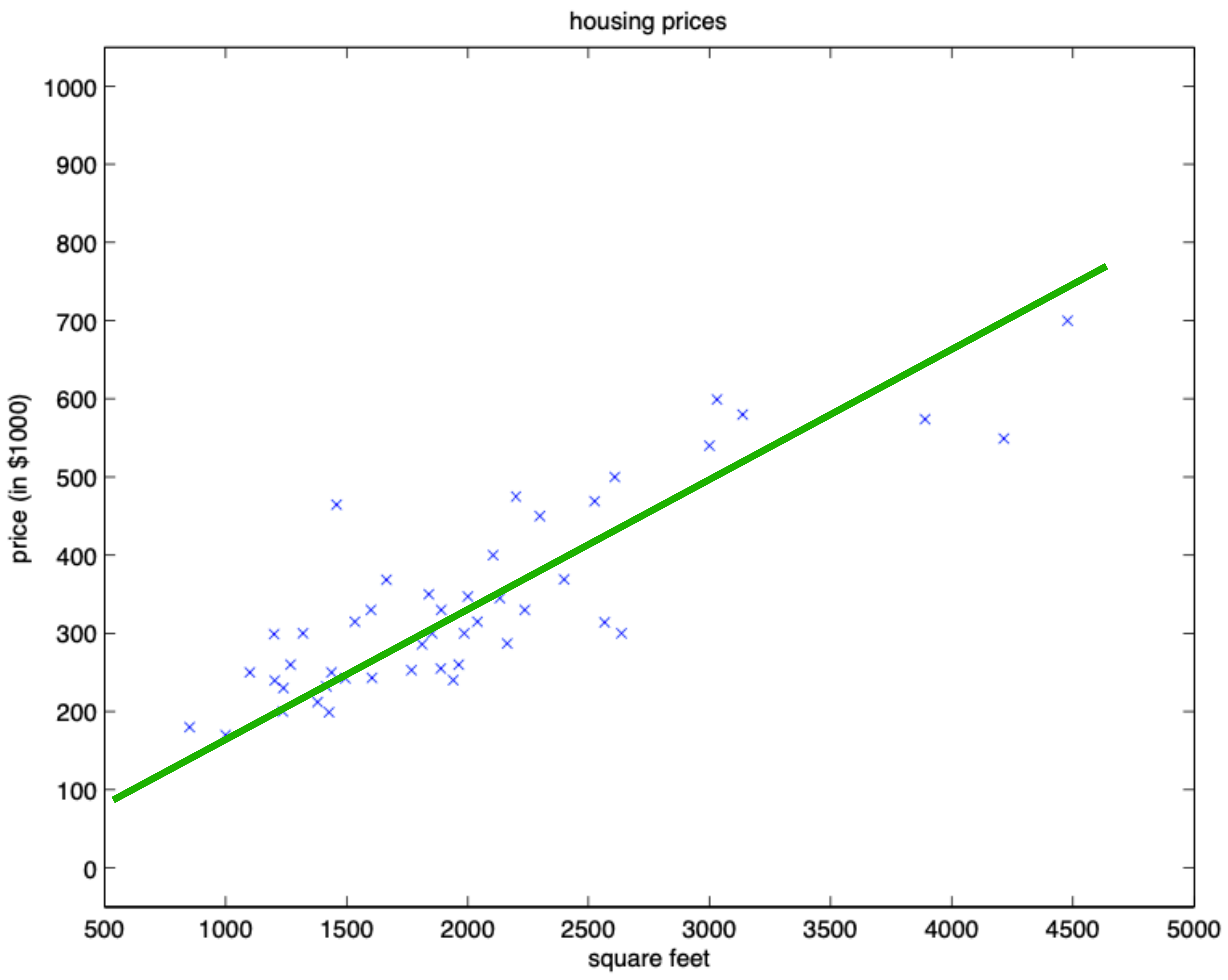


Given x_{test}, our prediction is:

$$\hat{y} = x_{test}^{\top}\hat{w}$$
$$= x_{test}^{\top}(XX^{\top})^{-1}XY$$

Prediction using linear regression

Once we learned $\hat{w}$, we can use it to make prediction on any new feature x



Given x_{test}, our prediction is:

$$\hat{y} = x_{test}^\top \hat{w}$$

$$= x_{test}^\top (XX^\top)^{-1} XY$$

$$= \sum_i \left(x_{test}^\top (XX^\top)^{-1} x_i \right) \cdot y_i$$

Outline for Today

1. Intro on Linear Regression

2. Normal equation for linear Regression

3. Interpretation of Linear Regression using MLE / MAP

Derive Linear regression via Maximum Likelihood Estimation

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Derive Linear regression via Maximum Likelihood Estimation

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Let's maximize the log-likelihood of the data, i.e.,

Derive Linear regression via Maximum Likelihood Estimation

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Let's maximize the log-likelihood of the data, i.e.,

$$\arg\max_{w} \sum_{i=1}^{n} \ln P(y_i \mid x_i; w)$$

Derive Linear regression via Maximum Likelihood Estimation

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Let's maximize the log-likelihood of the data, i.e.,

$$\arg\max_w \sum_{i=1}^{n} \ln P(y_i \mid x_i; w)$$

$$= \arg\max_w \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w^\top x_i - y_i)^2 - \ln(Z)$$

Derive Linear regression via Maximum Likelihood Estimation

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0,\sigma^2)$

Let's maximize the log-likelihood of the data, i.e.,

$$\arg\max_w \sum_{i=1}^{n} \ln P(y_i \mid x_i; w)$$

$$= \arg\max_w \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w^\top x_i - y_i)^2 - \ln(Z)$$

$$= \arg\min_w \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Derive Linear regression via MAP

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

To use MAP, we need to define a prior over w, we use Gaussian as well here:

$$w \sim \mathcal{N}(0, r^2 I)$$

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$$

MAP:

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2 \right)$$

MAP:

$$\arg\max_{w} \ln P(w \mid \mathcal{D})$$

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$$

MAP:

$$\arg\max_w \ln P(w \mid \mathcal{D})$$

$$= \arg\max_w \ln P(w) + \ln P(\mathcal{D} \mid w)$$

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z}\exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$$

MAP:

$$\arg\max_w \ln P(w \mid \mathcal{D})$$

$$= \arg\max_w \ln P(w) + \ln P(\mathcal{D} \mid w)$$

$$= \arg\max_w \frac{-w^\top w}{2r^2} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w^\top x_i - y_i)^2$$

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$$

MAP:

$$\arg\max_w \ln P(w \mid \mathcal{D})$$

$$= \arg\max_w \ln P(w) + \ln P(\mathcal{D} \mid w)$$

$$= \arg\max_w \frac{-w^\top w}{2r^2} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w^\top x_i - y_i)^2$$

$$= \arg\min_w \frac{\sigma^2}{r^2} w^\top w + \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$$

MAP:

$$\arg\max_w \ln P(w \mid \mathcal{D})$$

$$= \arg\max_w \ln P(w) + \ln P(\mathcal{D} \mid w)$$

$$= \arg\max_w \frac{-w^\top w}{2r^2} + \sum_{i=1}^n -\frac{1}{2\sigma^2}(w^\top x_i - y_i)^2$$

$$= \arg\min_w \frac{\sigma^2}{r^2} w^\top w + \sum_{i=1}^n (w^\top x_i - y_i)^2 \qquad = \arg\min_w \lambda \|w\|_2^2 + \sum_{i=1}^n (w^\top x_i - y_i)^2$$

Ridge Linear Regression

$$\arg\min_{w} \lambda \|w\|_2^2 + \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

In this case, we can derive a closed-form solution as well:

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY$$

Ridge Linear Regression

$$\arg\min_{w} \lambda \|w\|_2^2 + \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

In this case, we can derive a closed-form solution as well:

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY$$

Note that it works even $XX^\top$ is not full rank

Summary for today

1. Linear regression, Normal equation, and MLE / MAP interpretation

Summary for today

1. Linear regression, Normal equation, and MLE / MAP interpretation

2. Your take-home question: what is the SGD update rule for Linear regression? Is the update rule intuitively explainable?

Summary for today

1. Linear regression, Normal equation, and MLE / MAP interpretation

2. Your take-home question: what is the SGD update rule for Linear regression? Is the update rule intuitively explainable?

3. Next Tue: Support Vector Machine!