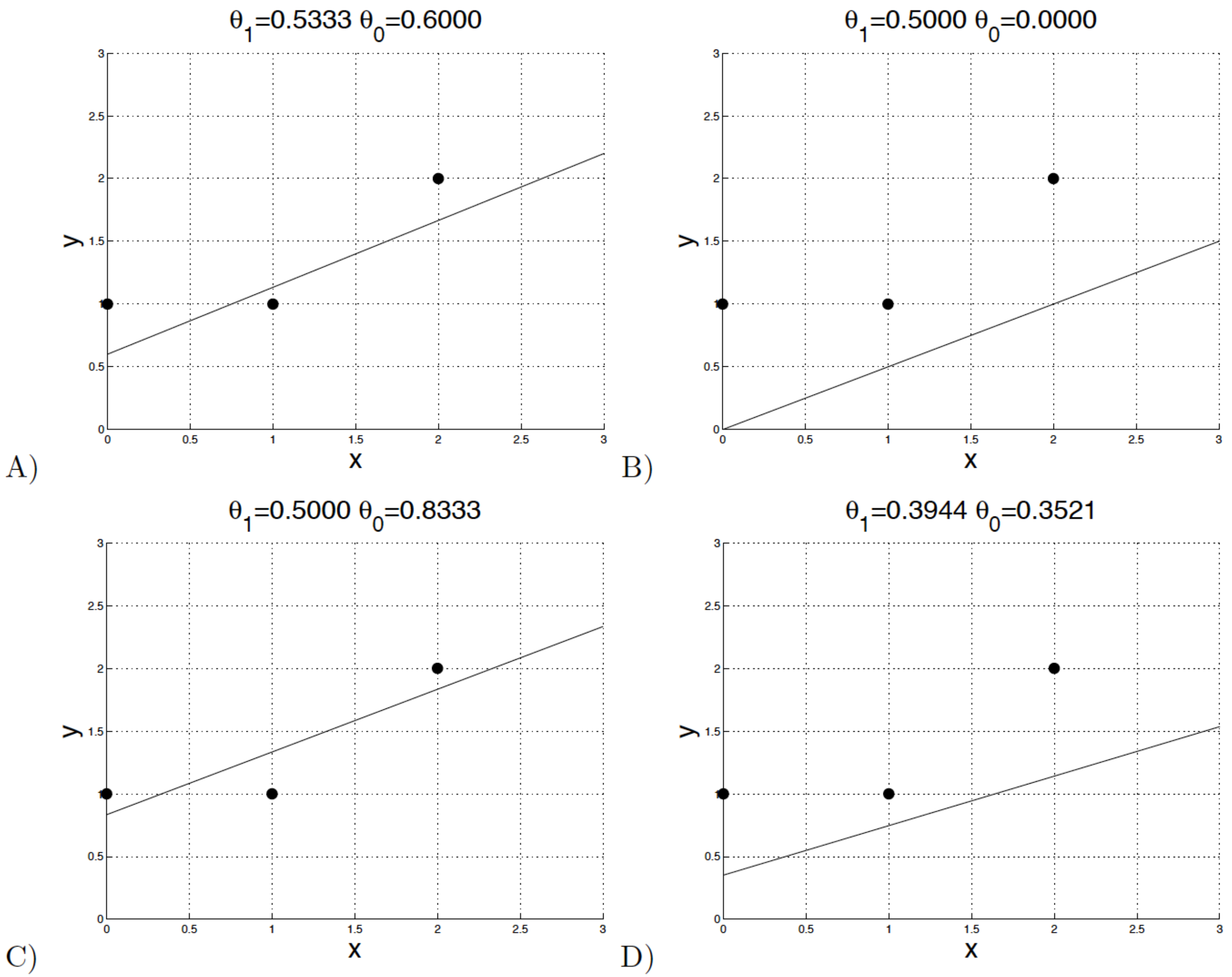
Exam preparation

Sandjai Bhulai
Vrije Universiteit Amsterdam

s.bhulai@vu.nl
19 October 2018

VU VRIJE UNIVERSITEIT AMSTERDAM
Faculty of Science

Linear regression



Linear regression

- Please assign each plot to one of the following regularization methods

- No regularization: $\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2$

- L2 regularization with $\lambda = 5$: $\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(\theta_1^2 + \theta_0^2)$

- L1 regularization with $\lambda = 5$: $\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(|\theta_1| + |\theta_0|)$

- L2 regularization with $\lambda = 1$: $\sum_{i=1}^{3}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(\theta_1^2 + \theta_0^2)$

Linear regression

- Please assign each plot to one of the following regularization methods

- No regularization: C

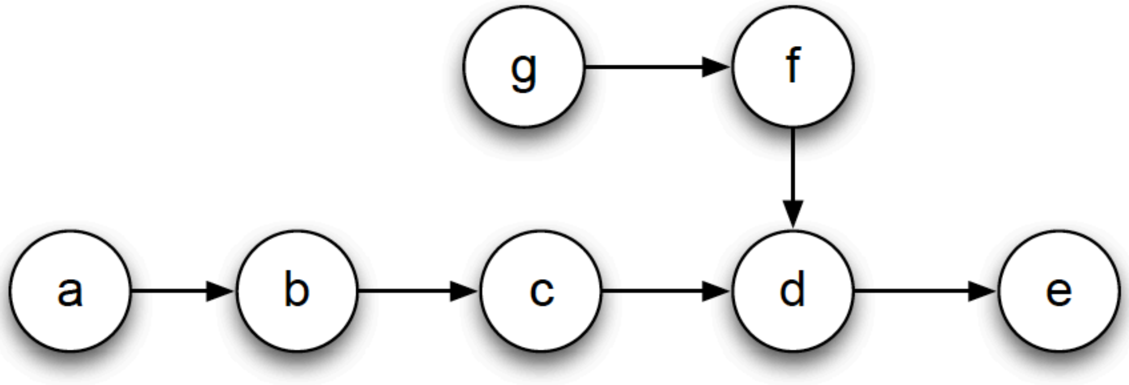
- L2 regularization with $\lambda = 5$: D

- L1 regularization with $\lambda = 5$: B

- L2 regularization with $\lambda = 1$: A

Graphical models

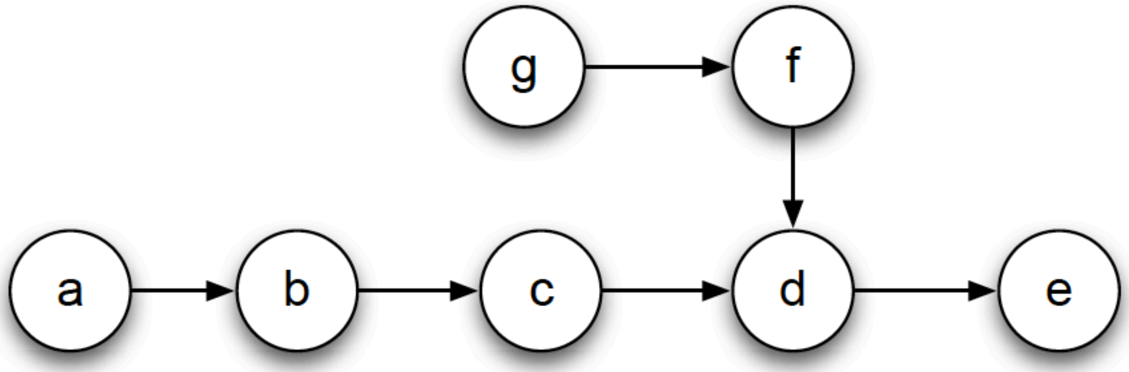
- Consider the following graphical model:



- Write the expression for the joint likelihood of the network in its factored form.
- Let $X = \{c\}, Y = \{b, d\}, Z = \{a, e, f, g\}$. Is X conditionally independent of Z given Y? If yes, explain why. If no, show a path that is not blocked.

Graphical models

- Consider the following graphical model:

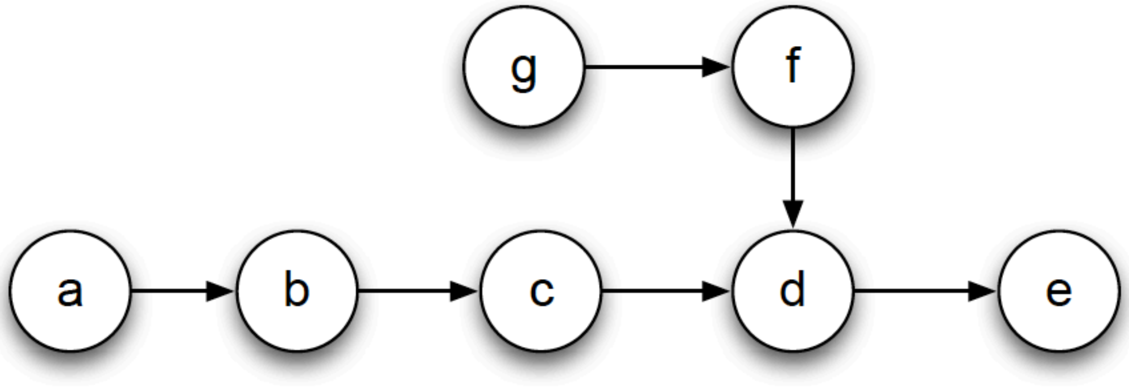


- Write the expression for the joint likelihood of the network in its factored form.

$$p(a,b,c,d,e,f,g) = p(a)\,p(b\,|\,a)\,p(c\,|\,b)\,p(g)\,p(f\,|\,g)\,p(d\,|\,c,f)\,p(e\,|\,d)$$

Graphical models

- Consider the following graphical model:



- Let $X = \{c\}, Y = \{b, d\}, Z = \{a, e, f, g\}$. Is X conditionally independent of Z given Y? If yes, explain why. If no, show a path that is not blocked.
- No, the path $c \rightarrow d \rightarrow f$ is not blocked.

Decision trees

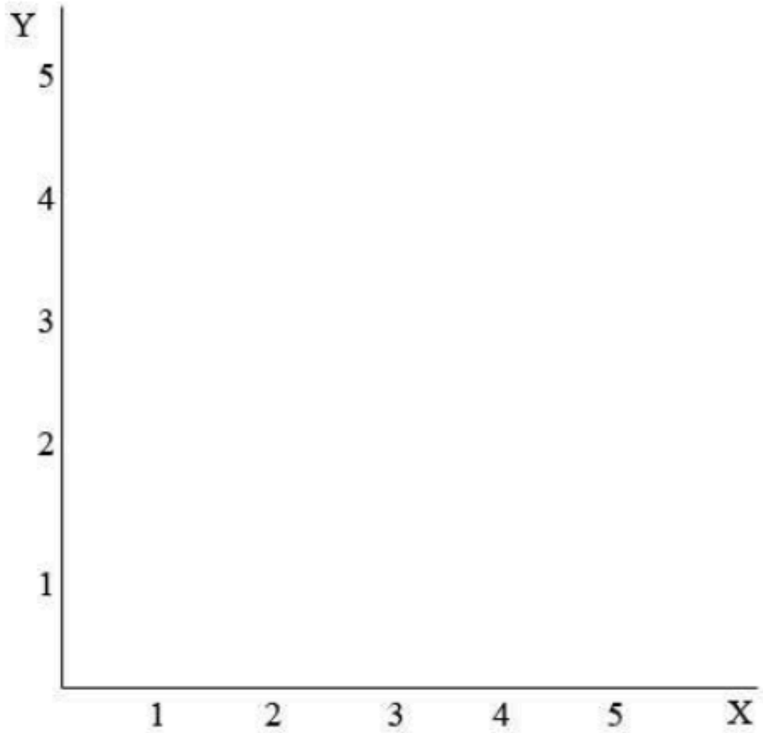
- In class, we discussed greedy algorithms for learning decision trees from training data.

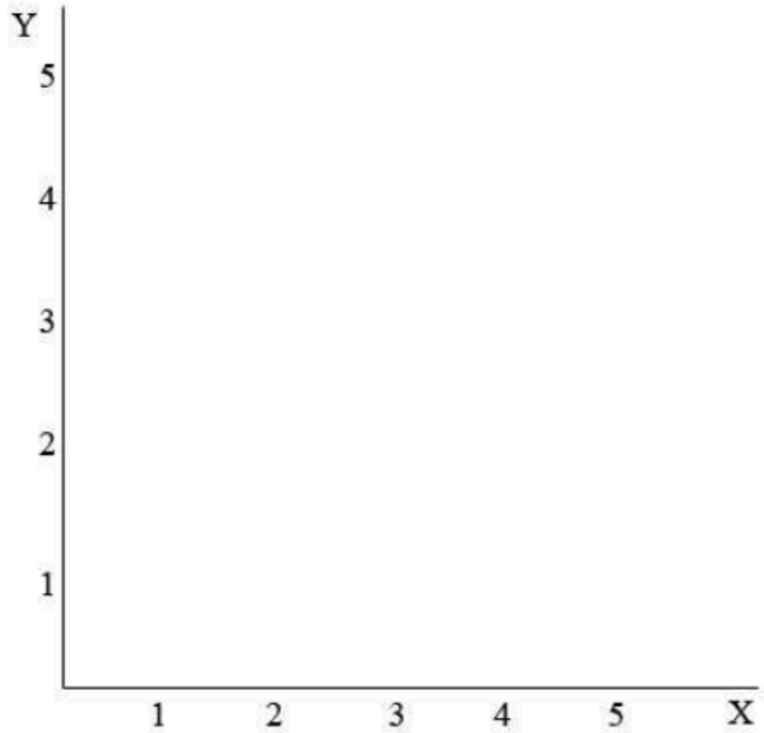
- In a **standard decision tree**, each level of the recursion will find one decision boundary that partitions the feature space into two regions. Each region is then partitioned recursively using the same procedure.

- In a **point-based look-ahead decision tree**, the feature space is partitioned into four regions by a single point. E.g., if the point is $(X, Y) = (3,4)$, this gives the regions $[X > 3, Y > 4]$, $[X > 3, Y \leq 4]$, $[X \leq 3, Y > 4]$, $[X \leq 3, Y \leq 4]$

Decision trees

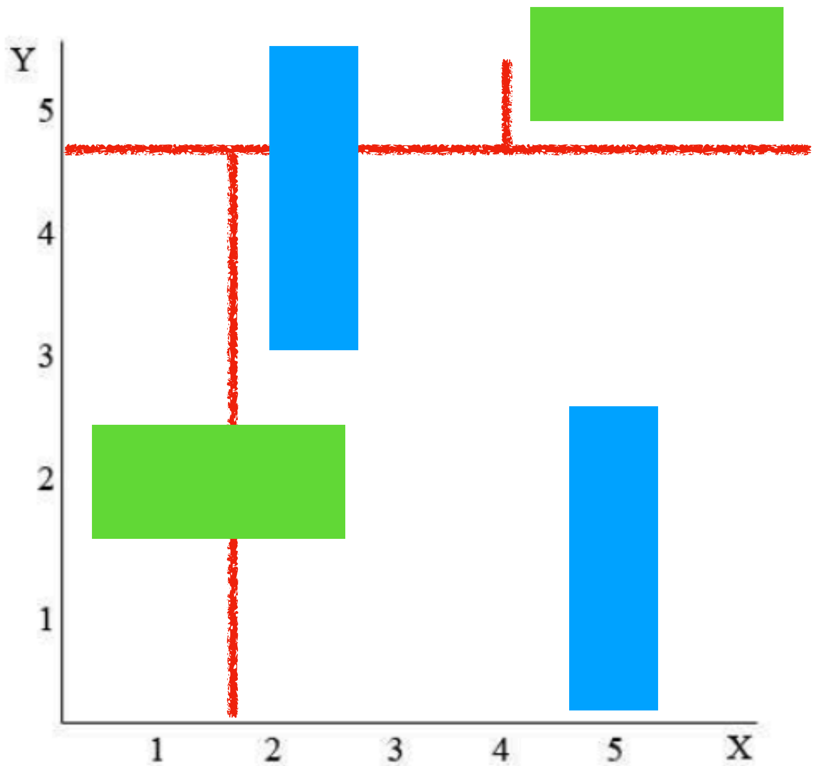
- Draw a dataset so that a standard decision tree with four regions will poorly classify the data, but the point-based look-ahead decision tree will perfectly classify the data.

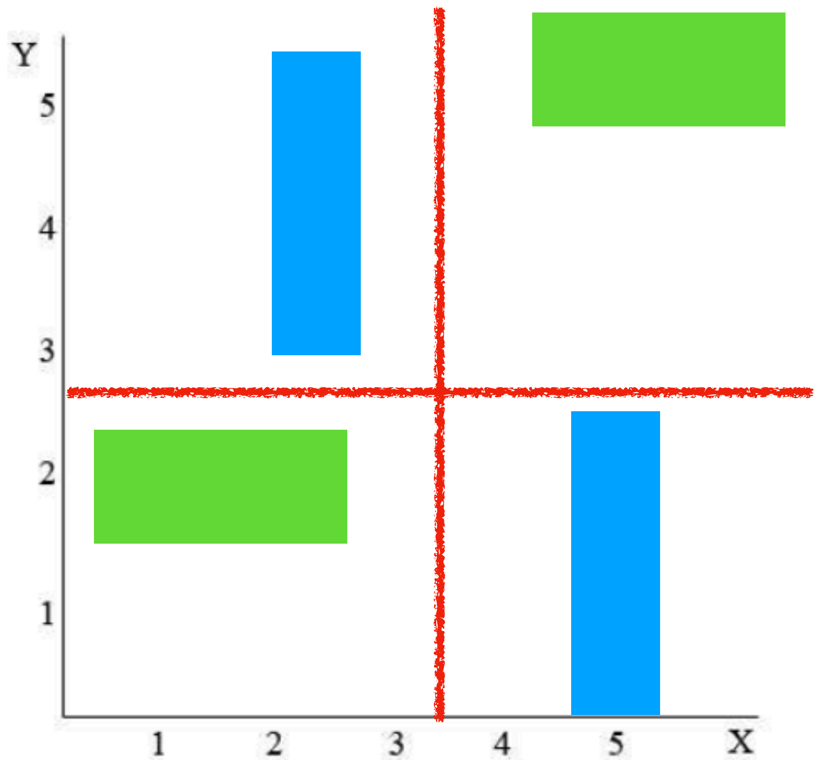




Decision trees

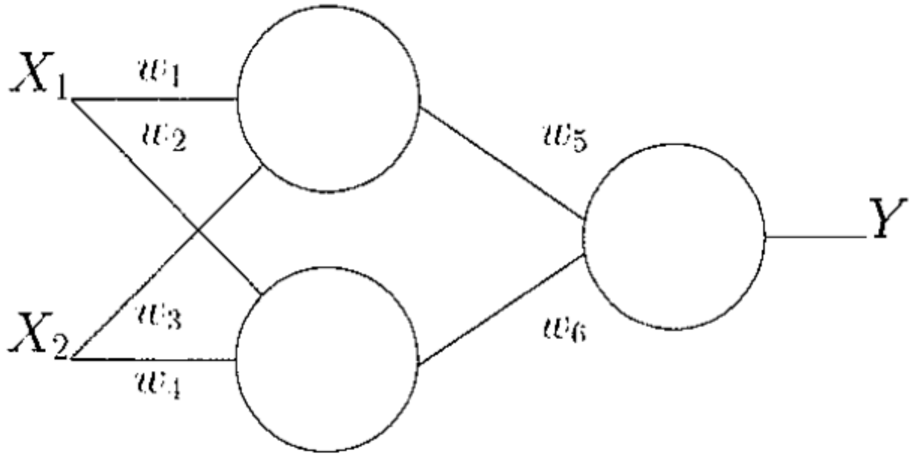
- Draw a dataset so that a standard decision tree with four regions will poorly classify the data, but the point-based look-ahead decision tree will perfectly classify the data.





Neural networks

- Suppose that we have a neural network (shown below) with linear activation units. In other words, the output of each unit is determined by the activation function $g(x) = cx$



- Can any function that is represented by the network also be represented by a single unit neural network? If so, please provide the weights and the activation function.

Neural networks

- Can any function that is represented by the network also be represented by a single unit neural network? If so, please provide the weights and the activation function.

- Yes, take as weights $w_1 w_5 + w_2 w_6$ and $w_3 w_5 + w_4 w_6$ with activation function $g(x) = c^2 x$

- Can the space of functions that is represented by the network be represented by linear regression?

Neural networks

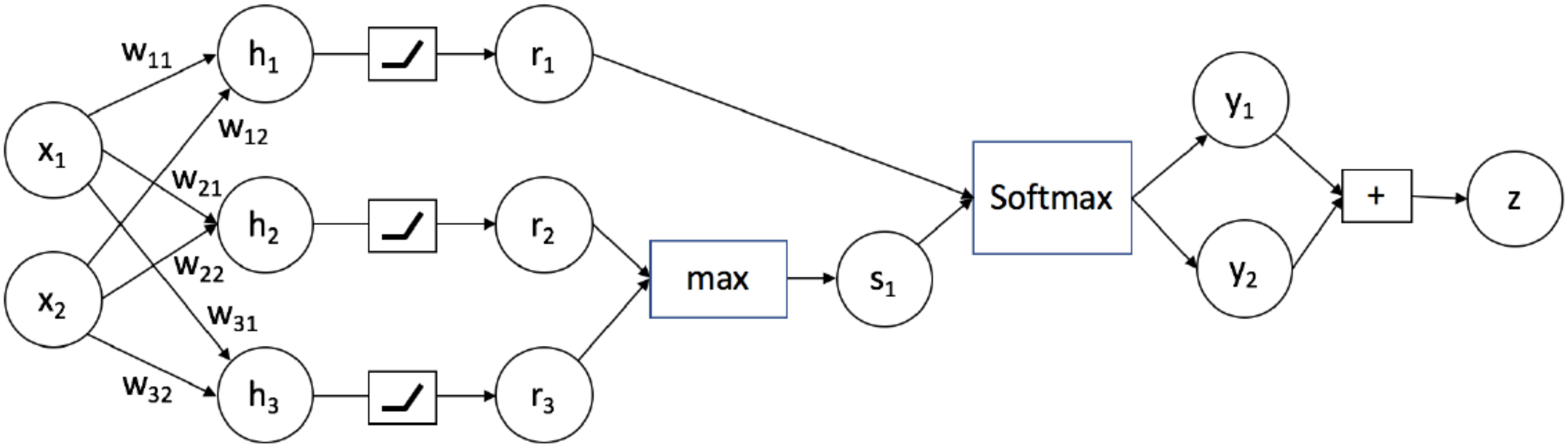
- Can the space of functions that is represented by the network be represented by linear regression?

- Yes, the functions in the network have the form

$$y = c^2(w_1w_5 + w_2w_6)x_1 + c^2(w_3w_5 + w_4w_6)x_2 = \beta_1x_1 + \beta_2x_2$$

Deep learning

- Consider the network below with inputs x_1, x_2



$$h_1 = w_{11}x_1 + w_{12}x_2 \qquad h_2 = w_{21}x_1 + w_{22}x_2 \qquad h_3 = w_{31}x_1 + w_{32}x_2$$

$$r_1 = \max(h_1, 0) \qquad r_2 = \max(h_2, 0) \qquad r_3 = \max(h_3, 0)$$

$$s_1 = \max(r_2, r_3)$$

$$y_1 = \frac{\exp(r_1)}{\exp(r_1) + \exp(s_1)} \qquad y_2 = \frac{\exp(s_1)}{\exp(r_1) + \exp(s_1)}$$

$$z = y_1 + y_2$$

Deep learning

- Compute the values of the internal nodes given

$$x_1 = 1, x_2 = -2, w_{11} = 6, w_{12} = 2, w_{21} = 4, w_{22} = 7, w_{31} = 5, w_{32} = 1$$

h_1	h_2	h_3	r_1	r_2

r_3	s	y_1	y_2	z

Deep learning

- Compute the values of the internal nodes given

$$x_1 = 1, x_2 = -2, w_{11} = 6, w_{12} = 2, w_{21} = 4, w_{22} = 7, w_{31} = 5, w_{32} = 1$$

h_1	h_2	h_3	r_1	r_2
2	-10	3	2	0

r_3	s	y_1	y_2	z
3	3	$\frac{1}{1+e}$	$\frac{e}{1+e}$	1

Deep learning

- Compute the following gradients analytically.

$\frac{\partial h_1}{\partial w_{12}}$	$\frac{\partial h_1}{\partial x_1}$	$\frac{\partial r_1}{\partial h_1}$	$\frac{\partial y_1}{\partial r_1}$

$\frac{\partial y_1}{\partial s_1}$	$\frac{\partial z}{\partial y_1}$	$\frac{\partial z}{\partial x_1}$	$\frac{\partial s_1}{\partial r_2}$

Deep learning

- Compute the following gradients analytically.

$\frac{\partial h_1}{\partial w_{12}}$	$\frac{\partial h_1}{\partial x_1}$	$\frac{\partial r_1}{\partial h_1}$	$\frac{\partial y_1}{\partial r_1}$
x_2	w_{11}	$1[h_1 > 0]$	$y_1(1 - y_1)$

$\frac{\partial y_1}{\partial s_1}$	$\frac{\partial z}{\partial y_1}$	$\frac{\partial z}{\partial x_1}$	$\frac{\partial s_1}{\partial r_2}$
$-y_1 y_2$	1	0	$1[r_2 > r_3]$

Support vector machines

- Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are (1,1) and (-1, -1). The negative examples are (1,-1) and (-1,1).

- Are the positive examples linearly separable from the negative examples in the original space?

Support vector machines

- Are the positive examples linearly separable from the negative examples in the original space?

- No

- Consider the feature transformation $\varphi(x) = [1, x_1, x_2, x_1 x_2]$, where x_1 and x_2 are the first and second coordinates of a general example. The prediction function is $y(x) = w^T \varphi(x)$ in this feature space. Give the coefficients w of a maximum-margin decision surface sparating the positive examples from the negative examples. You should be able to do this by inspection, without any significant computation.

Support vector machines

- Consider the feature transformation $\varphi(x) = [1, x_1, x_2, x_1 x_2]$, where x_1 and x_2 are the first and second coordinates of a general example. The prediction function is $y(x) = w^T \varphi(x)$ in this feature space. Can you linearly separate the examples now. If so, how?. You should be able to do this by inspection, without any significant computation.

- The product $x_1 x_2$ is 1 for the positive example and -1 for the negative examples.

- What kernel $K(x, x')$ does this feature transformation φ correspond to?

Support vector machines

- What kernel $K(x, x')$ does this feature transformation φ correspond to?

$$1 + x_1 x_1' + x_2 x_2' + x_1 x_2 x_1' x_2'$$

Reinforcement learning

- Imagine an unknown game which has only two states $\{A, B\}$ and in each state the agent has two actions to choose from: {Up, Down}. Suppose a game agent chooses actions according to some policy π and generate the following sequence of actions and rewards in the unknown game:

t	s_t	a_t	s_{t+1}	r_t
0	A	Down	B	2
1	B	Down	B	-4
2	B	Up	B	0
3	B	Up	A	3
4	A	Up	A	-1

Reinforcement learning

- Recall the update function of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

Assume that all Q-values are initialised as 0, the discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.5$. What are the following Q-values learned by running Q-learning with the experience sequence given by the table.

- Q(A, Down) = ?
- Q(B, Up) = ?

Reinforcement learning

- Recall the update function of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

Assume that all Q-values are initialised as 0, the discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.5$. What are the following Q-values learned by running Q-learning with the experience sequence given by the table.

- Q(A, Down) = 1
- Q(B, Up) = 7/4