Linear Regression

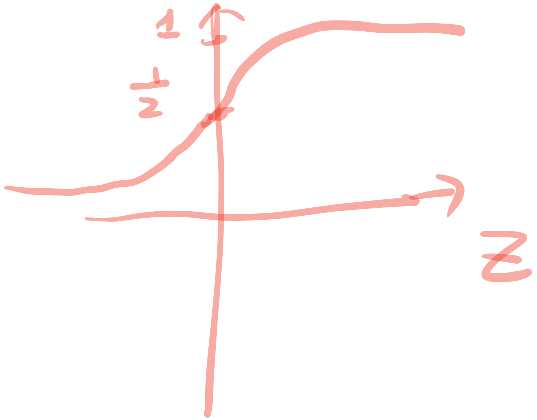
Announcements

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

$z := y(w^\top x)$



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 $(x, y) \sim \mathcal{D}$

$$\nabla_w \ln(1 + \exp(-y(w^\top x)))$$

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

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

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

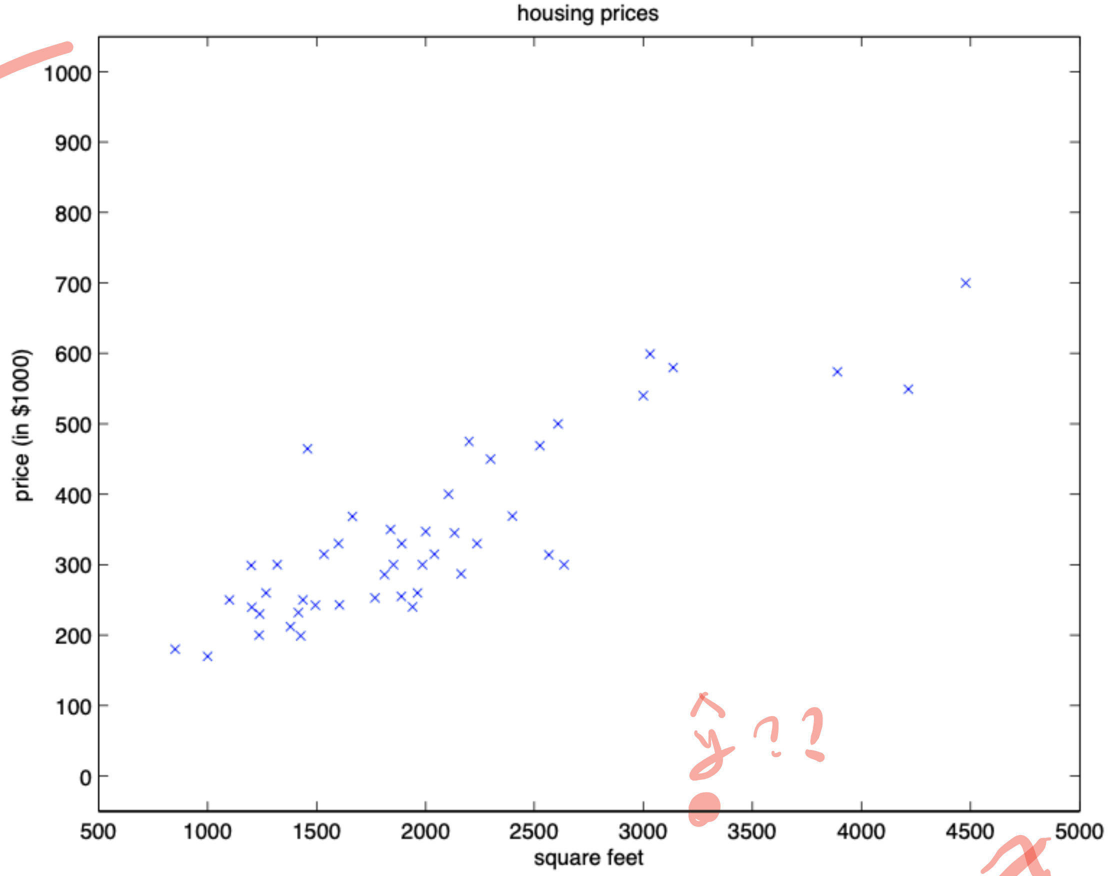
(Example from Stanford CS229)

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:



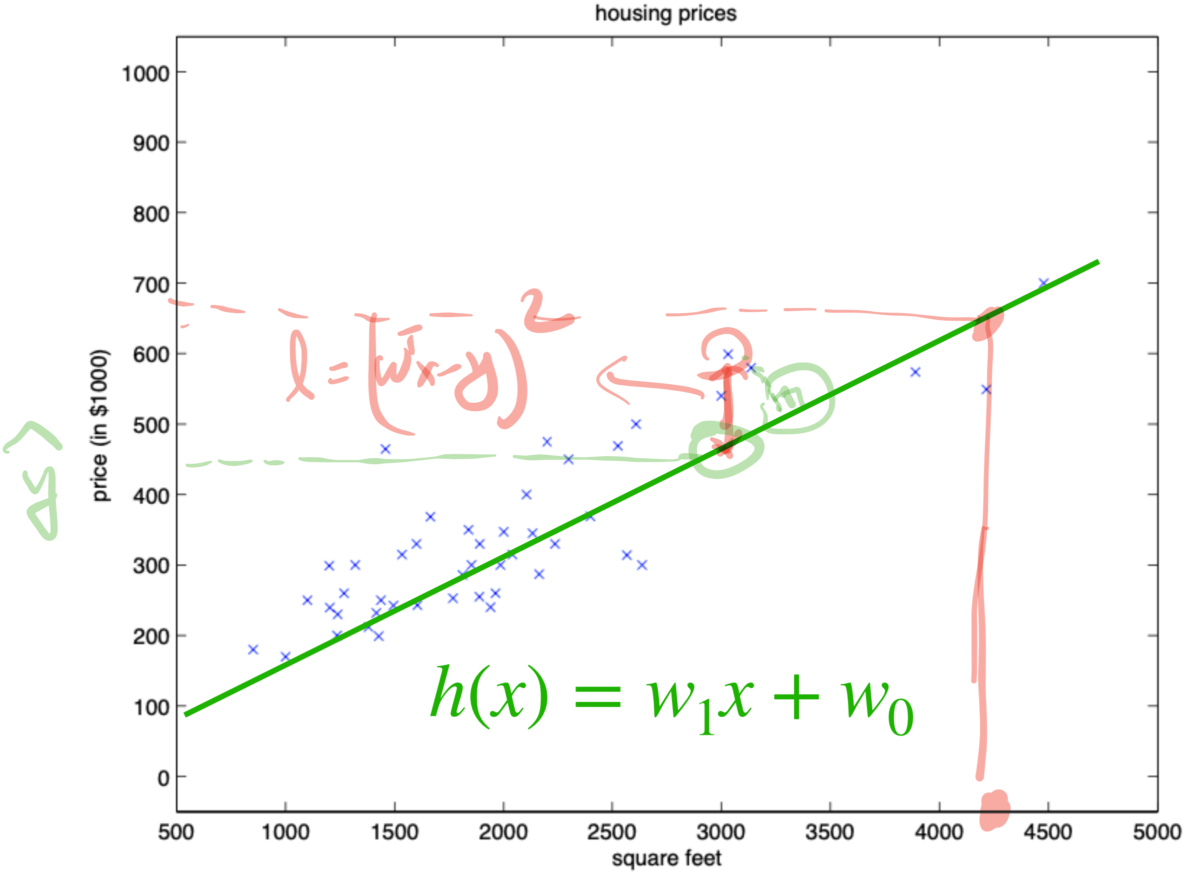
(Example from Stanford CS229)

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:

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

$$= \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}^T \begin{bmatrix} x(1) \\ x(2) \\ 1 \end{bmatrix}$$

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

$w \in \mathbb{R}^d$

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 $\ell(w^\top x, y) = (w^\top x - y)^2$

Diff between $w^T x$ & GT 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., $|w^\top x - y|$?

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 $(w^\top x - y)^3$?

$(w^\top x - y)^{10}$

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

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Linear regression solution

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}$$

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

$$\|w\|_2^2 = \sum_{i=1}^{d} w_i^2$$

Let's compute the closed-form solution:

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

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

$$\left\| \begin{bmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_n^\top & - \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right\|_2^2$$

Linear regression solution

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

Let's compute the closed-form solution:

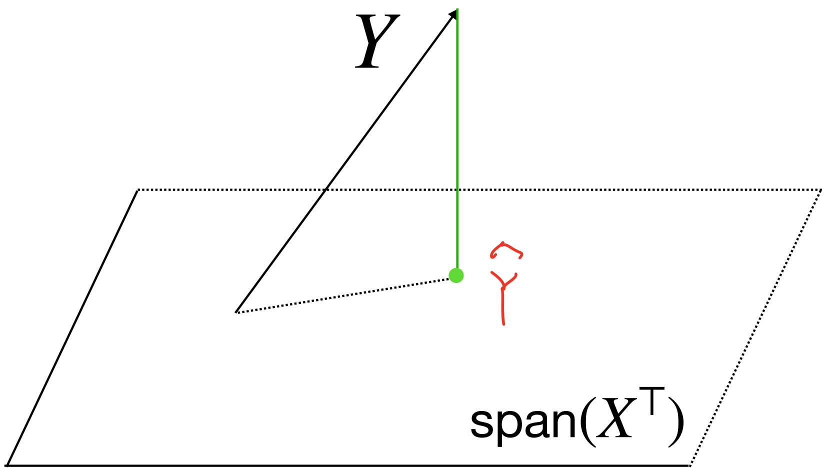
Define $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$

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

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

Linear regression solution

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



$$\min_{\hat{Y}} \|\hat{Y} - Y\|_2^2$$

$$\hat{Y} = X^\top w$$

Linear regression solution

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

Linear regression solution

$$(X^\top w - Y)^\top (X^\top w - Y)$$

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

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

$$2XX^\top w - 2XY = 0$$

$$\|X^\top w - Y\|_2^2 = w^\top X X^\top w - 2w^\top X Y + Y^\top Y$$

$$\nabla_w (w^\top X X^\top w) = XX^\top w + (XX^\top)^\top w = 2(XX^\top)w$$

$$\nabla_w (2w^\top X Y) = 2XY$$

Linear regression solution

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

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY = 0$$

$$XX^\top w = XY$$

Normal Eqn

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

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

$= I$

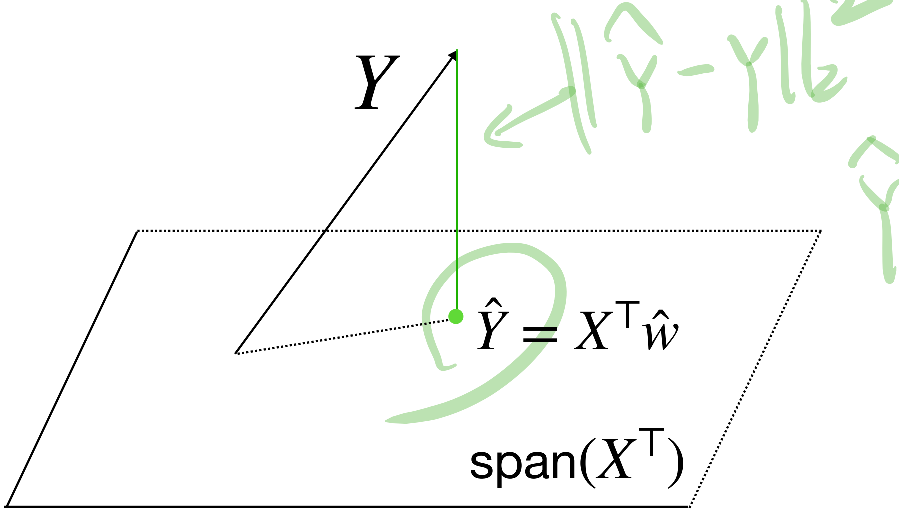
$$\Rightarrow w = (XX^\top)^{-1} XY$$

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



$\hat{Y} = X^\top \hat{w} = X^\top (XX^\top)^{-1} XY$

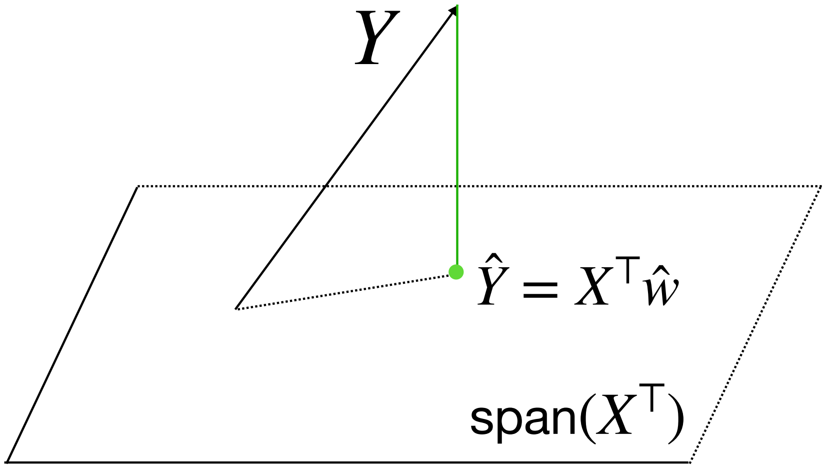
Linear regression solution

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

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

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

$X = [x_1\ x_2 \ldots x_n] \in \mathbb{R}^{d \times n}$



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

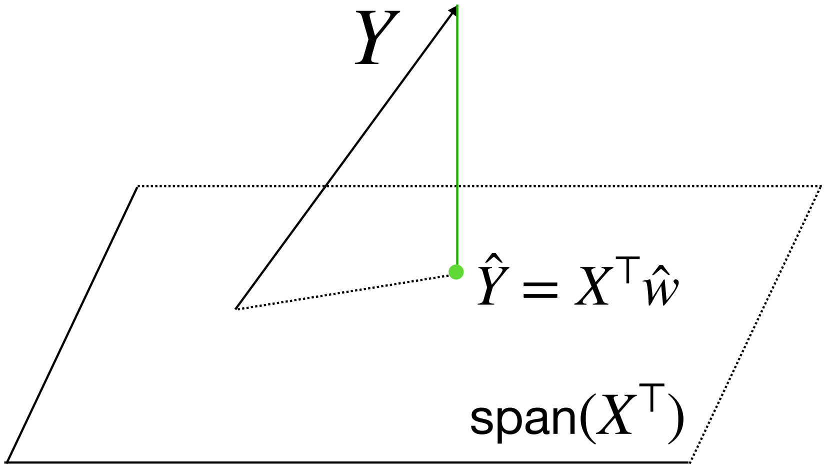
$XX^\top \in \mathbb{R}^{d \times d}$

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

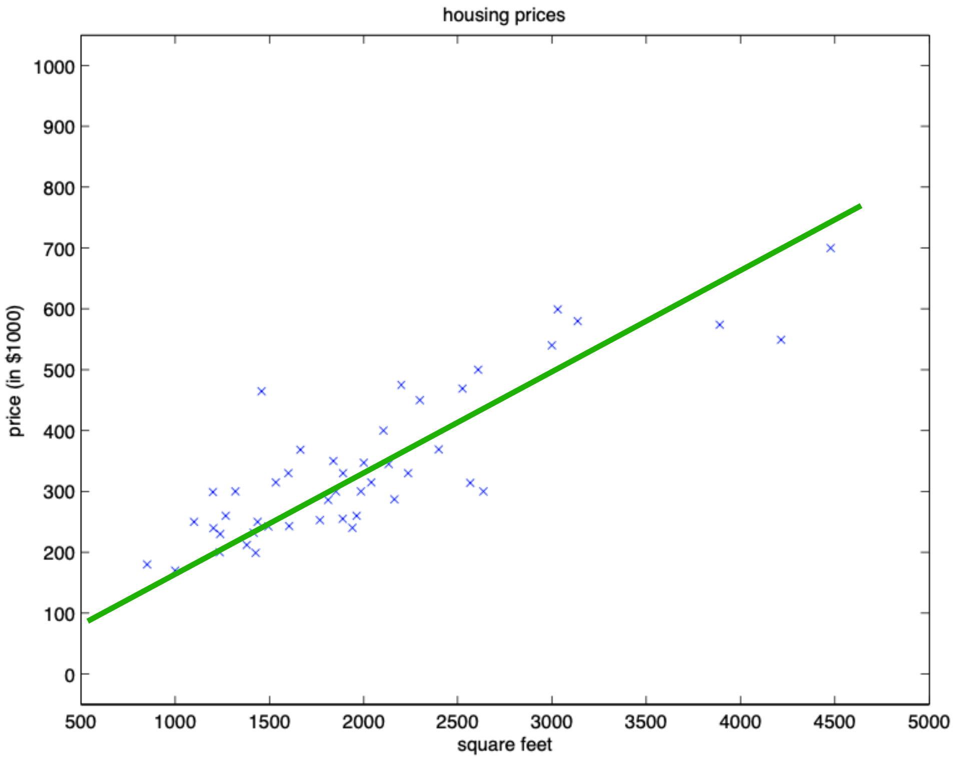


What if $XX^\top$ is not full rank?

(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

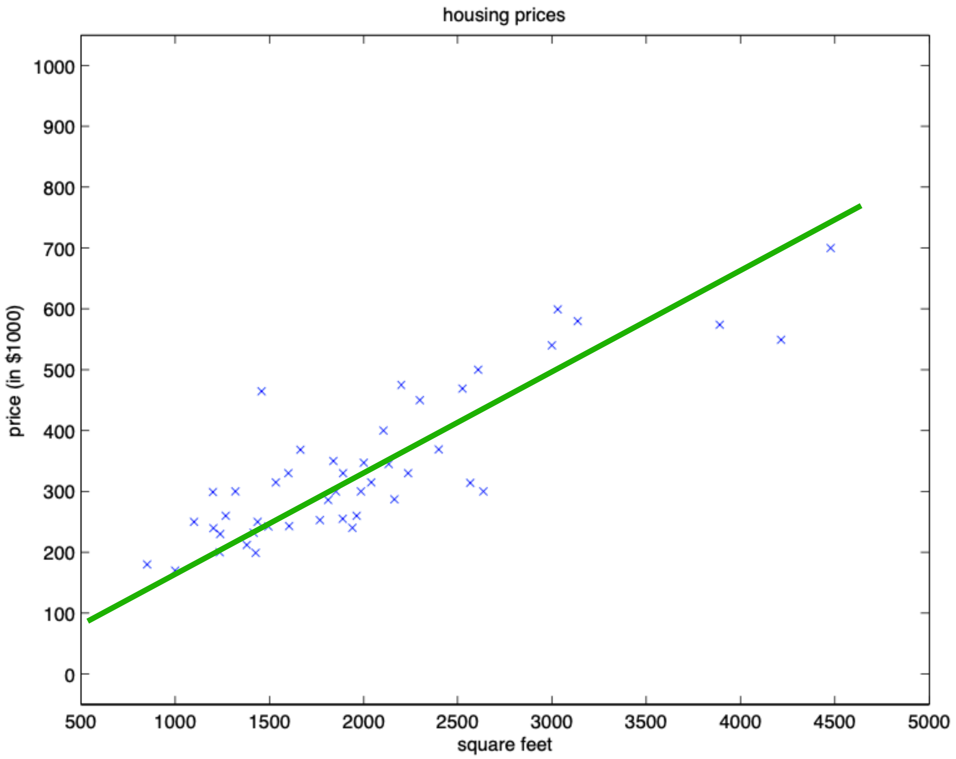


Given x_{test}, our prediction is:

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

Prediction using linear regression

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



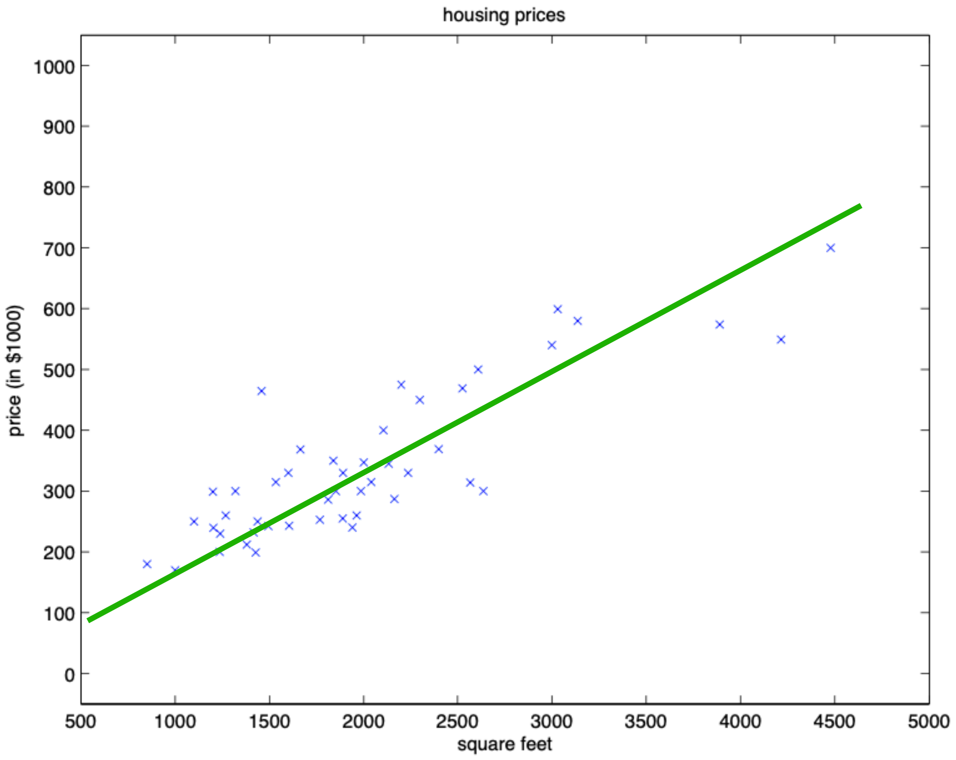
Given x_{test}, our prediction is:

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

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

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:

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

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

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

scalar

$$= \sum_{i=1}^{n} \alpha_i y_i$$

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., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

mean σ^2

$$\prod_{i=1}^{n} P(y_i \mid x_i; w)$$

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.,

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.,

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

$$\ln\left(\frac{1}{Z} \exp\left(-\frac{1}{2}(y_i - x_i^\top w)^2/\sigma^2\right)\right)$$

$$= -\ln(Z) - \frac{1}{2}(y_i - x_i^\top w)^2/\sigma^2$$

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.,

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

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

$$\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)$$

$$Z := \sigma\sqrt{2\pi}$$

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.,

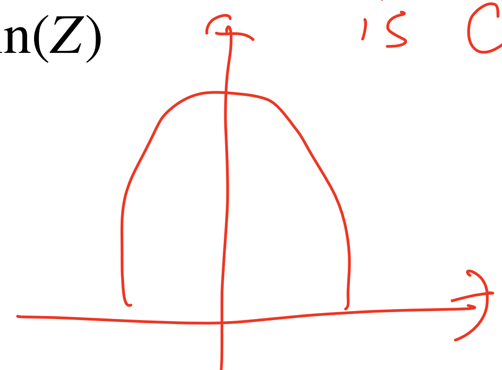
$$\arg\max_w \sum_{i=1}^{n} \ln P(y_i \mid x_i; w)$$

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

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

$\ln P(y \mid x; w)$

is concave wrt w



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:

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

r big number

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:

$$\arg\max_{w} \ln P(w \mid \mathcal{D}) \propto \text{prior} \times \text{likelihood}$$

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:

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

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

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

$\hookrightarrow \ln P(w)$ $\qquad \nwarrow \ln \prod_{i=1}^{n} P(y_i \mid x_i; w)$

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:

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

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

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

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

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:

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

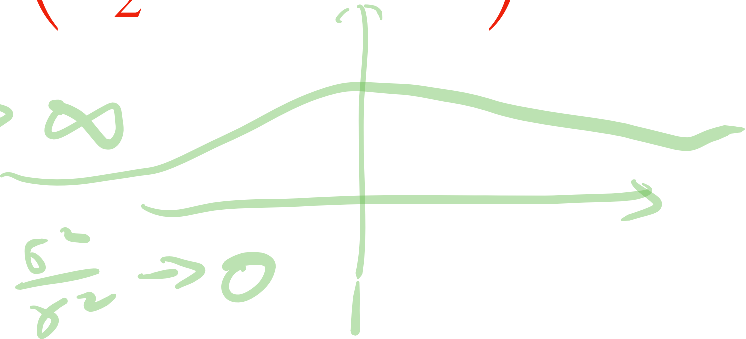
$$= \arg\max_w \ln P(w) + \ln P(\mathcal{D} \mid w)$$

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

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

set $r \to \infty$

$\Rightarrow \lambda = \frac{\sigma^2}{r^2} \to 0$



$\lambda := \frac{\sigma^2}{r^2}$

Regularization

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:

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

(Recall for normal Eqn: $(XX^T)^{-1}XY$

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

$XX^\top + \lambda I$ is always PD ($\lambda > 0$)

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!

$$f(x) = x^T A x \qquad A \in \mathbb{R}^{d \times d}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$$\nabla_x f(x) = A\cancel{x} + A^T x$$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \checkmark \\ \frac{\partial f(x)}{\partial x_2} \checkmark \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \checkmark \end{bmatrix} = Ax + A^T x$$

$$Ax = \sum_{i=1}^{d} A_i x_i \qquad A = \begin{bmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{bmatrix}$$

$$x^T A x = x^T (Ax)$$

$$= x^T \left(\sum_{i=1}^{d} a_i x_i \right)$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{d} x_j a_{ij} x_i$$

$$\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = a_1 x_1 + a_2 \cdot x_2$$

$$x^T A x = \sum_{i=1}^{d} \sum_{j=1}^{d} x_j A_{i,j} x_i$$

$A \in \mathbb{R}^{d \times d}$

$$i \rightarrow \begin{bmatrix} - A_{i,j} \\ \vdots \\ \end{bmatrix}$$
j

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$$\frac{\partial x^T A x}{\partial x_1}$$