LECTURE 8

SPRING 2021
APPLIED MACHINE LEARNING
CIHANG XIE

SLIDE CREDIT:
NARGES NOROUZI

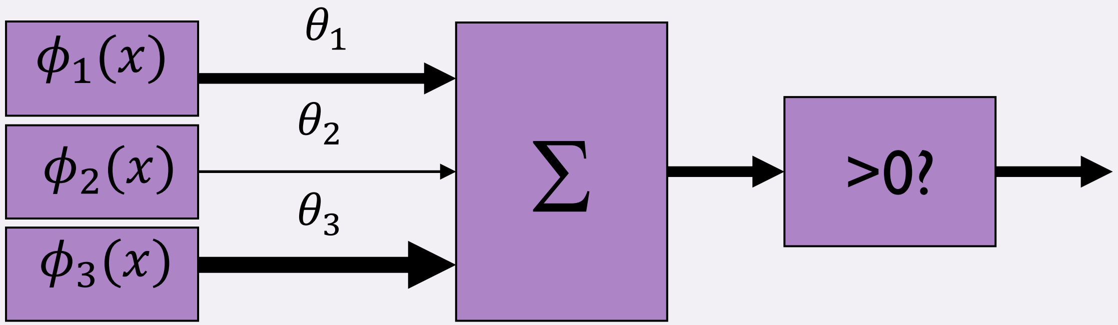
TODAY

- Review of Linear Classification
- Logistic Regression
  - Introduction
  - Maximum likelihood optimization
  - Softmax activation for multi-class classification

LINEAR CLASSIFIERS

- Inputs are feature values
- Each feature has a weight
- Weighted sum is the activation

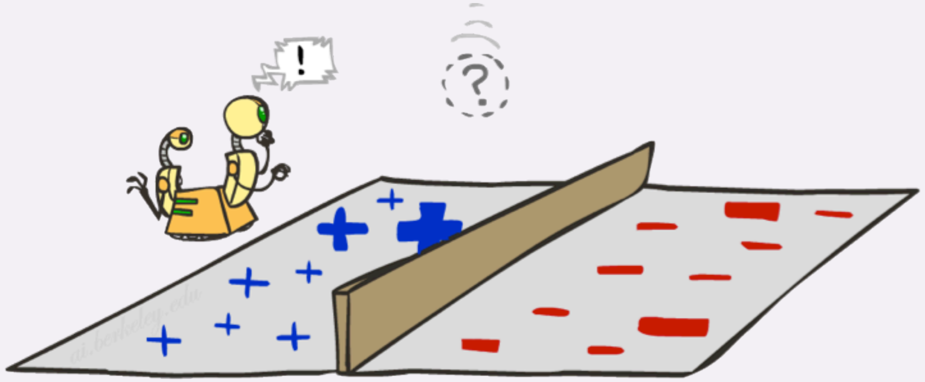
$$\text{activation}_\theta(x) = \sum_i \theta_i \phi_i(x) = \theta . \phi(x)$$



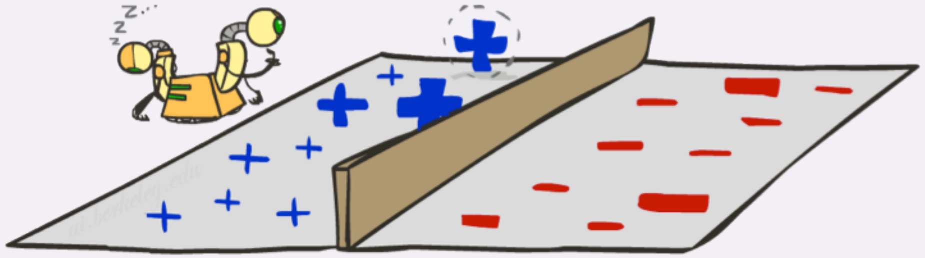
- If the activation is:
  - Positive, output +1
  - Negative, output -1

LEARNING: BINARY CLASSIFIER

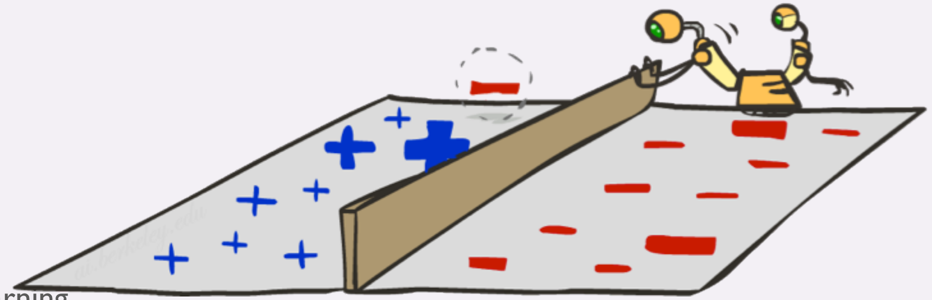
- Start with weights = 0
- For each training instance:
  - Classify with current weights



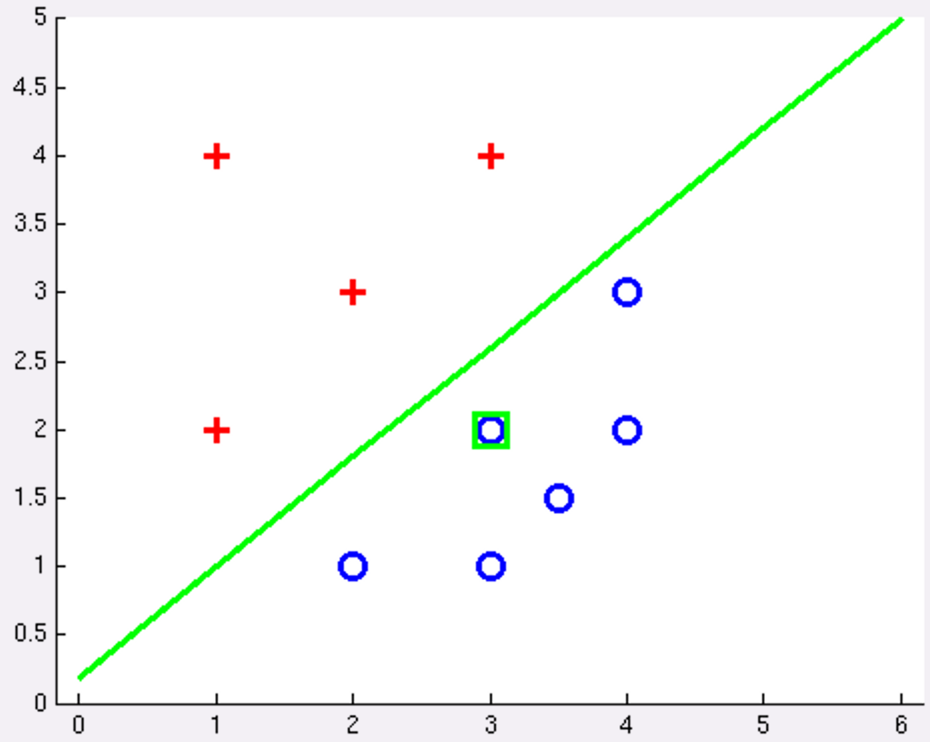
  - If correct (i.e., Y=Y*), no change!



  - If wrong: adjust the weight vector

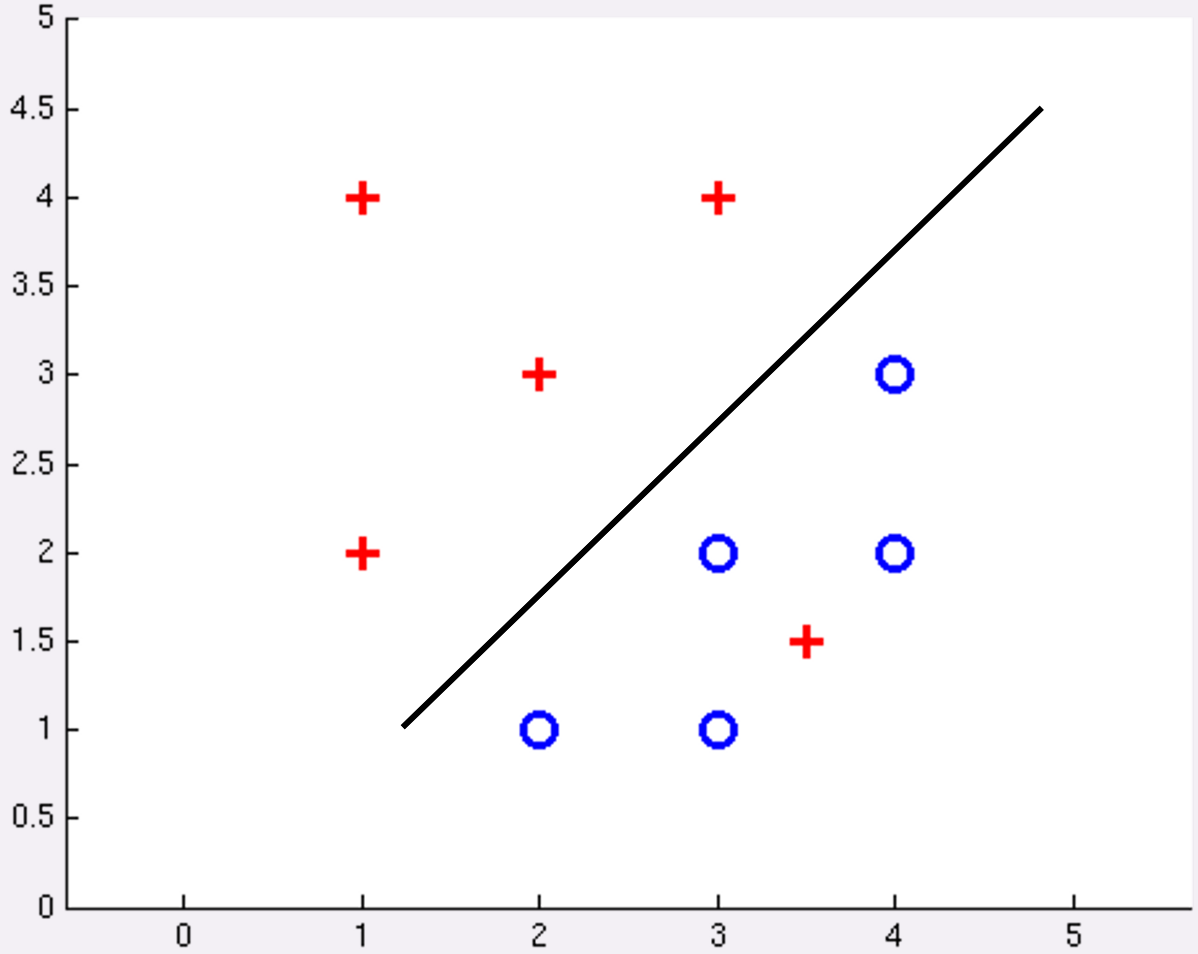


SEPARABLE CASE: DETERMINISTIC DECISION

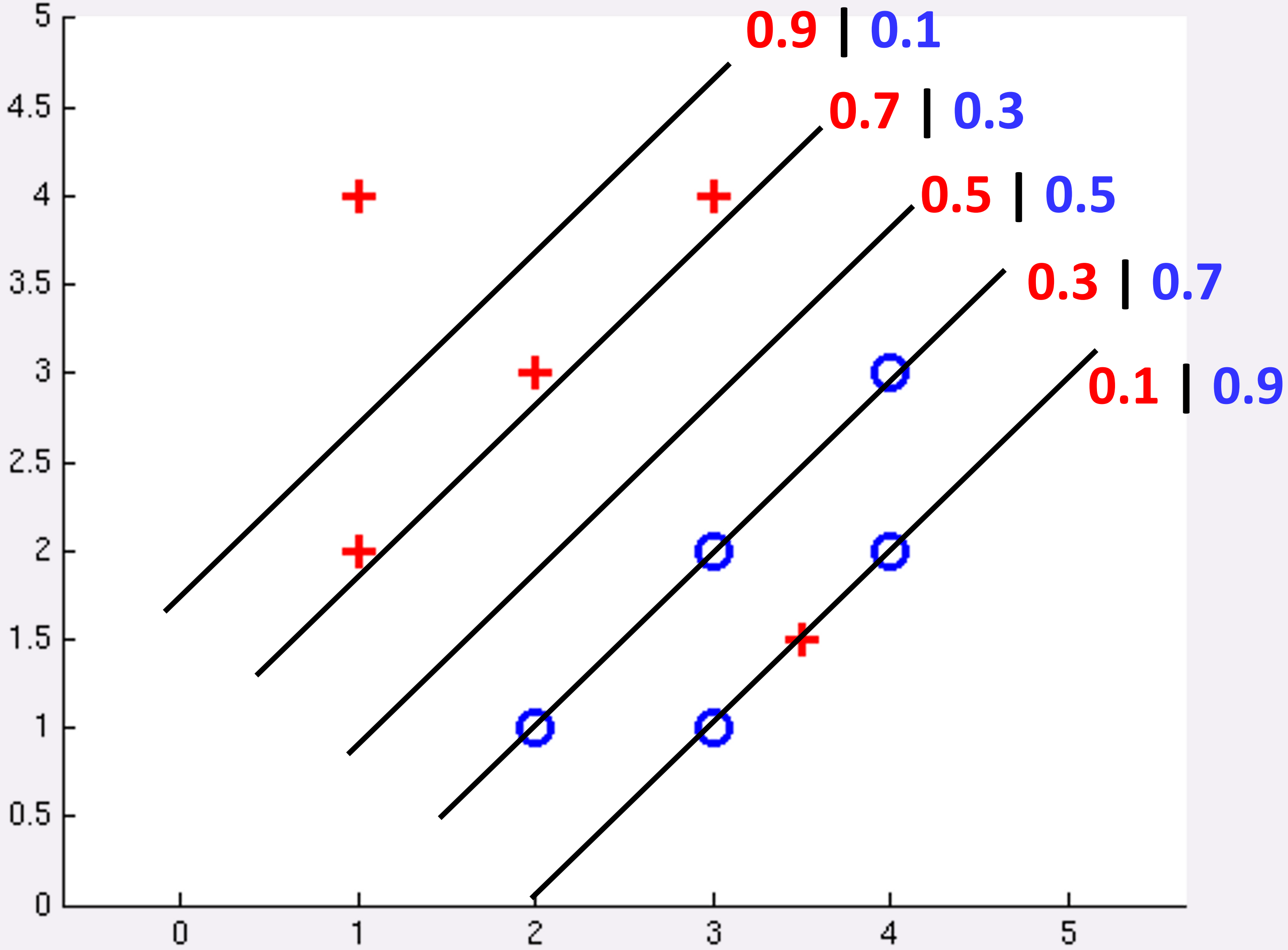


NON-SEPARABLE CASE: DETERMINISTIC DECISION

Even the best linear boundary makes at least one mistake



PROBABILISTIC DECISION



PROBABILISTIC CLASSIFICATION

PROBABILITY REVIEW

FROM PROBABILITY TO ODDS

- Another way of thinking about probabilities is to transform them using the function:

$$odds = \frac{p}{1 - p}$$

- This is the probability of something happening divided by the probability of it not happening.
- Similarly, if we were told that the odds of an event E are x to y, then

$$odds(E) = \frac{x}{y}$$

Which means

$$p(E) = \frac{x}{x+y}$$

FROM PROBABILITY TO ODDS

- If odds in favor of X solving a problem are 4 to 3 and odds **against** Y solving the same problem are 2 to 6. Find probability for:

  --- (i) X solving the problem

  --- (ii) Y solving the problem

- What's the range of possible values for the odds ratio?

THE LOGIT FUNCTION

- With one more transformation we can get a value that is unbounded over the real numbers.
- This is the logit function, it takes a value between 0 to 1 and maps it to a value between $-\infty$ and $+\infty$:

$$z = \log\left(\frac{p}{1-p}\right)$$

LOGISTIC FUNCTION

- Logit function:

$$z = \log_{\mathrm{e}}\left(\frac{p}{1-p}\right)$$

$$e^{z} = \frac{p}{1-p}$$

- Logistic function (inverse logit function):

$$p = \frac{1}{1+\ e^{-z}}$$

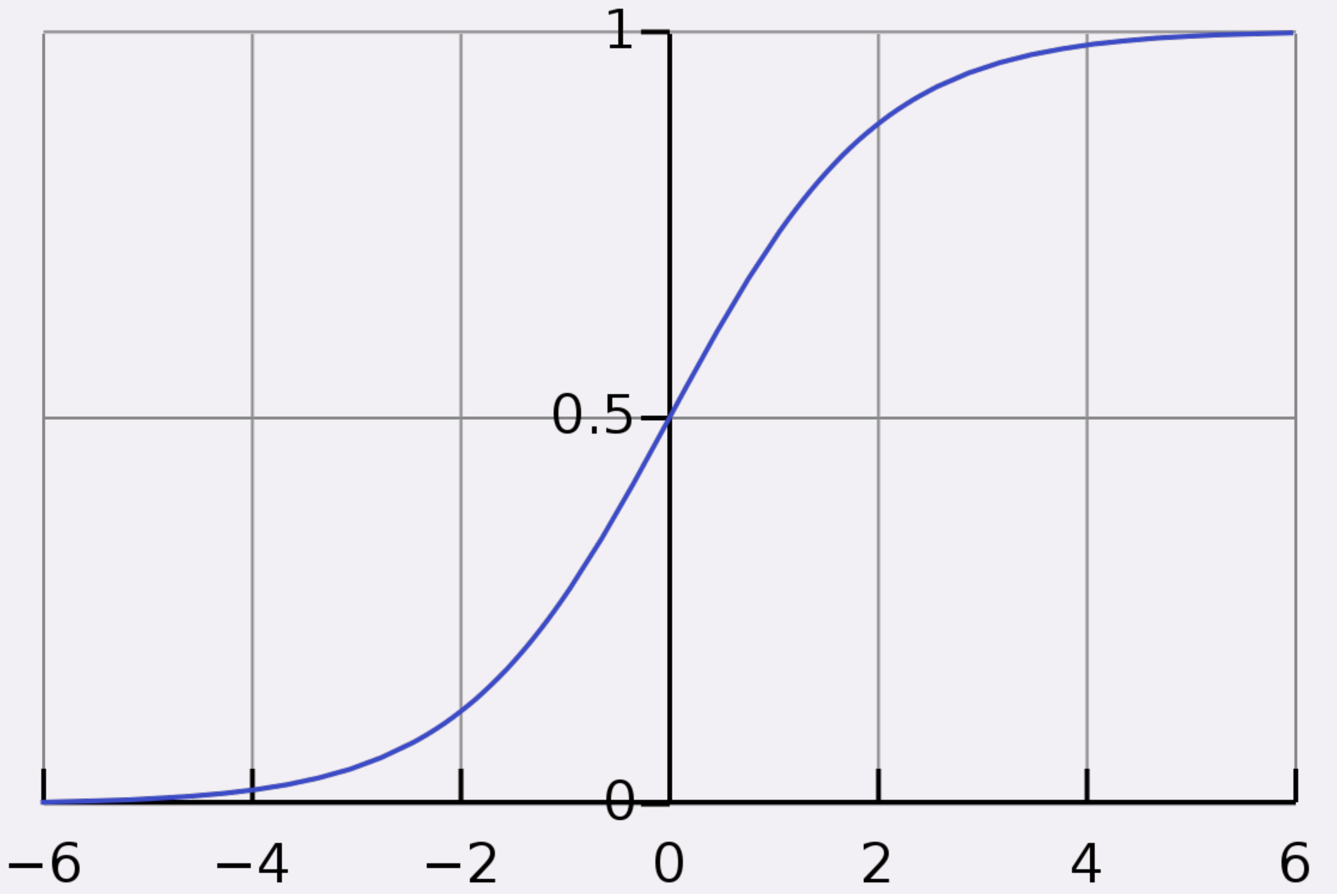
- The logistic function takes a value between $-\infty$ and $+\infty$ and maps it to a value between 0 and 1.

DIFFERENT WAYS OF EXPRESSING PROBABILITY

- Consider a two-outcome probability space, where:
  - $p(O_1) = p$
  - $p(O_2) = 1 - p = q$
- Can express probability O_1 as:

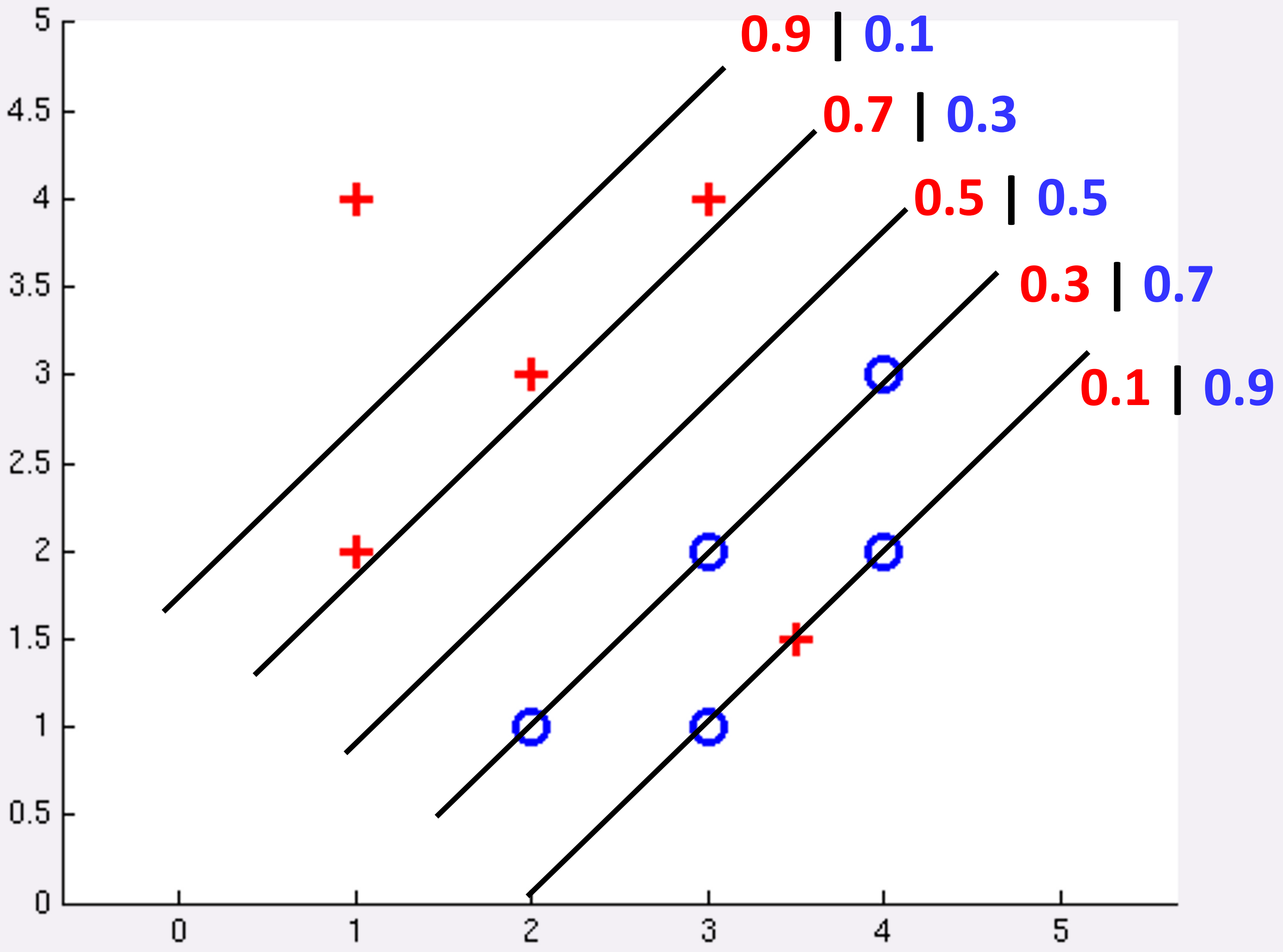
	notation	Range Equivalents		
Standard probability	p	0	0.5	1

LOGISTIC FUNCTION OR SIGMOID



LOGISTIC REGRESSION

PROBABILISTIC DECISION



LOGISTIC REGRESSION

- Name is somewhat misleading. Really a technique for classification, not regression

- Involves a more probabilistic view of classification.

USING A LOGISTIC REGRESSION MODEL

- Model consists of a vector θ in (d+1)-dimensional feature space
- For a point x in feature space, project it onto θ to convert it into a real number z in the range in the range $-\infty$ to $+\infty$

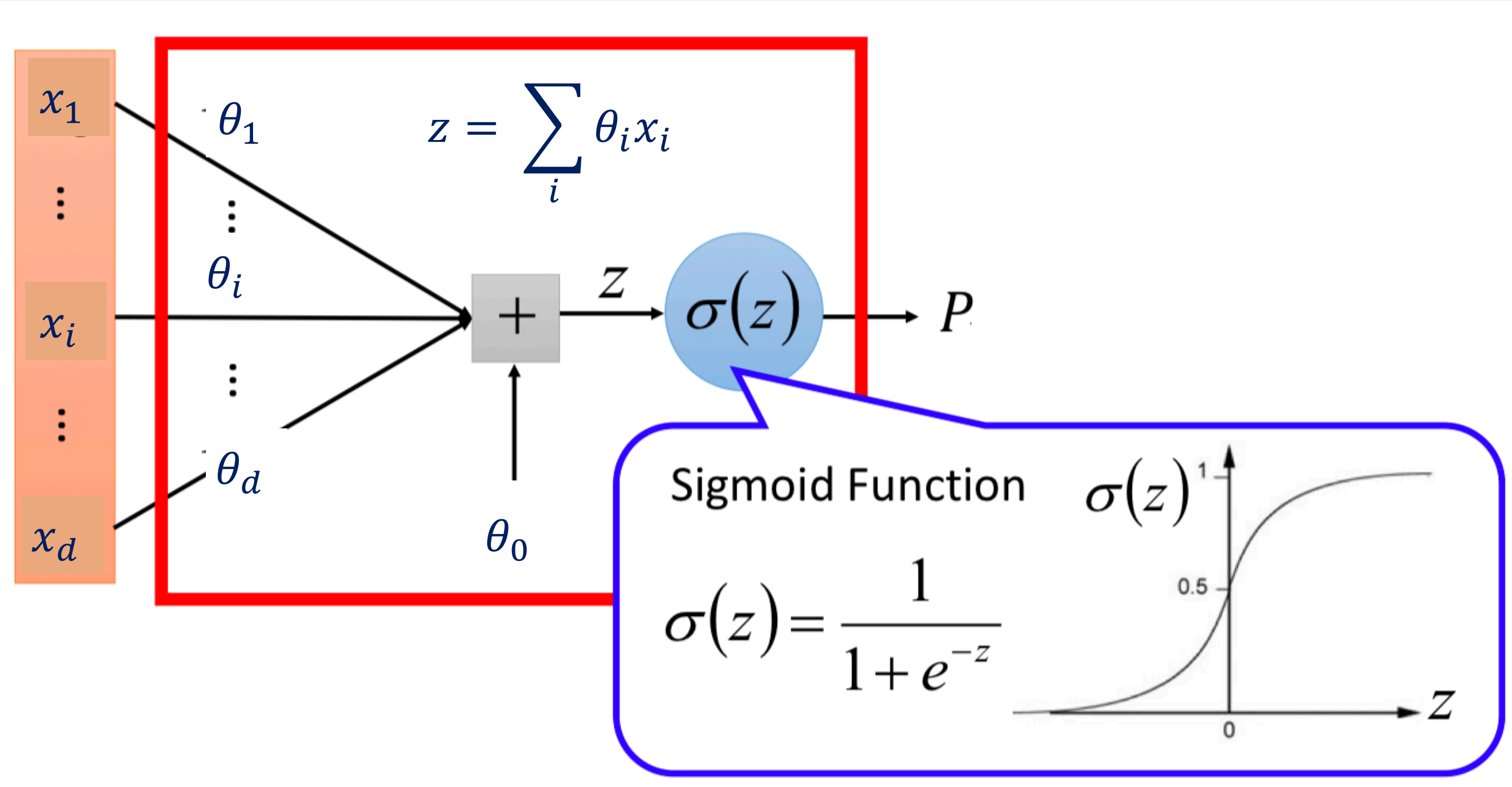
$$z = \theta_0 + \theta_1 x_1 + \ \dots + \theta_d x_d$$
$$z = \theta . x = \theta^T x$$

- Map z to the range 0 to 1 using the logistic function (sigmoid function)

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

- Overall, logistic regression maps a point in d-dimensional space to a value in the range 0 to 1.

LOGISTIC FUNCTION

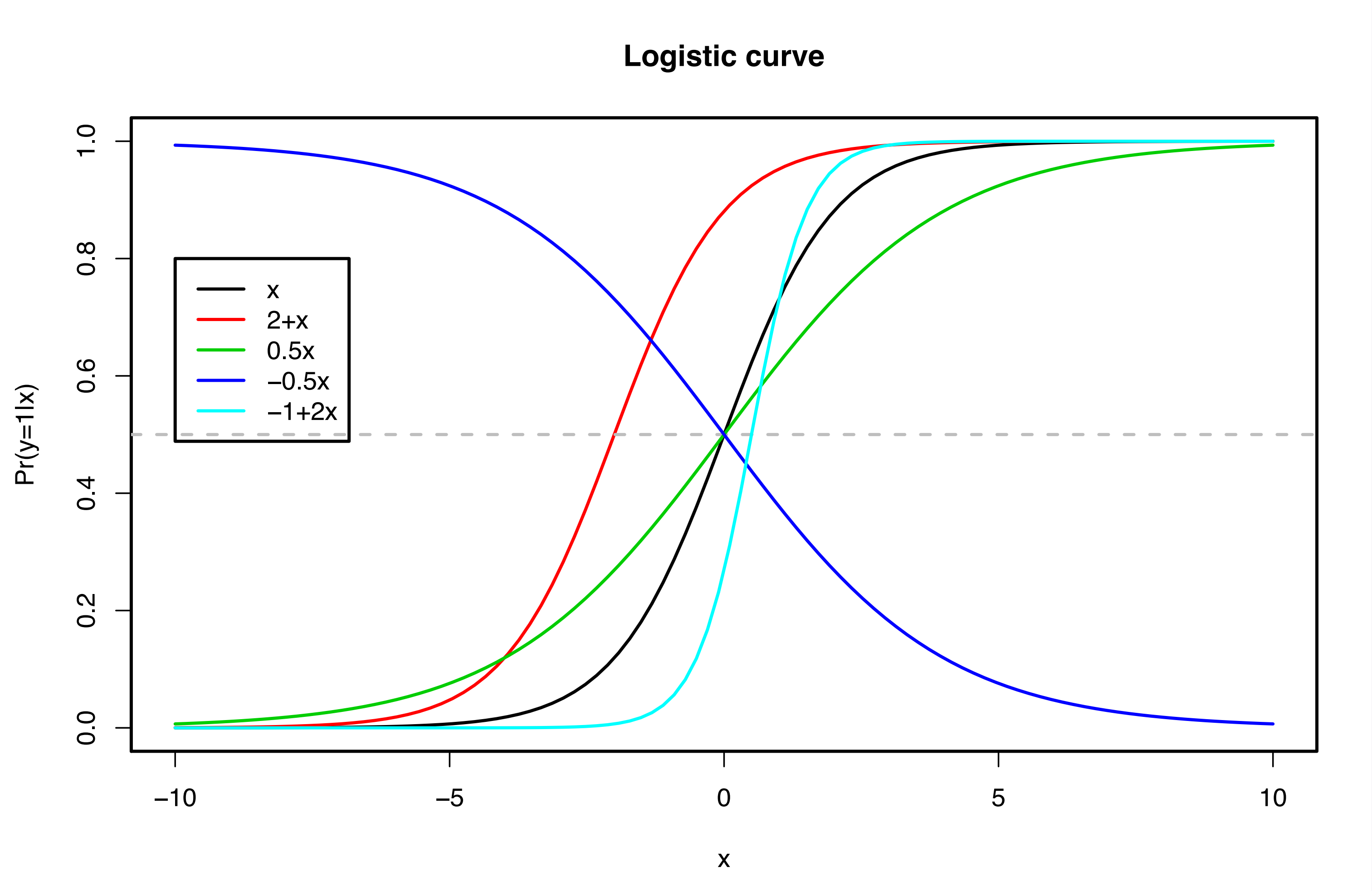


https://walkccc.github.io/CS/ML/5/

PROPERTIES OF LR

- One parameter per data dimension (feature) and a bias
- Features can be discrete or continuous
- Output of the model $y \in [0, 1]$
- Allows for gradient-based learning of parameters

SHAPE OF LOGISTIC FUNCTION



PROBABILISTIC INTERPRETATION

- If we have a value between 0 and 1, let's use it to model class probability:

$$p(C = 1|x) = \sigma(\theta^T x) \;\; with\, \sigma(z) = \frac{1}{1 + e^{-z}}$$

- Substituting we have

$$p(C = 1|x) = \frac{1}{1 + e^{-\theta^T x}}$$

- Suppose we have two classes, how can I compute $p(C = 0|x)$?
- Use the marginalization property of probability

$$p(C = 0|x) + p(C = 1|x) = 1$$

- Thus

$$p(C = 0|x) = 1 - \frac{1}{1 + e^{-\theta^T x}} = \frac{e^{-\theta^T x}}{1 + e^{-\theta^T x}}$$

BEST θ?

- Maximum likelihood estimation:

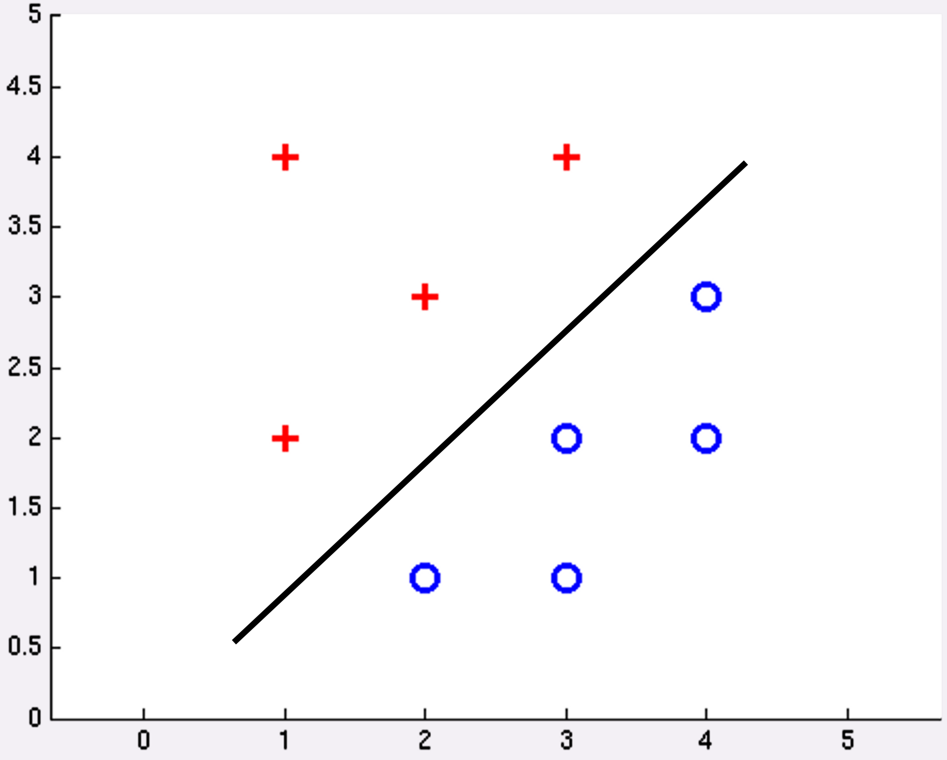
$$\max_{\theta} \quad ll(w) = \max_{\theta} \quad \sum_{i} \log P(y^{(i)} | x^{(i)}; \theta)$$

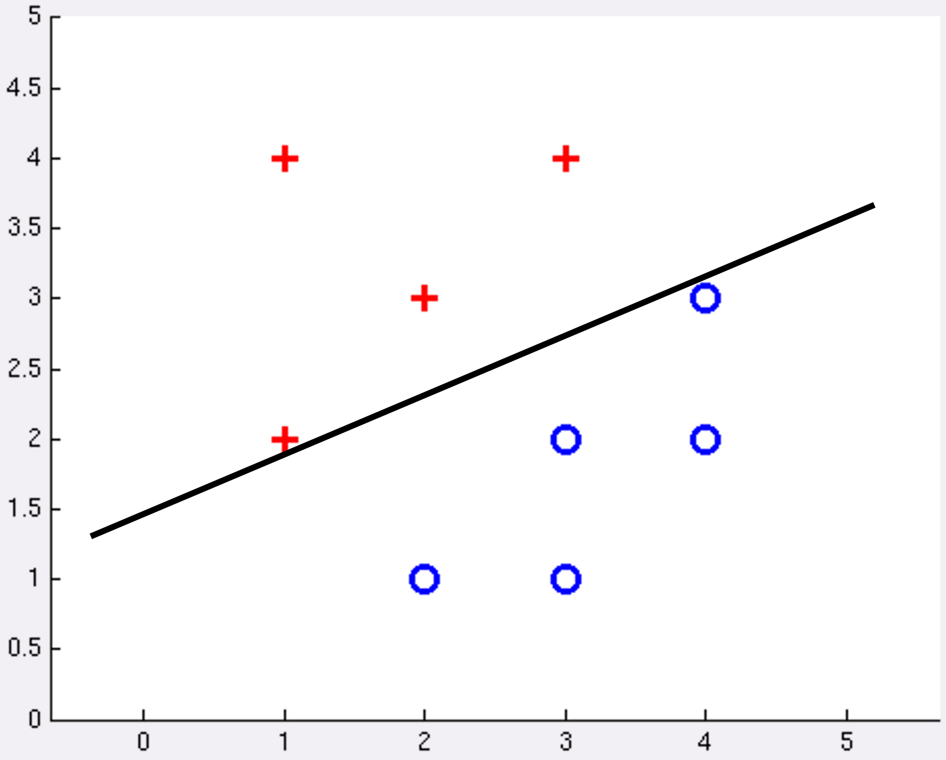
with:

$$P(y^{(i)} = +1 | x^{(i)}; \theta) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

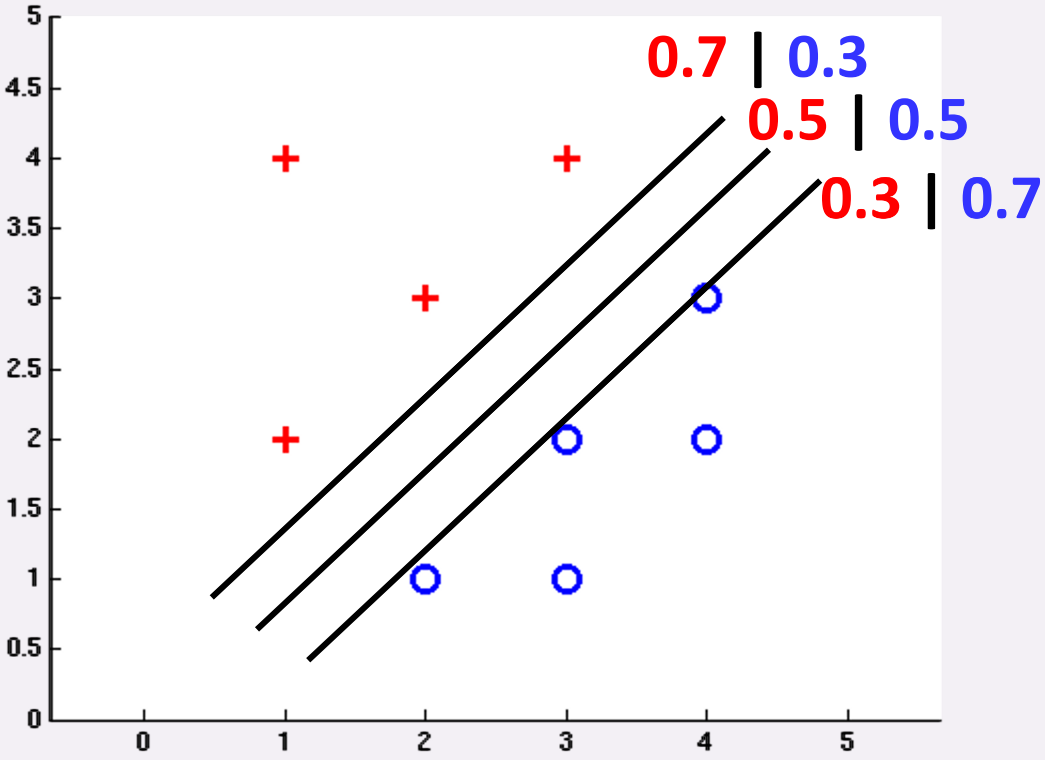
$$P(y^{(i)} = -1 | x^{(i)}; \theta) = 1 - \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

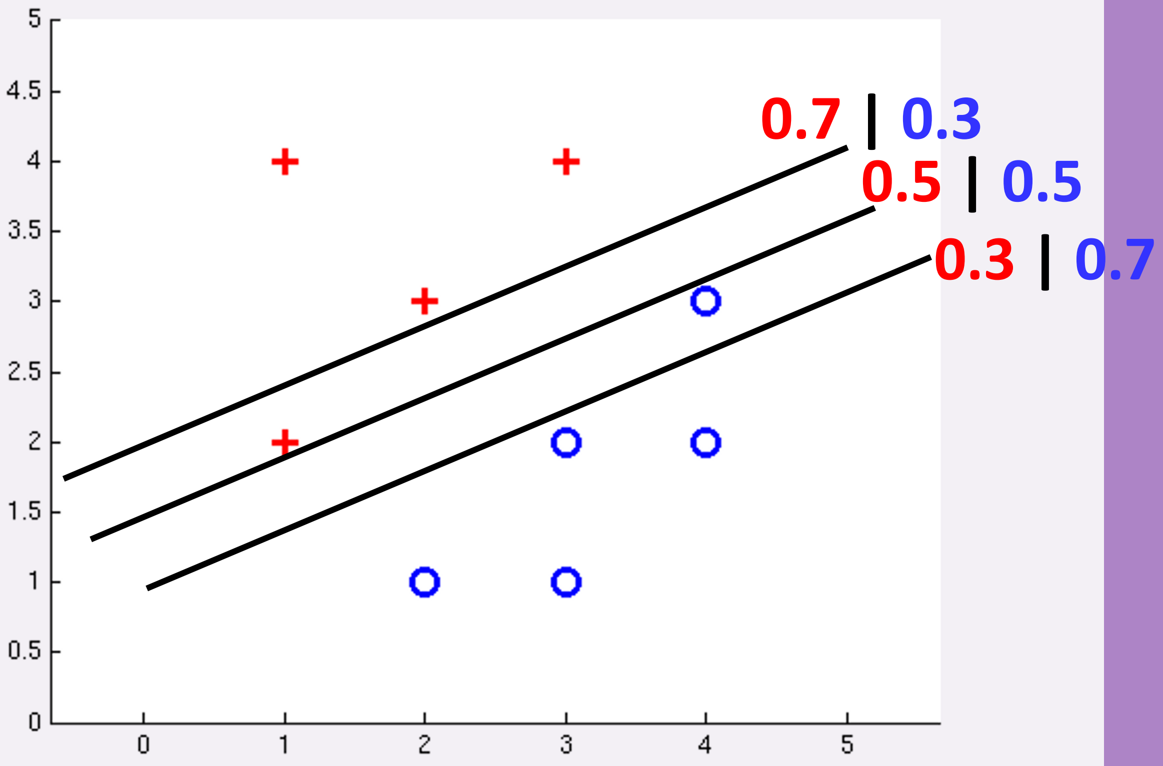
SEPARABLE CASE: DETERMINISTIC DECISION – MANY OPTIONS





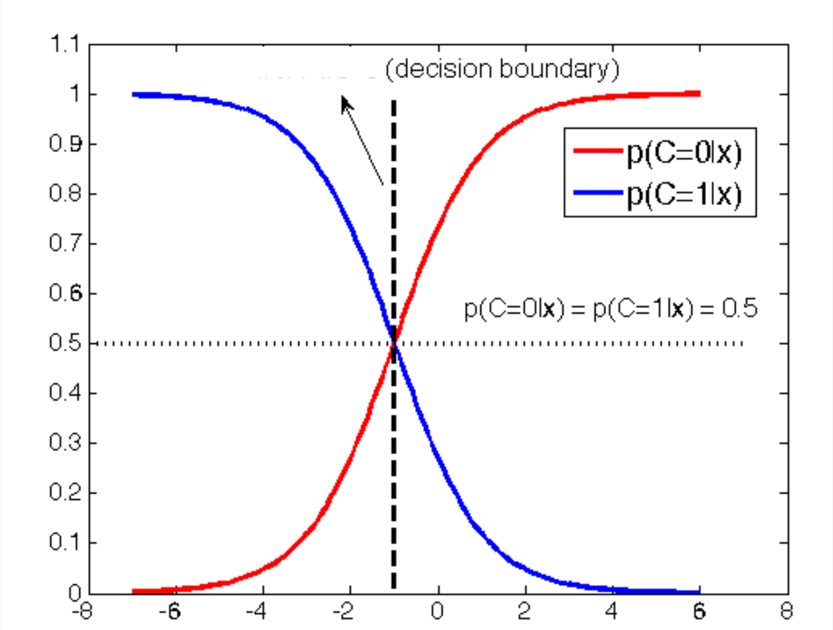
SEPARABLE CASE: PROBABILISTIC DECISION – CLEAR PREFERENCE



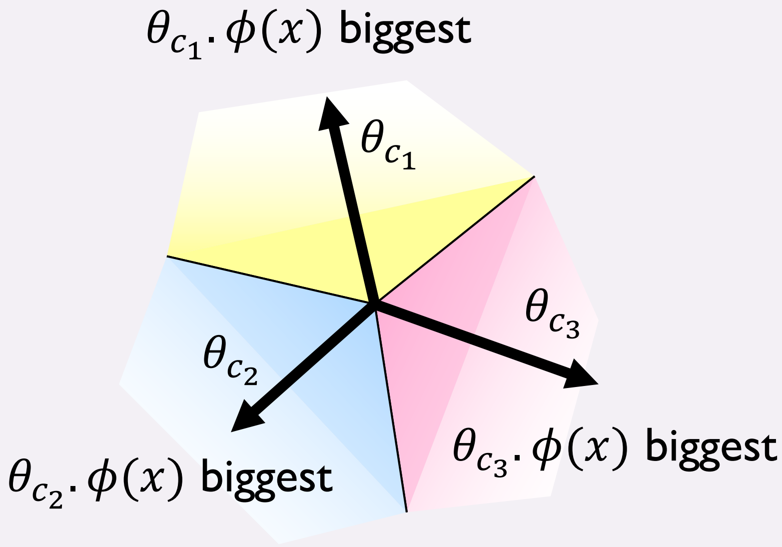


DECISION BOUNDARY FOR LR

- What is the *decision boundary* for logistic regression?
- $p(C = 1|x, \theta) = \frac{1}{1+e^{-\theta^T x}} = 0.5$
- $p(C = 0|x, \theta) = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} = 0.5$
- Decision boundary: $\theta^T x = 0$
- Logistic regression has a linear decision boundary



MULTICLASS PROBABILISTIC REGRESSION



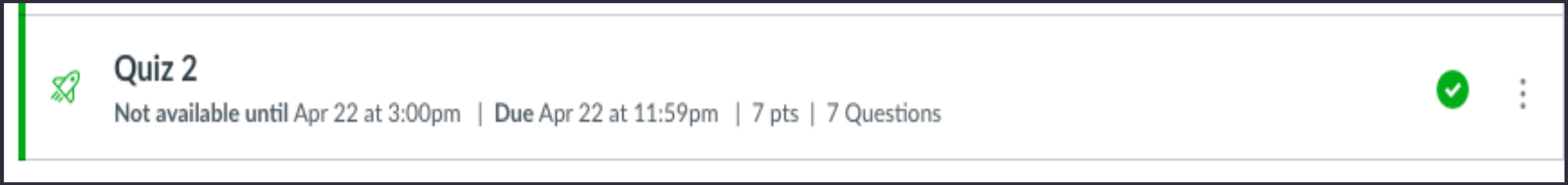
- Recall:
  - A weight vector for each class: θ_c
  - Score (activation) of a class c: $z_c = \theta_c.\phi(x)$
  - Prediction highest score wins

$$y = \underset{c}{\operatorname{argmax}}\, \theta_c.\phi(x)$$

- How to make the scores into probabilities?

$$\underbrace{z_1, z_2, z_3}_{\text{original activations}} \rightarrow \underbrace{\frac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}}, \frac{e^{z_2}}{e^{z_1}+e^{z_2}+e^{z_3}}, \frac{e^{z_3}}{e^{z_1}+e^{z_2}+e^{z_3}}}_{\text{softmax activations}}$$

QUIZ 2 (DUE THURSDAY)



QUESTIONS?