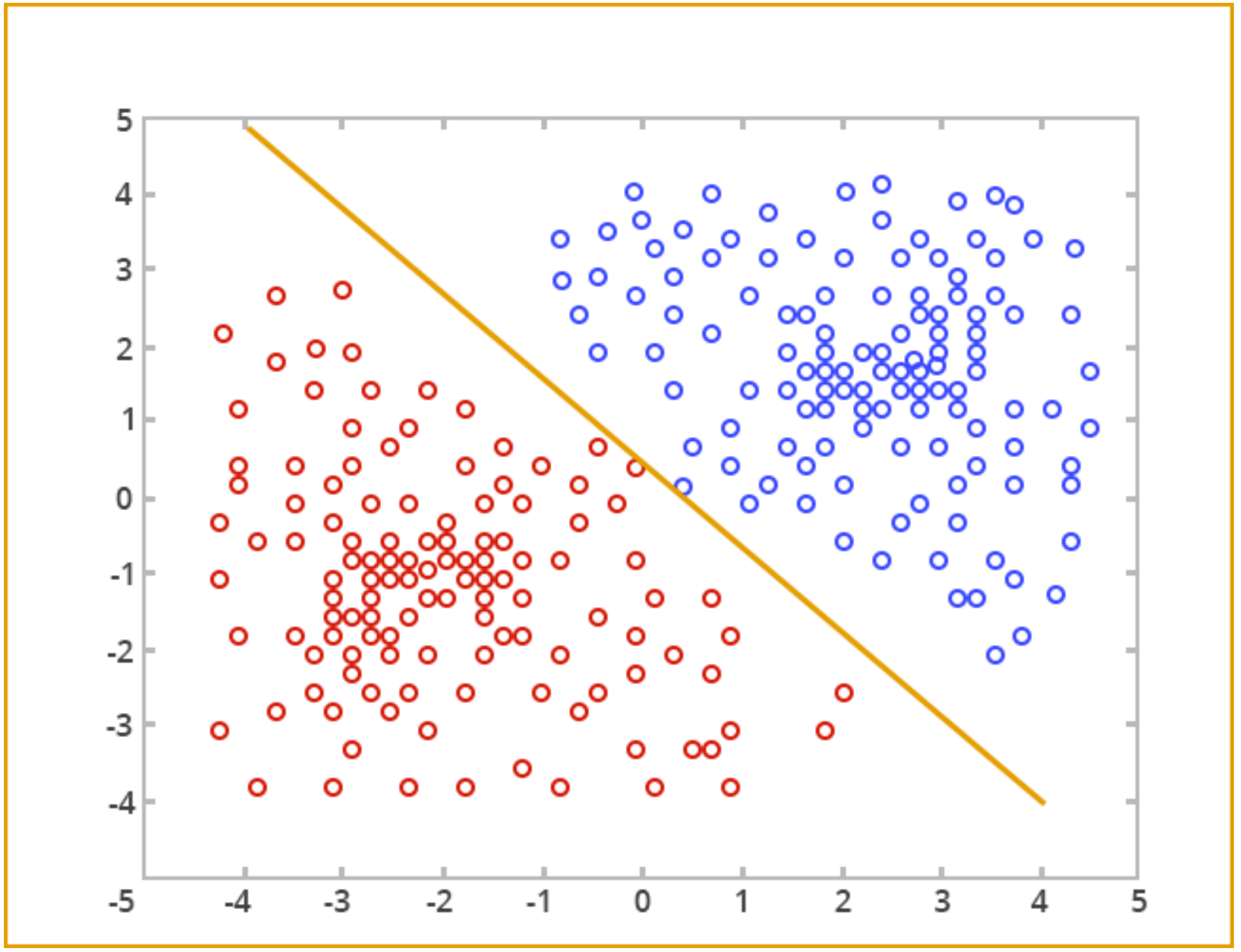


Logistic Regression

Leandro L. Minku

General Idea

- In Naïve Bayes, we explicitly model the class conditional probability distributions and prior probabilities of the classes.
    - We then use the Bayes Theorem together with the conditional independence assumption to determine the probability of an instance belonging to a given class.
- In Logistic Regression, we will model the probability (actually the odds) of an instance belonging to a given class directly, as a linear combination of the independent variables.
    - Then, we can predict the class based on such probabilities.
- We will start with binary classification problems.

The Need for the Logit Function

- Consider that we wish to model $p(c_1 | \mathbf{x})$ as a function of the independent variables:

$$p(c_1 | \mathbf{x}) = w_0 + w_1 x_1 + \cdots + w_n x_n$$

$$p(c_1 | \mathbf{x}) = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n, \text{ where } x_0 = 1$$

$$p(c_1 | \mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

$$p_1 = \mathbf{w}^T \mathbf{x}$$

- If that was possible, we would be able to treat this classification problem in a similar way to a regression problem, by learning the coefficients $\mathbf{w}$.

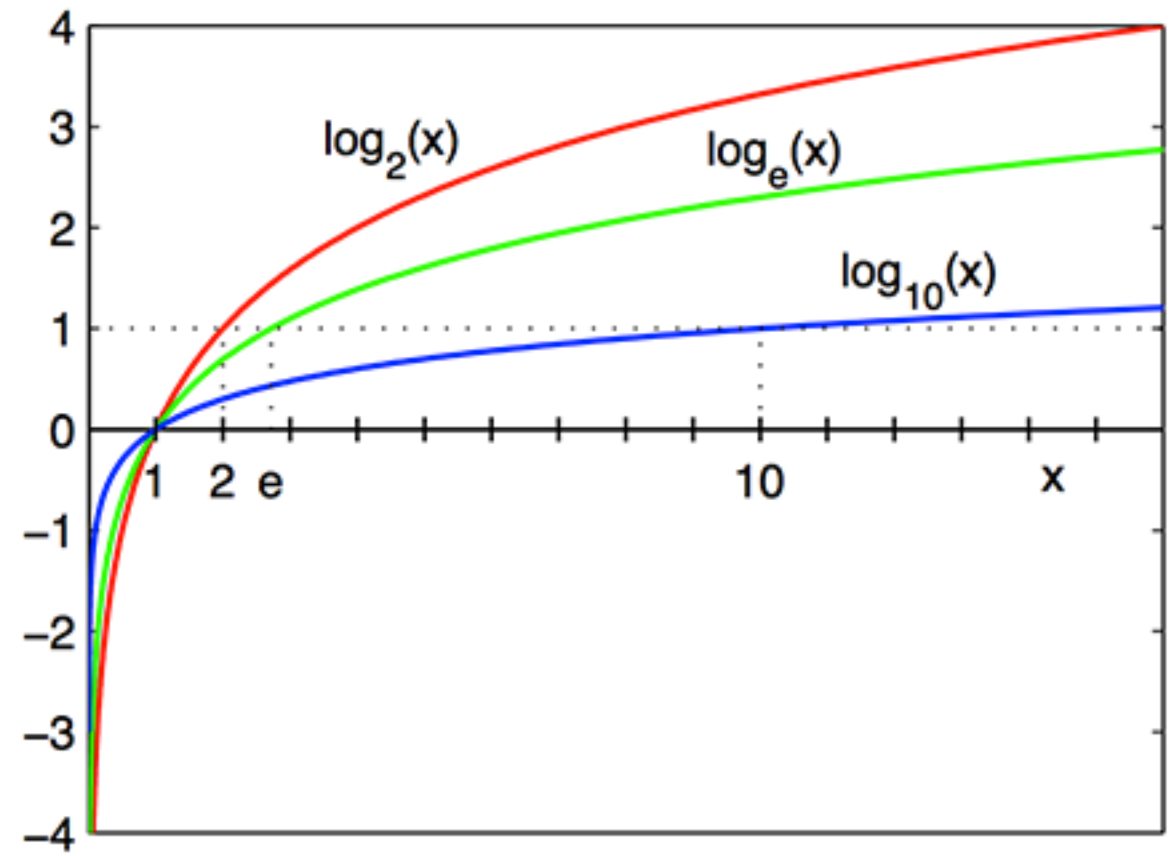
- However, $\mathbf{w}^T \mathbf{x}$ could assume any values in [-∞,∞], whereas p_1 should be in [0,1].

The Need for the Logit Function

- To fix that, one might think of modelling $\ln(p_1)$ instead of p_1:

$$\ln(p_1) = \mathbf{w}^T \mathbf{x}$$

- However, logarithms are unbounded only from one direction and linear functions are not.

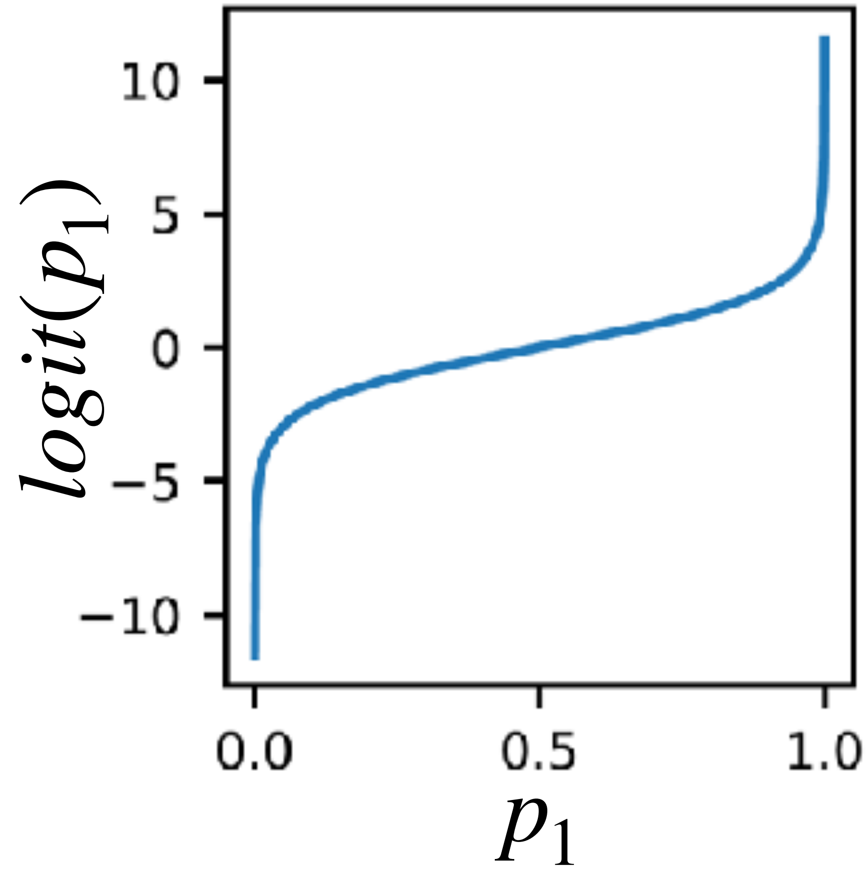


The Need for the Logit Function

- A solution would be to model logit$(p_1) = \mathbf{w}^T\mathbf{x}$, where

  logit(p_1) = ln $\left(\frac{p_1}{1 - p_1}\right)$

- Logit enables us to map from [0,1] to [-∞,∞].



The Odds

$$\text{logit}(p_1) = \ln\left(\frac{p_1}{1 - p_1}\right)$$

- Odds: ratio of probabilities of two possible outcomes:

$$o_1 = \frac{p_1}{p_0} = \frac{p_1}{1 - p_1}$$

- For example,

  If $p_1 = 0.7$ and $p_0 = 0.3$, $o_1 \approx 2.33$
  If $p_1 = 0.3$ and $p_0 = 0.7$, $o_1 \approx 0.43$
  If $p_1 = 0.5$ and $p_0 = 0.5$, $o_1 = 1$

- If $o_1 \geq 1$, predict class c_1.
- If $o_1 < 1$, predict class c_0.

Logit

- Logit: logarithm of the odds.

$$\text{logit}(p_1) = \ln\left(\frac{p_1}{1 - p_1}\right)$$

- For example,

If $p_1 = 0.7$ and $p_0 = 0.3$, $\text{logit}(p_1) \approx 0.85$
If $p_1 = 0.3$ and $p_0 = 0.7$, $\text{logit}(p_1) \approx -0.85$
If $p_1 = 0.5$ and $p_0 = 0.5$, $\text{logit}(p_1) = 0$

- If $\text{logit}(p_1) = \mathbf{w}^T\mathbf{x} \geq 0$, predict class c_1.
- If $\text{logit}(p_1) = \mathbf{w}^T\mathbf{x} < 0$, predict class c_0.

This is the key idea behind logistic regression!

A Linear Classifier

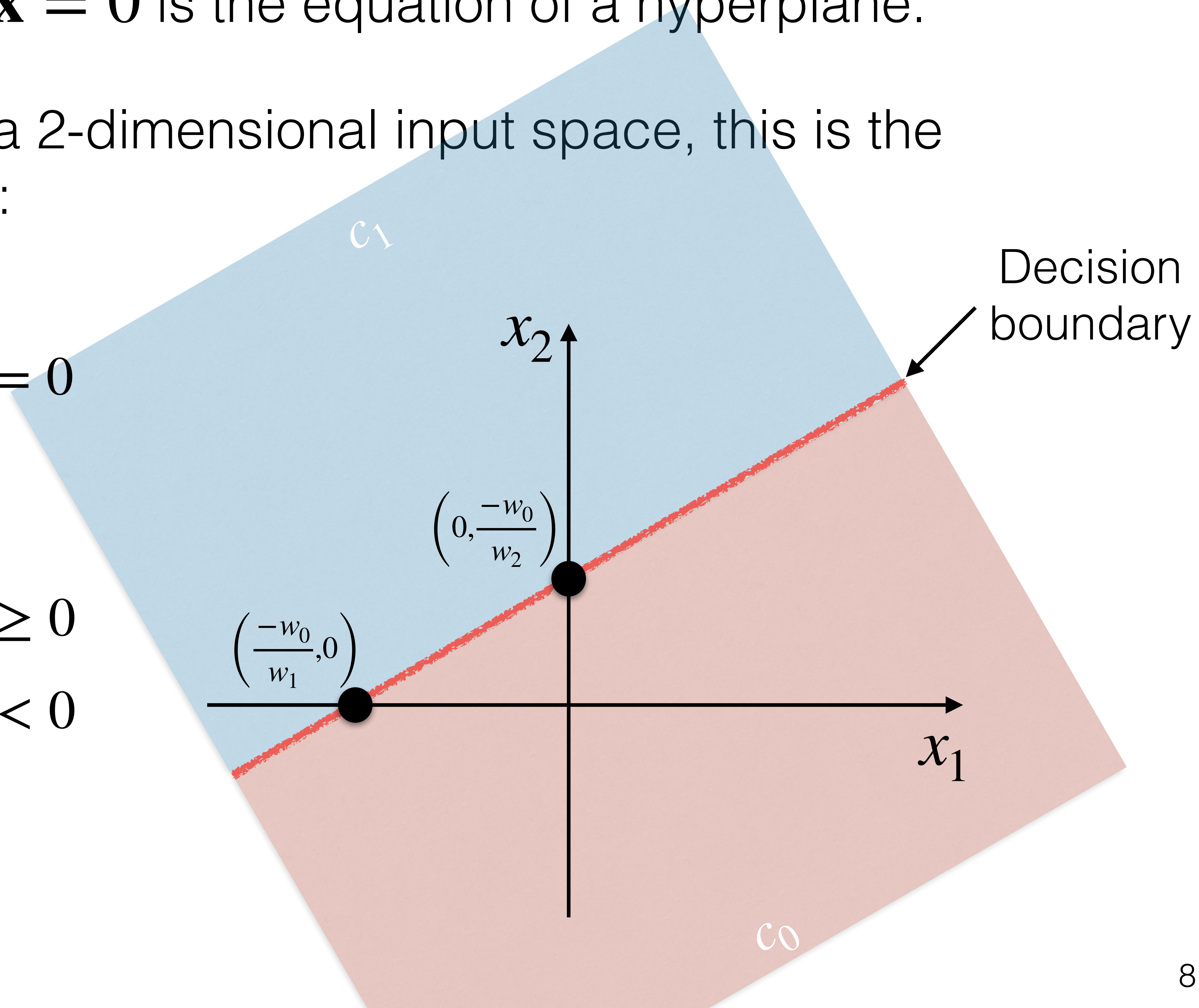
- The equation $\mathbf{w}^T\mathbf{x} = 0$ is the equation of a hyperplane.
- For example, for a 2-dimensional input space, this is the equation of a line:

$$w_0x_0 + w_1x_1 + w_2x_2 = 0$$

$$w_1x_1 + w_2x_2 = -w_0$$

$$w_0x_0 + w_1x_1 + w_2x_2 \geq 0$$

$$w_0x_0 + w_1x_1 + w_2x_2 < 0$$



Computing the Probabilities p_1 and p_0

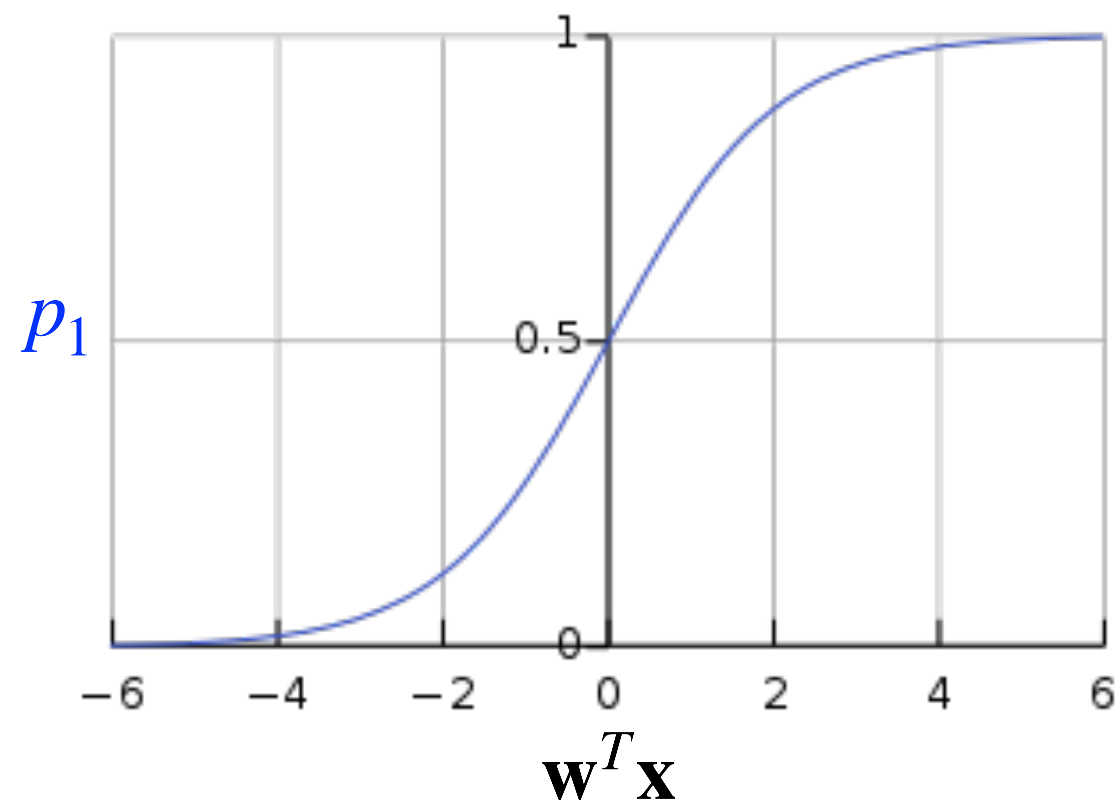
- $\text{logit}(p_1) = \mathbf{w}^T\mathbf{x}$ $\begin{cases} \mathbf{w}^T\mathbf{x} \geq 0 \rightarrow c_1 \\ \mathbf{w}^T\mathbf{x} < 0 \rightarrow c_0 \end{cases}$

- If we solve $\text{logit}(p_1) = \mathbf{w}^T\mathbf{x}$ for p_1 we get:

$$p_1 = \frac{e^{(\mathbf{w}^T\mathbf{x})}}{1 + e^{(\mathbf{w}^T\mathbf{x})}}$$

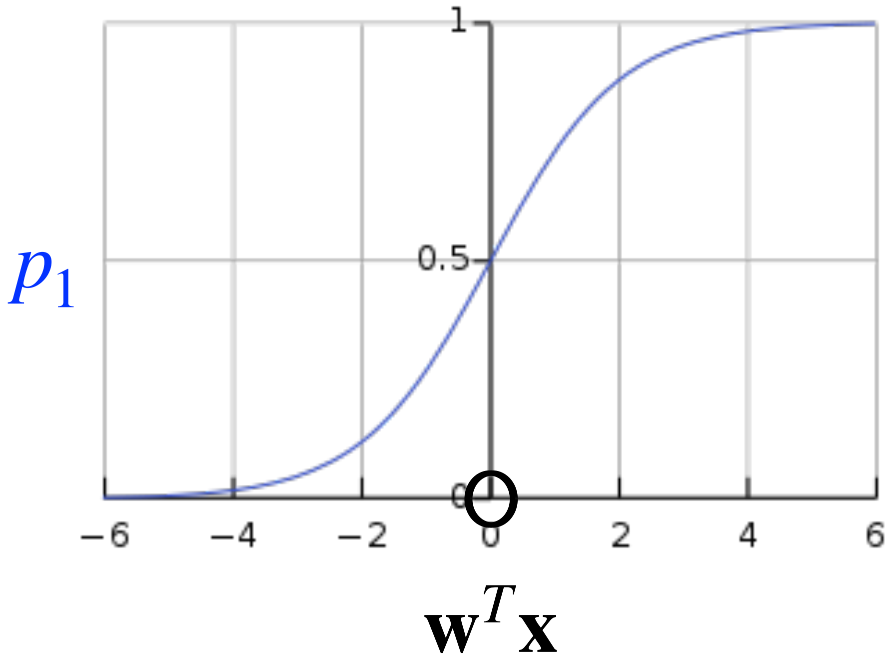
$$p_0 = 1 - p_1 = \frac{1}{1 + e^{(\mathbf{w}^T\mathbf{x})}}$$

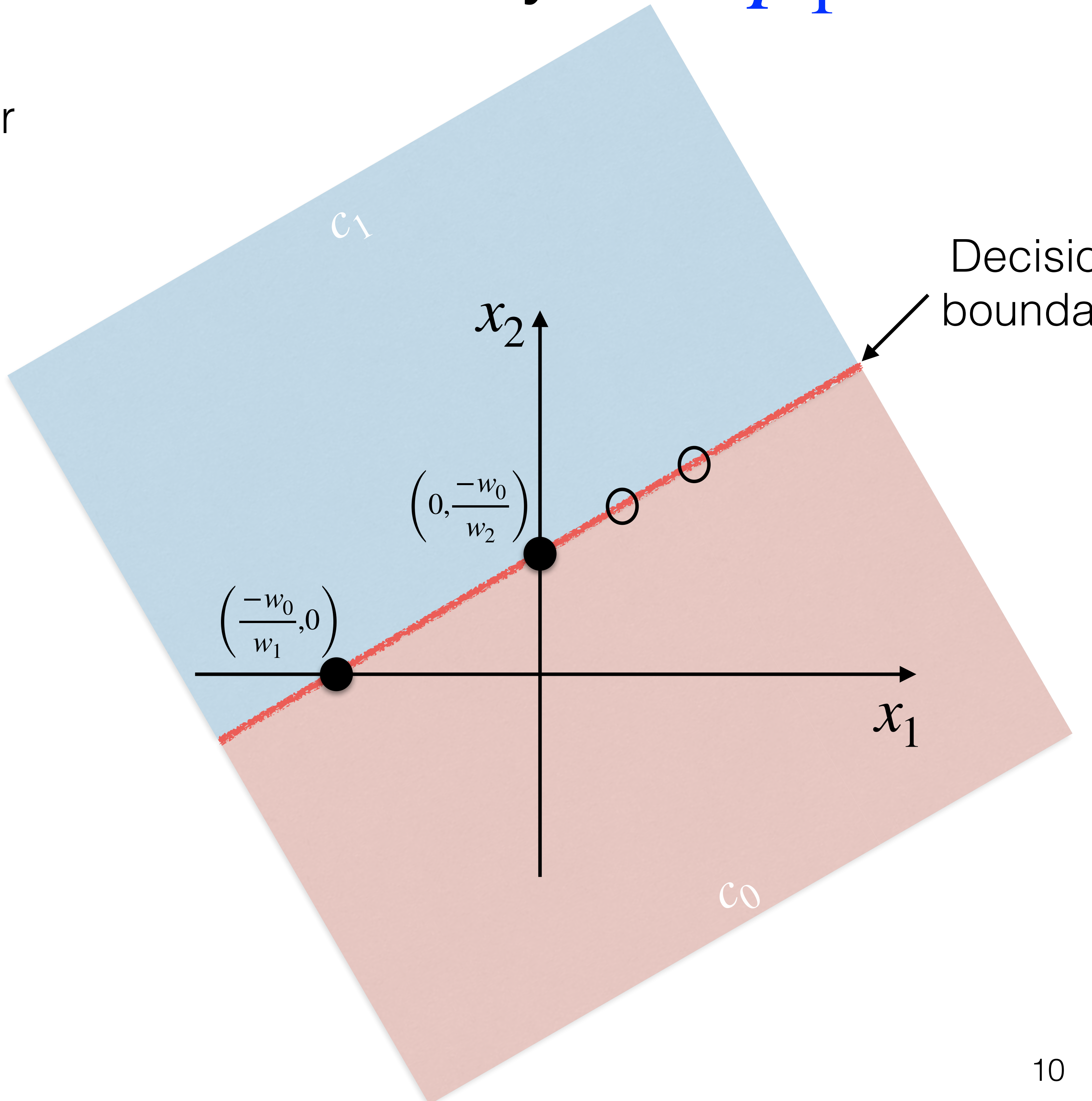
Sigmoid logistic function



The Relationship Between the Distance To The Decision Boundary and p_1

- The larger $|\mathbf{w}^T\mathbf{x}|$, the further away from the decision boundary the example $\mathbf{x}$ is.
- The larger $\mathbf{w}^T\mathbf{x}$, the higher p_1.
- The more negative $\mathbf{w}^T\mathbf{x}$, the smaller the p_1 (and the larger the p_0).





How to Learn **w**?

$$\text{logit}(p_1) = \mathbf{w}^T\mathbf{x} \begin{cases} \mathbf{w}^T\mathbf{x} \geq 0 \rightarrow c_1 \\ \mathbf{w}^T\mathbf{x} < 0 \rightarrow c_0 \end{cases}$$

- Given a training set

$$\mathcal{T} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

- Find **w** that maximises the following likelihood:

$$p(\mathcal{T}\,|\,\mathbf{w}) = \mathcal{L}(\mathbf{w}) = \prod_{i=1}^{N} p_1(\mathbf{x}^{(i)}, \mathbf{w})^{y^{(i)}} p_0(\mathbf{x}^{(i)}, \mathbf{w})^{1-y^{(i)}}$$

Here, we are explicitly writing that p_1 depends on $\mathbf{x}^{(i)}$ and $\mathbf{w}$ by writing it as $p_1(\mathbf{x}^{(i)}, \mathbf{w})$.

$y^{(i)} = 1 \qquad 1 - y^{(i)} = 0$

$y^{(i)} = 0 \qquad 1 - y^{(i)} = 1$

How to Learn **w**?

- Equivalent to finding **w** that maximises the log-likelihood:

$$\ln(\mathscr{L}(\mathbf{w})) = \sum_{i=1}^{N} y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)} - \ln(1 + \exp(\mathbf{w}^T\mathbf{x}^{(i)}))$$

- We can write finding the optimal weights **w*** that maximise the log-likelihood as:

$$\mathbf{w}^* = \text{argmax}_{\mathbf{w}} \left[\sum_{i=1}^{N} y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)} - \ln(1 + \exp(\mathbf{w}^T\mathbf{x}^{(i)})) \right]$$

- An algorithm called Iterative Reweighted Least Squares can be used to find **w*** (out of scope of this module).

How To Deal With Multiple Classes?

- Perform multiple logistic regressions against a pivot class c_M.

- Predict the class associated to the highest probability.

$$\ln\left(\frac{p_1}{p_M}\right) = \mathbf{w}_1^{*T}\mathbf{x}$$

$$\ln\left(\frac{p_2}{p_M}\right) = \mathbf{w}_2^{*T}\mathbf{x}$$

$$\ldots$$

$$\ln\left(\frac{p_{M-1}}{p_M}\right) = \mathbf{w}_{M-1}^{*T}\mathbf{x}$$

Solving this we get:

$$p_i = \frac{e^{(\mathbf{w_i^{*T}}\mathbf{x})}}{1 + \sum_{i=1}^{M-1} e^{(\mathbf{w}_i^{*T}\mathbf{x})}}$$

for $i = \{1,2,\cdots,M\}$

- How to find $\mathbf{w}_i^*$? Apply Iterative Reweighted Least Squares, for each class c_i

Advantages vs Disadvantages

- Advantage:
  - No need to choose a probability density function.

- Disadvantages:
  - Less efficient than Naïve Bayes for large training sets.
  - Requires larger training sets to perform well.

Applications

- Logistic regression is a traditional applied statistics approach, and has been used for many problems.
  - Medical problems, e.g., predicting mortality in injured patients, predicting the risk of developing a certain disease.
  - Marketing problems, e.g., predict a customer's propensity to buy a product or halt a subscription.
  - Economics problems, e.g., predict the likelihood of a homeowner defaulting on a mortgage.
  - Etc.

Quiz

- Consider a binary classification problem with 2 independent variables.
- Consider that Iterative Reweighted Least Squares learnt that

  $\mathbf{w}*^{\mathbf{T}} = (0.1, 0.2, 0.6)$

- What class would logistic regression predict for a new instance $\mathbf{x}^T = (1,3)$?

Further Reading

- Essential:
  - Iain Styles’ notes on Logistic Regression.

- Recommended:
  - Bishop’s book on “Machine Learning and Pattern Recognition”, chapter 4.3.2 (Logistic Regression) and 4.3.4 (Multiclass Logistic Regression).