Classification problem

Overfitting and regularization

Khiem Nguyen

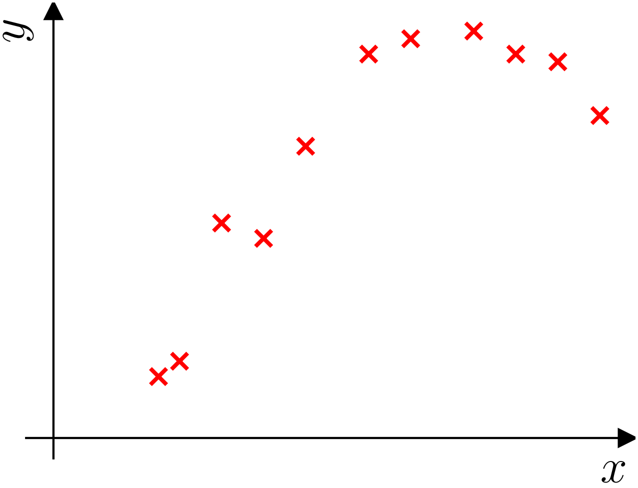
Email	khiem.nguyen@glasgow.ac.uk
MS Teams	khiem.nguyen@glasgow.ac.uk
Whatsapp	+44 7729 532071 (Emergency only)

May 18, 2025

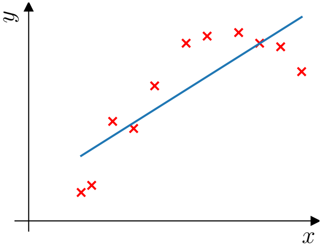
Table of Contents

1. Underfitting (high bias) versus Overfitting (high varianace)
2. Addressing overfitting
3. Regularization: Intuition and Formulation

Underfitting versus overfitting



Underfit versus overfit



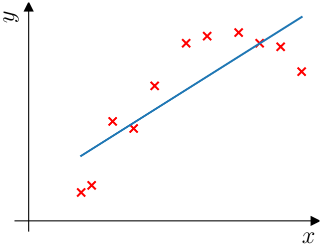
Underfitting

$w_1 x + b$

Does not fit the training set well

high bias

Underfit versus overfit

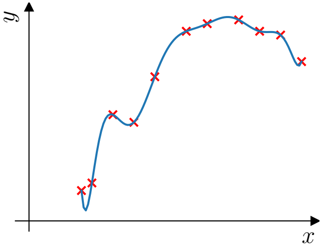


Underfitting

$w_1 x + b$

Does not fit the training set well

high bias



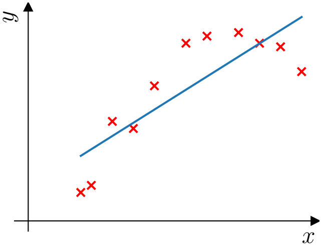
Overfitting

$w_1 x + w_2 x^2 + \cdots + w_{10} x^{10} + b$

Fit the training set extremely well

high variance

Underfit versus overfit

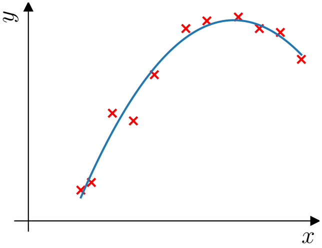


Underfitting

$w_1 x + b$

Does not fit the training set well

high bias

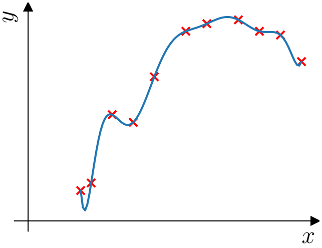


Just right

$w_1 x + w_2 x^2 + b$

Fit training set pretty well

generalization



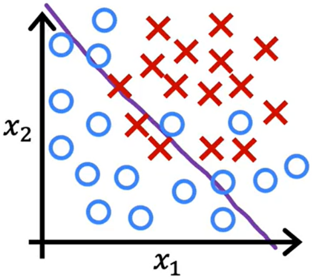
Overfitting

$w_1 x + w_2 x^2 + \cdots + w_{10} x^{10} + b$

Fit the training set extremely well

high variance

Underfitting versus overfitting: Classification



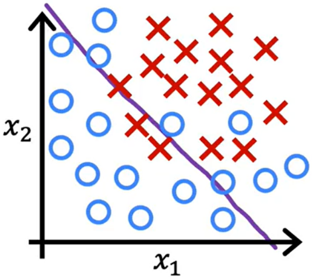
Underfit – high bias

$z = w_1 x_1 + w_2 x_2 + b$

$f_{\vec{w},b}(\vec{x}) = g(z)$

g is the sigmoid function

Underfitting versus overfitting: Classification

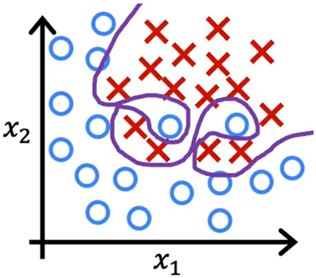


Underfit – high bias

$$z = w_1 x_1 + w_2 x_2 + b$$

$$f_{\vec{w},b}(\vec{x}) = g(z)$$

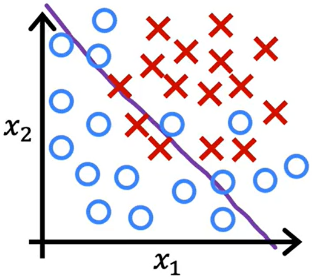
g is the sigmoid function



Overfit - high variance

$$\begin{aligned} z = {} & w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 \\ & + w_4 x_1^2 x_2^2 + w_5 x_1^2 x_2^3 \\ & + w_6 x_1^3 x_2 + \cdots + b \end{aligned}$$

Underfitting versus overfitting: Classification

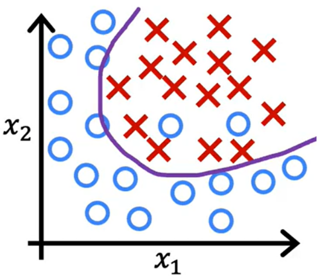


Underfit – high bias

$$z = w_1x_1 + w_2x_2 + b$$

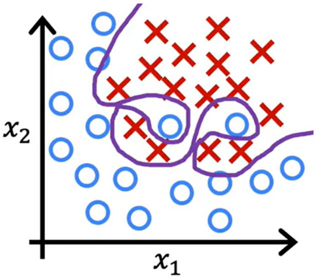
$$f_{\vec{w},b}(\vec{x}) = g(z)$$

g is the sigmoid function



Just right

$$z = w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x^2 + w_5x_1x_2 + b$$



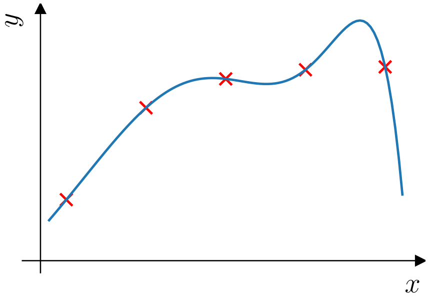
Overfit - high variance

$$z = w_1x_1 + w_2x_2 + w_3x_1^2x_2 + w_4x_1^2x_2^2 + w_5x_1^2x_2^3 + w_6x_1^3x_2 + \cdots + b$$

Table of Contents

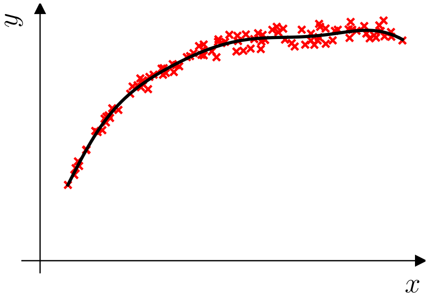
1. Underfitting (high bias) versus Overfitting (high varianace)
2. Addressing overfitting
3. Regularization: Intuition and Formulation

Addressing overfitting: Collect more training examples



Overfit

Maybe reduce the order of fitting?



Collect more training examples

It is not always easier to harvest more data

Select features to include/exculde

size	bedrooms	floors	age	avg. income	$\cdots$	distance to center	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

Select features to include/exculde

size	bedrooms	floors	age	avg. income	$\cdots$	distance to center	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

all features

$+$

insufficient data

$\downarrow$

overfit

Select features to include/exculde

size	bedrooms	floors	age	avg. income	$\cdots$	distance to center	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

all features

$+$

insufficient data

$\downarrow$

overfit

selected features

size

bedrooms

age

distance to center

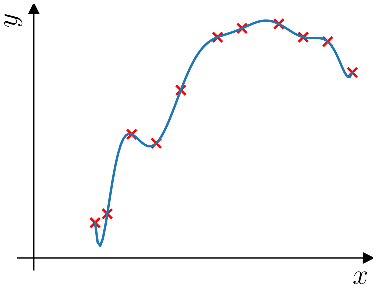
just right feature selection

disadvantage

$\downarrow$

useful features could be lost

Regularization

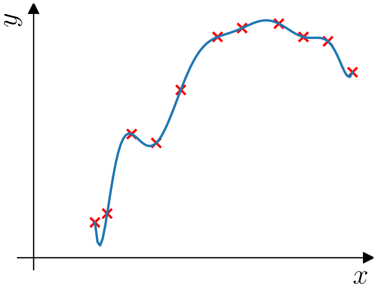


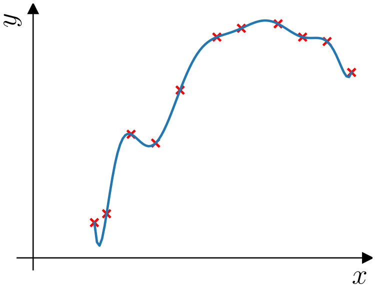
$$f_{\vec{w},b}(\vec{x}) = -8246.12x + 0.1351x^2 - \cdots + 33781x^8 - 542x^9 + 33.92x^{10} + 974.89$$

large values for model parameters $\vec{w}, b$

Produced by by using hand-written code with regularization on $w_3, \ldots, w_{10}$

Regularization



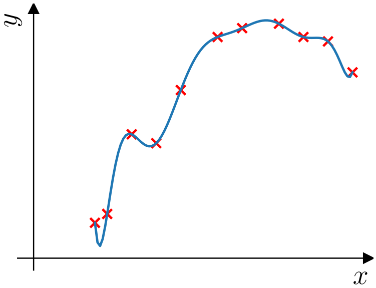


$$f_{\vec{w},b}(\vec{x}) = -8246.12x + 0.1351x^2 - \cdots + 33781x^8 - 542x^9 + 33.92x^{10} + 974.89$$

large values for model parameters $\vec{w}, b$

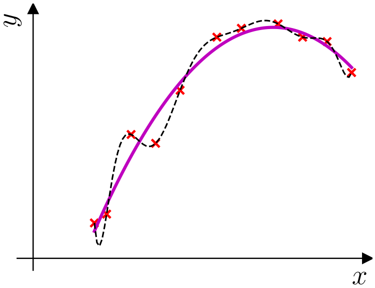
Produced by by using hand-written code with regularization on $w_3, \ldots, w_{10}$

Regularization



$$f_{\vec{w},b}(\vec{x}) =$$
$-8246.12x + 0.1351x^2 - \cdots + 33781x^8 - 542x^9 + 33.92x^{10} + 974.89$

large values for model parameters $\vec{w}, b$



$$f_{\vec{w},b}(\vec{x}) =$$
$2.971x_{\text{scaled}} - 2.35x^2_{\text{scaled}} + \text{small-no.}x^3_{scaled} + \cdots \text{small-no.}x^4_{scaled} + \text{small-no.}x^{10}_{scaled} + 1.845$

smaller values for model parameters $\vec{w}, b$

Produced by by using hand-written code with regularization on $w_3, \ldots, w_{10}$

Address overfitting: Summary

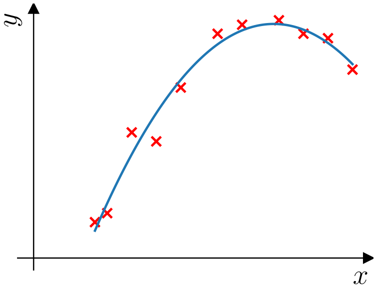
Options

1. Collect more data
2. Select features (feature selection)
3. Reduce size of parameters (**regularization**!) $\rightarrow$ We study it now!

Table of Contents

1. Underfitting (high bias) versus Overfitting (high varianace)
2. Addressing overfitting
3. Regularization: Intuition and Formulation

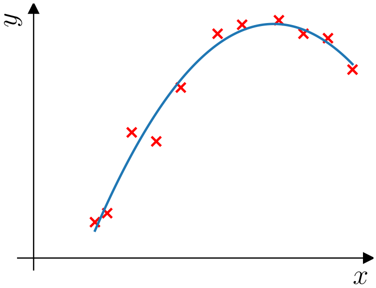
Regularization: Intuition



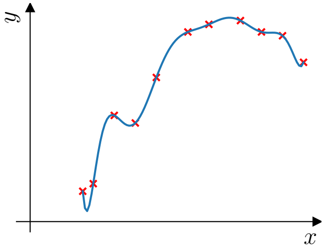
$w_1 x + w_2 x^2 + b$

Idea: Make $w_3, w_4, \ldots, w_{10}$ really small (≈ 0)

Regularization: Intuition



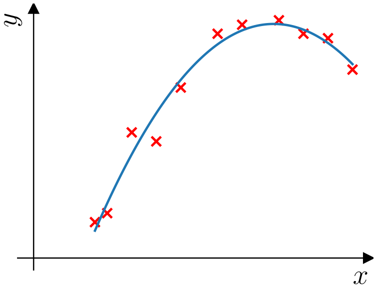
$$w_1x + w_2x^2 + b$$



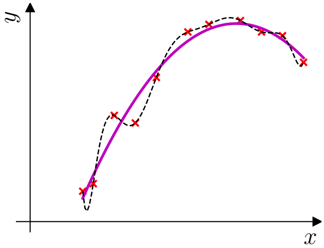
$$w_1x + w_2x^2$$
$$+w_3 \times x^3 + \cdots + w_{10} \times x^{10} + b$$

<u>Idea</u>: Make $w_3, w_4, \ldots, w_{10}$ really small (≈ 0)

Regularization: Intuition



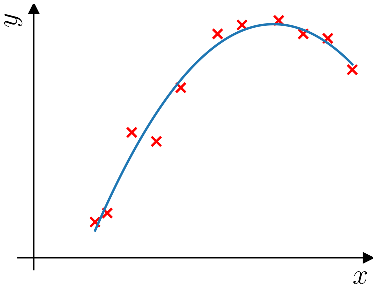
$$w_1 x + w_2 x^2 + b$$



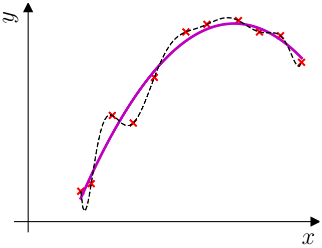
$$w_1 x + w_2 x^2$$
$$+\cancelto{\approx 0}{w_3} \times x^3 + \cdots + \cancelto{\approx 0}{w_{10}} \times x^{10} + b$$

<u>**Idea**</u>: Make $w_3, w_4, \ldots, w_{10}$ really small (≈ 0)

Regularization: Intuition



$$w_1 x + w_2 x^2 + b$$

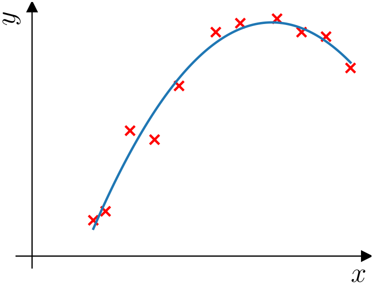


$$w_1 x + w_2 x^2$$
$$+\cancelto{\approx 0}{w_3} \times x^3 + \cdots + \cancelto{\approx 0}{w_{10}} \times x^{10} + b$$

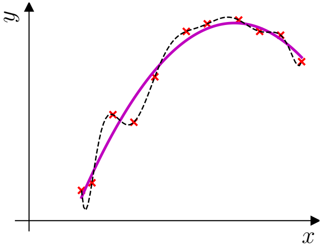
<u>Idea</u>: Make $w_3, w_4, \ldots, w_{10}$ really small (≈ 0)

$$\min_{\vec{w},b} \left\{ \frac{1}{2m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)})^2 \right\}$$

Regularization: Intuition



$$w_1 x + w_2 x^2 + b$$



$$w_1 x + w_2 x^2$$
$$+\cancelto{\approx 0}{w_3} \times x^3 + \cdots + \cancelto{\approx 0}{w_{10}} \times x^{10} + b$$

<u>**Idea**</u>: Make $w_3, w_4, \ldots, w_{10}$ really small (≈ 0)

$$\min_{\vec{w},b} \frac{1}{2m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)})^2 + \text{Large-number} \times w_3^2 + \cdots + \text{Large-number} \times w_{10}^2$$

To make the cost function small, $w_3, \ldots, w_{10}$ should become smaller ($\to 0$).

Regularization: Cost function

size	bedrooms	floors	age	avg. income	$\cdots$	distance to center	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

Smaller values $w_1, w_2, \dots, w_n, b$ $\rightarrow$ simpler model $\rightarrow$ less likely to overfit

$$J(\vec{w}, b) = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right)^2 + \underbrace{\frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2}_{\text{regularization term}}$$

λ – **regularization parameter** Remember: $\lambda > 0$

Regularization

$$\min_{\vec{w},b} J(\vec{w},b) = \min_{\vec{w},b} \left\{ \frac{1}{2} \underbrace{\frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right)^2}_{\text{mean squared error}} + \underbrace{\frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2}_{\text{regularization term}} \right\}$$

λ balances both goals

1. fit the data
2. keep w_j small (if needed)

But

choosing large λ may lead to simple fit. For example, with $\lambda = 10^{10}$

$$f_{\vec{w},b}(\vec{x}) = \cancel{w_1}^{\approx 0} x + \cancel{w_2}^{\approx 0} x^2 + \cdots + \cancel{w_{10}}^{\approx 0} x^{10} + b \quad \rightarrow \quad f_{\vec{w},b}(\vec{x}) = 0$$

Regularized linear regression

$$\min_{\vec{w},b} J(\vec{w},b) = \min_{\vec{w},b} \left\{ \underbrace{\frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2}_{1/2\ \times \text{MSE}} + \underbrace{\frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2}_{\text{regularization term}} \right\}$$

Gradient descent

Repeat

$$\begin{cases} w_1 = w_1 - \alpha \dfrac{\partial J}{\partial w_1}(\vec{w},b) \\ \quad\vdots \\ w_n = w_n - \alpha \dfrac{\partial J}{\partial w_n}(\vec{w},b) \\ b = b - \alpha \dfrac{\partial J}{\partial b}(\vec{w},b) \end{cases}$$

$$\frac{\partial J}{\partial w_j} = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)x_j^{(i)}}_{\text{from } 1/2\ \times \text{MSE}} + \frac{\lambda}{m} w_j$$

$$j = 1, \ldots, n$$

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)$$

Don't have to regularize b

Implementing gradient descent

Repeat

$$\begin{cases} w_j = w_j - \alpha \left\{ \frac{1}{m} \sum_{i=1}^{m} \left[\left(f_{\vec{w},b}(x^{(i)}) - y^{(i)} \right) x_j^{(i)} \right] + \frac{\lambda}{m} w_j \right\} \\ \\ b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(x^{(i)}) - y^{(i)} \right) \end{cases}$$

$$w_j = \underbrace{w_j - \alpha \frac{\lambda}{m} w_j}_{w_j \left(1 - \alpha \frac{\lambda}{m}\right)} \; \underbrace{- \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right) x_j^{(i)}}_{\text{usual update as without regularization}}$$

The term $w_j\left(1 - \alpha \frac{\lambda}{m}\right)$ shrinks w_j a bit in each iteration

Example:

$$\alpha \frac{\lambda}{m} = 0.1 \times \frac{1}{1000} = 0.0001$$

$$\rightarrow \quad w_j \underbrace{(1 - 0.0001)}_{0.9999}$$

Derivation of gradient of the cost function $J(\vec{w}, b)$ (optional)

$$\frac{\partial J}{\partial w_1} J(\vec{w}, b) = \frac{\partial}{\partial w_1} \left\{ \frac{1}{2m} \sum_{i=1}^{m} \Big(\underbrace{\vec{w} \cdot \vec{x}^{(i)} + b - y^{(i)}}_{w_1 x_1^{(i)} + \cdots + w_n x_n^{(i)} + b - y^{(i)}} \Big)^2 + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2 \right\}$$

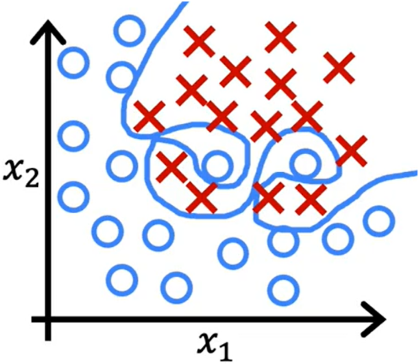
$$= \frac{1}{\cancel{2}m} \sum_{i=1}^{m} \cancel{2} \left[\left(\vec{w} \cdot \vec{x}^{(i)} + b - y^{(i)} \right) x_1^{(i)} \right] + \frac{\lambda}{\cancel{2}m} \cancel{2} w_1$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left[\Big(\underbrace{\vec{w} \cdot \vec{x}^{(i)} + b}_{f_{\vec{w},b}(\vec{x}^{(i)})} - y^{(i)} \Big) x_1^{(i)} \right] + \frac{\lambda}{m} w_1 = \frac{1}{m} \sum_{i=1}^{m} \left[\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right) x_1^{(i)} \right] + \frac{\lambda}{m} w_1$$

Generalization ($\partial J / \partial b$ the same as before):

$$\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left[\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right) x_j^{(i)} \right] + \frac{\lambda}{m} w_j$$

Regularized logistic regression

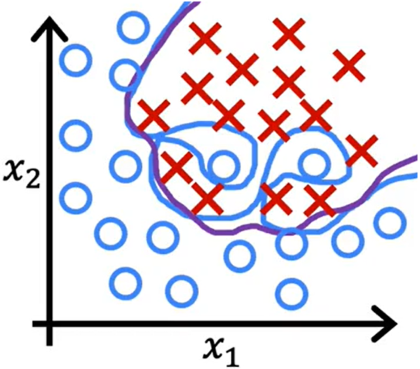


$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2$$

$$+ w_4 x_1^2 x_2^2 + w_5 x_1^2 x_2^3 + \cdots + b$$

$$f_{\vec{w},b} = \frac{1}{1 + \mathrm{e}^{-z}}$$

Regularized logistic regression



$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2$$

$$+ w_4 x_1^2 x_2^2 + w_5 x_1^2 x_2^3 + \cdots + b$$

$$f_{\vec{w},b} = \frac{1}{1 + \mathrm{e}^{-z}}$$

Cost function

$$J(\vec{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(f_{\vec{w},b}(\vec{x}^{(i)}) \right) + (1 - y^{(i)}) \log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)}) \right) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2$$

$$\min_{\vec{w},b} J(\vec{w}, b) \quad \rightarrow \quad w_j \downarrow$$

Regularized logistic regression:

$$J(\vec{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log\left(f_{\vec{w},b}(\vec{x}^{(i)}) \right) + (1 - y^{(i)}) \log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)}) \right) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2$$

Gradient descent

Repeat

$$\begin{cases} w_j = w_j - \alpha \dfrac{\partial J}{\partial w_j}(\vec{w}, b) \\ \qquad j = 1, \dots, n \\ b = b - \alpha \dfrac{\partial J}{\partial b}(\vec{w}, b) \end{cases}$$

$$\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} w_j$$

$$j = 1, \dots, n$$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)} \right)$$

Looks same as for linear regression! BUT

$f_{\vec{w},b}(\vec{x}) = g(z(\vec{x}; \vec{w}, b))$ is logistic regression function model