Module 3.3: Learning Parameters: (Infeasible) guess work
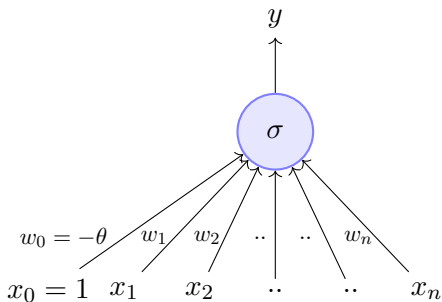
$$y$$

$$\sigma$$

$$w_0 = -\theta \quad w_1 \quad w_2 \quad .. \quad .. \quad w_n$$

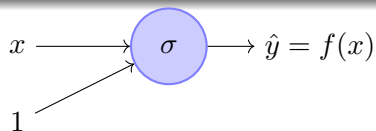$$x_0 = 1 \quad x_1 \quad x_2 \quad .. \quad .. \quad x_n$$

$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**
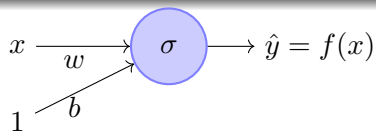
y

σ

$w_0 = -\theta$ w_1 w_2 .. .. w_n

$x_0 = 1$ x_1 x_2 .. .. x_n
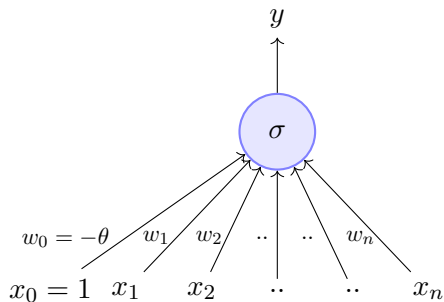
$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**

- σ stands for the sigmoid function (logistic function in this case)
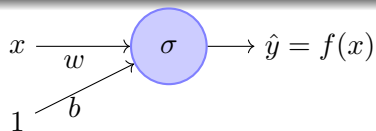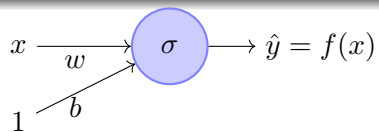
$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**

- σ stands for the sigmoid function (logistic function in this case)

- For ease of explanation, we will consider a very simplified version of the model having just 1 input

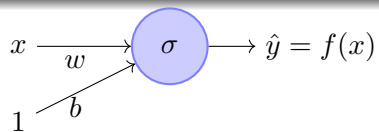$x \xrightarrow{w} \sigma \rightarrow \hat{y} = f(x)$

$1 \xrightarrow{b}$

$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**

- σ stands for the sigmoid function (logistic function in this case)

- For ease of explanation, we will consider a very simplified version of the model having just 1 input

- Further to be consistent with the literature, from now on, we will refer to w_0 as b (bias)

$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

- Keeping this supervised ML setup in mind, we will now focus on this **model** and discuss an **algorithm** for learning the **parameters** of this model from some given **data** using an appropriate **objective function**

- σ stands for the sigmoid function (logistic function in this case)

- For ease of explanation, we will consider a very simplified version of the model having just 1 input

- Further to be consistent with the literature, from now on, we will refer to w_0 as b (bias)

- Lastly, instead of considering the problem of predicting like/dislike, we will assume that we want to predict $criticsRating(y)$ given $imdbRating(x)$ (for no particular reason)

$$x \xrightarrow{w} \sigma \longrightarrow \hat{y} = f(x)$$

$$1 \xrightarrow{b}$$

$$f(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

$$x \xrightarrow{w} \sigma \rightarrow \hat{y} = f(x)$$

$$1 \xrightarrow{b}$$

$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

Training objective

Find w and b such that:
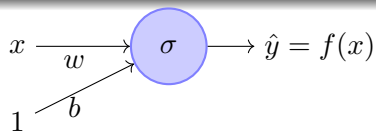
$$\underset{w,b}{\text{minimize}} \, \mathscr{L}(w, b) = \sum_{i=1}^{N}(y_i - f(x_i))^2$$
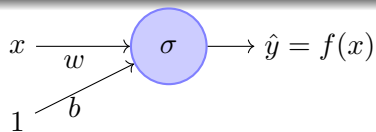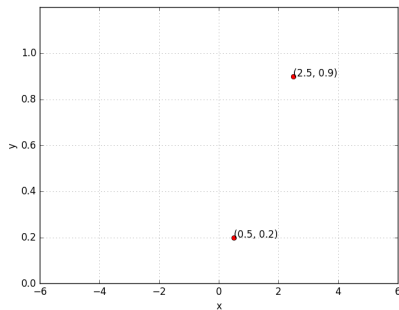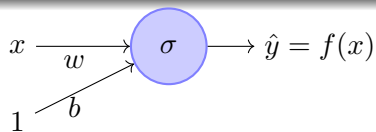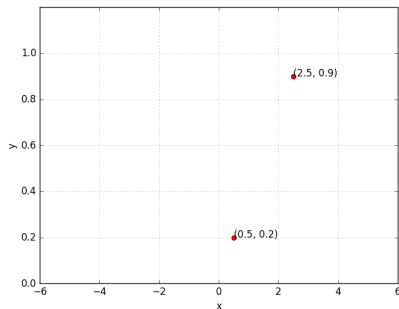
$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

$\{x_i, y_i\}_{i=1}^{N} \to N$ pairs of (x, y)

Training objective

Find w and b such that:

$$\underset{w,b}{\text{minimize}} \, \mathscr{L}(w, b) = \sum_{i=1}^{N}(y_i - f(x_i))^2$$

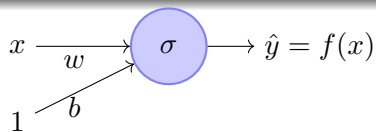$x \xrightarrow{w} \sigma \longrightarrow \hat{y} = f(x)$

$1 \xrightarrow{b}$

$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$

What does it mean to train the network?

- Suppose we train the network with $(x, y) = (0.5, 0.2)$ and $(2.5, 0.9)$

$$f(x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

- Suppose we train the network with $(x, y) = (0.5, 0.2)$ and $(2.5, 0.9)$
- At the end of training we expect to find w*, b* such that:
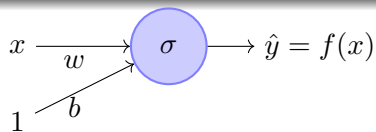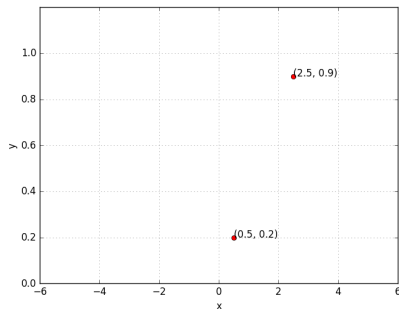
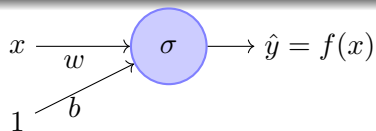$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$
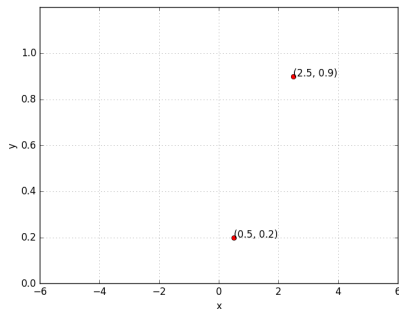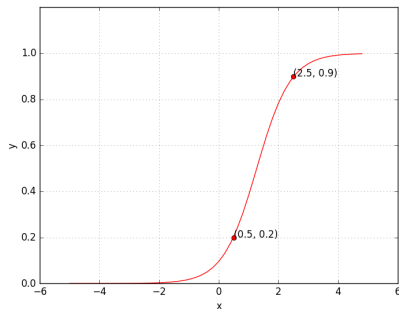

What does it mean to train the network?

- Suppose we train the network with $(x, y) = (0.5, 0.2)$ and $(2.5, 0.9)$
- At the end of training we expect to find w*, b* such that:
- $f(0.5) \to 0.2$ and $f(2.5) \to 0.9$
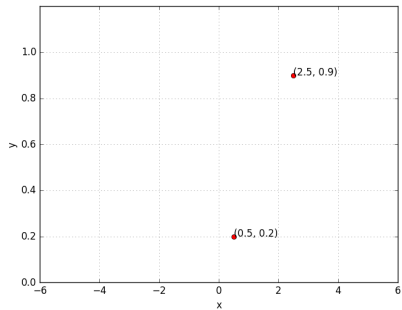
$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

What does it mean to train the network?

- Suppose we train the network with $(x, y) = (0.5, 0.2)$ and $(2.5, 0.9)$
- At the end of training we expect to find w*, b* such that:
- $f(0.5) \to 0.2$ and $f(2.5) \to 0.9$

In other words...

- We hope to find a sigmoid function such that $(0.5, 0.2)$ and $(2.5, 0.9)$ lie on this sigmoid

$$f(x) = \frac{1}{1+e^{-(w \cdot x + b)}}$$

What does it mean to train the network?

- Suppose we train the network with $(x, y) = (0.5, 0.2)$ and $(2.5, 0.9)$
- At the end of training we expect to find w*, b* such that:
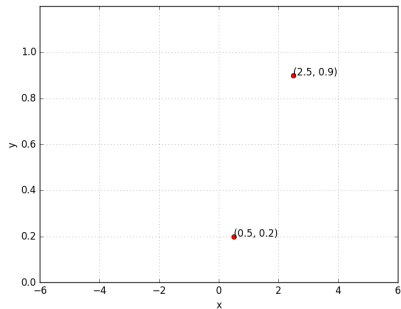- $f(0.5) \rightarrow 0.2$ and $f(2.5) \rightarrow 0.9$

In other words...

- We hope to find a sigmoid function such that $(0.5, 0.2)$ and $(2.5, 0.9)$ lie on this sigmoid
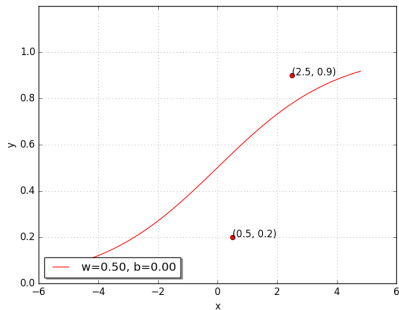
Let us see this in more detail....

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

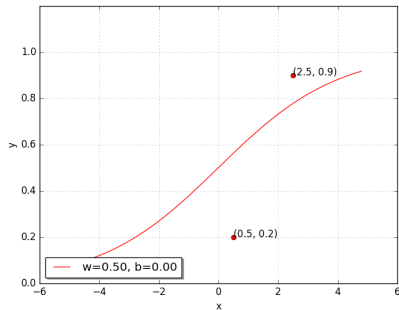- Can we try to find such a w^*, b^* manually

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

- Can we try to find such a w^*, b^* manually
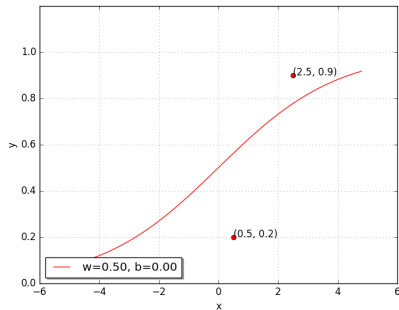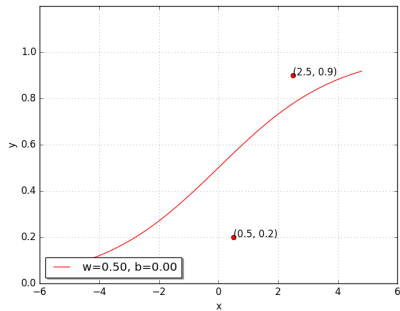- Let us try a random guess.. (say, $w = 0.5, b = 0$)

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

- Can we try to find such a w^*, b^* manually
- Let us try a random guess.. (say, $w = 0.5, b = 0$)
- Clearly not good, but how bad is it ?

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

- Can we try to find such a w^*, b^* manually
- Let us try a random guess.. (say, $w = 0.5, b = 0$)
- Clearly not good, but how bad is it ?
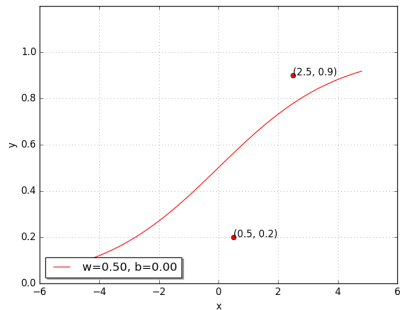- Let us revisit $\mathscr{L}(w, b)$ to see how bad it is ...
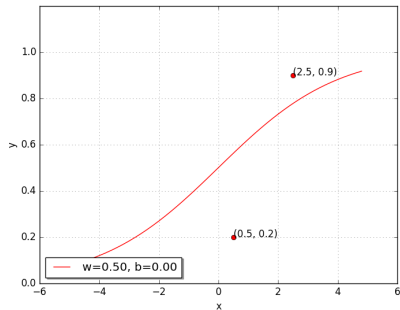
$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

$$\mathscr{L}(w, b) = \frac{1}{2} * \sum_{i=1}^{N}(y_i - f(x_i))^2$$
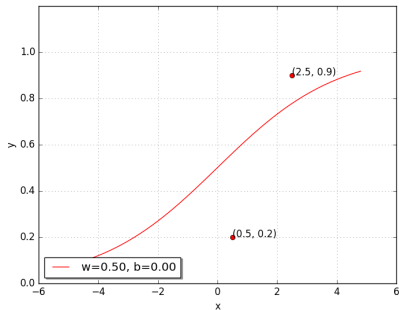
$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

$$\mathscr{L}(w,b) = \frac{1}{2} * \sum_{i=1}^{N}(y_i - f(x_i))^2$$
$$= \frac{1}{2} * (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2$$
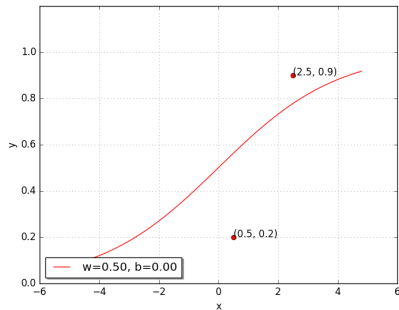
$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

$$\mathscr{L}(w, b) = \frac{1}{2} * \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \frac{1}{2} * (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2$$

$$= \frac{1}{2} * (0.9 - f(2.5))^2 + (0.2 - f(0.5))^2$$

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

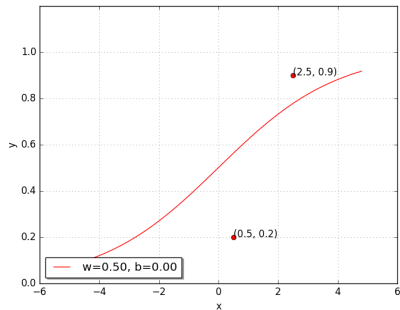Mitesh M. Khapra   CS7015 (Deep Learning) : Lecture 3

$$\mathscr{L}(w, b) = \frac{1}{2} * \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \frac{1}{2} * (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2$$

$$= \frac{1}{2} * (0.9 - f(2.5))^2 + (0.2 - f(0.5))^2$$

$$= 0.073$$

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Mitesh M. Khapra   CS7015 (Deep Learning) : Lecture 3

$$\mathscr{L}(w, b) = \frac{1}{2} * \sum_{i=1}^{N} (y_i - f(x_i))^2$$
$$= \frac{1}{2} * (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2$$
$$= \frac{1}{2} * (0.9 - f(2.5))^2 + (0.2 - f(0.5))^2$$
$$= 0.073$$

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

We want $\mathscr{L}(w, b)$ to be as close to 0 as possible

Mitesh M. Khapra   CS7015 (Deep Learning) : Lecture 3

Let us try some other values of w, b

w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Let us try some other values of w, b
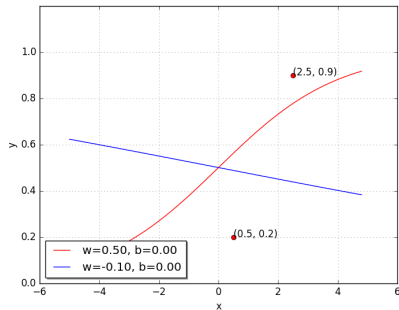
w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
-0.10	0.00	0.1481

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Oops!! this made things even worse...
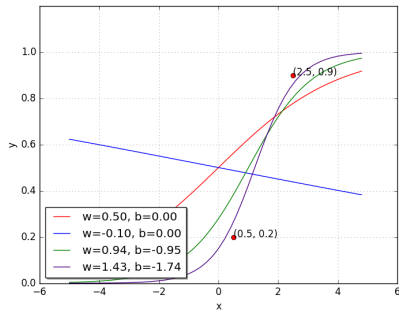
Let us try some other values of w, b
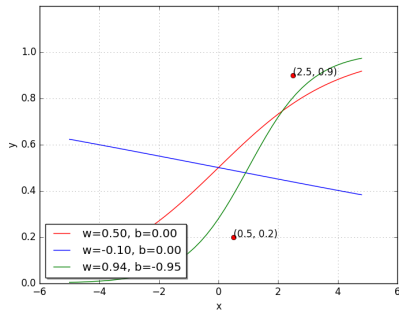
w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
-0.10	0.00	0.1481
0.94	-0.94	0.0214

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Perhaps it would help to push w and b in the other direction...

Let us try some other values of w, b

w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
-0.10	0.00	0.1481
0.94	-0.94	0.0214
1.42	-1.73	0.0028

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Let us keep going in this direction, *i.e.*, increase w and decrease b

Let us try some other values of w, b
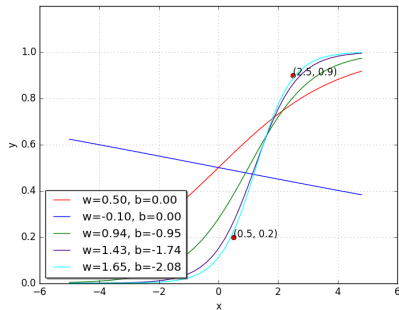
w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
-0.10	0.00	0.1481
0.94	-0.94	0.0214
1.42	-1.73	0.0028
1.65	-2.08	0.0003

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Let us keep going in this direction, *i.e.*, increase w and decrease b

Let us try some other values of w, b
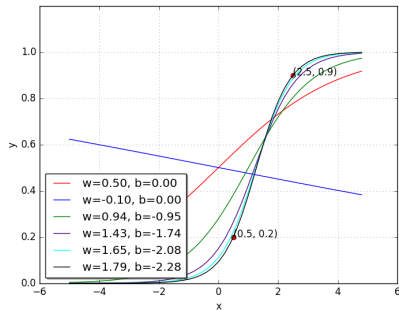
w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
-0.10	0.00	0.1481
0.94	-0.94	0.0214
1.42	-1.73	0.0028
1.65	-2.08	0.0003
1.78	-2.27	0.0000

$$\sigma(x) = \frac{1}{1 + e^{-(wx+b)}}$$

With some guess work and intuition we were able to find the right values for w and b

Let us look at something better than our "guess work" algorithm....

- Since we have only 2 points and 2 parameters (w, b) we can easily plot $\mathscr{L}(w, b)$ for different values of (w, b) and pick the one where $\mathscr{L}(w, b)$ is minimum
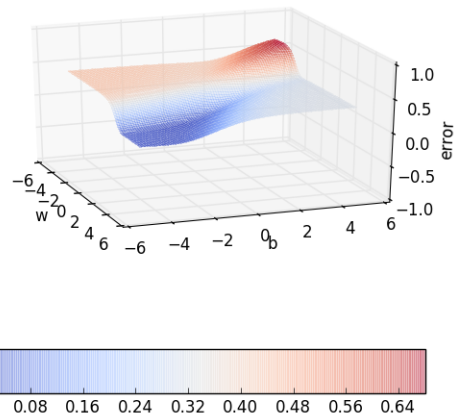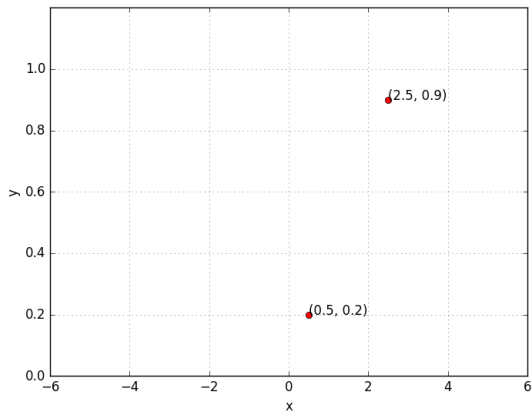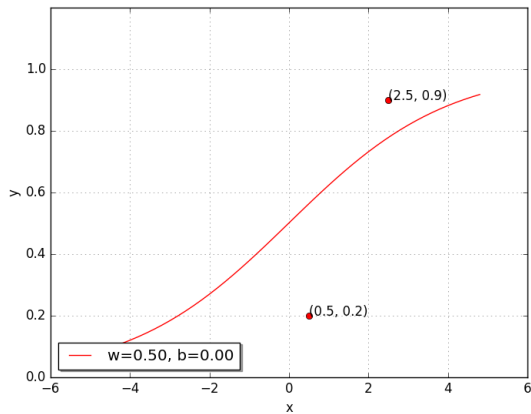
Random search on error surface

- Since we have only 2 points and 2 parameters (w, b) we can easily plot $\mathscr{L}(w, b)$ for different values of (w, b) and pick the one where $\mathscr{L}(w, b)$ is minimum
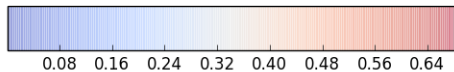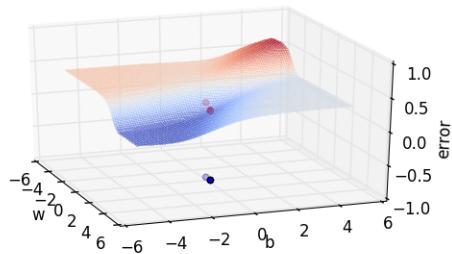
Random search on error surface
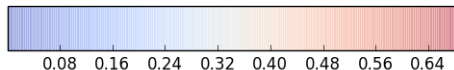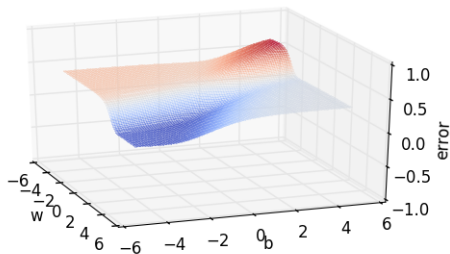
- Since we have only 2 points and 2 parameters (w, b) we can easily plot $\mathscr{L}(w, b)$ for different values of (w, b) and pick the one where $\mathscr{L}(w, b)$ is minimum

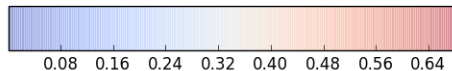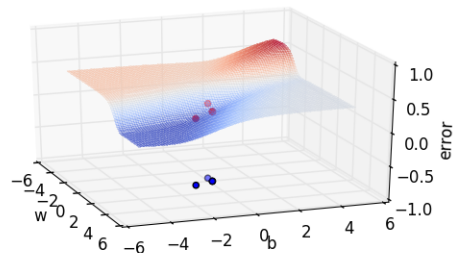- But of course this becomes intractable once you have many more data points and many more parameters !!

Random search on error surface
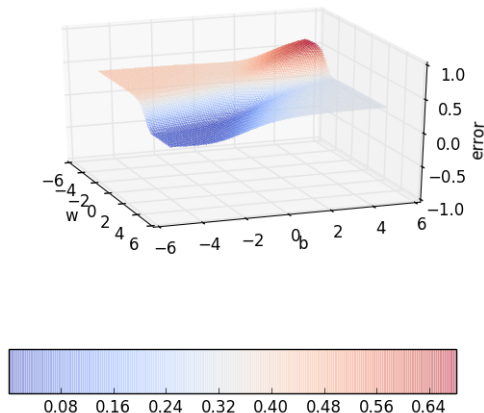
- Since we have only 2 points and 2 parameters (w, b) we can easily plot $\mathscr{L}(w, b)$ for different values of (w, b) and pick the one where $\mathscr{L}(w, b)$ is minimum
- But of course this becomes intractable once you have many more data points and many more parameters !!
- Further, even here we have plotted the error surface only for a small range of (w, b) [from $(-6, 6)$ and not from $(-\inf, \inf)$]

*Let us look at the geometric interpretation of our
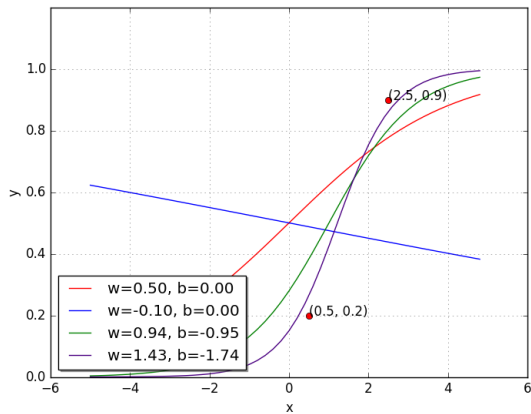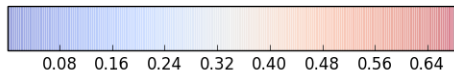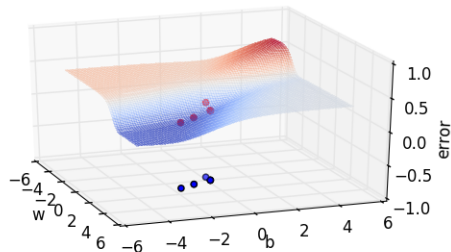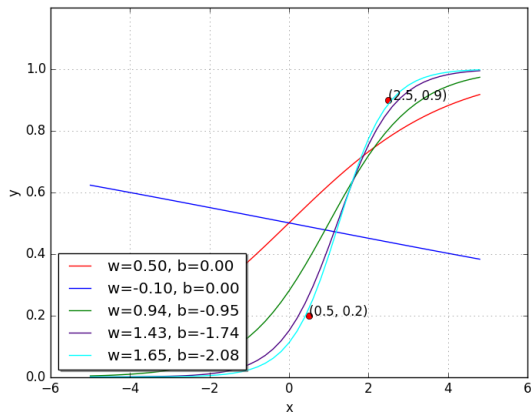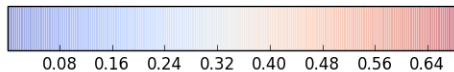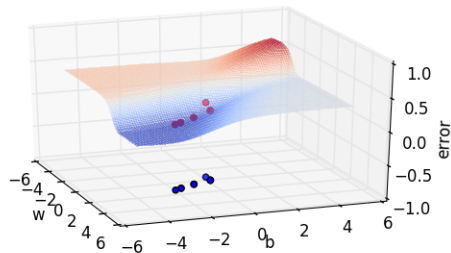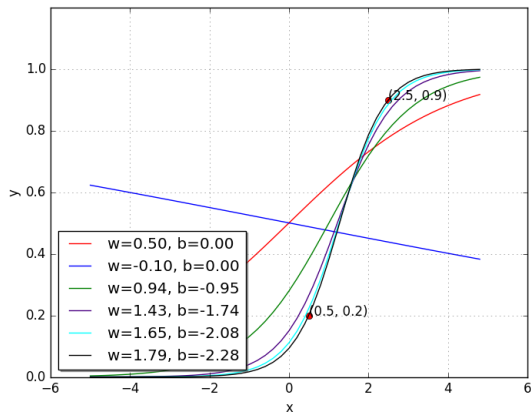"guess work" algorithm in terms of this error surface*

Random search on error surface

Random search on error surface

Random search on error surface

Random search on error surface

Random search on error surface

Random search on error surface

Random search on error surface