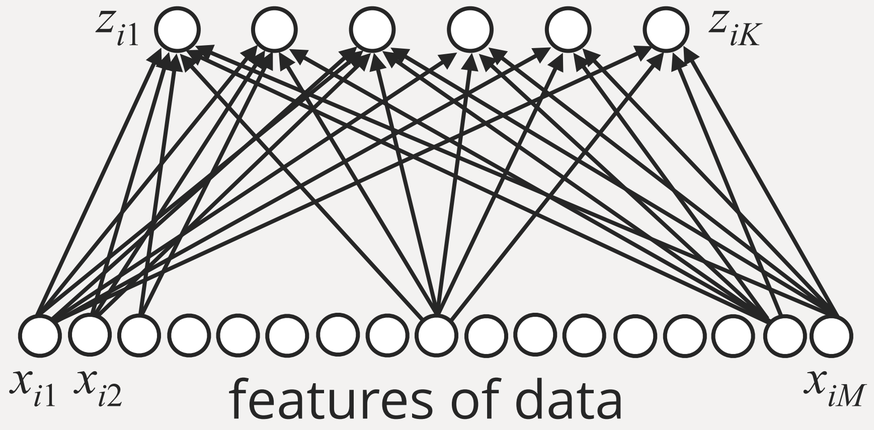
Interpretation of Multilayer Perceptron



topics

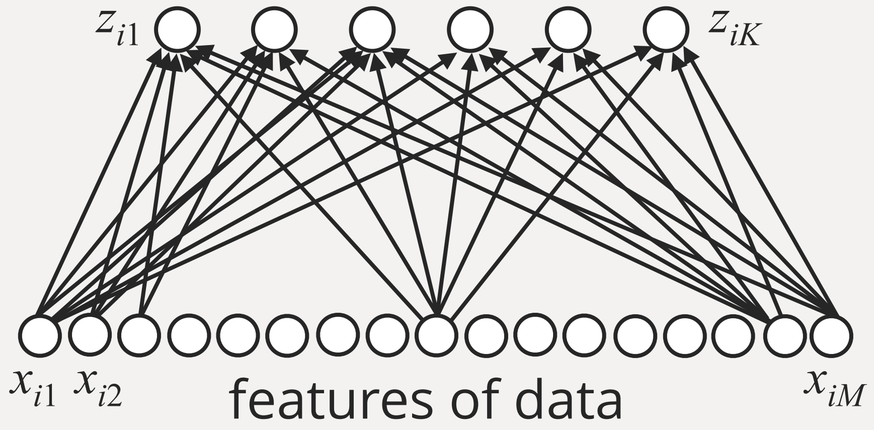
b_1	b_2	•	•	b_K

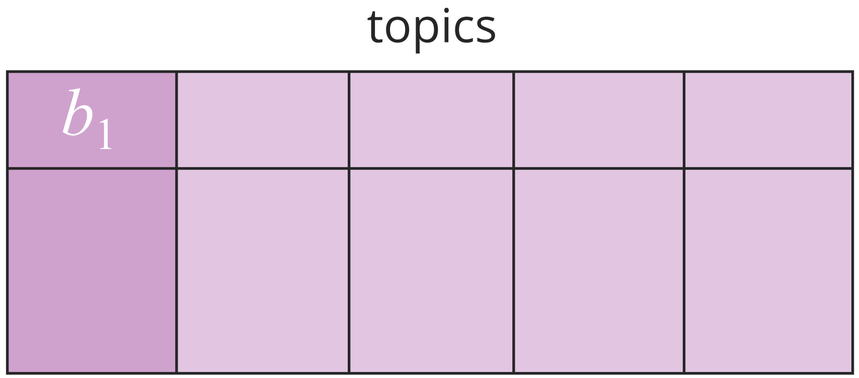
$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$\vdots$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$



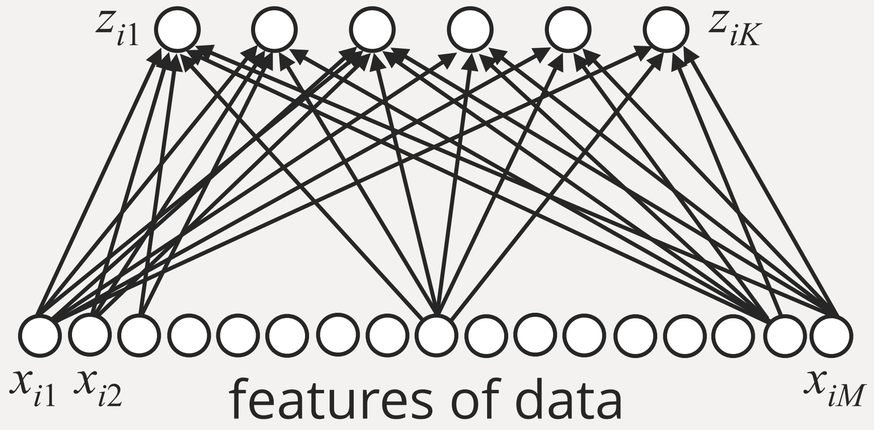


$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$\vdots$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$



topics

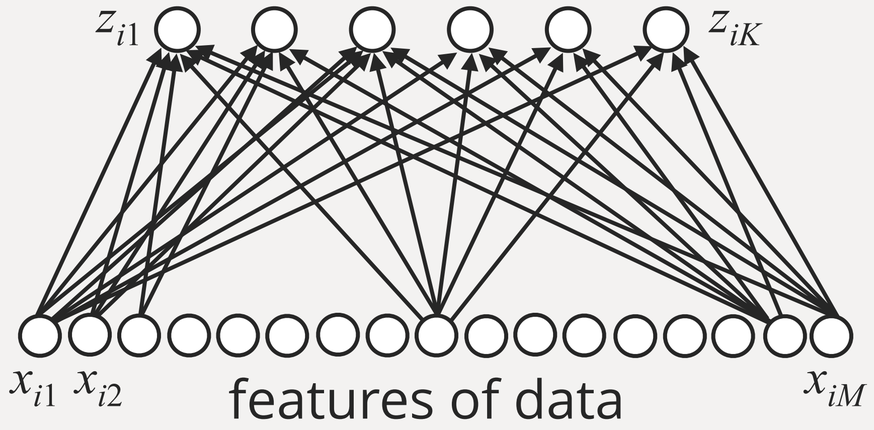
b_1				
sports				

$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$\vdots$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$



topics

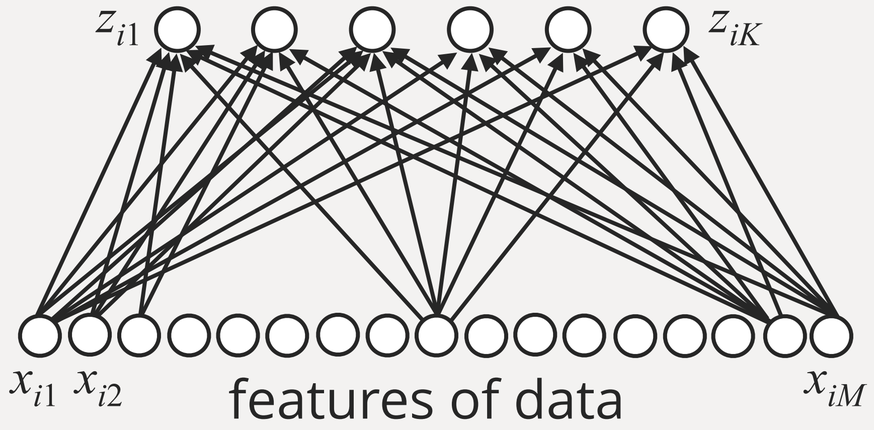
b_1	b_2			
sports	history			

$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$\vdots$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$



topics

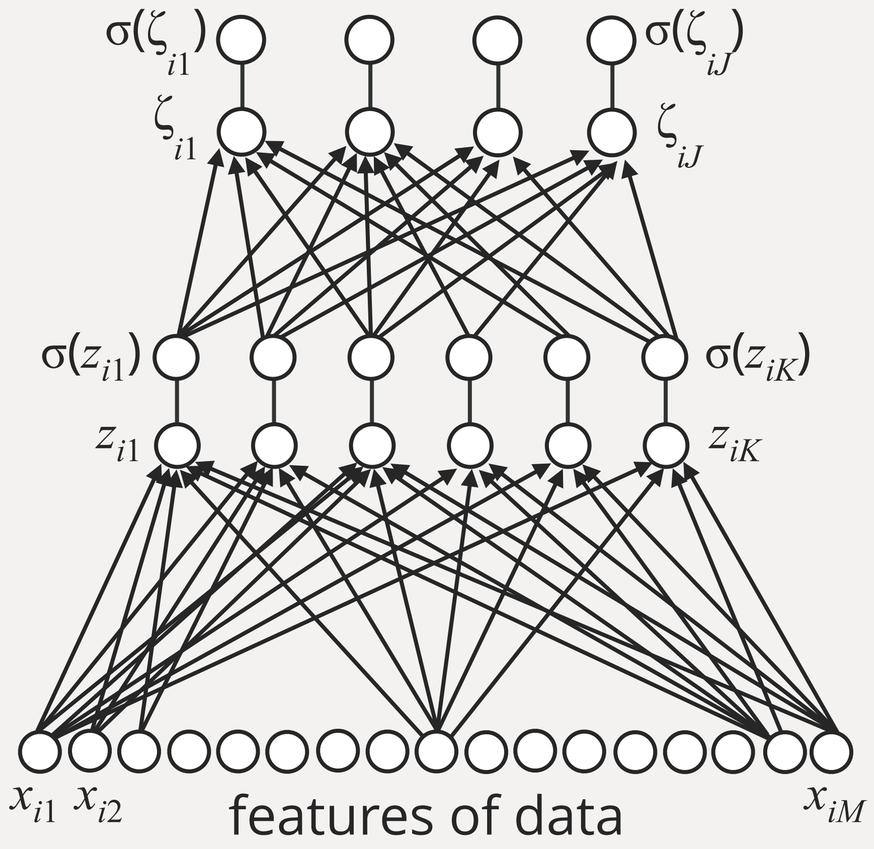
b_1	b_2	•	•	b_K
sports	history	•	•	politics

$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$\vdots$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$



meta-topics

c_1	c_2	•	•	c_J

topics

b_1	b_2	•	•	b_K
sports	history	•	•	politics

$$\zeta_{i1} = c_{01} + \sigma(z_i) \odot c_1$$

$$\zeta_{i2} = c_{02} + \sigma(z_i) \odot c_2$$

$$\vdots$$

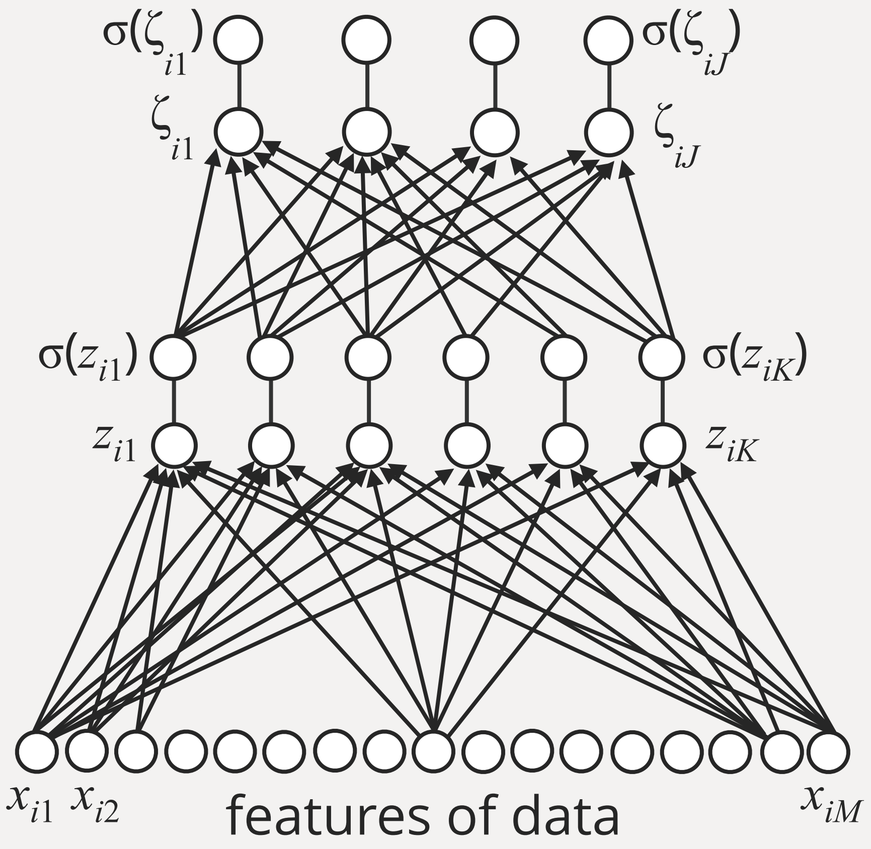
$$\zeta_{iJ} = c_{0J} + \sigma(z_i) \odot c_J$$

$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$\vdots$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$



meta-topics

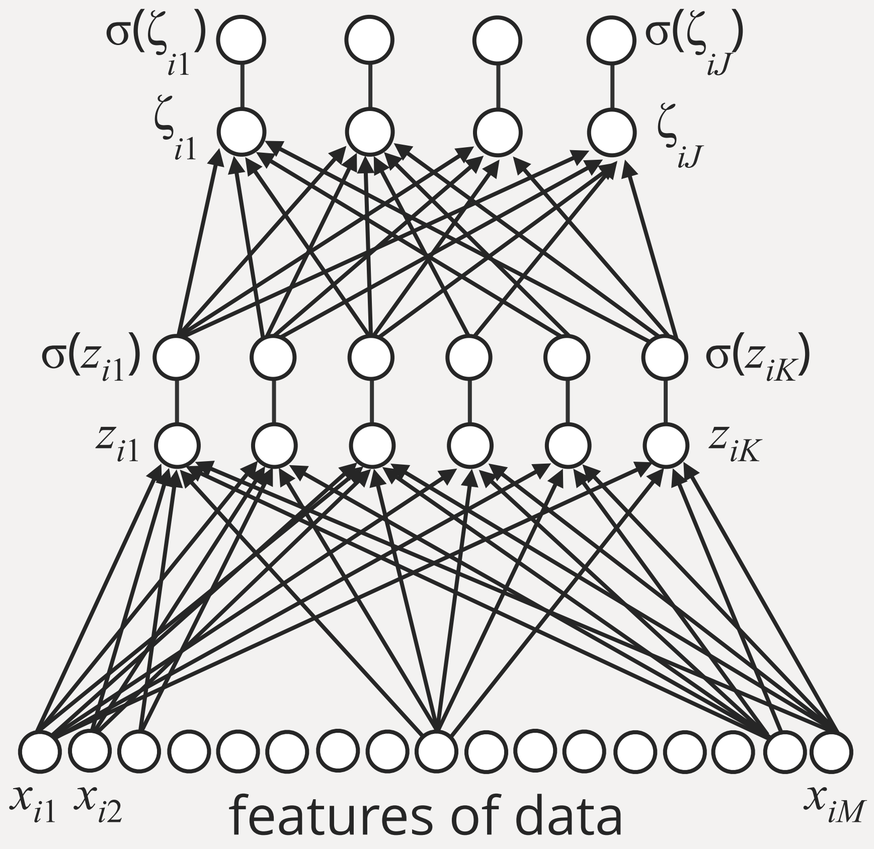
c_1				
sports + history				

$$\zeta_{i1} = c_{01} + \sigma(z_i) \odot c_1$$
$$\zeta_{i2} = c_{02} + \sigma(z_i) \odot c_2$$
$$\vdots$$
$$\zeta_{iJ} = c_{0J} + \sigma(z_i) \odot c_J$$

topics

b_1	b_2	•	•	b_K
sports	history	•	•	politics

$$z_{i1} = b_{01} + x_i \odot b_1$$
$$z_{i2} = b_{02} + x_i \odot b_2$$
$$\vdots$$
$$z_{iK} = b_{0K} + x_i \odot b_K$$



meta-topics

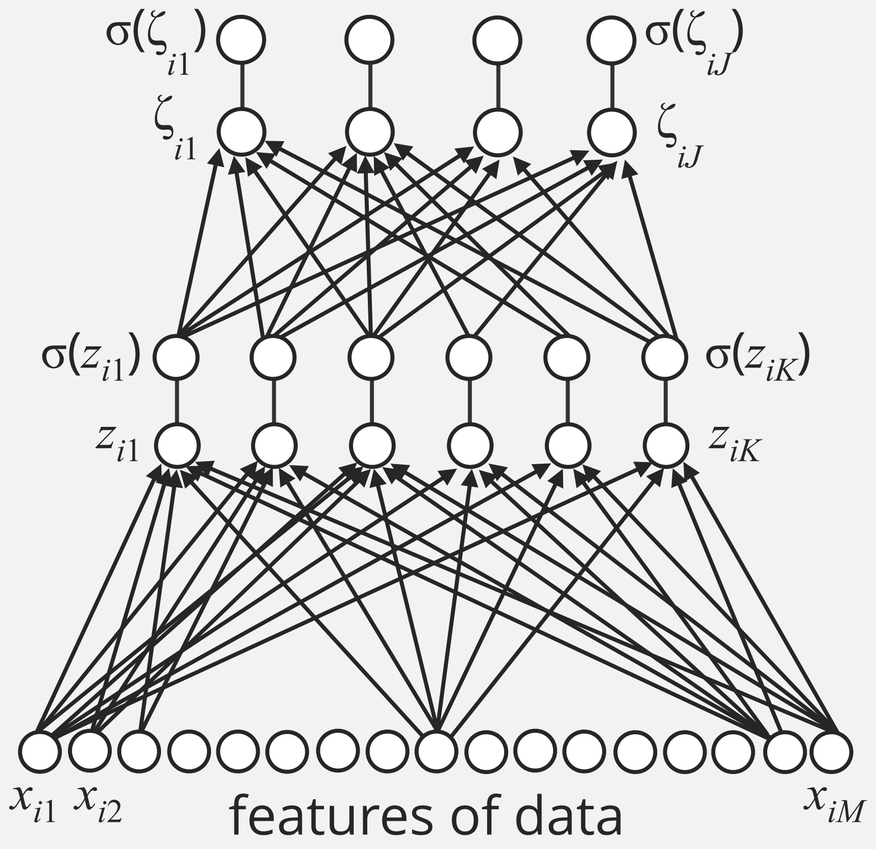
c_1	c_2			
sports + history	politics + sports			

$$\zeta_{i1} = c_{01} + \sigma(z_i) \odot c_1$$
$$\zeta_{i2} = c_{02} + \sigma(z_i) \odot c_2$$
$$\vdots$$
$$\zeta_{iJ} = c_{0J} + \sigma(z_i) \odot c_J$$

topics

b_1	b_2	•	•	b_K
sports	history	•	•	politics

$$z_{i1} = b_{01} + x_i \odot b_1$$
$$z_{i2} = b_{02} + x_i \odot b_2$$
$$\vdots$$
$$z_{iK} = b_{0K} + x_i \odot b_K$$



meta-topics

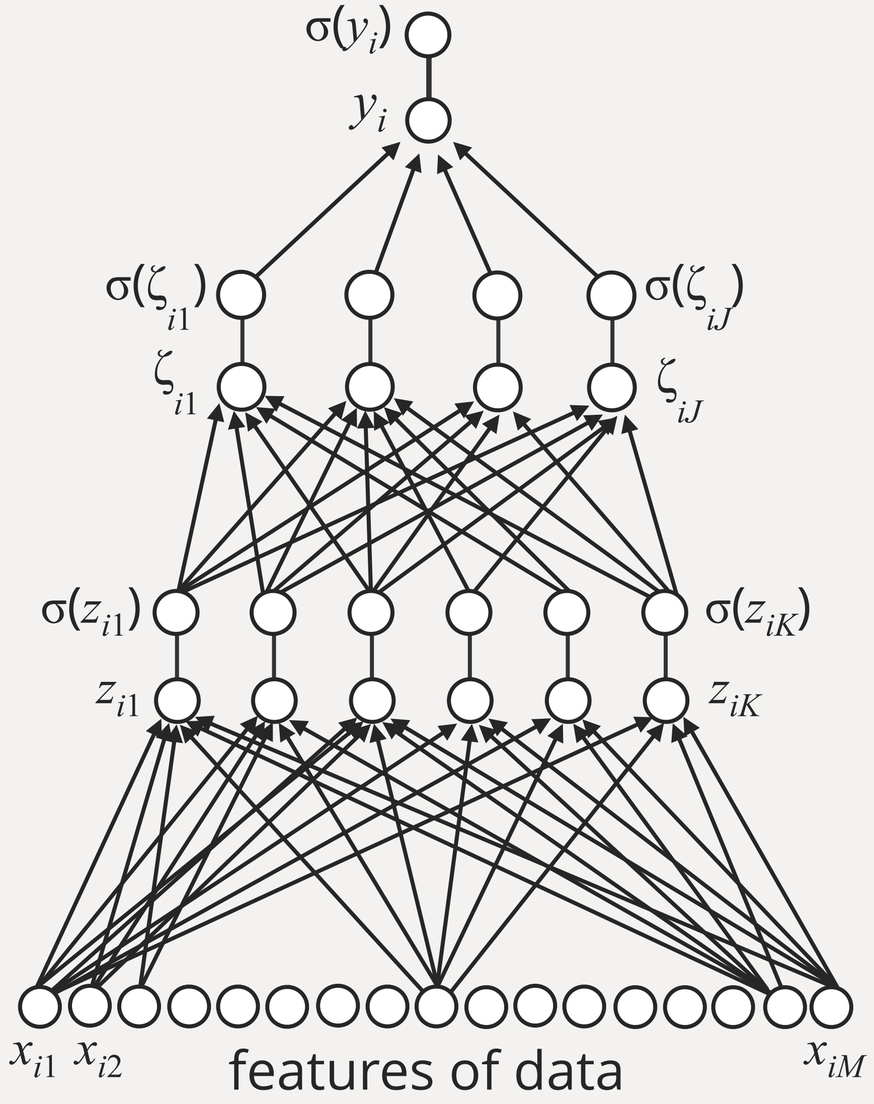
c_1	c_2	•	•	c_J
sports + history	politics + sports	•	•	politics + sports + history

$$\zeta_{i1} = c_{01} + \sigma(z_i) \odot c_1$$
$$\zeta_{i2} = c_{02} + \sigma(z_i) \odot c_2$$
$$\vdots$$
$$\zeta_{iJ} = c_{0J} + \sigma(z_i) \odot c_J$$

topics

b_1	b_2	•	•	b_K
sports	history	•	•	politics

$$z_{i1} = b_{01} + x_i \odot b_1$$
$$z_{i2} = b_{02} + x_i \odot b_2$$
$$\vdots$$
$$z_{iK} = b_{0K} + x_i \odot b_K$$



prediction

$$y_i = d_0 + \sigma(\zeta_i) \odot d$$

meta-topics

c_1	c_2	•	•	c_J
sports + history	politics + sports	•	•	politics + sports + history

$$\zeta_{i1} = c_{01} + \sigma(z_i) \odot c_1$$

$$\zeta_{i2} = c_{02} + \sigma(z_i) \odot c_2$$

$$\zeta_{iJ} = c_{0J} + \sigma(z_i) \odot c_J$$

topics

b_1	b_2	•	•	b_K
sports	history	•	•	politics

$$z_{i1} = b_{01} + x_i \odot b_1$$

$$z_{i2} = b_{02} + x_i \odot b_2$$

$$z_{iK} = b_{0K} + x_i \odot b_K$$