ECE 515
Information Theory

Logistic Regression

Cross Entropy

- The **cross entropy** between the probability distributions p(X) and q(X) is defined as

$$\mathrm{H}(\mathrm{p},\mathrm{q}) = \mathrm{H}(\mathrm{p}(X))+\mathrm{D}(\mathrm{p}(X)||\mathrm{q}(X))$$

$$\mathrm{H}(\mathrm{p},\mathrm{q}) = \mathrm{E}_{\mathrm{p}}[-\log(\mathrm{q}(X)]$$

$$\mathrm{H}(\mathrm{p},\mathrm{q}) = -\sum_{i=1}^{n} \mathrm{p}(y_i)\log \mathrm{q}(y_i)$$

Linear Regression

- Training data: $(x_i, y_i), i = 1,2, \cdots, n$
- Model: $\hat{y} = wx + b$
- Loss function: Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

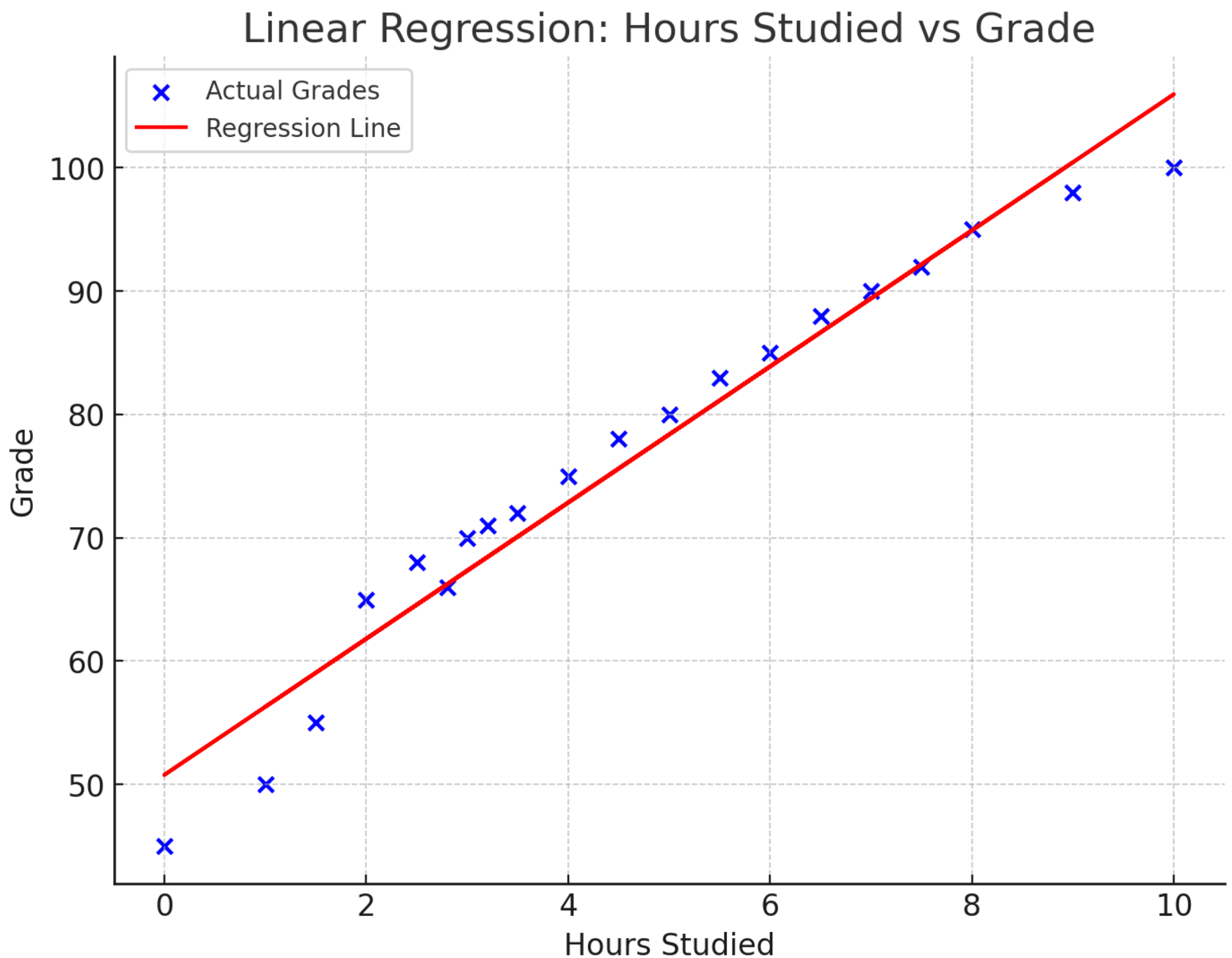
- Optimization function:

$$\min_{w,b}\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \min_{w,b}\frac{1}{n}\sum_{i=1}^{n}\big(y_i - (wx_i + b)\big)^2$$

ECE 515 Student Data

Hours Studied (x)	Grade (*y*)	Hours Studied (x)	Grade (*y*)
2	65	10	100
3	70	3.5	72
5	80	2.5	68
1	50	4.5	78
4	75	6.5	88
6	85	1.5	55
8	95	2.8	66
7	90	3.2	71
9	98	7.5	92
0	45	5.5	83



Grade: 50.78+5.52×(Hours Studied)

Intercept: b=50.78

Slope: w=5.52

Logistic Regression

- Important analytical tool in natural and social sciences
- Key supervised machine learning tool for classification
- The foundation of neural networks

Binary Outcomes are Common and Important

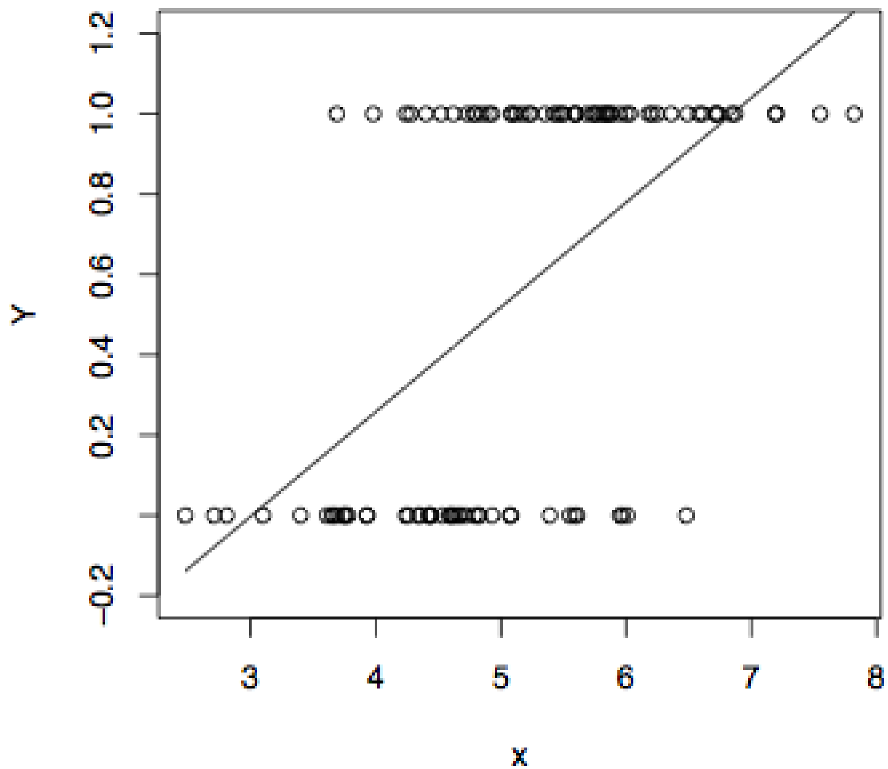
- The patient survives the operation or does not
- The accused is convicted or is not
- The customer makes a purchase or does not
- The marriage lasts at least five years or does not
- The student passes the exam or does not

ECE 515 Student Data

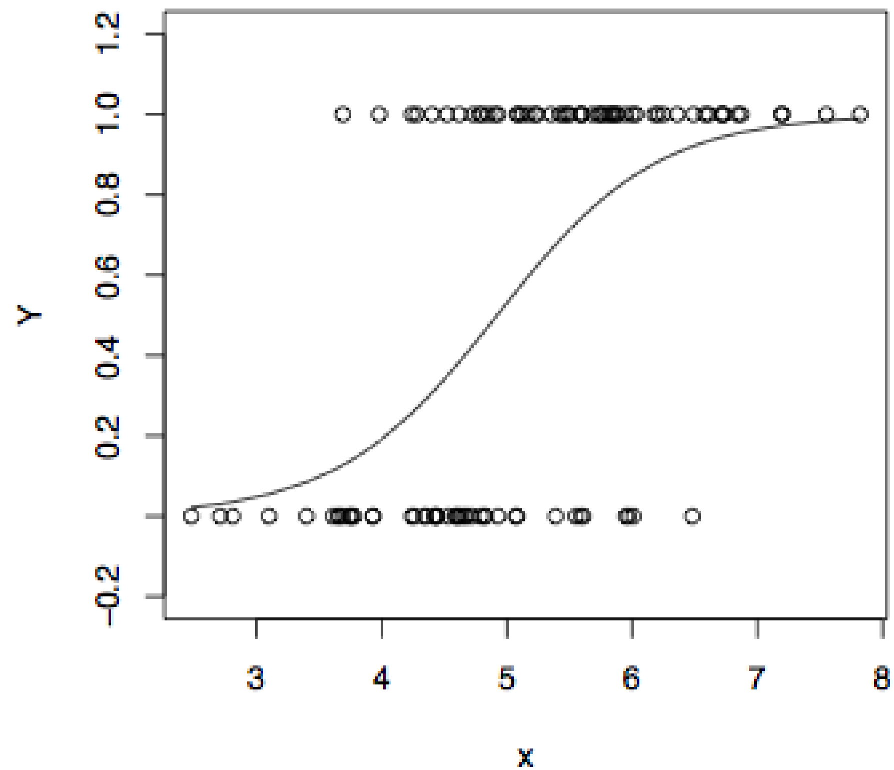
Hours Studied (x)	Passed Exam (*y*)	Hours Studied (*x*)	Passed Exam (*y*)
1	0	16	0
2	0	17	1
3	0	18	1
4	0	19	1
5	0	20	1
6	0	21	0
7	1	22	1
8	0	23	1
9	1	24	1
10	1	25	1
11	1	26	1
12	1	27	1
13	1	28	1
14	0	29	1
15	1	30	1

Linear Regression vs. Logistic Regression

Linear Regression Curve



Logistic Regression Curve



Logistic Regression

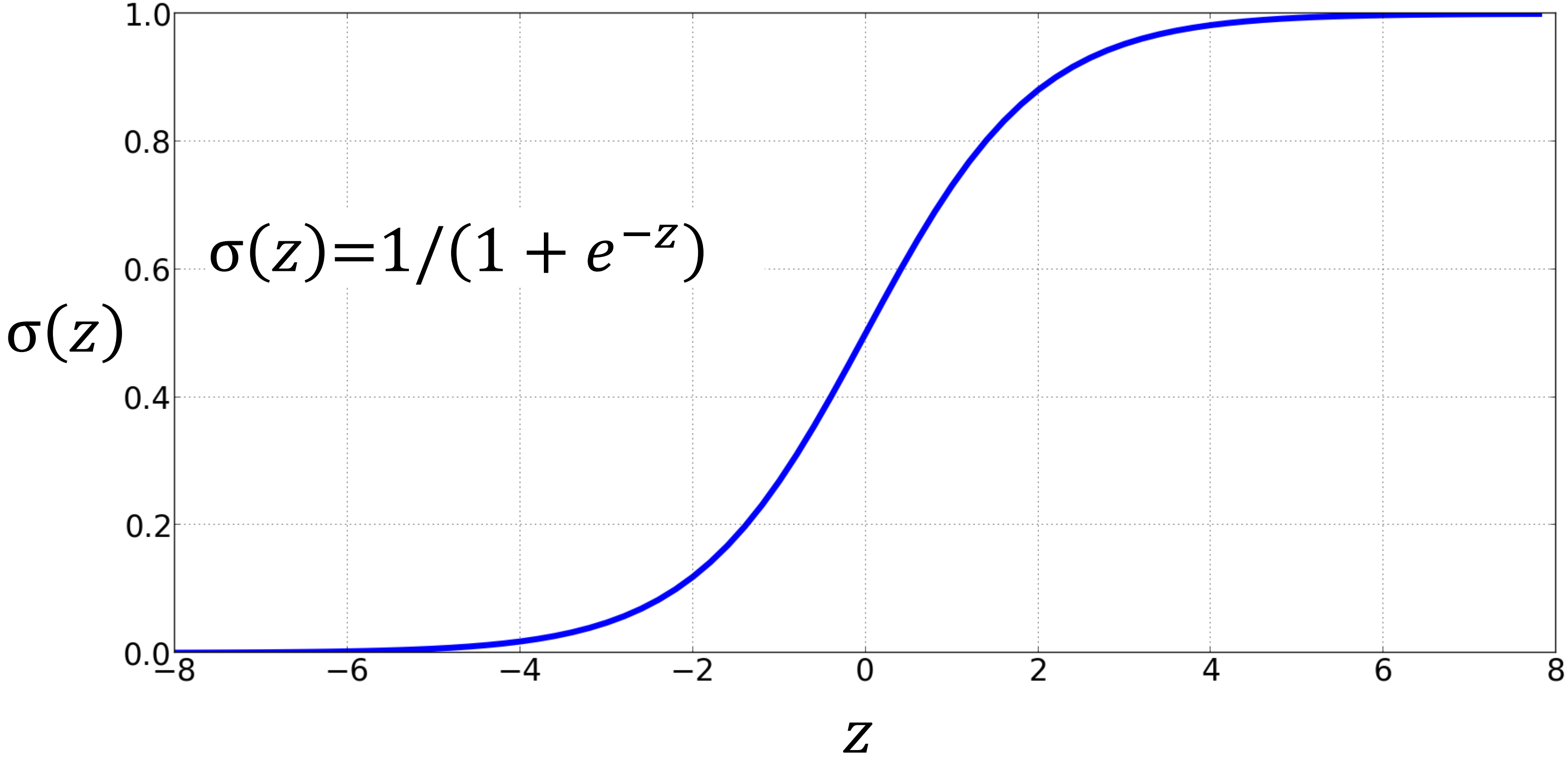
- Output is binary

  y=1 (pass) or y=0 (fail)
- A function is required that is restricted to [0,1]
- A common choice is the Logistic (Sigmoid) function

$$\sigma(z) = \frac{1}{(1+e^{-z})}$$

- Can be treated as a probability

The Logistic (Sigmoid) Function

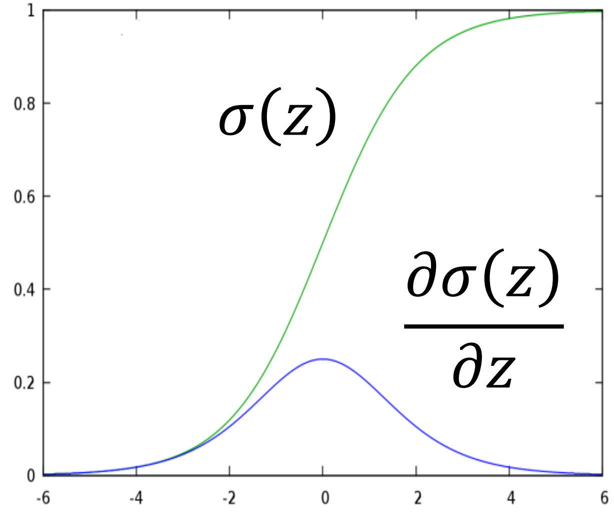


Logistic Function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$1 - \sigma(z) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{z}} = \sigma(-z)$$

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z))$$



Idea of Logistic Regression

- Compute $z = wx + b$
- Pass it through the logistic function

$$\hat{y} = \sigma(z) = \sigma(wx + b)$$

- Then treat it as a probability

$$\mathrm{p}(y = 1|x) = \sigma(z) = \frac{1}{1+e^{-(wx+b)}}$$

$$\mathrm{p}(y = 0|x) = 1 - \sigma(z) = 1 - \mathrm{p}(y = 1|x)$$

- Combining these gives

$$\begin{aligned}\mathrm{p}(y|x) &= \mathrm{p}(y = 1|x)^{y} \times \mathrm{p}(y = 0|x)^{1-y} \\ &= \mathrm{p}(y = 1|x)^{y} \times (1 - \mathrm{p}(y = 1|x))^{1-y}\end{aligned}$$

Loss Function

- Given *w* and *b*, the probability of generating the training data is the likelihood function

$$L = \prod_{i=1}^{n} p(y_i|x_i)$$

- We want to find the parameters *w* and *b* that maximize *L*

$$\underset{w,b}{\operatorname{argmax}} L = \underset{w,b}{\operatorname{argmax}} \prod_{i=1}^{n} \mathrm{p}(y_i|x_i)$$

- For simplicity, the log is typically used

$$\log L = \sum_{i=1}^{n} \log \mathrm{p}(y_i|x_i)$$

- This is called the log-likelihood function

Loss Function

- Maximizing the log-likelihood is the same as minimizing the negative log-likelihood

$$\min_{w,b} -\log L = \min_{w,b} -\sum_{i=1}^{n} \log \mathrm{p}(y_i|x_i)$$

- Substituting the probability for logistic regression gives

$$-\sum_{i=1}^{n} \log \mathrm{p}(y_i|x_i)$$

$$= -\sum_{i=1}^{n} y_i \log \mathrm{p}(y_i|x_i) + (1 - y_i)\log\big(1 - \mathrm{p}(y_i|x_i)\big)$$

Cross Entropy

$$-\sum_{i=1}^{n} y_i \log \mathrm{p}(y_i|x_i) + (1 - y_i)\log\big(1 - \mathrm{p}(y_i|x_i)\big)$$

$$\mathrm{H}(\mathrm{p}, \mathrm{q}) = -\sum_{i=1}^{n} \mathrm{p}(y_i) \log \mathrm{q}(y_i)$$

Distribution p(Y):

$$\mathrm{p}(y = 1) = y$$
$$\mathrm{p}(y = 0) = 1 - y$$

Distribution q(Y):

$$\mathrm{q}(y = 1) = \sigma(\mathrm{z})$$
$$\mathrm{q}(y = 0) = 1 - \sigma(\mathrm{z})$$

Optimization Function

$$\min_{w,b} - \sum_{i=1}^{n} y_i \log \mathrm{p}(y_i|x_i) + (1 - y_i) \log\big(1 - \mathrm{p}(y_i|x_i)\big)$$

- Because the loss function is convex, a simple optimization algorithm such as gradient descent can be used

Example

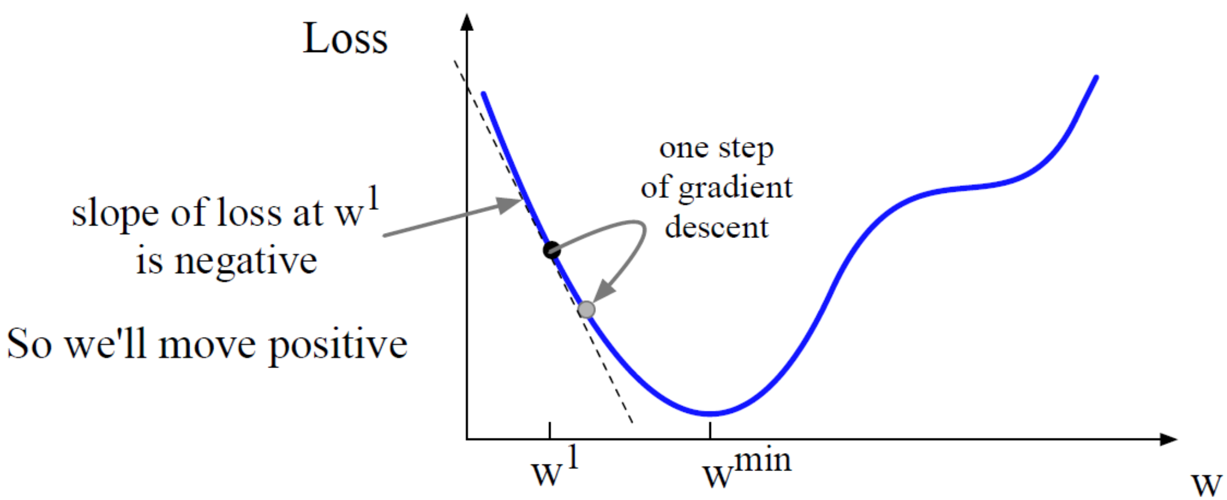
- Passing ECE 515 Final Exam
- Feature: Hours of Study
- Output: pass or fail
- $\mathrm{p(pass)} = \frac{1}{1+e^{-(wx+b)}}$
- $\mathrm{p(fail)} = 1 - \mathrm{p(pass)} = 1 - \frac{1}{1+e^{-(wx+b)}}$
- *x* is the number of hours studied
- *w* is the weight (slope)
- *b* is the bias (intercept)

ECE 515 Student Data

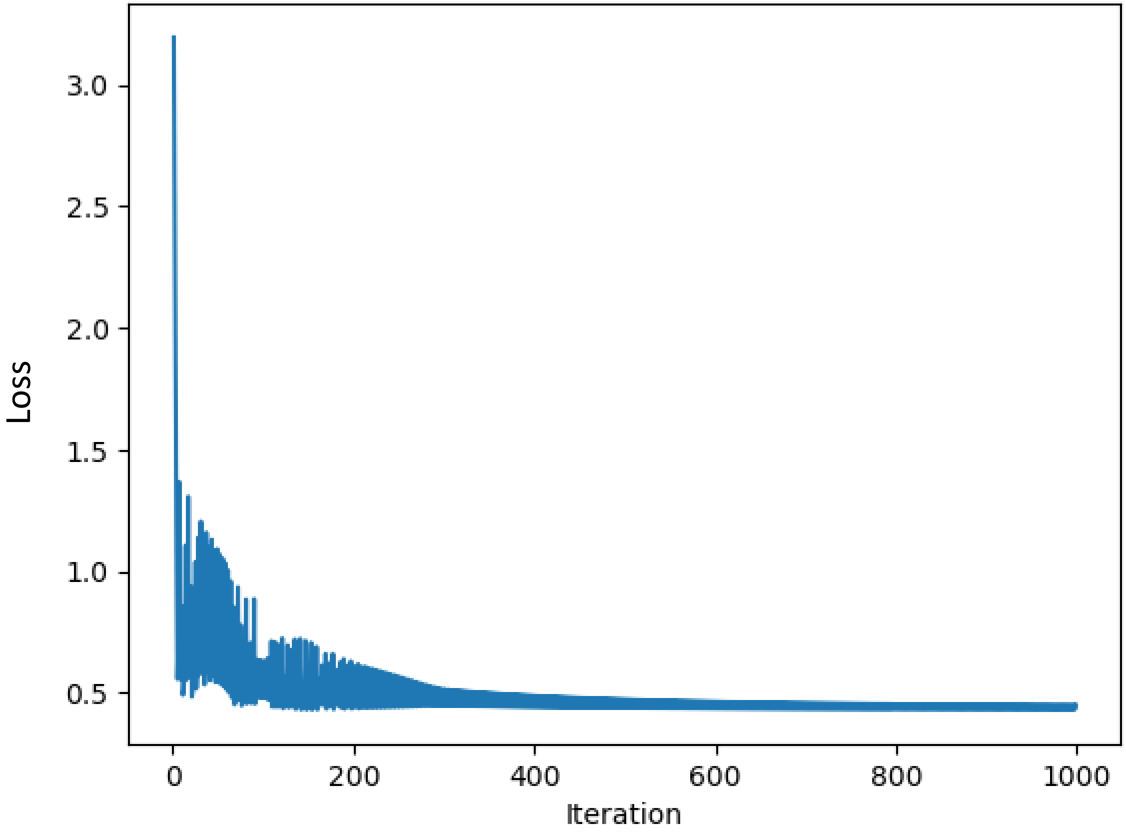
Hours Studied (x)	Passed Exam (*y*)	Hours Studied (*x*)	Passed Exam (*y*)
1	0	16	0
2	0	17	1
3	0	18	1
4	0	19	1
5	0	20	1
6	0	21	0
7	1	22	1
8	0	23	1
9	1	24	1
10	1	25	1
11	1	26	1
12	1	27	1
13	1	28	1
14	0	29	1
15	1	30	1

Gradients

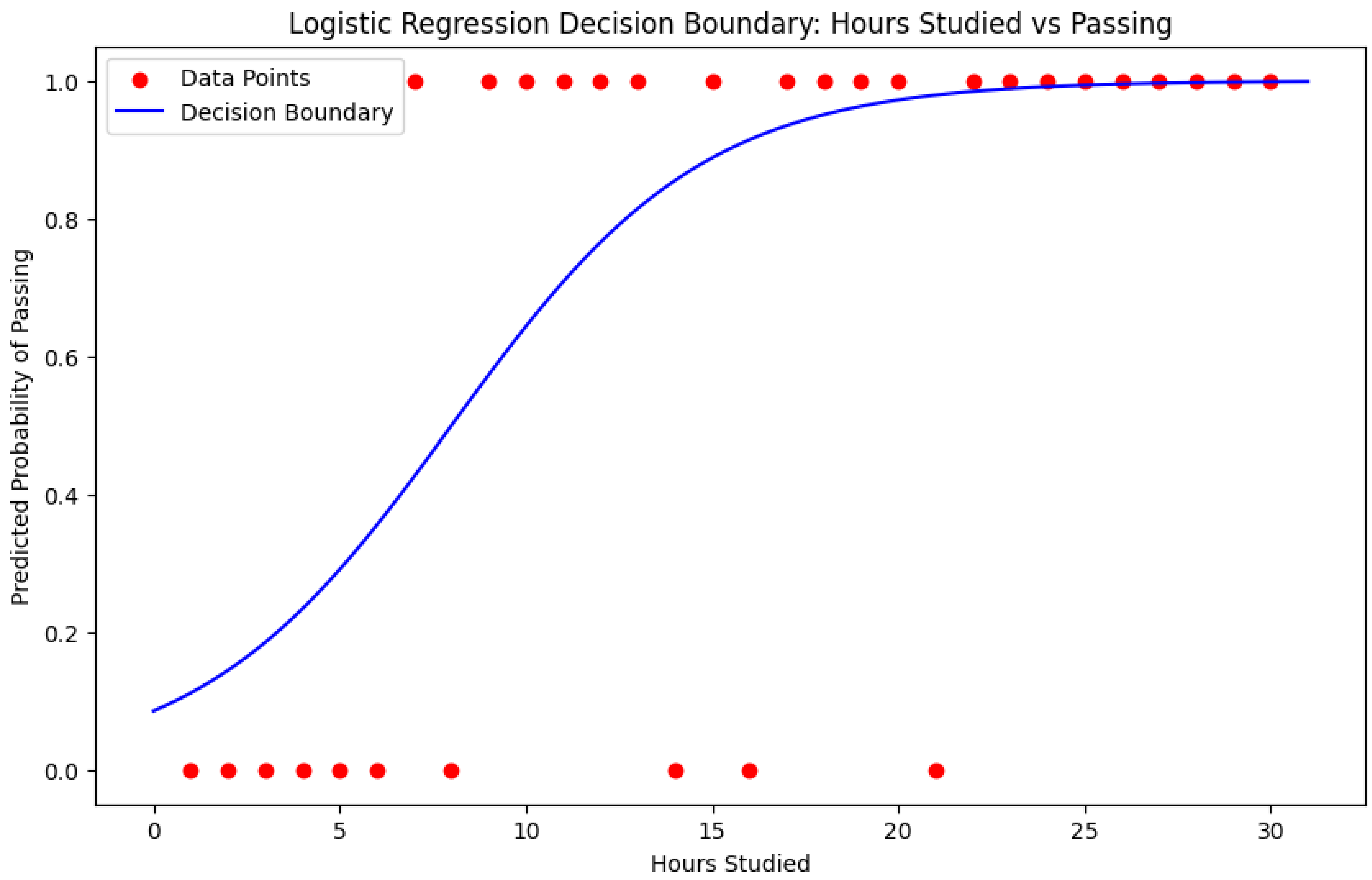
- The **gradient** of a function of many variables is a vector pointing in the direction of the greatest increase in a function.
- **Gradient Descent**: Find the gradient of the loss function at the current point and move in the **opposite** direction.
- For a scalar *w*



Loss Function



Optimal weight *w* = 0.296
Optimal bias *b* = -2.373


Logistic Regression Decision Boundary: Hours Studied vs Passing
Data Points
Decision Boundary
Predicted Probability of Passing
Hours Studied
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

Logistic Regression

- Training data: $(x_i, y_i), i = 1,2,\cdots,n$
- Model: $\hat{y} = \sigma(z) = \sigma(wx + b)$
- Loss function: cross entropy

$$-\sum_{i=1}^{n} y_i \log \mathrm{p}(y_i|x_i) + (1 - y_i)\log\big(1 - \mathrm{p}(y_i|x_i)\big)$$

- Optimization function:

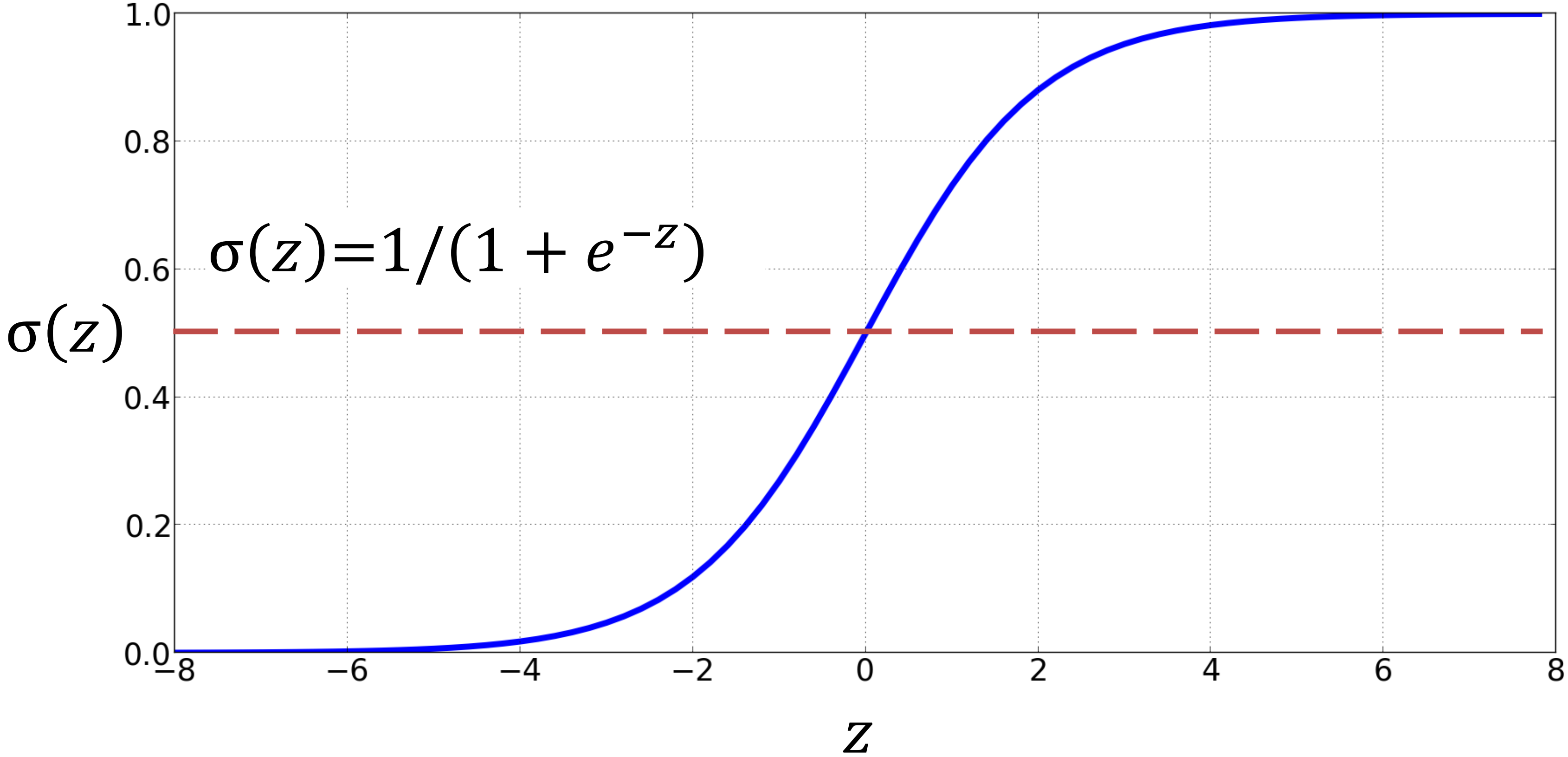
$$\min_{w,b} -\sum_{i=1}^{n} y_i \log \mathrm{p}(y_i|x_i) + (1 - y_i)\log\big(1 - \mathrm{p}(y_i|x_i)\big)$$

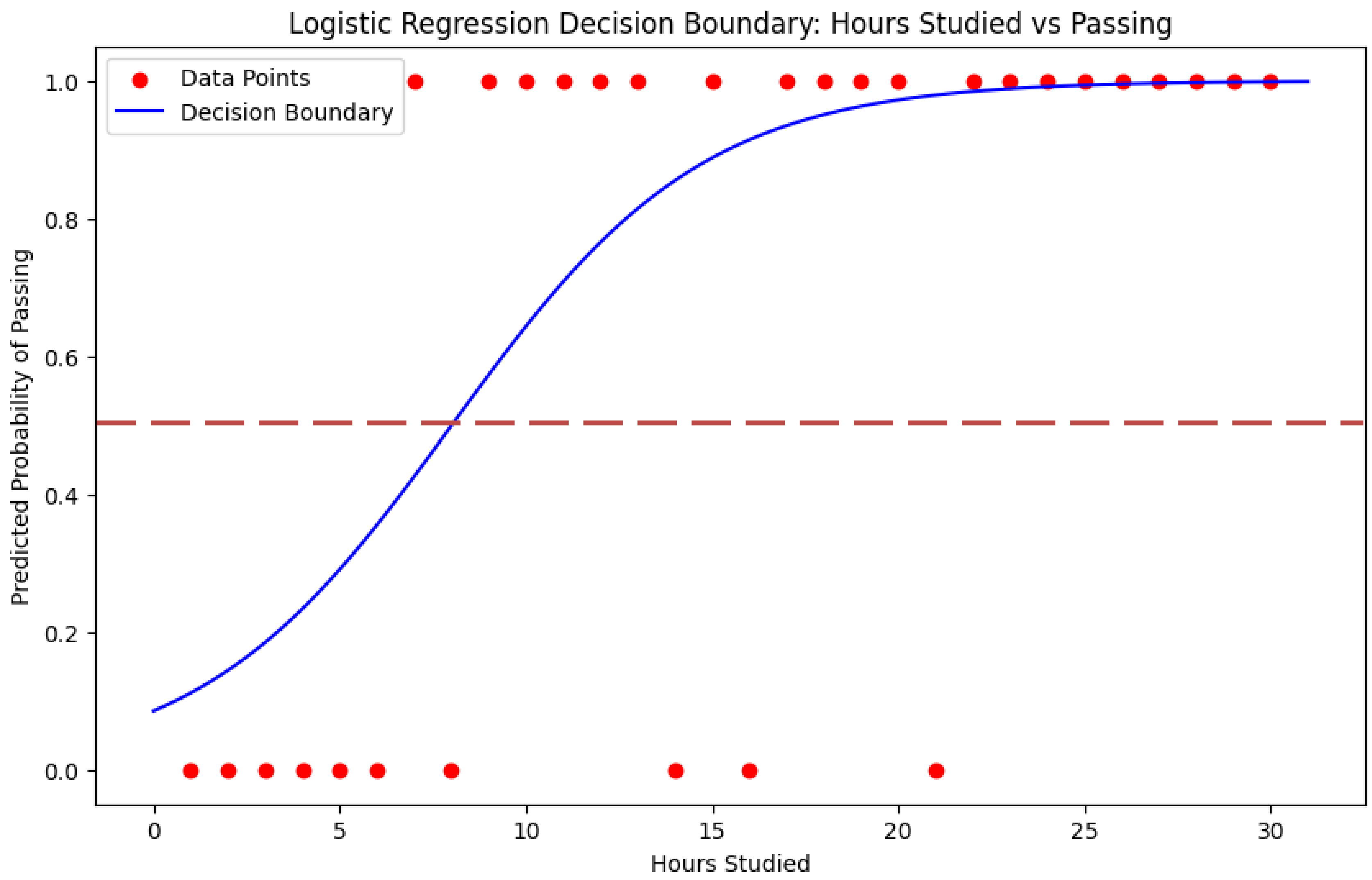
Turning a Probability Into a Classifier

$$\hat{y} = \begin{cases} 1 \text{ if } \mathrm{p}(y = 1|x) > 0.5 \\ 0 \text{ otherwise} \end{cases}$$

0.5 is called the decision boundary

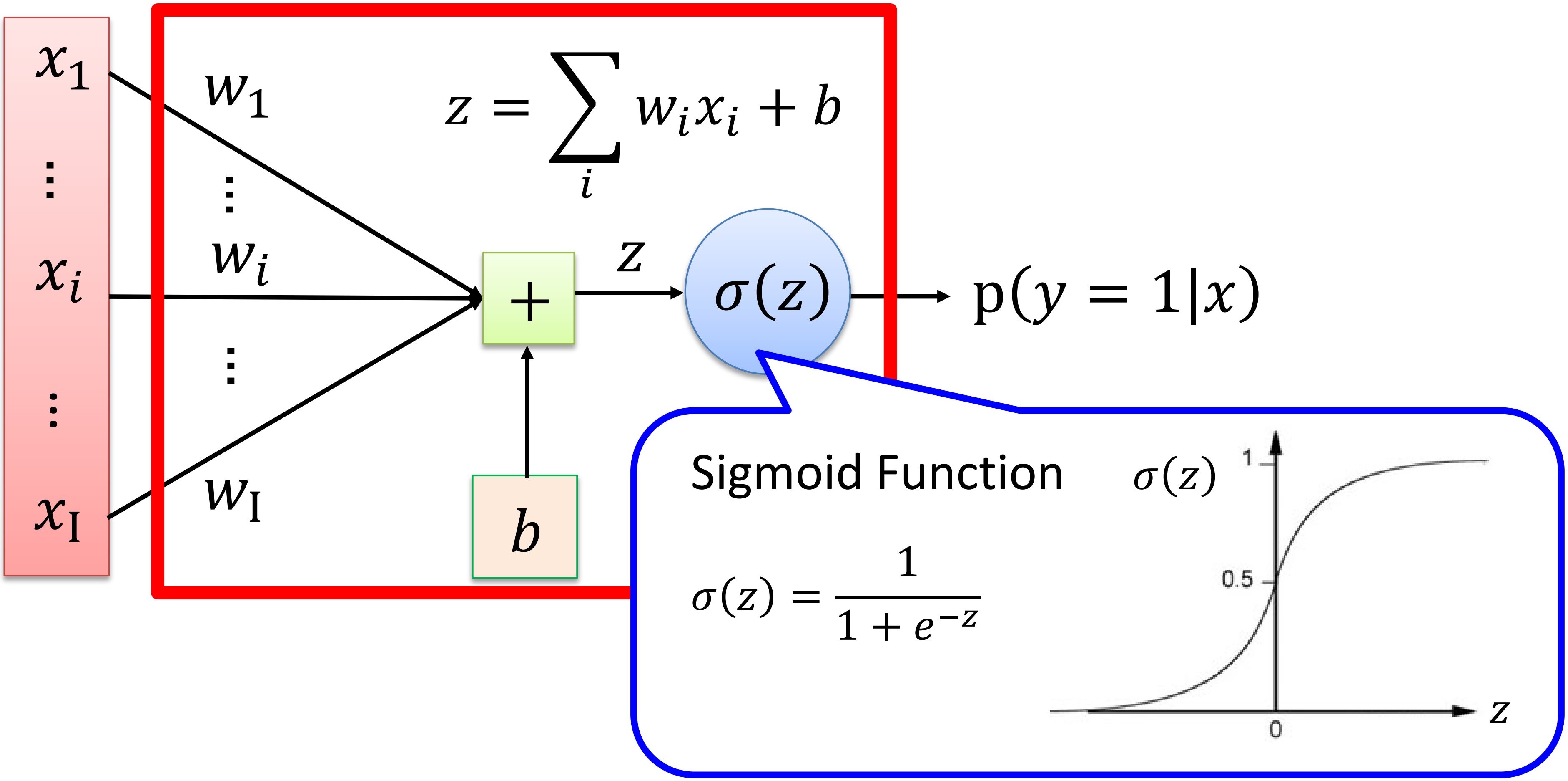
Probabilistic Classifier





Logistic Regression Decision Boundary: Hours Studied vs Passing
Data Points
Decision Boundary
Predicted Probability of Passing
Hours Studied
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30

Neuron Structure



MSE Loss

- Left: linear regression MSE loss
- Right: logistic regression MSE loss