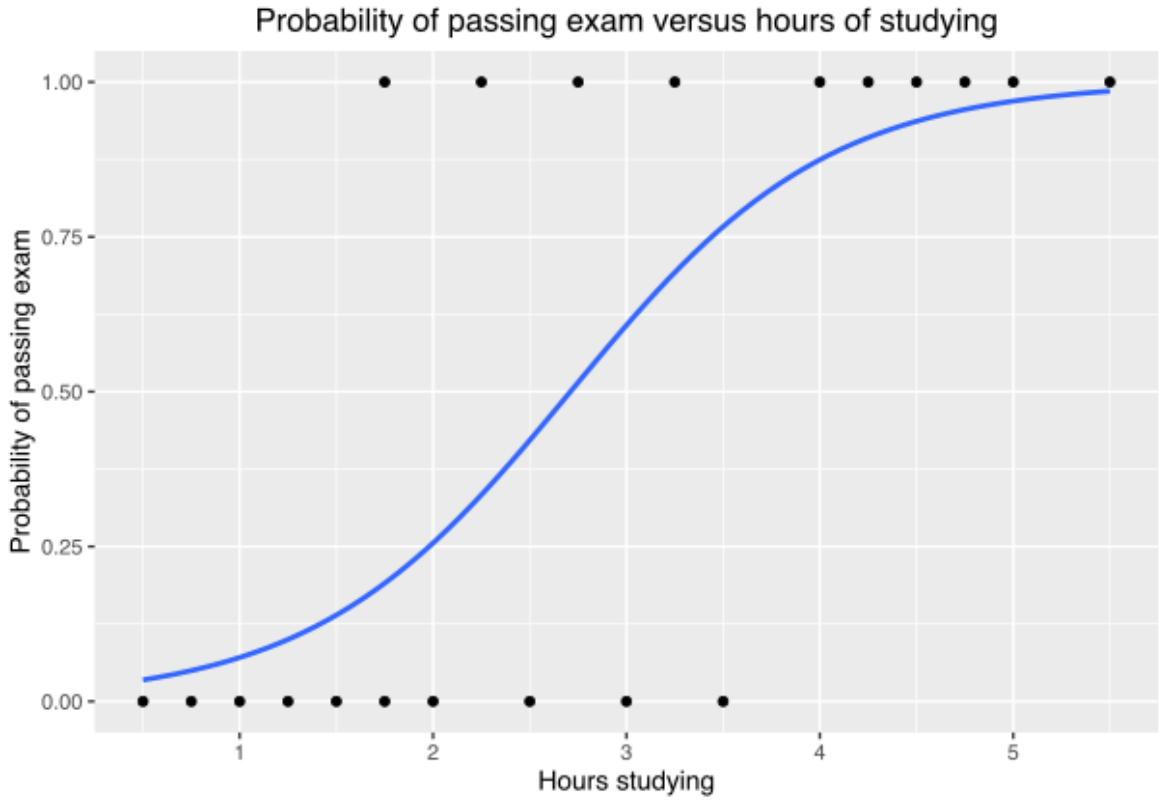
CS440/ECE448 Lecture 8: Logistic Regression

Mark Hasegawa-Johnson, 2/2025



CC-SA 4.0, https://commons.wikimedia.org/wiki/File:Exam_pass_logistic_curve.svg

Outline

- Linear regression: Review
- Logistic regression: Output is a probability
- Derivative of the sigmoid
- Derivative of the log sigmoid

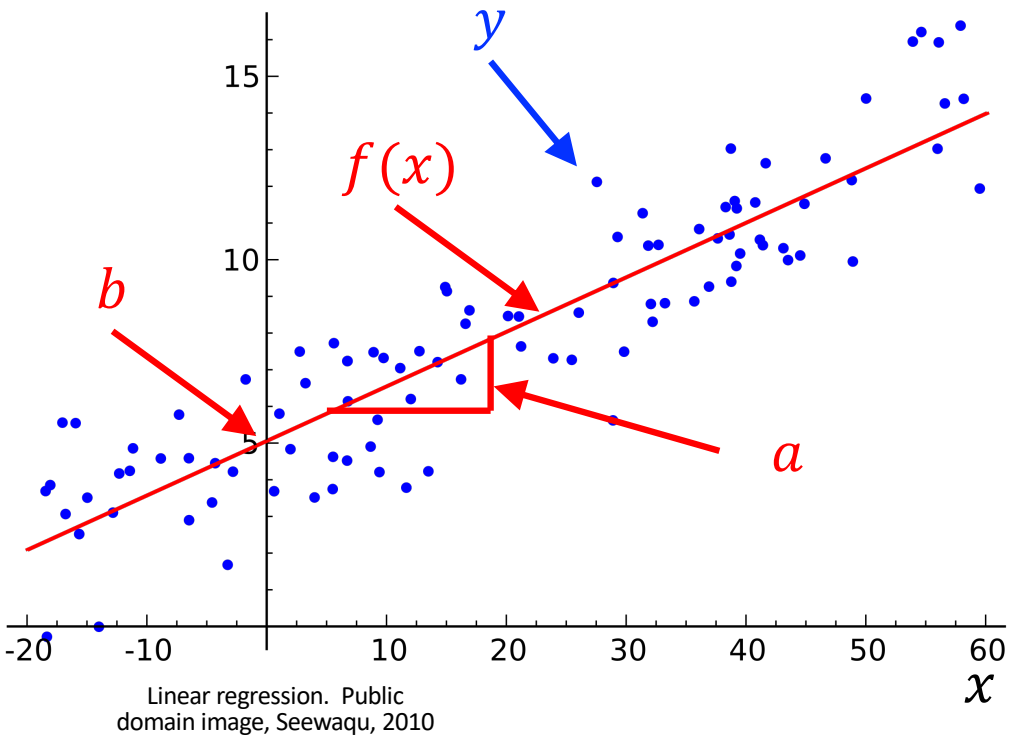
Linear regression

Linear regression is used to estimate a real-valued target variable, y, using a linear function of another variable, x:

$$f(x) = ax + b$$

… so that …

$$f(x) \approx y$$



Linear regression. Public domain image, Seewaqu, 2010

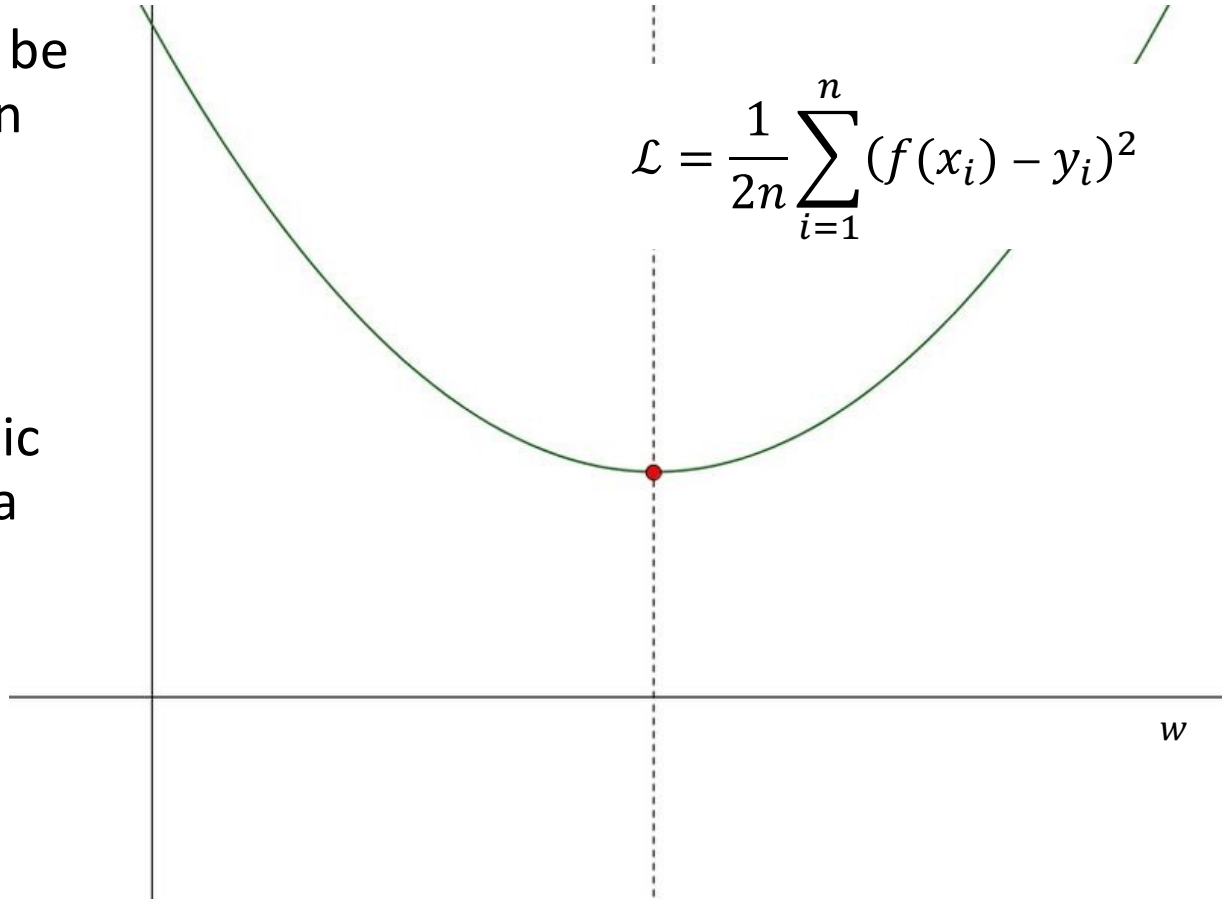
MSE = Parabola

A good closed-form solution can be achieved by minimizing the mean squared error,

$$\mathcal{L} = \frac{1}{2n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

Since it's a non-negative quadratic function of $\boldsymbol{w} = [a \quad b]^T$, it has a unique minimum, given by the closed-form solution

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$



Gradient descent and SGD

Often, closed-form solution is too computationally expensive. In those situations, we choose a random initial guess for the value of $\boldsymbol{w}$, and then improve it using either gradient descent:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}}, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i$$

Or stochastic gradient descent:

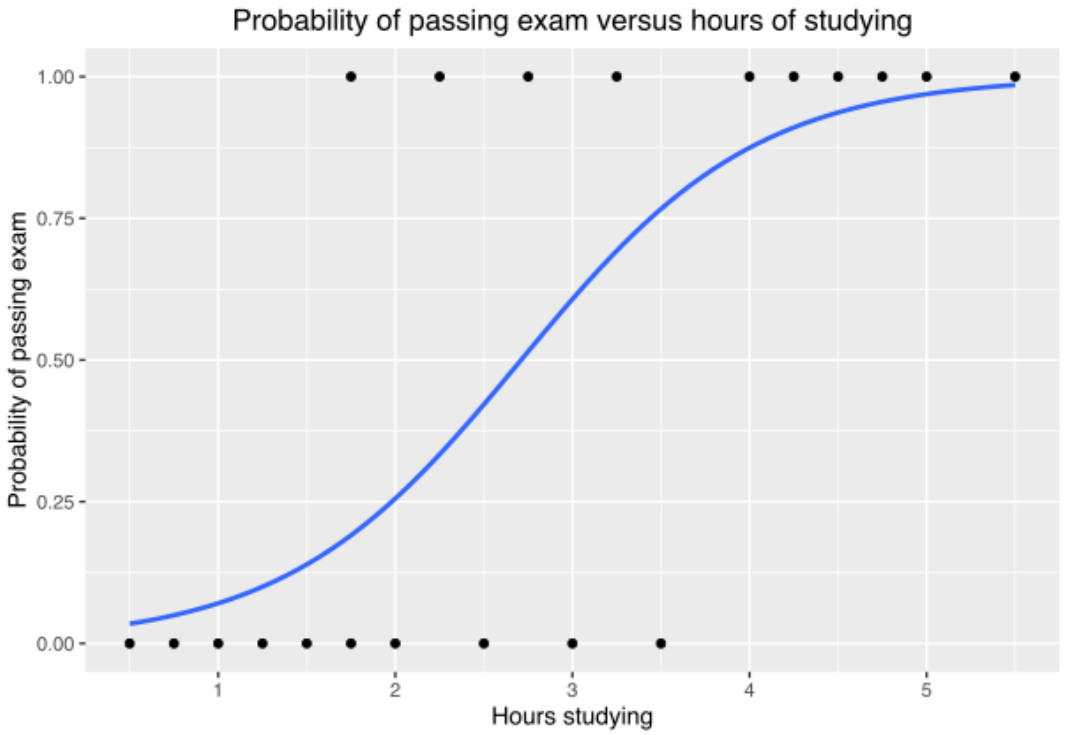
$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}}, \qquad \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}} = \epsilon_i \boldsymbol{x}_i$$

Outline

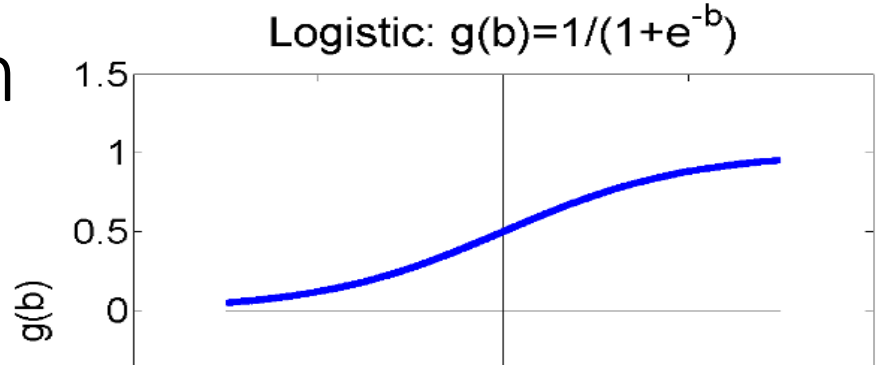
- Linear regression: Review
- Logistic regression: Output is a probability
- Derivative of the sigmoid
- Derivative of the log sigmoid

Logistic regression: Output is a probability

- Logistic regression was invented by psychologists in the early 20th century
- They wanted to model binary outcomes, like "Did student i pass or fail the test?" In other words, every output is either $y_i = 1$ or $y_i = 0$
- Instead of modeling $y_i = \boldsymbol{x}_i^T \boldsymbol{w}$ as a real number, it makes more sense to try to model $P(y_i = 1 | \boldsymbol{x}_i)$ as some kind of function of $\boldsymbol{x}_i$.



The logistic sigmoid function



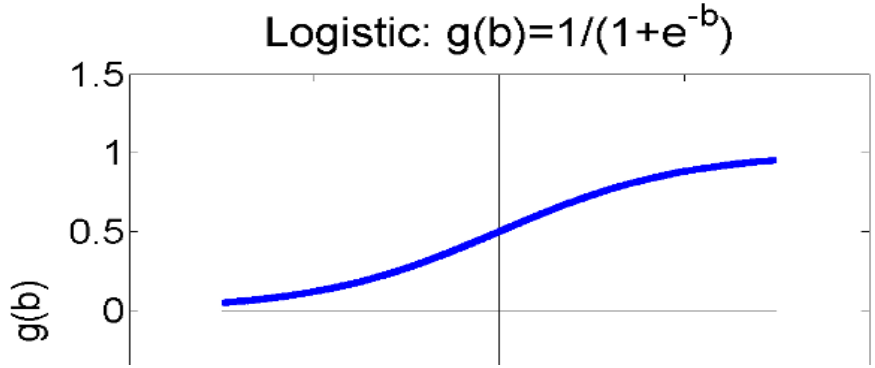
- To model $P(y_i = 1|\boldsymbol{x}_i)$ as some kind of function of $\boldsymbol{x}_i$, we need some kind of nonlinear function that squashes the linear output $\boldsymbol{x}_i^T \boldsymbol{w}$ down to the range $0 \leq f(\boldsymbol{x}_i) \leq 1$.
- Psychologists studied many possibilities, but the one most often used today is the logistic sigmoid function:

$$f(\boldsymbol{x}_i) = \sigma(\boldsymbol{x}_i^T \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{w}}}$$

This function is called sigmoid because it is S-shaped.

$$\sigma(z) = \begin{cases} 1 & z \to \infty \\ 0.5 & z = 0 \\ 0 & z \to -\infty \end{cases}$$

Interpretation as a probability



- Since $0 < f(\boldsymbol{x}) < 1$, we can interpret $f(\boldsymbol{x})$ as a probability
- Specifically, we interpret it as $f(\boldsymbol{x}) = P(Y = 1|X = \boldsymbol{x})$.

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{x}^T\boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{x}^T\boldsymbol{w}}}$$

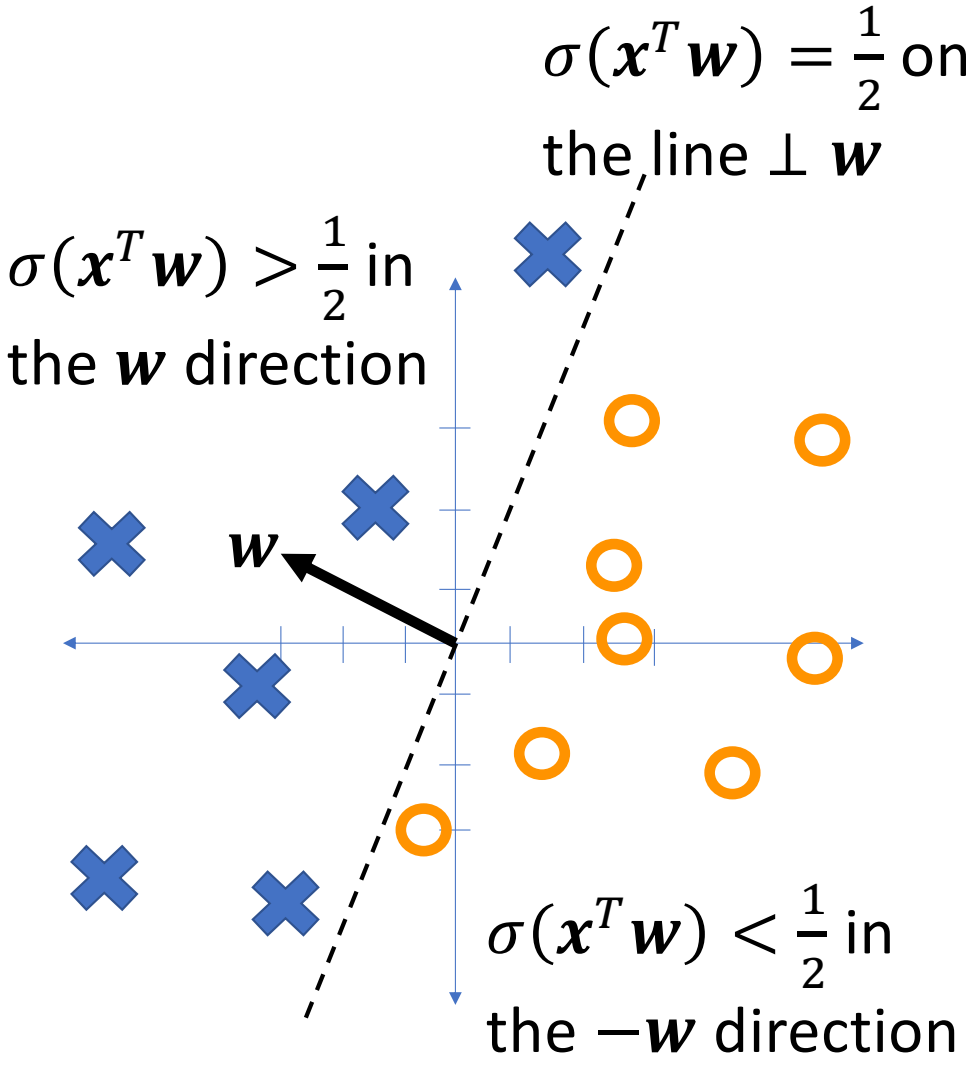
- The argument of the sigmoid, $\boldsymbol{x}^T\boldsymbol{w}$, is called the "logit." Notice that there is a straightforward relationship between the logit and the probability:

$$\sigma(z) = \begin{cases} 1 & z \to \infty \\ 1/2 & z = 0 \\ 0 & z \to -\infty \end{cases}$$

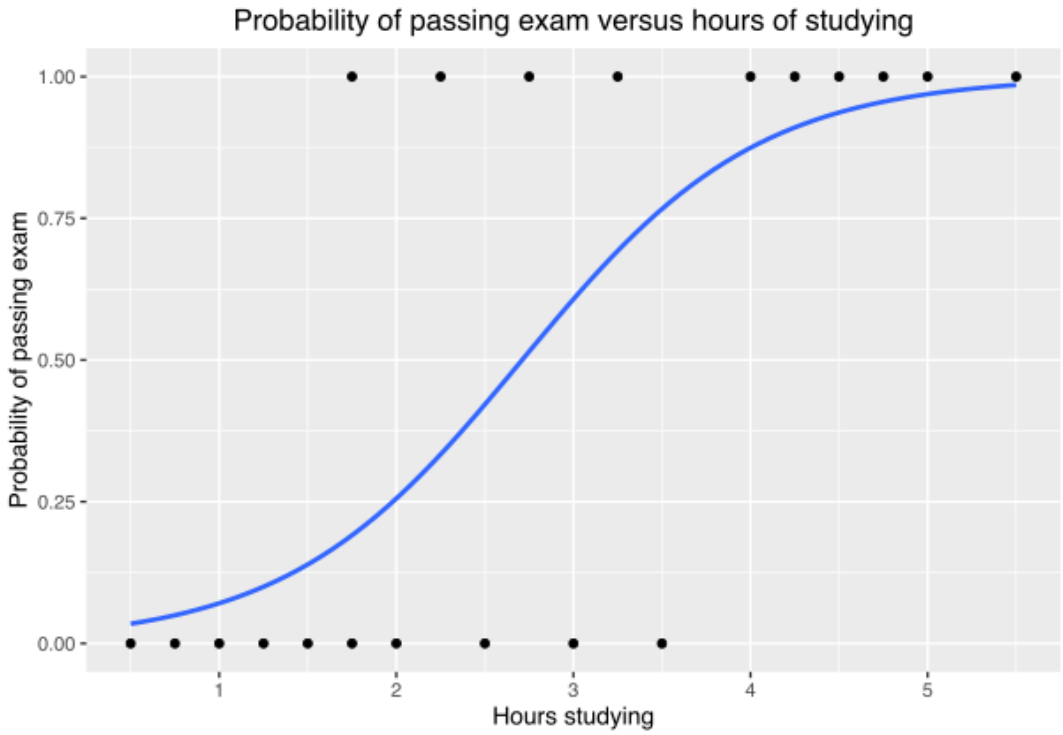
Geometric interpretation

Suppose $\boldsymbol{x}$ is a 2d vector. Then:

- $\boldsymbol{x}^T\boldsymbol{w} = 0$, and $\sigma(\boldsymbol{x}^T\boldsymbol{w}) = \frac{1}{2}$, on the line where $\boldsymbol{x} \perp \boldsymbol{w}$
- $\boldsymbol{x}^T\boldsymbol{w} > 0$, and $\sigma(\boldsymbol{x}^T\boldsymbol{w}) > \frac{1}{2}$, in the $\boldsymbol{w}$ direction
- $\boldsymbol{x}^T\boldsymbol{w} < 0$, and $\sigma(\boldsymbol{x}^T\boldsymbol{w}) < \frac{1}{2}$, in the $-\boldsymbol{w}$ direction



Example



For example, the blue line here shows

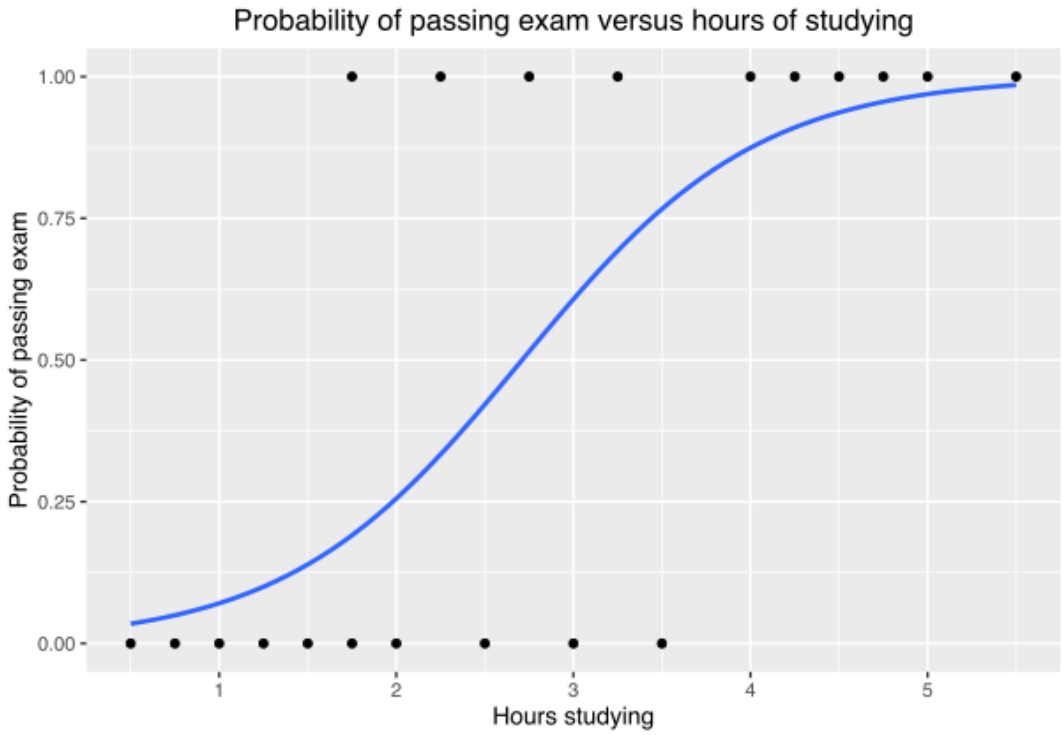
$$P(Y = 1|x) = \sigma(3x - 2.75)$$

$$= \begin{cases} 1 & x \to \infty \\ 1/2 & x = 2.75 \\ 0 & x \to -\infty \end{cases}$$

Outline

- Linear regression: Review
- Logistic regression: Output is a probability
- Derivative of the sigmoid
- Derivative of the log sigmoid

How is logistic regression trained?



The blue line is the trained classifier.

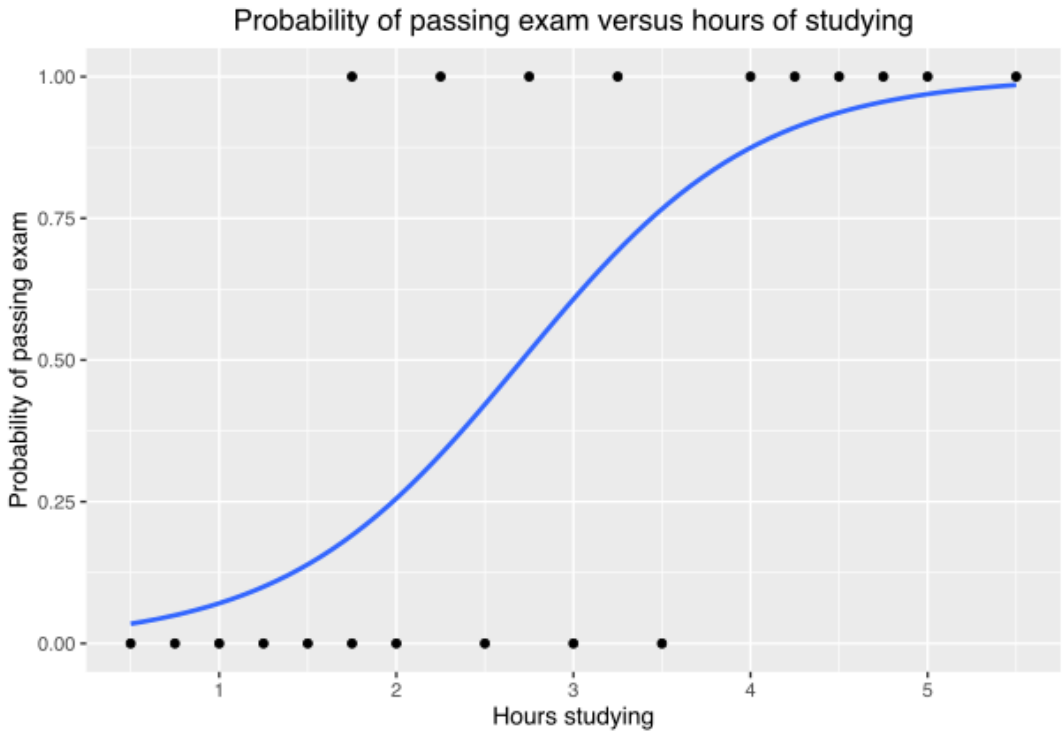
The training data are the black dots:

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_{20}, y_{20})\}$$

$$= \left\{ \begin{array}{c} (0.5,0), (0.75,0), (1.0,0), (1.25,0), \\ (1.5,0), (1.75,0), (2.0,0), (2.5,0), \\ (3.0,0), (3.5,0), (1.75,1), (2.25,1), \\ (2.75,1), (3.25,1), (4.0,1), (4.25,1), \\ (4.5,1), (4.75,1), (5.0,1), 5.5,1) \end{array} \right\}$$

Given that training data, how did we learn the model $P(Y = 1|x) = \sigma(3x - 2.75)$?

First try: SGD



Consider a method based on stochastic gradient descent:

1. Randomly choose a training datum $(\boldsymbol{x}_i, y_i)$
2. If $y_i = 1$: adjust $\boldsymbol{w}$ to increase $\sigma(\boldsymbol{x}_i^T \boldsymbol{w})$:

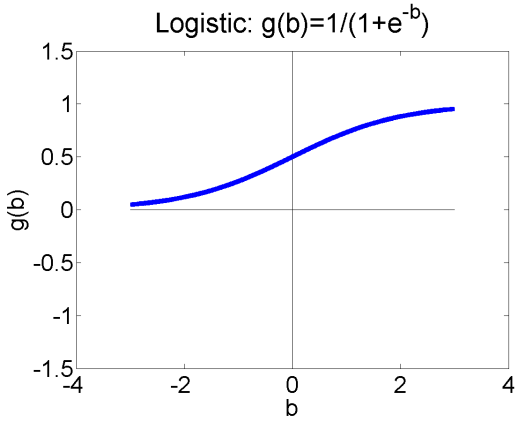
$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta \frac{\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$$

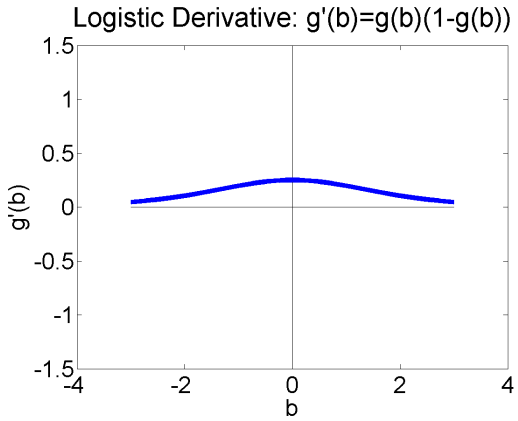
3. If $y_i = 0$: adjust $\boldsymbol{w}$ to decrease $\sigma(\boldsymbol{x}_i^T \boldsymbol{w})$:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$$

4. Repeat

OK, what's $\dfrac{\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$?



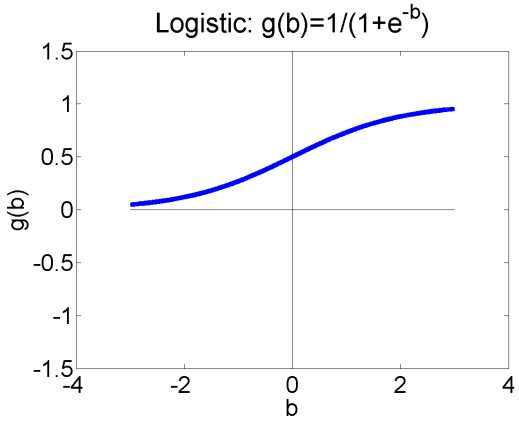


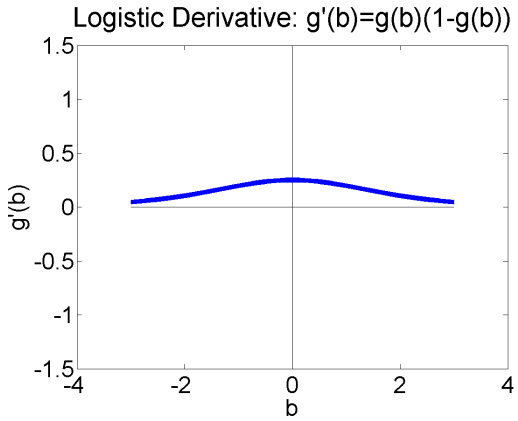
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\frac{\partial \sigma(z)}{\partial z} = \left(-\frac{1}{(1 + e^{-z})^2}\right)(-e^{-z}) = \left(\frac{1}{1 + e^{-z}}\right)\left(\frac{e^{-z}}{1 + e^{-z}}\right)$$

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

OK, what's $\dfrac{\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$?

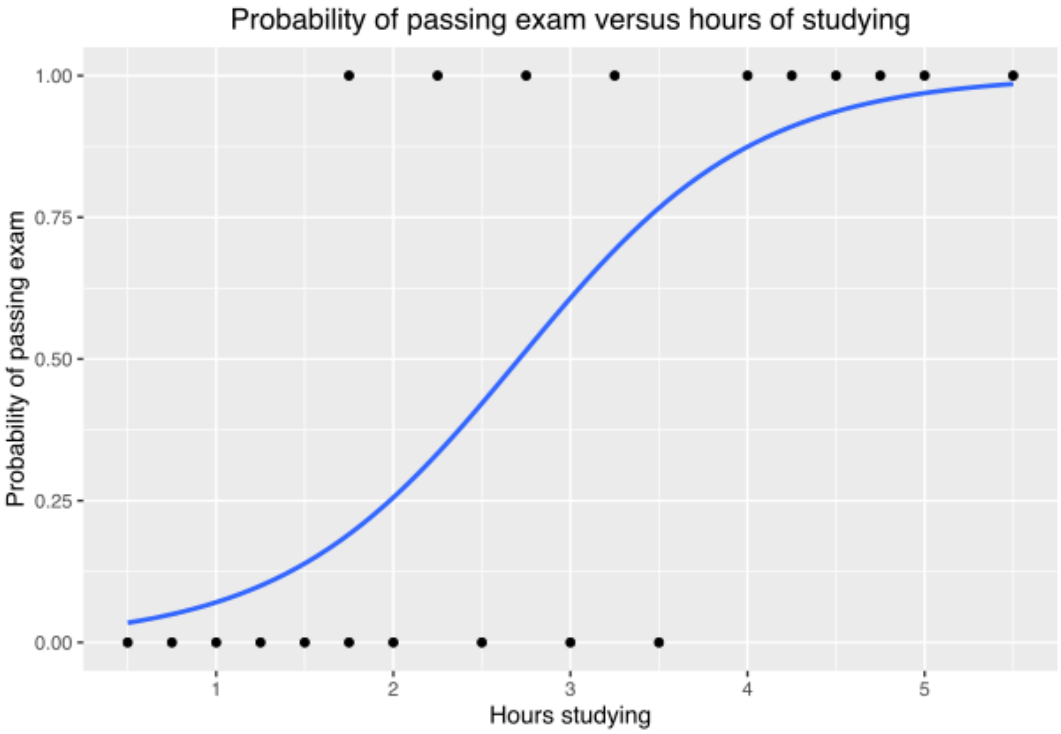




$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))$$

$$\frac{\partial \sigma(z)}{\partial \boldsymbol{w}} = \frac{\partial \sigma(z)}{\partial z}\frac{\partial z}{\partial \boldsymbol{w}} = \sigma(z)(1 - \sigma(z))\boldsymbol{x}_i$$

SGD using probabilities



Consider a method based on stochastic gradient descent:

1. Randomly choose a training datum $(\boldsymbol{x}_i, y_i)$
2. If $y_i = 1$:
   $\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta\sigma(z)(1 - \sigma(z))\boldsymbol{x}_i$
3. If $y_i = 0$:
   $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta\sigma(z)(1 - \sigma(z))\boldsymbol{x}_i$
4. Repeat

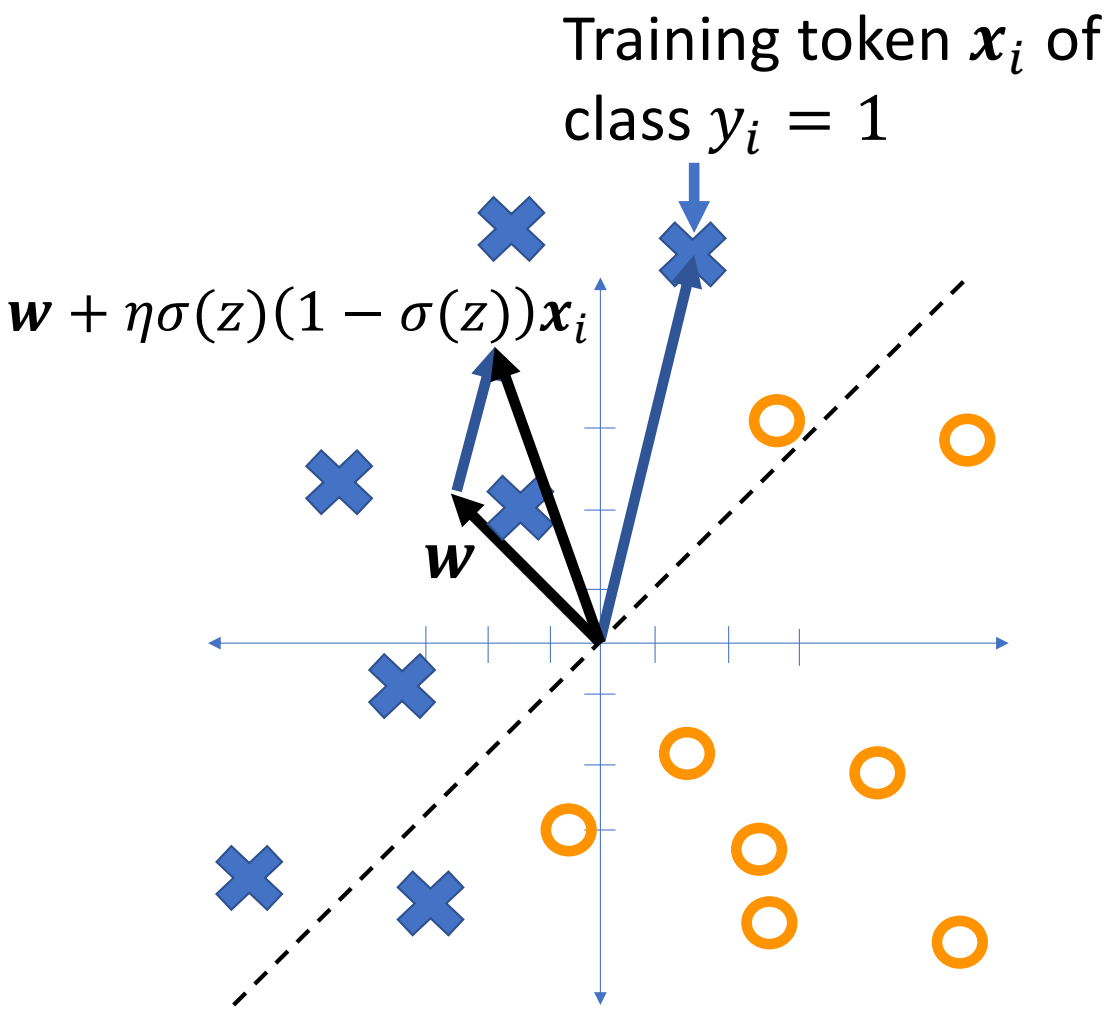
Try the quiz!

- Go to PrairieLearn, try the quiz!

1. Randomly choose a training datum $(\boldsymbol{x}_i, y_i)$
2. If $y_i = 1$: adjust $\boldsymbol{w}$ to increase $\sigma(\boldsymbol{x}_i^T \boldsymbol{w})$:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta \frac{\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$$

$$= \boldsymbol{w} + \eta \sigma(z)(1 - \sigma(z))\boldsymbol{x}_i$$



Outline

- Linear regression: Review
- Logistic regression: Output is a probability
- Derivative of the sigmoid
- Derivative of the log sigmoid

Maximize probability of the training dataset?

- SGD with probabilities is not very systematic
- Here's a systematic approach: Let's try to maximize the probability of the training dataset.
- The training dataset is $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$
- Let's suppose all those examples are chosen independently. Then

$$P(\mathcal{D}) = \prod_{i=1}^{n} P(Y = y_i | \boldsymbol{x}_i)$$

- What is $\frac{\partial}{\partial \boldsymbol{w}} \prod_{i=1}^{n} P(Y = y_i | \boldsymbol{x}_i)$? Answer: UGLY!!!!!

No! Maximize its log probability!

- Notice that whatever value of $\boldsymbol{w}$ maximizes this:

$$P(\mathcal{D}) = \prod_{i=1}^{n} P(Y = y_i | \boldsymbol{x}_i)$$

… will also maximize this:

$$\log P(\mathcal{D}) = \sum_{i=1}^{n} \log P(Y = y_i | \boldsymbol{x}_i)$$

$$= \sum_{y_i=1} \log \sigma(\boldsymbol{x}_i^T \boldsymbol{w}) + \sum_{y_i=0} \log\left(1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)$$

OK, what's $\frac{\partial \log \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$?

$$\frac{\partial \log \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{1}{\sigma(\boldsymbol{x}_i^T \boldsymbol{w})} \frac{\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$$

$$= \frac{1}{\sigma(\boldsymbol{x}_i^T \boldsymbol{w})} \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\big(1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\big)\boldsymbol{x}_i$$

$$= \big(1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\big)\boldsymbol{x}_i$$

OK, what's $\frac{\partial \log\left(1-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)}{\partial \boldsymbol{w}}$?

$$\frac{\partial \log\left(1-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)}{\partial \boldsymbol{w}} = \frac{1}{\left(1-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)} \frac{-\partial \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}}$$

$$= \frac{1}{\left(1-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)} \left(-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\left(1-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)\right) \boldsymbol{x}_i$$

$$= \left(0-\sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right) \boldsymbol{x}_i$$

OK, what's $\frac{\partial \log P(\mathcal{D})}{\partial \boldsymbol{w}}$?

$$\frac{\partial \log P(\mathcal{D})}{\partial \boldsymbol{w}} = \sum_{y_i=1} \frac{\partial \log \sigma(\boldsymbol{x}_i^T \boldsymbol{w})}{\partial \boldsymbol{w}} + \sum_{y_i=0} \frac{\partial \log\big(1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\big)}{\partial \boldsymbol{w}}$$

$$= \sum_{y_i=1} \big(1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\big)\boldsymbol{x}_i + \sum_{y_i=0} \big(0 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\big)\boldsymbol{x}_i$$

Cross-entropy

This loss function:

$$\mathcal{L}_i = -\ln \Pr(Y = y_i | \boldsymbol{x}_i)$$

is called cross-entropy. The term comes from physics, where "entropy" is our degree of uncertainty about whether or not something will happen.

CC-SA 4.0, https://en.wikipedia.org/wiki/File:Ultra_slow-motion_video_of_glass_tea_cup_smashed_on_concrete_floor.webm

Gradient descent

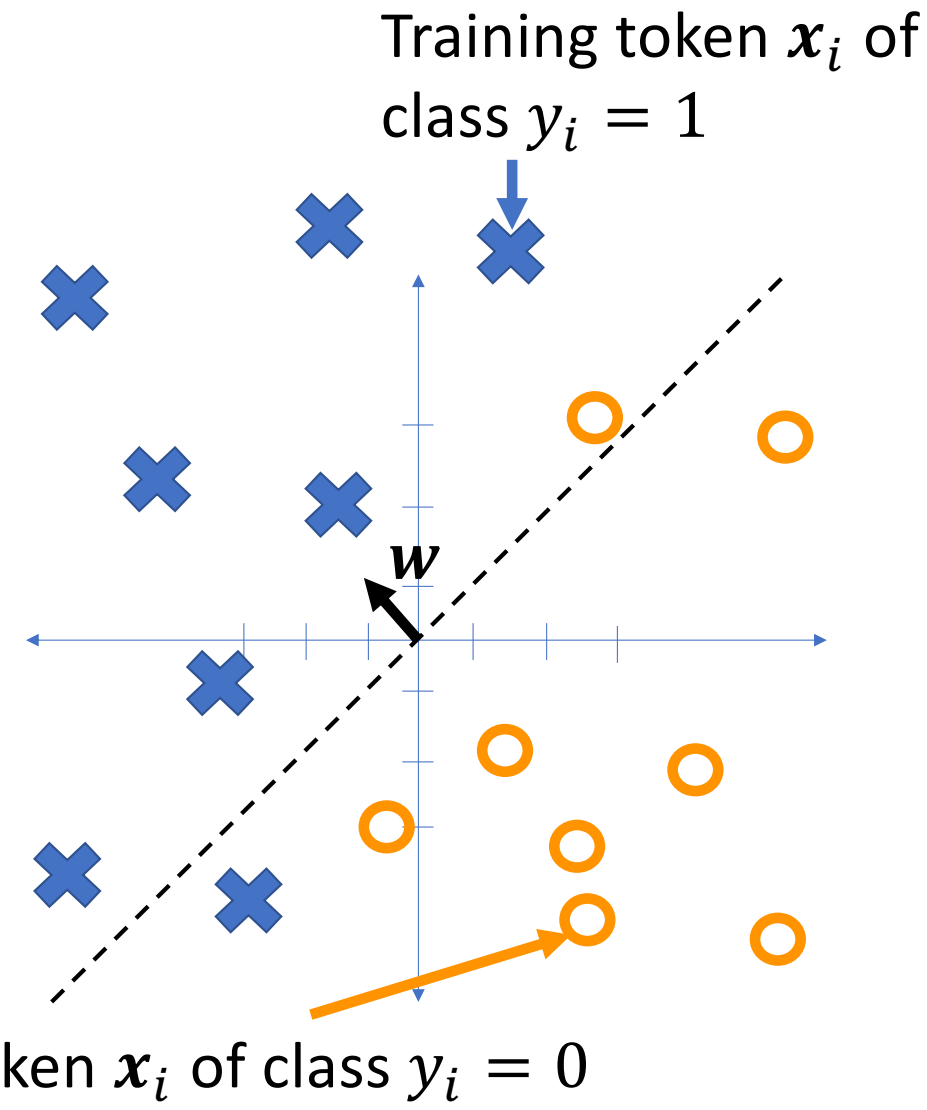
Suppose we have training tokens $(\boldsymbol{x}_i, y_i)$, and we have some initial weight vector $\boldsymbol{w}$. Then we can update it as

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}}$$

…where…

$$\mathcal{L} = -\log P(\mathcal{D})$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \sum_{i=1}^{n} \left(\sigma\left(\boldsymbol{x}_i^T \boldsymbol{w}\right) - y_i\right) \boldsymbol{x}_i$$



Conclusions

- Linear regression

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}}, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i$$

- Logistic regression

$$f(\boldsymbol{x}_i) = \sigma(\boldsymbol{x}_i^T \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{w}}}$$

- Derivative of the sigmoid

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

- Derivative of the negative log sigmoid

$$\mathcal{L} = -\sum_{y_i=1} \log \sigma\left(\boldsymbol{x}_i^T \boldsymbol{w}\right) - \sum_{y_i=0} \log\left(1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{w})\right)$$