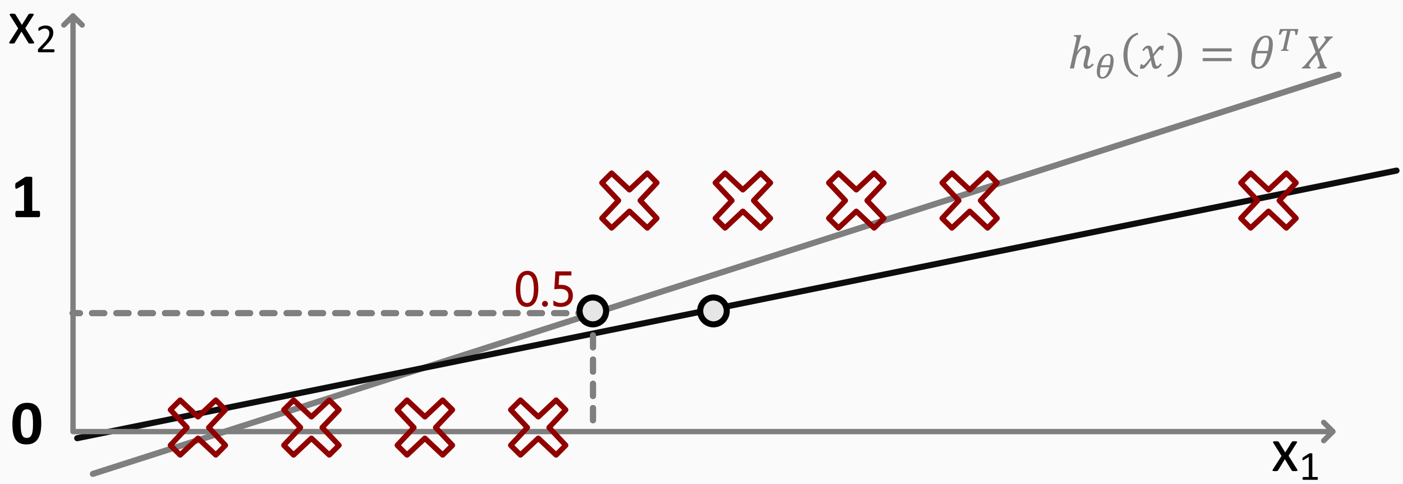
logistic regression ▪ basics

classification and representation

classification

the assignments in a single class classification problem are typically as follows:

$$y \in \{0,1\} \quad \rightarrow \quad \begin{array}{l} 0\text{: "negative class"} \\ 1\text{: "positive class"} \end{array}$$



although circumstances could allow linear regression to predict a proper output for the predicted values, Applying linear regression to classification problems is generally not effective; it does not fit outliers and will result in false negatives and positives:

threshold classifier output $h_\theta(x)$ at 0.5:

if $h_\theta(x) \geq 0.5$, predict "$y = 1$"

if $h_\theta(x) < 0.5$, predict "$y = 0$"

Which of the following statements is true?

- If linear regression doesn't work on a classification task as in the previous example shown in the video, applying feature scaling may help.
- If the training set satisfies $0 \leq y^{(i)} \leq 1$ for every training example $(x^{(i)}, y^{(i)})$, then linear regression's prediction will also satisfy $0 \leq h_\theta(x) \leq 1$ for all values of x.
- If there is a feature x that perfectly predicts y, i.e. if $y = 1$ when $x \geq c$ and $y = 0$ whenever $x < c$ (for some constant c), then linear regression will obtain zero classification error.
- None of the above statements are true.

Correct Response

additionally, linear regression often experiences values where $h_\theta(x)$ can predict values > 1 or < 0; the latter is outside of the classifications of $y = 0$ or $y = 0$

hypothesis representation

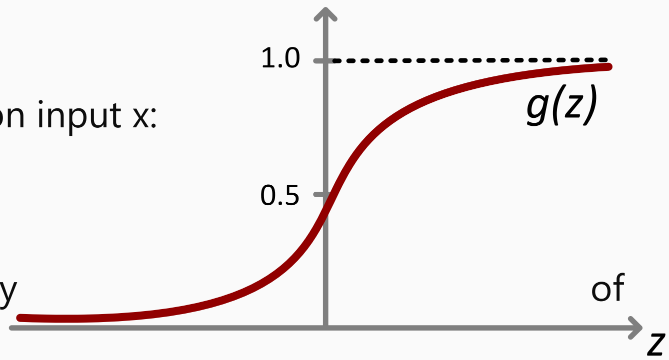
the logistic regression models requires outputs within the range $0 \leq h_\theta(x) \leq 1$

therefore, the linear regression function is altered: $h_\theta(x) = \theta^T X \rightarrow g(\theta^T X)$

intuition expansion: $h_\theta(x) = g(\theta^T X) \qquad g(z) = \frac{1}{1+e^{-z}}$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T X}} \rightarrow \text{sigmoid/logistic function}$$

interpretation of the hypothesis' output



$h_\theta(x) =$ the estimated probability that $y = 1$ on input x:

if $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{some measured value} \end{bmatrix}$

and $h_\theta(x) = 0.7$ then there is a 70% probability of the measured value is the **positive class** indicated in the problem set

$h_\theta(x) = P(y = 1|x{:}\,\theta)$; "probability that $y = 1$ given x, parameterized by θ"

concretely, because logistic regression can only return outputs of $y = 0$ and $y = 1$:

$$P(y = 0|x{:}\,\theta) + P(y = 1|x{:}\,\theta) = 1$$

$$P(y = 0|x{:}\,\theta) = 1 - P(y = 1|x{:}\,\theta)$$

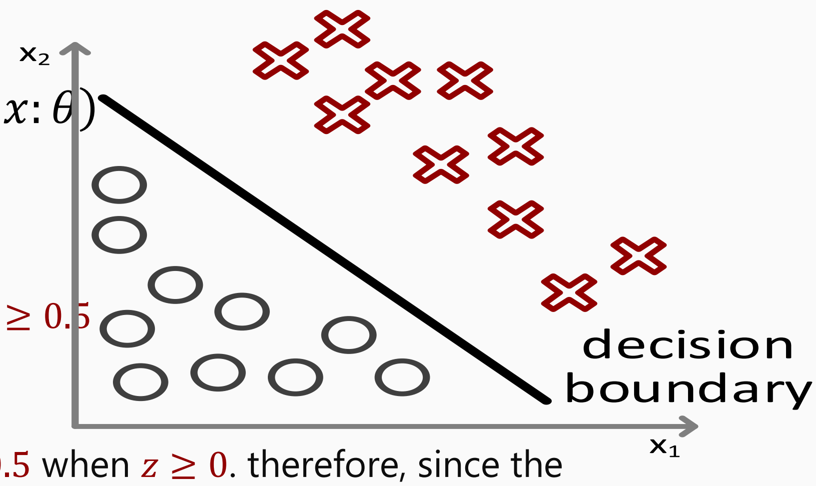
Suppose we want to predict, from data x about a tumor, whether it is malignant ($y = 1$) or benign ($y = 0$). Our logistic regression classifier outputs, for a specific tumor, $h_\theta(x) = P(y = 1|x;\theta) = 0.7$, so we estimate that there is a 70% chance of this tumor being malignant. What should be our estimate for $P(y = 0|x;\theta)$, the probability the tumor is benign?

- $P(y = 0|x;\theta) = 0.3$

Correct Response

- $P(y = 0|x;\theta) = 0.7$
- $P(y = 0|x;\theta) = 0.7^2$
- $P(y = 0|x;\theta) = 0.3 \times 0.7$

decision boundary



$$h_\theta(x) = g(\theta^T X) = P(y = 1|x\colon\theta)$$

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

assuming a prediction of "$y = 1$" if $h_\theta(x) \geq 0.5$

and a prediction of "$y = 0$" if $h_\theta(x) < 0.5$

looking at the sigmoid function; $g(z) \geq 0.5$ when $z \geq 0$. therefore, since the hypothesis for logistic regression is $h_\theta(x) = g(\theta^T X)$, then $h_\theta(x) = g(\theta^T X) \geq 0.5$, whenever $\theta^T X \geq 0$ because $\theta^T X$ effectively taken on the value of z in the sigmoid function $g(z)$

conversely, when $h_\theta(x) < 0.5$ then $g(z) \leq 0.5$ considering that $h_\theta(x) = g(\theta^T X)$ illustrated above. Therefore when $h_\theta(x) = g(\theta^T X) < 0.5$, whenever $\theta^T X < 0$ because $\theta^T X$ effectively taken on the value of z in the sigmoid function $g(z)$

determination of the decision boundary example

given the function $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$ with the following parameters:

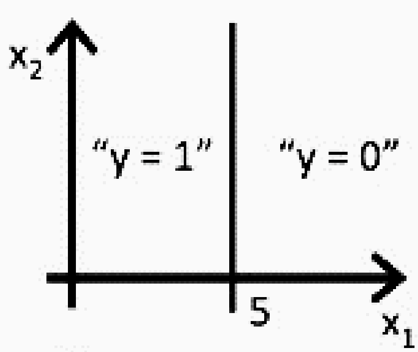
$\theta_0 = -3, \theta_1 = 1, \theta_2 = 1$ produces the parameter vector $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$

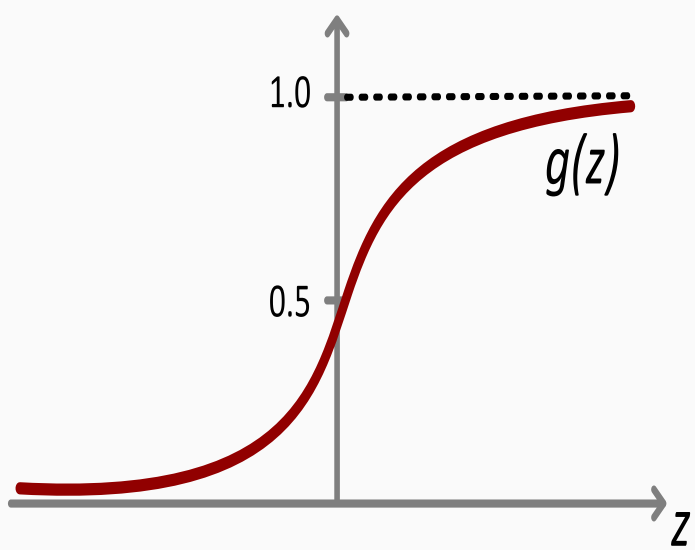
referring to the above formulas, "$y = 1$" will be predicted if $\theta^T X = -3 + x_1 + x_2 \geq 0$

any example of (x_a, x_2) that satisfies the equation $-3 + x_1 + x_2 \geq 0$ will predict "$y = 1$"

additional rule notation is as follows: $(-3 + x_1 + x_2 \geq 0) = (x_1 + x_2 \geq 3)$

Consider logistic regression with two features x_1 and x_2. Suppose $\theta_0 = 5, \theta_1 = -1, \theta_2 = 0$, so t hat $h_\theta(x) = g(5 - x_1)$. Which of these shows the decision boundary of $h_\theta(x)$?





nonlinear decision boundaries

adding additional higher order polynomial terms can adapt to fitting nonlinear datasets

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

determination of the decision boundary example

given the function $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$; with the parameters:

$\theta_0 = -1, \theta_1 = 0, \theta_2 = 0, \theta_3 = 1, \theta_4 = 1$ produces the parameter vector $\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$

with the above formulas, "$y = 1$" will be predicted if $\theta^T X = -1 + x_1^2 + x_2^2 \geq 0$

simplified rule notation is as follows: $(-1 + x_1^2 + x_2^2 \geq 0) = (x_1^2 + x_2^2 \geq 1)$

adding more complex polynomial features allows the algorithm to fit more complex decision boundaries. an important principal of logistic regression is that the **decision boundary** is a property **not** of the **training set**, but of the **hypothesis** under the parameters. therefore, as long as parameter vector θ is known, it will define the decision boundary. the training set does not define the decision boundary but can be used to fit the parameters θ

finally, the function determines the complexity:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \cdots)$$