CS 540 Introduction to Artificial Intelligence
Classification - Naive Bayes

University of Wisconsin–Madison
Fall 2025, Section 3
October 3, 2025

Announcements

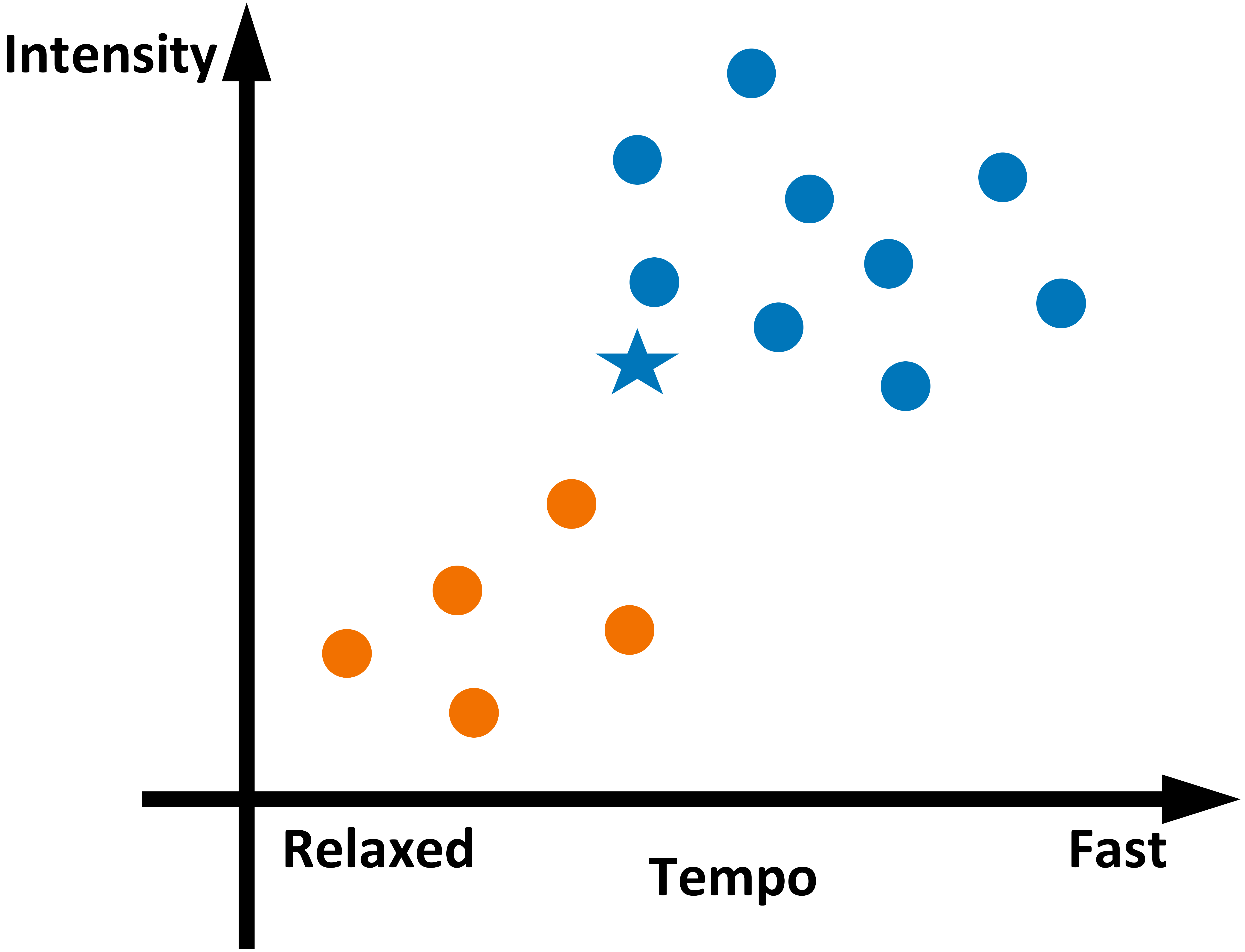
- HW3 due today, 10/3 at 11:59 PM
- HW4 out; build a clustering algorithm
- Class roadmap:

ML: Unsupervised Learning
ML Linear Regression
Machine Learning: K - Nearest Neighbors & Naive Bayes
Machine Learning: Neural Networks I (Perceptron)
Machine Learning: Neural Networks II

Supervised Learning

Last Class: k-Nearest Neighbor Classifier

User Sharon



Dislike

Like

Last Class: k-Nearest Neighbor Classifier

- **Input**: **Training data** $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$

  Distance function $d(\mathbf{x}_i, \mathbf{x}_j)$; **number of neighbors** k; **test data** $\mathbf{x}^*$

1. Find the k training instances $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$ closest to $\mathbf{x}^*$ under $d(\mathbf{x}_i, \mathbf{x}_j)$
2. Output y^*, the majority class of $y_{i_1}, \ldots, y_{i_k}$. Break ties randomly.

Last Class: Maximum Likelihood Estimation

- MLE solves

$$\underset{\theta}{\operatorname{argmax}}\, p(x_1, \dots, x_n \mid \theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i \mid \theta)$$

- Rewrite the problem in an equivalent form

$$\underset{\theta}{\operatorname{argmax}}\, p(x_1, \dots, x_n \mid \theta) = \underset{\theta}{\operatorname{argmin}}(-\log p(x_1, \dots, x_n \mid \theta))$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

Connecting MLE and Loss Minimization

- MLE solves

$$\operatorname*{argmax}_{\theta} p(\, x_1, \dots, x_n \mid \theta \,) = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta)$$

- Rewrite the problem in an equivalent form

$$\operatorname*{argmax}_{\theta} p(\, x_1, \dots, x_n \mid \theta \,) = \operatorname*{argmin}_{\theta}(-\log p(\, x_1, \dots, x_n \mid \theta \,))$$

$$= \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} -\log p(x_i \mid \theta)$$

Connecting MLE and Loss Minimization

- We call "$-\log p(x_i \mid \theta)$" the **negative log likelihood**

- May define $\ell(\theta; x_i) := -\log p(x_i \mid \theta)$

- Maximum likelihood estimation is **loss minimization.** Different notation, same computation.

$$\underset{\theta}{\operatorname{argmax}}\, p(x_1, \ldots, x_n \mid \theta) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(\theta; x_i)$$

Naïve Bayes Classifier

Example 1: Play outside or not?

- If weather is sunny, will my 2-year-old daughter want to play outside?

Posterior probability p(Yes | ) vs. p(No | )

Example 1: Play outside or not?

- If weather is sunny, will my 2-year-old daughter want to play outside?

Posterior probability p(Yes | ) vs. p(No | )

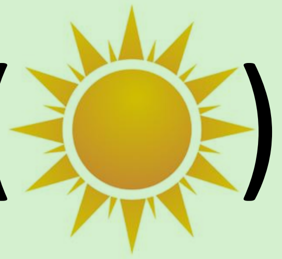
- Weather = {Sunny, Rainy, Overcast}
- Play = {Yes, No}
- Observed data {Weather, play on day *m*}, m={1,2,…,N}

Example 1: Play outside or not?

- If weather is sunny, would you like to play outside?

Posterior probability p(Yes | ) vs. p(No | )

- Weather = {Sunny, Rainy, Overcast}
- Play = {Yes, No}
- Observed data {Weather, play on day *m*}, m={1,2,…,N}

p(Play | ) = p( | Play) p(Play) / p()

Bayes rule

Example 1: Play outside or not?

- **Step 1**: Convert the data to a frequency table of Weather and Play

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Example 1: Play outside or not?

- **Step 1**: Convert the data to a frequency table of Weather and Play
- **Step 2**: Based on the frequency table, calculate **likelihoods** and **priors**

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

p(Play = Yes) = 0.64

p(| Yes) = 3/9 = 0.33

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Example 1: Play outside or not?

- **Step 3**: Based on the likelihoods and priors, calculate posteriors

P(Yes|)

=P( |Yes)*P(Yes)/P()

?

P(No|)

=P( |No)*P(No)/P()

?

Example 1: Play outside or not?

- **Step 3**: Based on the likelihoods and priors, calculate posteriors

P(Yes| ☀)

=P(☀ |Yes)*P(Yes)/P(☀)

=0.33*0.64/0.36

=0.6

P(No|☀)

=P(☀ |No)*P(No)/P(☀)

=0.4*0.36/0.36

=0.4

P(Yes|☀) > P(No|☀) go outside and play!

Bayesian classification

$$\hat{y} = \arg\max p(y \mid \mathbf{x})$$ (Posterior)

(Prediction)

$$= \arg\max \frac{p(\mathbf{x} \mid y) \cdot p(y)}{p(\mathbf{x})}$$ (by Bayes' rule)

$$= \arg\max \; p(\mathbf{x} \mid y) p(y)$$

Bayesian classification

What if **x** has multiple attributes $\mathbf{x} = \{X_1, \ldots, X_k\}$

$$\hat{y} = \arg\max_y p(y \mid X_1, \ldots, X_k) \quad \text{(Posterior)}$$

(Prediction)

Bayesian classification

What if **x** has multiple attributes $\mathbf{x} = \{X_1, \ldots, X_k\}$

$$\hat{y} = \arg\max_y p(y \mid X_1, \ldots, X_k) \quad \text{(Posterior)}$$

(Prediction)

$$= \arg\max_y \frac{p(X_1, \ldots, X_k \mid y) \cdot p(y)}{p(X_1, \ldots, X_k)} \quad \text{(by Bayes' rule)}$$

Independent of y

Bayesian classification

What if **x** has multiple attributes $\mathbf{x} = \{X_1, \ldots, X_k\}$

$$\hat{y} = \arg\max_y p(y \mid X_1, \ldots, X_k)$$ (Posterior)

(Prediction)

$$= \arg\max_y \frac{p(X_1, \ldots, X_k \mid y) \cdot p(y)}{p(X_1, \ldots, X_k)}$$ (by Bayes' rule)

$$= \arg\max_y \; p(X_1, \ldots, X_k \mid y) \; p(y)$$

Class conditional likelihood

Class prior

Naïve Bayes Assumption

Conditional independence of feature attributes

$$p(X_1, \ldots, X_k \mid y)p(y) = \Pi_{i=1}^{k} p(X_i \mid y)p(y)$$

Easier to estimate

(using MLE!)

Example 2: Classify emails as spam

- Features = words in vocabulary
- One parameter θ_w for each word w
- Classify new emails as spam or not spam

Dear Valued Winner,

Congratulations! Your email address has been randomly selected as the GRAND PRIZE WINNER of **$5,000,000 USD** in our International Lottery Promotion.

Reply immediately with your credit card information to claim your prize…

$$p(\text{ spam } | \text{ email }) \propto p(\text{ email } | \text{ spam })p(\text{spam})$$

use naïve Bayes assumption to simplify this term

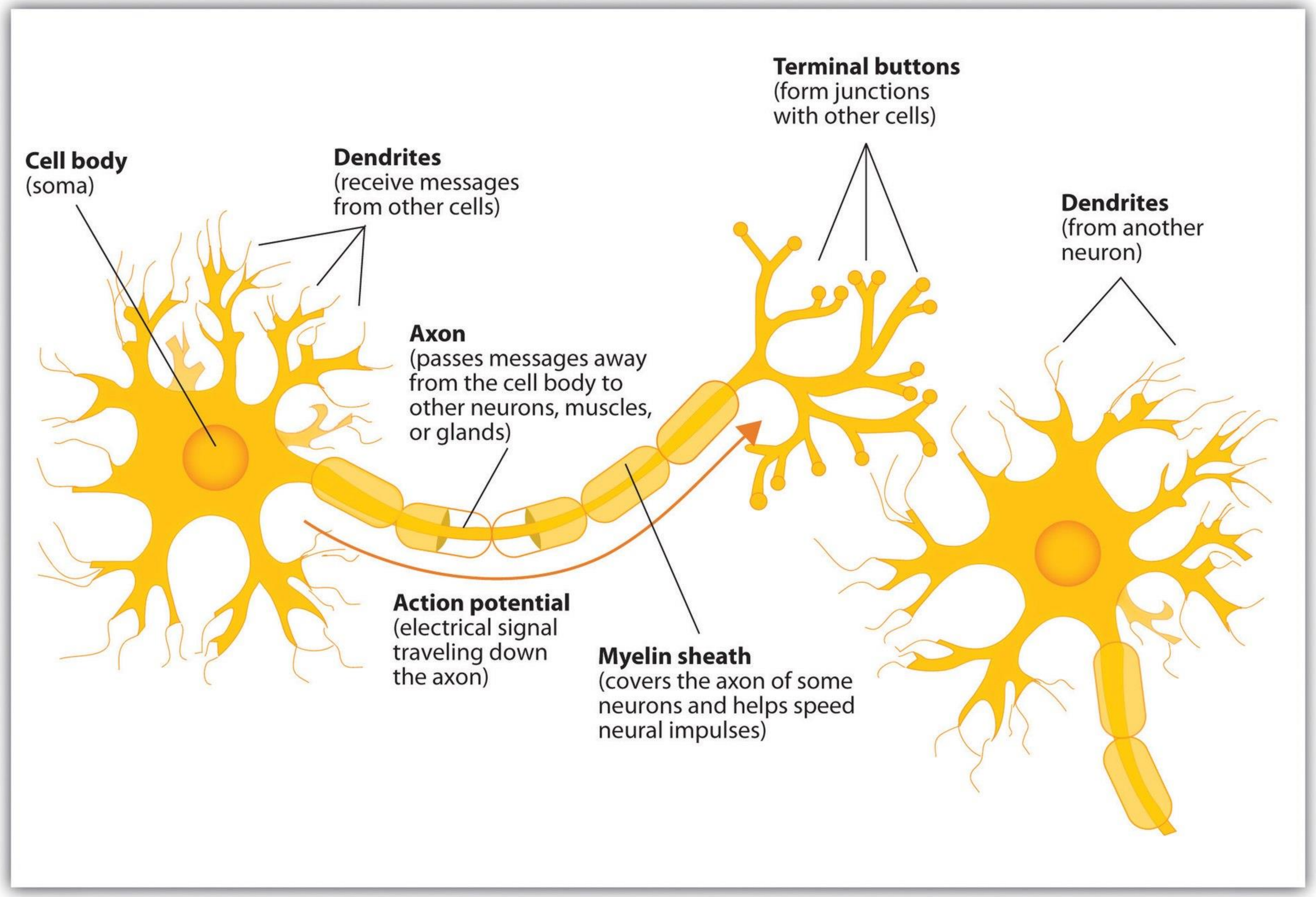
$$p(\text{ ``dear'' } | \text{ spam })p(\text{ ``valued'' } | \text{ spam })p(\text{ ``winner'' } | \text{ spam }) \cdots p(\text{ ``prize'' } | \text{ spam })$$

Looking Ahead: Single-layer Neural Network

Inspiration from neuroscience

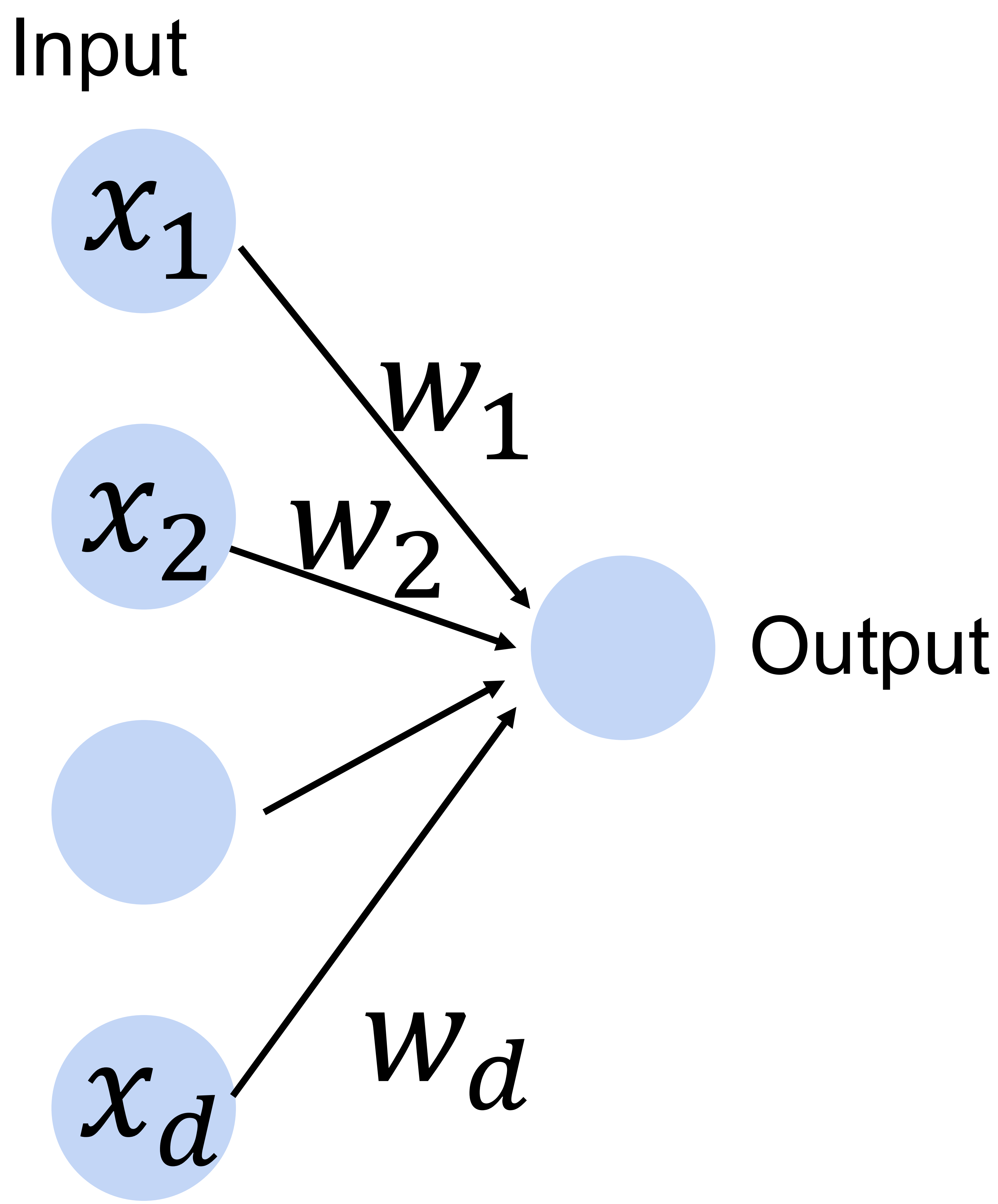
- Inspirations from human brains
- Networks of simple and homogenous units

(wikipedia)



Perceptron

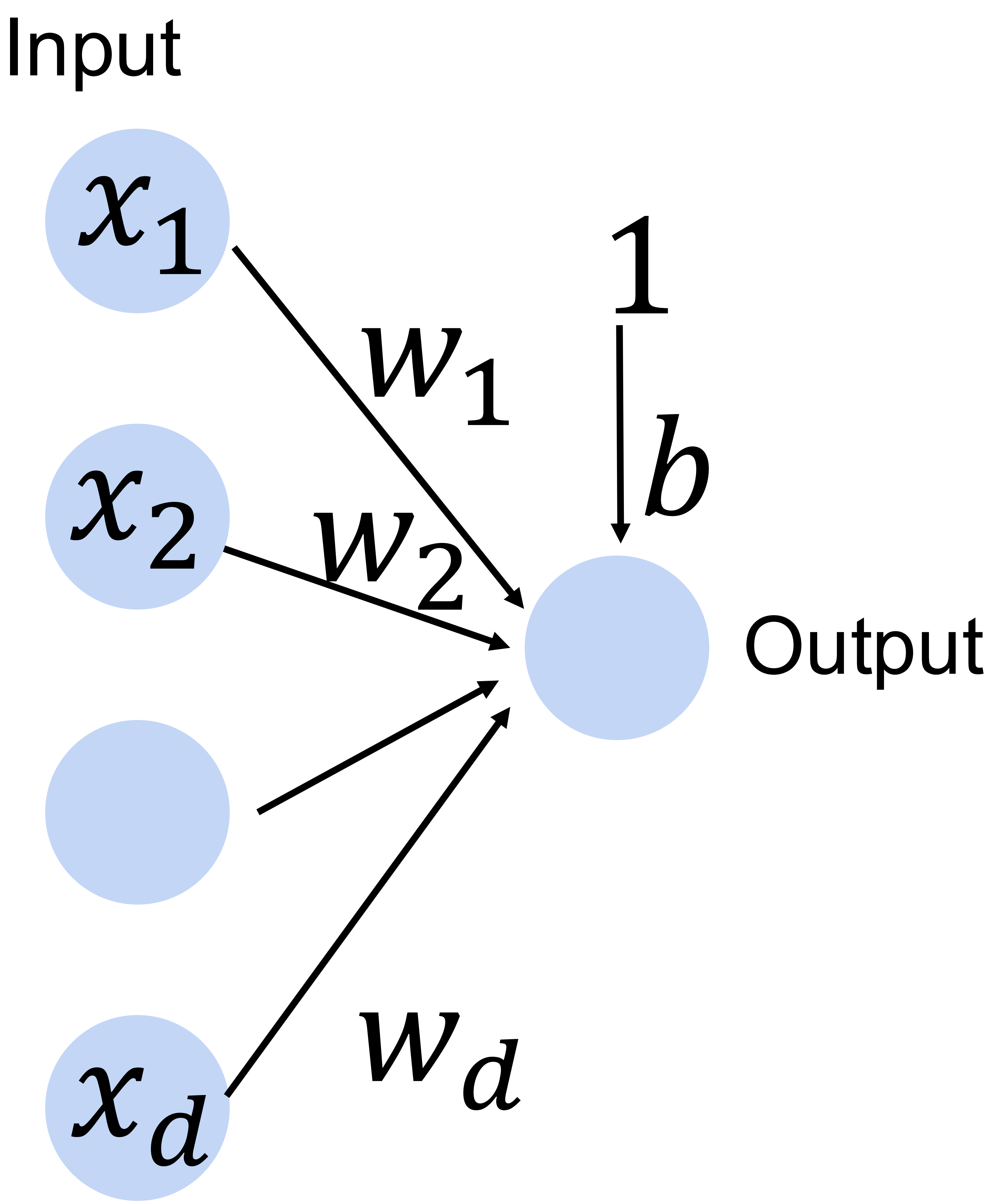
Cats vs. dogs?



Linear Perceptron

Given input $\mathbf{x}$, weight $\mathbf{w}$ and bias b, perceptron outputs:

$$f = \langle \mathbf{w}, \mathbf{x} \rangle + b$$



Cats vs. dogs?

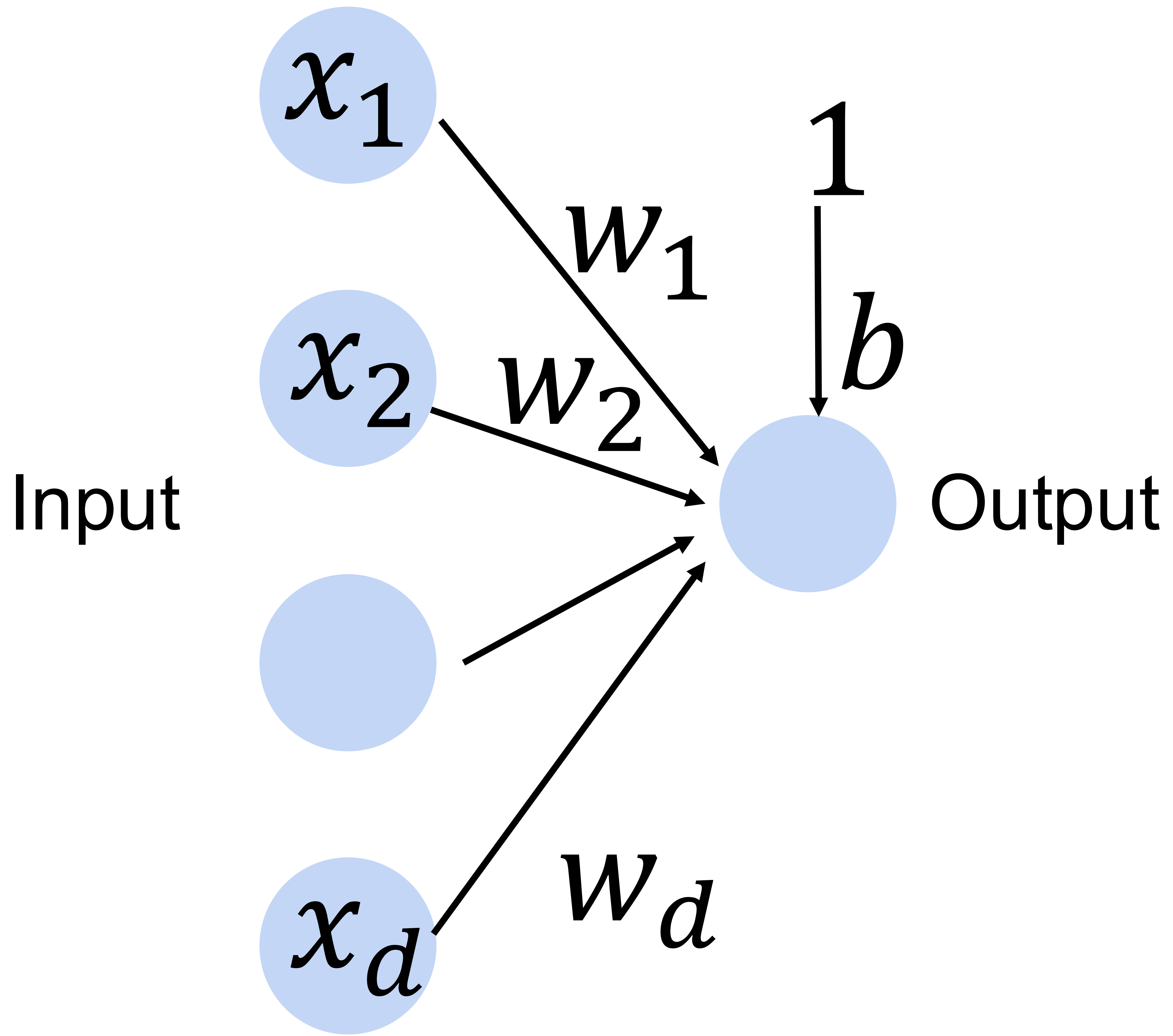
Perceptron

Given input $\mathbf{x}$, weight $\mathbf{w}$ and bias b, perceptron outputs:

$$o = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b) \qquad \sigma(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Activation function

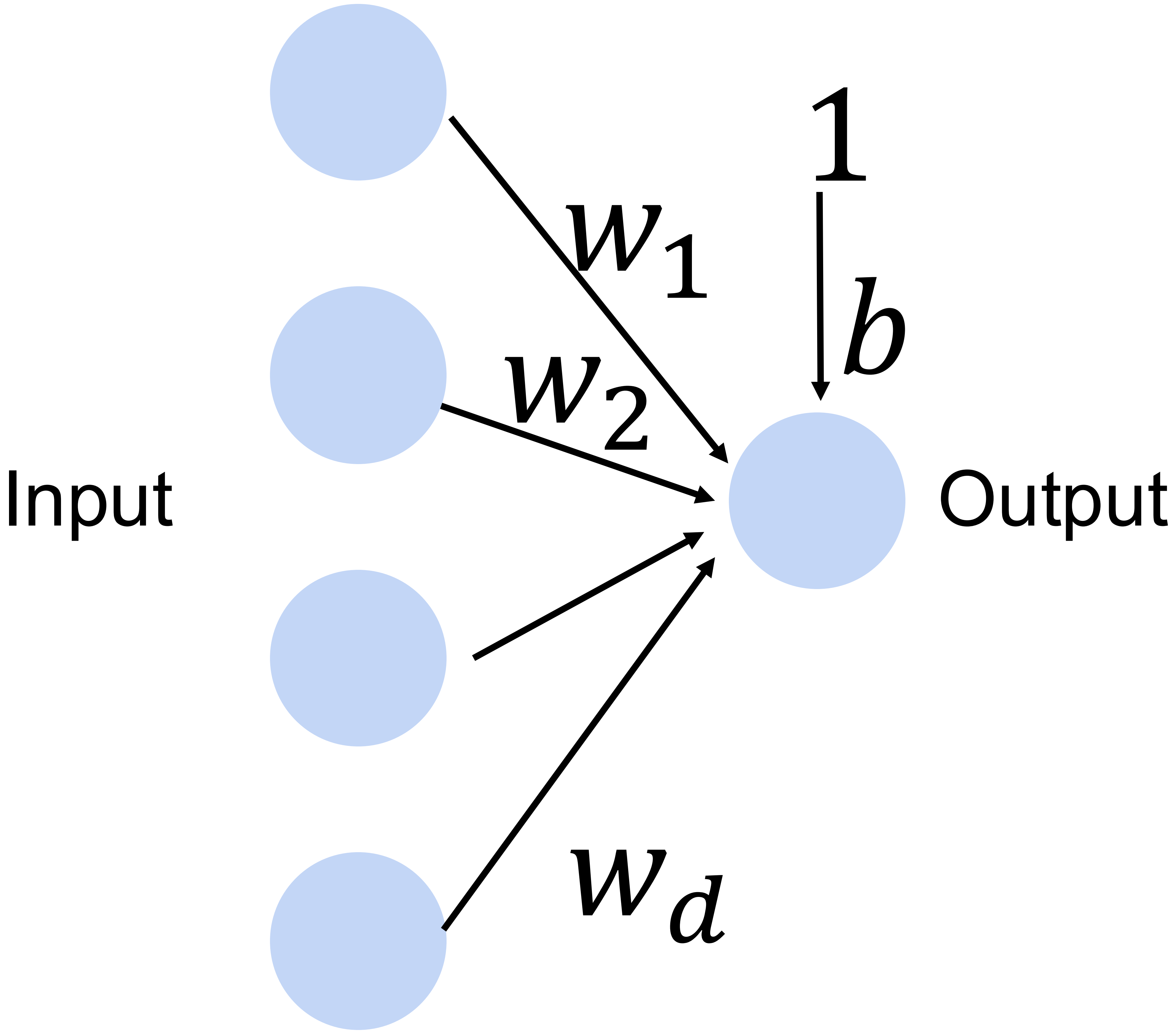
Cats vs. dogs?



Perceptron

Goal: learn parameters $\mathbf{w} = \{w_1, w_2, \ldots, w_d\}$ and b to minimize the classification error

Cats vs. dogs?



Thanks!