



浙江大学伊利诺伊大学厄巴纳香槟校区联合学院  
Zhejiang University-University of Illinois at Urbana Champaign Institute

## ECE448: Artificial Intelligence

### Lecture 14: Linear Classifiers-Bayesian, Perceptron, Logistic Regression

Prof. Hongwei Wang [hongweiwang@intl.zju.edu.cn](mailto:hongweiwang@intl.zju.edu.cn)

Prof. Mark Hasegawa-Johnson [jhasegaw@illinois.edu](mailto:jhasegaw@illinois.edu)

Spring 2023

1. **Naïve Bayes/BoW classifiers**
2. **Linear Classifiers in General**
3. **Perceptron**
4. **Differential Perceptron/Neural Net**

1. Naïve Bayes/BoW classifiers
2. Linear Classifiers in General
3. Perceptron
4. Differential Perceptron/Neural Net

- Model parameters: feature likelihoods  $P(\text{word} \mid \text{class})$  and priors  $P(\text{class})$ 
  - How do we obtain the values of these parameters?
  - Need *training set* of labeled samples from both classes

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

- This is the *maximum likelihood* (ML) estimate, or estimate that maximizes the likelihood of the training data:

$$\prod_{d=1}^D \prod_{i=1}^{n_d} P(w_{d,i} \mid \text{class}_{d,i})$$

$d$ : index of training document,  $i$ : index of a word

- Indexing the training dataset: TOKENS
  - $i$  = document token index,  $1 \leq i \leq m$  (there are  $m$  document tokens in the training dataset)
  - $j$  = word token index,  $1 \leq j \leq n$  (there are  $n$  word tokens in each document)
- Indexing the dictionary: TYPES
  - $c$  = class type,  $1 \leq c \leq C$  (there are a total of  $C$  different class types)
  - $w$  = word type,  $1 \leq w \leq V$  (there are a total of  $V$  words in the dictionary, i.e.,  $V$  different word types)

- One bit per document, per word type:
  - $F_{iw} = 1$  if word “w” occurs anywhere in the i’t h document
  - $F_{iw} = 0$  otherwise
- One bit per word token, per word type:
  - $F_{jw} = 1$  if the j’t h word token is “w”
  - $F_{jw} = 0$  otherwise

Example: “who saw who with who?”

$$F_{i, \text{"who"}} = 1$$

$$F_{j, \text{"who"}} = \{1, 0, 1, 0, 1\}$$

- Features:
  - $F_{iw} = 1$  if word “w” occurs anywhere in the i’t h document
- Parameters:
  - $\lambda_{cw} \equiv P(F_{iw} = 1|C = c)$
  - Note this means that  $P(F_{iw} = 0|C = c) = 1 - \lambda_{cw}$
- Parameter Learning:

$$\lambda_{cw} = \frac{(1 + \# \text{ documents containing } w)}{(1 + \# \text{ documents containing } w) + (1 + \# \text{ documents NOT containing } w)}$$

- Features:

- $F_{jw} = 1$  if the  $j$ 'th word token is word “w”

- Parameters:

- $\lambda_{cw} \equiv P(F_{jw} = 1 | C = c) = P(W_j = w | C = c)$
- Note this means that  $P(F_{jw} = 0 | C = c) = \sum_{v \neq w} \lambda_{cv}$

- Parameter Learning:

$$\lambda_{cw} = \frac{(1 + \# \text{ tokens of } w \text{ in the training database})}{\sum_{v=1}^V (1 + \# \text{ tokens of } v \text{ in the training database})}$$



Classification:

$$C^* = \operatorname{argmax} P(C=c | \text{document})$$

$$= \operatorname{argmax} P(C=c) P(\text{Document} | C=c)$$

$$= \operatorname{argmax}_c \left( \pi_c \prod_{w: f_{cw}=1} \lambda_{cw} \prod_{w: f_{cw}=0} (1 - \lambda_{cw}) \right)$$

$$P(C=c) * \prod_{\text{words that occurred}} P(\text{word occurs} | C=c) * \prod_{\text{didn't occur}} P(\text{didn't occur} | C=c)$$

Classification:

$$C^* = \operatorname{argmax} P(C=c | \text{document})$$

$$= \operatorname{argmax} P(C=c) P(\text{Document} | C=c)$$

$$= \operatorname{argmax}_c \left( \pi_c \prod_{j=1}^n \lambda_{cw_j} \right)$$

$$P(C=c) \prod_{\text{words in the document}} P(\text{get that particular word} \mid C=c)$$

Classification:

$$C^* = \arg \max_c \left( \pi_c \prod_{w=1}^V \left( \frac{\lambda_{cw}}{1 - \lambda_{cw}} \right)^{f_{cw}} (1 - \lambda_{cw}) \right)$$

$$C^* = \arg \max_c \left( \beta_c + \sum_{w=1}^V \alpha_{cw} f_{cw} \right)$$

In a 2-dimensional feature space  $(f_{c1}, f_{c2})$ , this is the equation for a line, with intercept  $-\beta_c$ , and with slope given by  $\alpha_{c1}/\alpha_{c2}$

$$\alpha_{cw} = \log \left( \frac{\lambda_{cw}}{1 - \lambda_{cw}} \right), \quad \beta_c = \log \left( \pi_c \prod_{w=1}^V (1 - \lambda_{cw}) \right)$$

Classification:

$$C^* = \arg \max_c \left( \pi_c \prod_{w=1}^V \lambda_{cw}^{s_w} \right)$$

Where  $s_w$  = number of times  $w$  occurred in the document!! So...

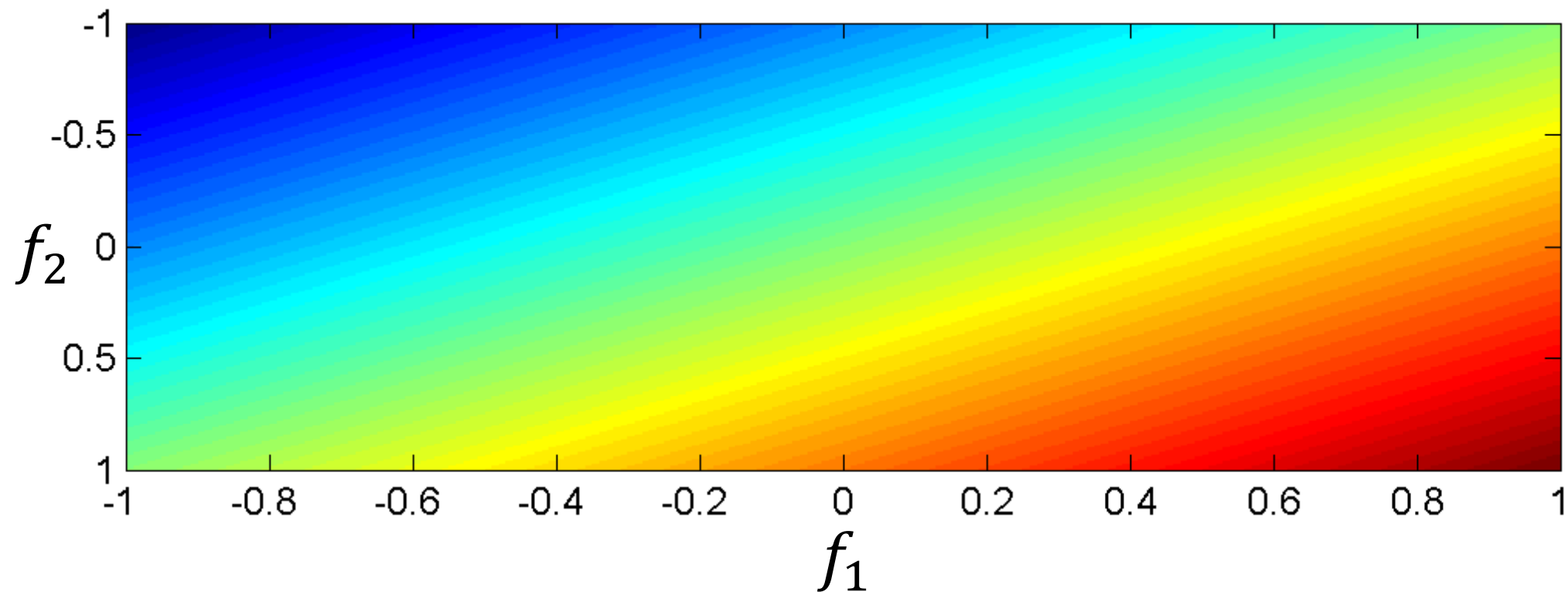
$$C^* = \arg \max_c \left( \beta_c + \sum_{w=1}^V \alpha_{cw} s_{cw} \right)$$

$$\alpha_{cw} = \log \lambda_{cw}, \quad \beta_c = \log \pi_c$$

In a 2-dimensional feature space  $(f_{c1}, f_{c2})$ , this is the equation for a line, with intercept  $-\beta_c$ , and with slope given by  $\alpha_{c1}/\alpha_{c2}$

1. Naïve Bayes/BoW classifiers
2. Linear Classifiers in General
3. Perceptron
4. Differential Perceptron/Neural Net

The function  $\beta_c + \sum_{w=1}^V \alpha_{cw} f_{cw}$  is an affine function of the features  $f_{cw}$ . That means that its contours are all straight lines. Here is an example of such a function, plotted as variations of color in a two-dimensional space  $f_1$  by  $f_2$ :

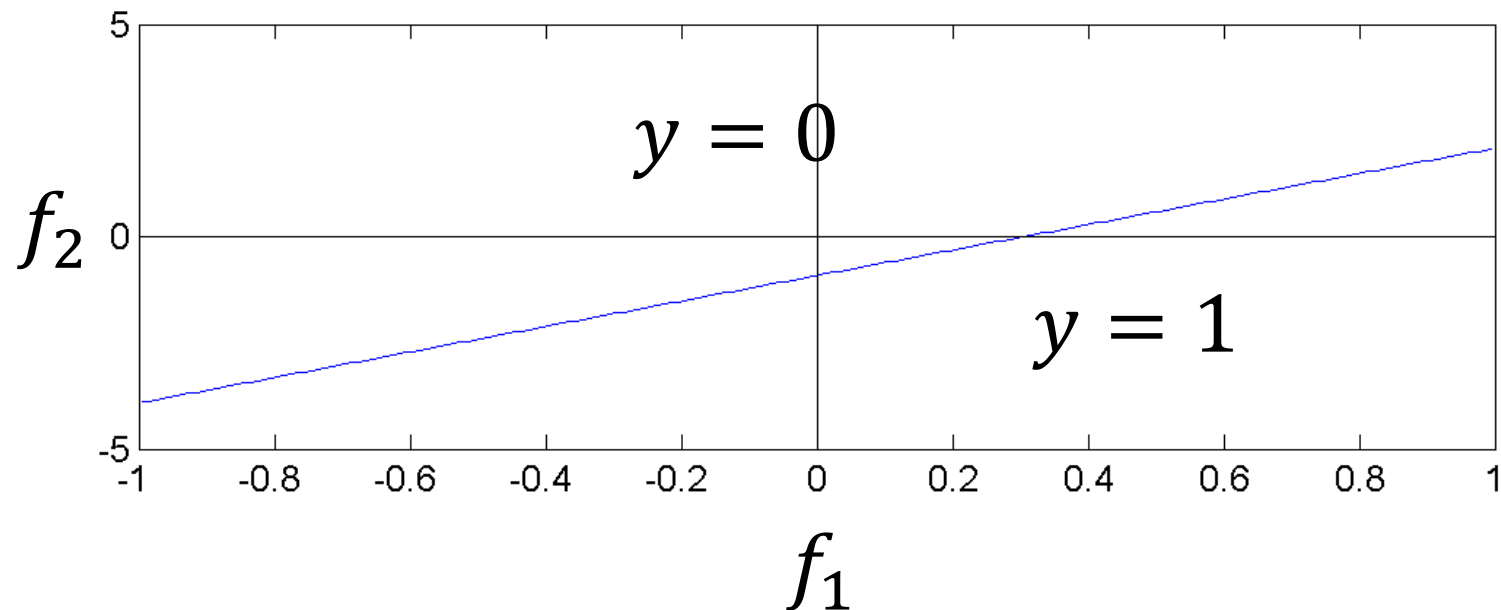


Consider the classifier

$$y = 1 \quad \text{if} \quad \beta_c + \sum_{w=1}^V \alpha_{cw} f_{cw} > 0$$

$$y = 0 \quad \text{if} \quad \beta_c + \sum_{w=1}^V \alpha_{cw} f_{cw} < 0$$

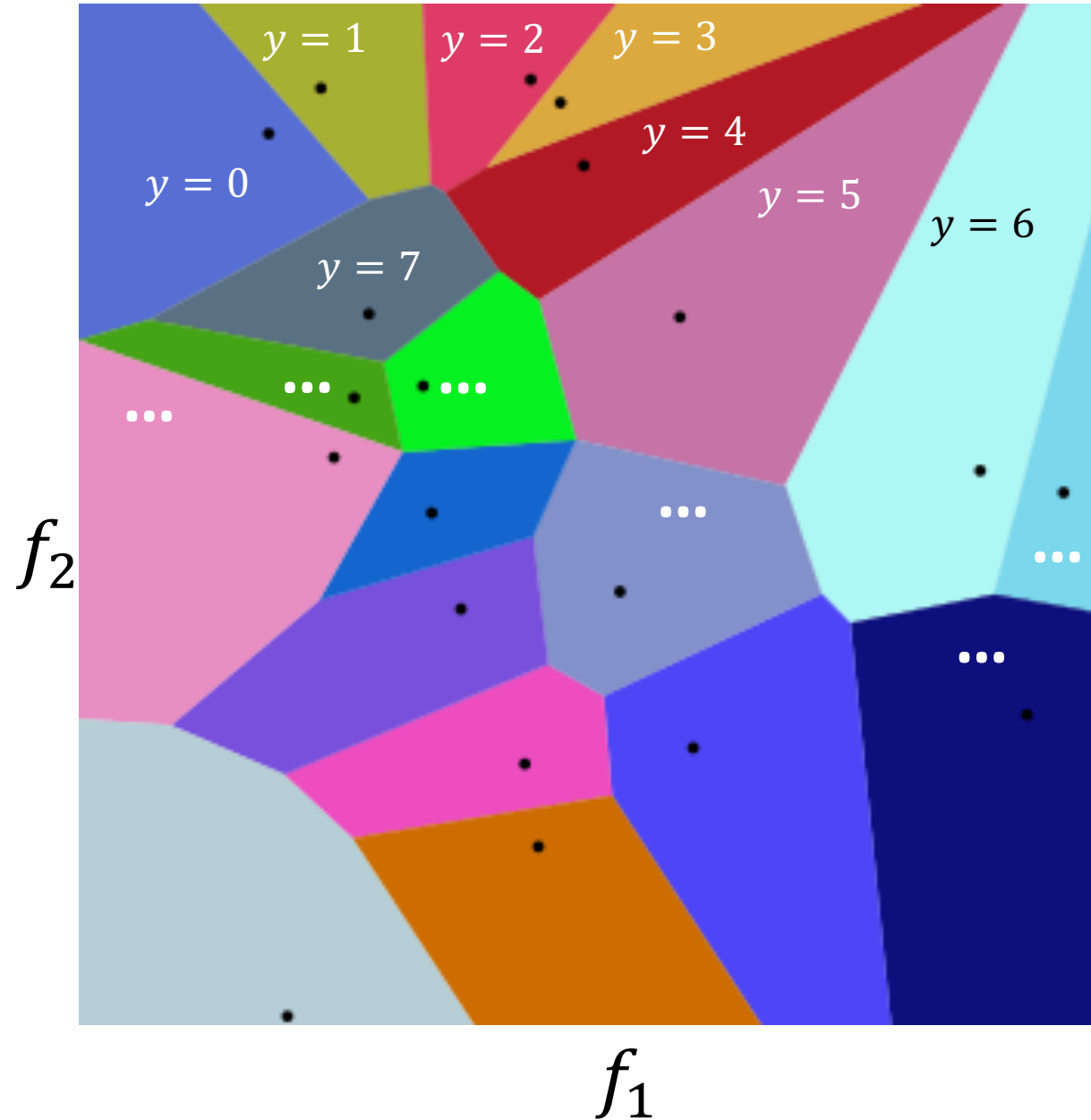
This is called a “linear classifier” because the boundary between the two classes is a line. Here is an example of such a classifier, with its boundary plotted as a line in the two-dimensional space  $f_1$  by  $f_2$ :



Consider the classifier

$$y = \arg \max_c \left( \beta_c + \sum_{w=1}^V \alpha_{cw} f_{cw} \right)$$

- This is called a “multi-class linear classifier.”
- The regions  $y = 0, y = 1, y = 2$  etc. are called “Voronoi regions.”
- They are regions with piece-wise linear boundaries. Here is an example from Wikipedia of Voronoi regions plotted in the two-dimensional space  $f_1$  by  $f_2$ :



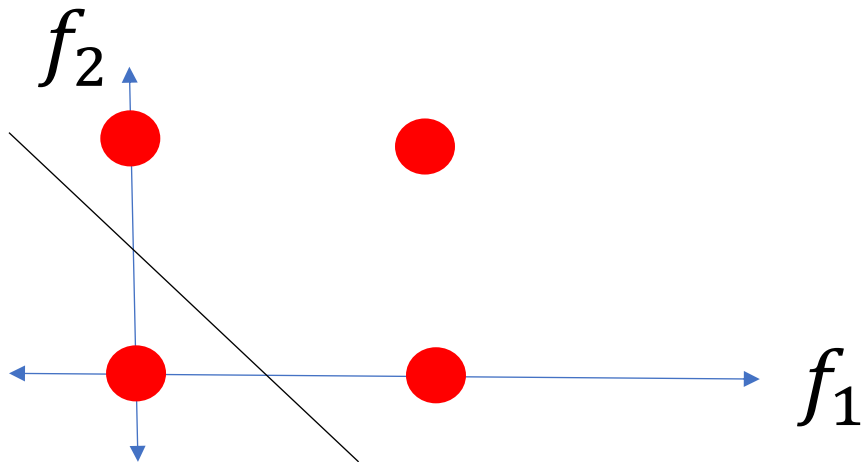


When the features are binary ( $f_w \in \{0,1\}$ ), many (but not all!) binary functions can be re-written as linear functions. For example, the function

$$y = (f_1 \vee f_2)$$

can be re-written as

$$y=1 \text{ iff } f_1 + f_2 - 0.5 > 0$$

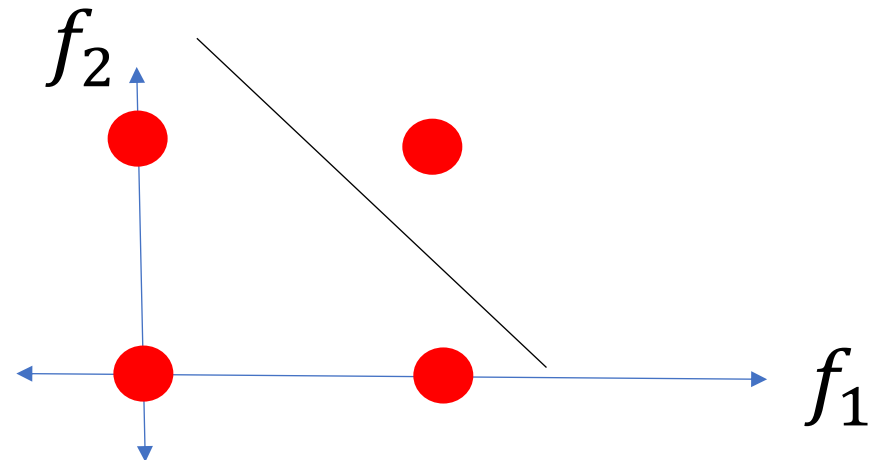


Similarly, the function

$$y = (f_1 \wedge f_2)$$

can be re-written as

$$y=1 \text{ iff } f_1 + f_2 - 1.5 > 0$$



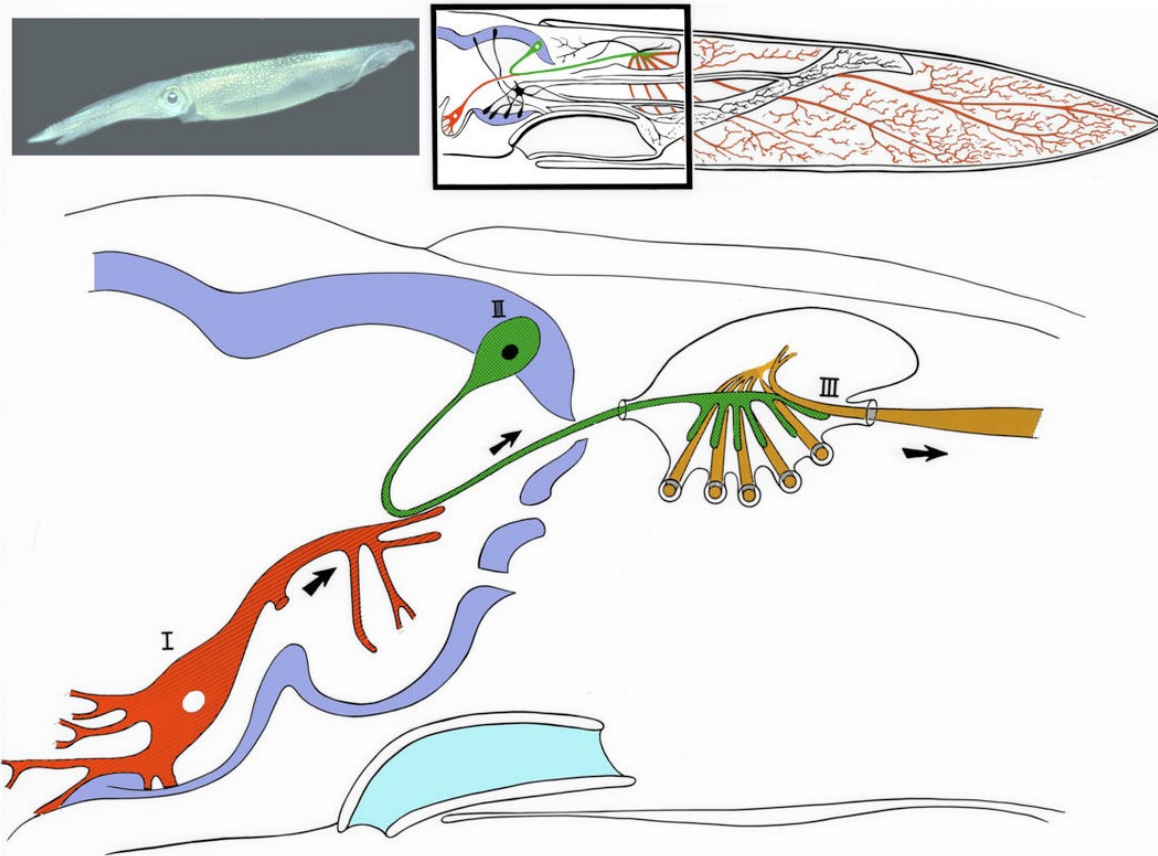
- Not all logical functions can be written as linear classifiers!
- Minsky and Papert wrote a book called *Perceptrons* in 1969. Although the book said many other things, the only thing most people remembered about the book was that:
  - **“A linear classifier cannot learn an XOR function.”**
- Because of that statement, most people gave up working on neural networks from about 1969 to about 2006.
- Minsky and Papert also proved that a two-layer neural net can learn an XOR function. But most people didn't notice.

Classification:

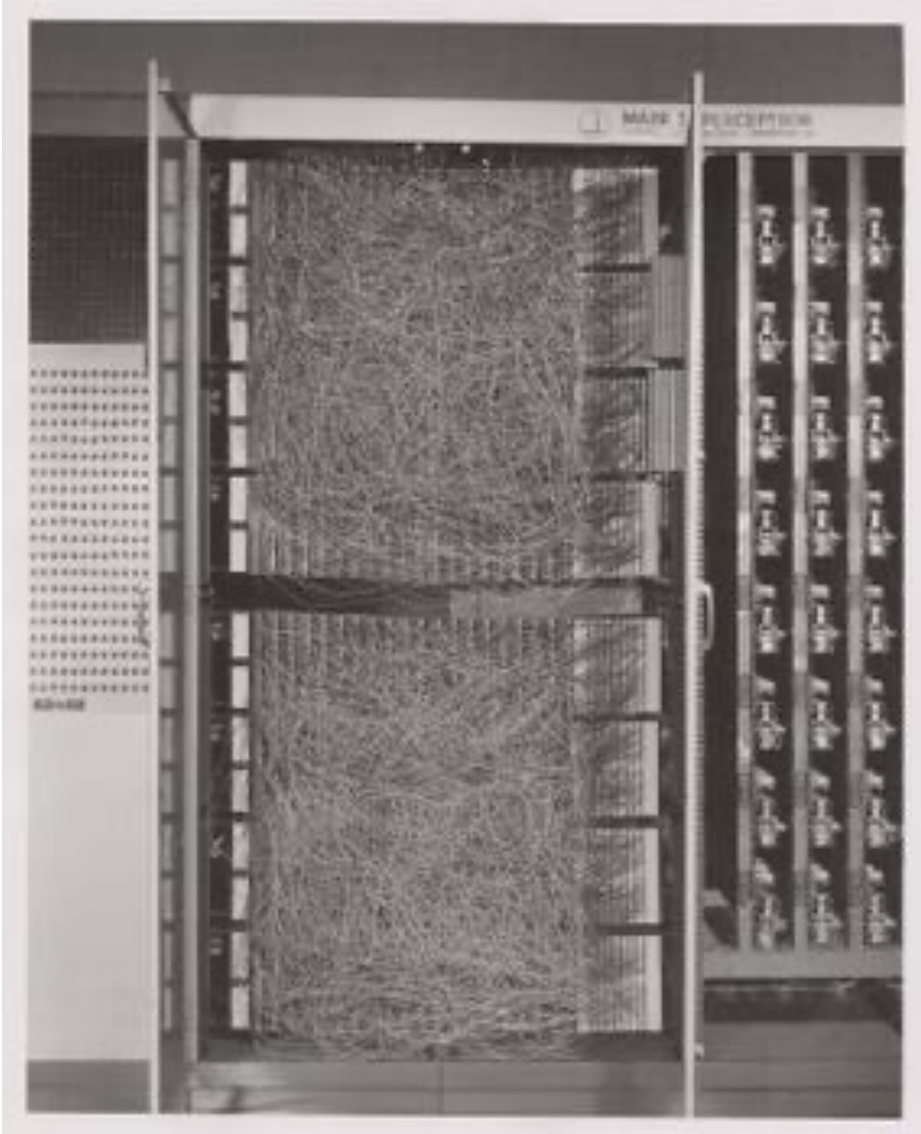
$$y = \arg \max_c \left( \beta_c + \sum_{w=1}^V \alpha_{cw} f_{cw} \right)$$

- Where  $f_{cw}$  are the features (binary, integer, or real),  $\alpha_{cw}$  are the feature weights, and  $\beta_c$  is the offset

1. Naïve Bayes/BoW classifiers
2. Linear Classifiers in General
3. Perceptron
4. Differential Perceptron/Neural Net



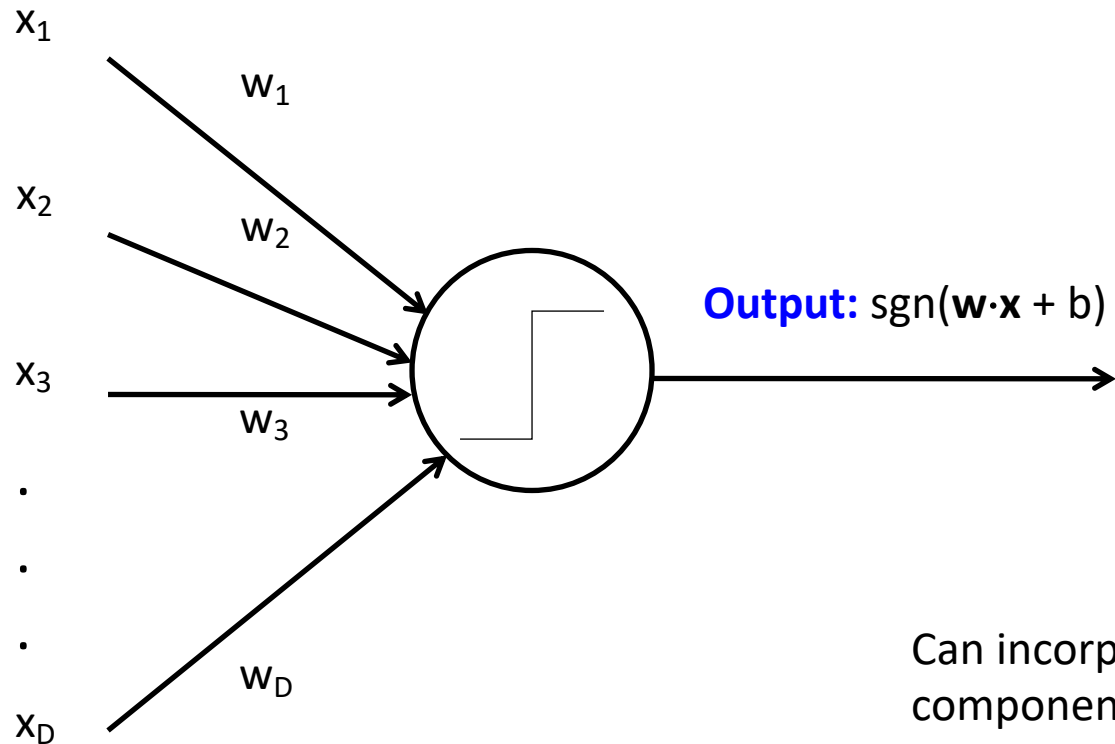
- 1909: Williams discovers that the giant squid has a giant neuron (axon 1mm thick)
- 1939: Young finds a giant synapse (fig. shown: Llinás, 1999, via Wikipedia). Hodgkin & Huxley put in voltage clamps.
- 1952: Hodgkin & Huxley publish an electrical current model for the generation of binary action potentials from real-valued inputs.



- 1959: Rosenblatt is granted a patent for the “perceptron,” an electrical circuit model of a neuron.

Input

Weights



Perceptron model: action potential = signum(affine function of the features)

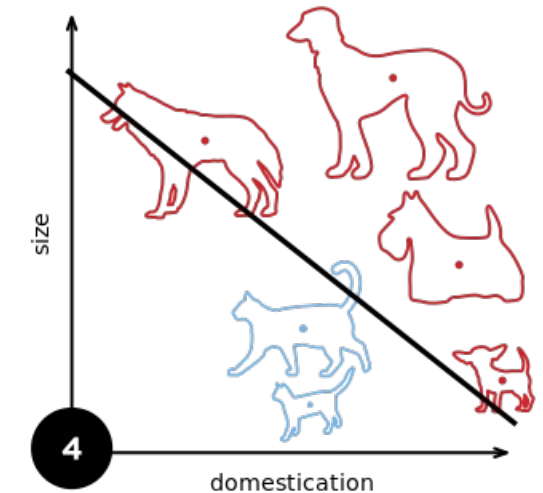
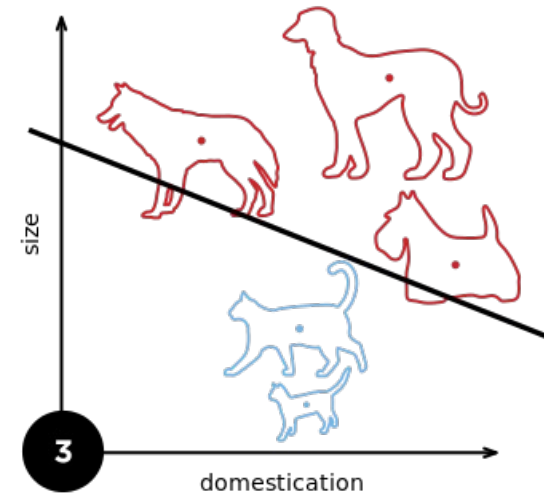
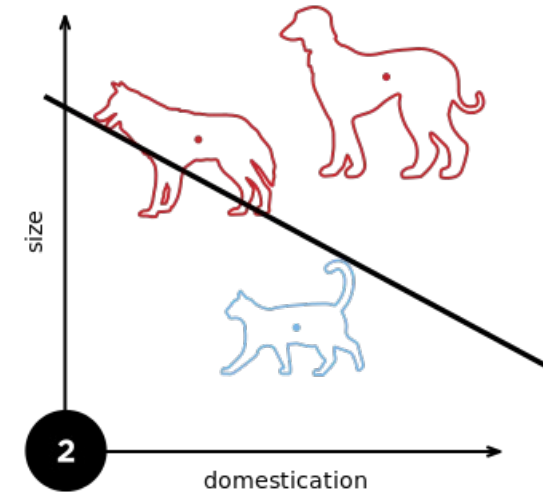
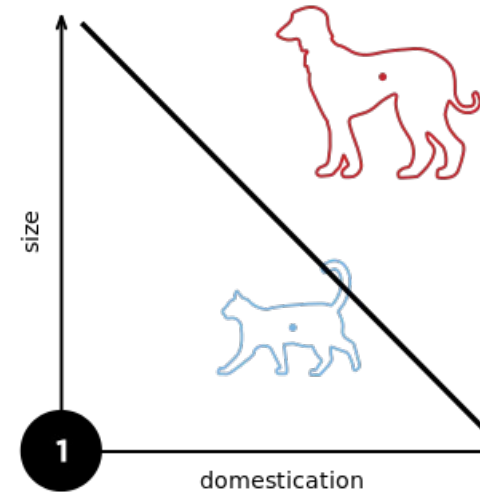
$$y = \text{sgn}(\alpha_1 f_1 + \alpha_2 f_2 + \dots + \alpha_V f_V + \beta) = \text{sgn}(\vec{w}^T \vec{f})$$

Where  $\vec{w} = [\alpha_1, \dots, \alpha_V, \beta]^T$   
and  $\vec{f} = [f_1, \dots, f_V, 1]^T$

Can incorporate bias as component of the weight vector by always including a feature with value set to 1

Rosenblatt's big innovation: the perceptron learns from examples.

- Initialize weights randomly
- Cycle through training examples in multiple passes (*epochs*)
- For each training example:
  - If classified correctly, do nothing
  - If classified incorrectly, update weights





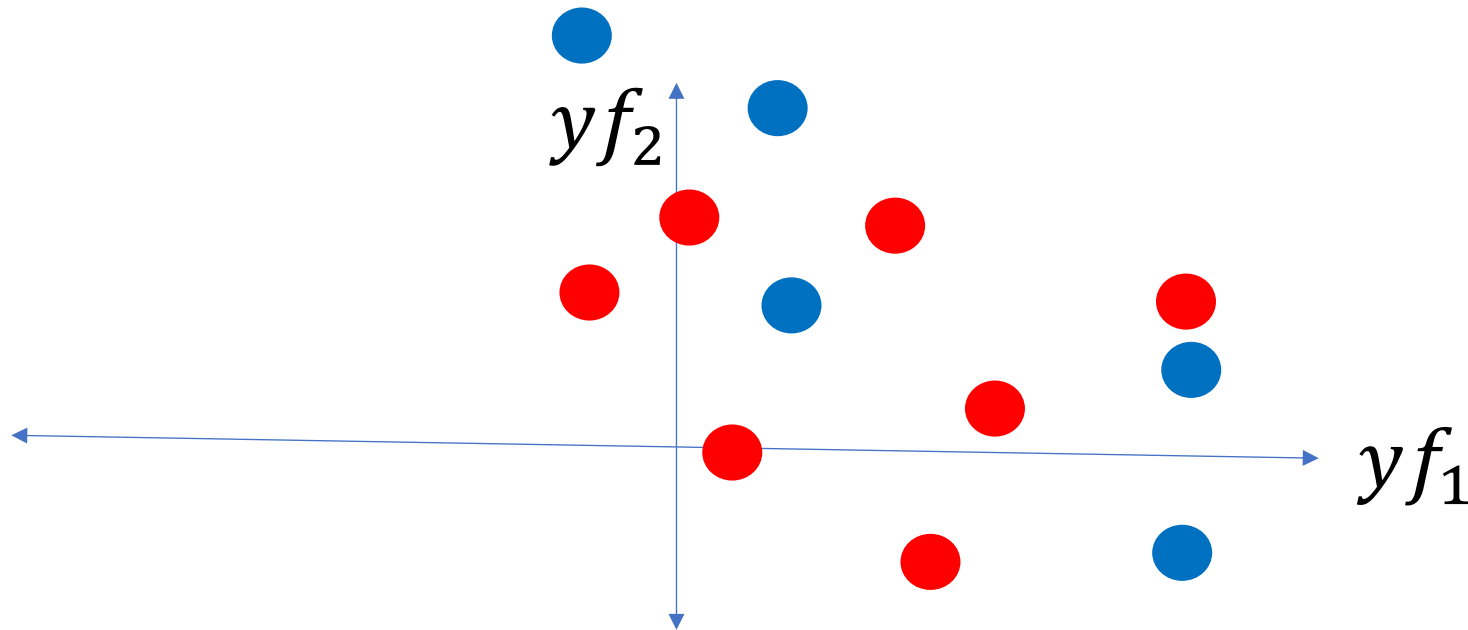
For each training instance  $\vec{f}$  with label  $y \in \{-1, 1\}$ :

- Classify with current weights:  $y' = \text{sgn}(\vec{w}^T \vec{f})$ 
  - Notice  $y' \in \{-1, 1\}$  too.
- Update weights:
  - if  $y = y'$  then do nothing
  - if  $y \neq y'$  then  $\vec{w} = \vec{w} + \eta y \vec{f}$
  - $\eta$  (eta) is a “learning rate.” More about that later.

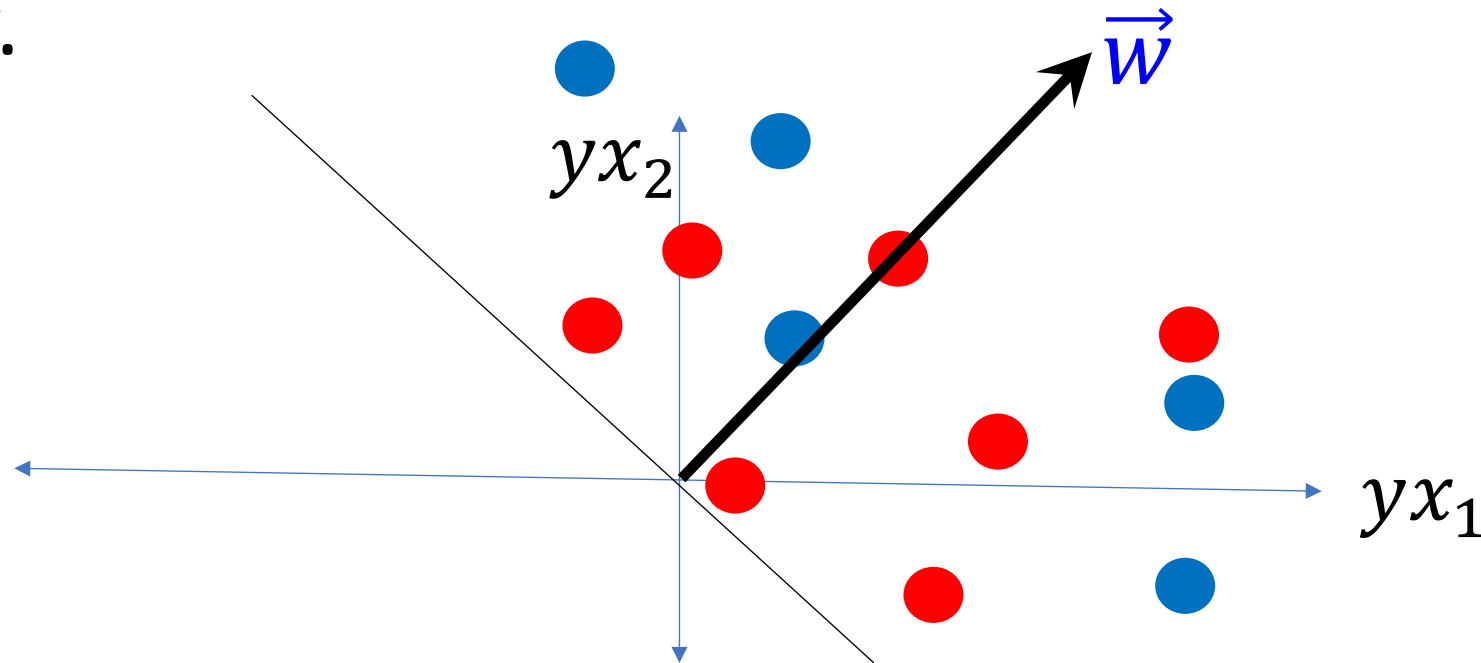
- If the data are linearly separable (if there exists a  $\vec{w}$  vector such that the true label is given by  $y' = \text{sgn}(\vec{w}^T \vec{f})$ ), then the perceptron algorithm is guarantee to converge, even with a constant learning rate, even  $\eta=1$ .
- In fact, training a perceptron is often the fastest way to find out if the data are linearly separable. If  $\vec{w}$  converges, then the data are separable; if  $\vec{w}$  diverges toward infinity, then no.
- If the data are not linearly separable, then perceptron converges iff the learning rate decreases, e.g.,  $\eta=1/n$  for the  $n$ 'th training sample.

A 2D scatter plot with axes labeled  $f_1$  (horizontal) and  $f_2$  (vertical). The plot displays two classes of data points: red circles and blue circles. There are 8 red points and 7 blue points scattered across the plane. The red points are located at approximately (0.5, 0.8), (0.8, 0.5), (1.0, 0.2), (1.2, 0.8), (1.5, 0.5), (1.8, 0.2), (0.2, 0.5), and (0.5, 0.8). The blue points are located at approximately (-0.5, 0.5), (-0.5, -0.5), (0.2, -0.8), (0.5, -0.5), (0.8, -0.2), (1.0, -0.8), and (1.2, -0.5).

- Instead of plotting  $\vec{f}$ , plot  $y \times \vec{f}$ . The red dots are unchanged; the blue dots are multiplied by -1.
- Since the original data were linearly separable, the new data are all in the same half of the feature space.



- Remember the perceptron training rule: if any example is misclassified, then we use it to update  $\vec{w} = \vec{w} + y \vec{f}$ .
- So eventually,  $\vec{w}$  becomes just a weighted average of  $y\vec{f}$ .
- ... and the perpendicular line,  $\vec{w}^T \vec{f} = 0$ , is the classifier boundary.



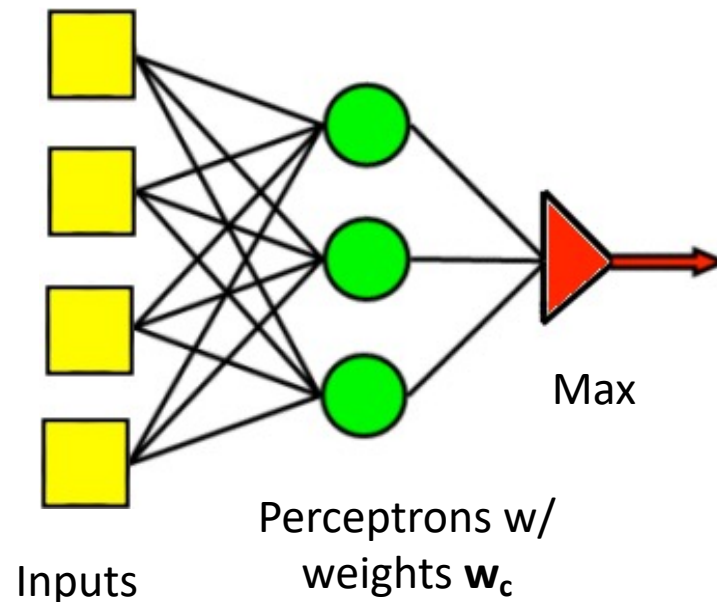
- If the data are linearly separable, then the perceptron will eventually find the equation for a line that separates them.
- If the data are NOT linearly separable, then perceptron converges iff the learning rate decreases, e.g.,  $\eta = 1/n$  for the  $n$ 'th training sample. .... In this case, convergence is trivially obvious, because  $y$  and  $\vec{f}$  are finite, therefore the weight updates  $\eta y \vec{f}$  approach 0 as  $\eta$  approaches 0.

- Bias (add feature dimension with value fixed to 1) vs. no bias
- Initialization of weights: all zeros vs. random
- Learning rate decay function
- Number of epochs (passes through the training data)
- Order of cycling through training examples (random)

- *One-vs-others* framework: Need to keep a weight vector  $\mathbf{w}_c$  for each class  $c$
- Decision rule:  $y = \operatorname{argmax}_c \mathbf{w}_c \cdot \mathbf{f}$
- Update rule: suppose example from class  $c$  gets misclassified as  $c'$ 
  - Update for  $c$ :  $\mathbf{w}_c \leftarrow \mathbf{w}_c + \eta \mathbf{f}$
  - Update for  $c'$ :  $\mathbf{w}_{c'} \leftarrow \mathbf{w}_{c'} - \eta \mathbf{f}$
  - Update for all classes other than  $c$  and  $c'$ : no change



- *One-vs-others* framework: Need to keep a weight vector  $\mathbf{w}_c$  for each class  $c$
- Decision rule:  $y = \operatorname{argmax}_c \mathbf{w}_c \cdot \mathbf{f}$



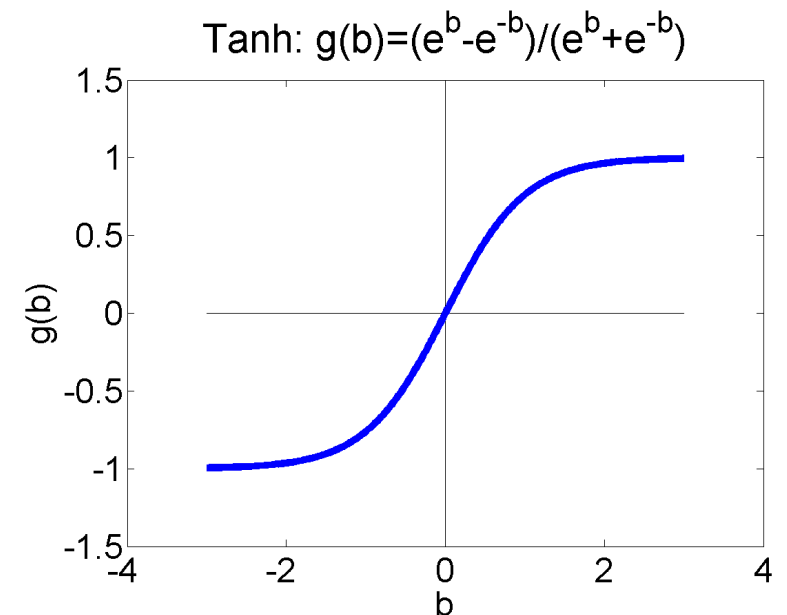
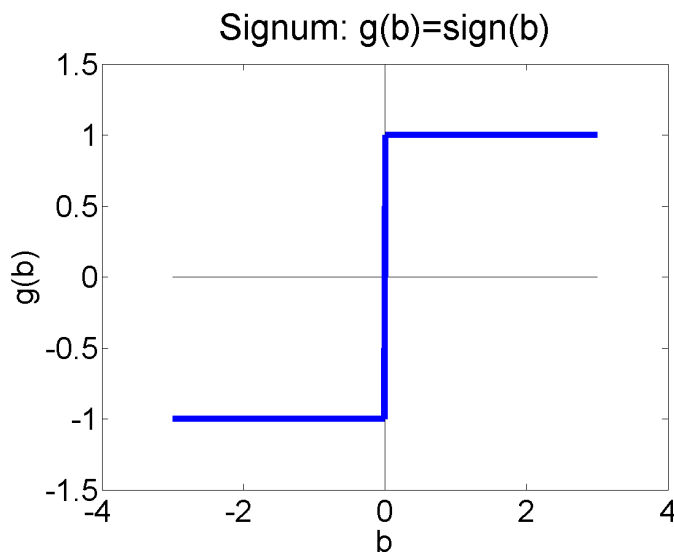
1. Naïve Bayes/BoW classifiers
2. Linear Classifiers in General
3. Perceptron
4. Differential Perceptron/Neural Net

- Also known as a “one-layer feedforward neural network,” also known as “logistic regression.” Has been re-invented many times by many different people.
- Basic idea: replace the non-differentiable decision function

$$y' = \text{sign}(\vec{w}^T \vec{f})$$

with a differentiable decision function

$$y' = \tanh(\vec{w}^T \vec{f})$$



- Suppose we have  $n$  training vectors,  $\vec{f}_1$  through  $\vec{f}_n$ . Each one has an associated label  $y_i \in \{-1, 1\}$ . Then we replace the true loss function,

$$L(y_1, \dots, y_n, \vec{f}_1, \dots, \vec{f}_n) = \sum_{i=1}^n (y_i - \text{sign}(\vec{w}^T \vec{f}_i))^2$$

with a differentiable error

$$L(y_1, \dots, y_n, \vec{f}_1, \dots, \vec{f}_n) = \sum_{i=1}^n (y_i - \tanh(\vec{w}^T \vec{f}_i))^2$$

- Why do we want a differentiable loss function?

$$L(y_1, \dots, y_n, \vec{f}_1, \dots, \vec{f}_n) = \sum_{i=1}^n (y_i - \tanh(\vec{w}^T \vec{f}_i))^2$$

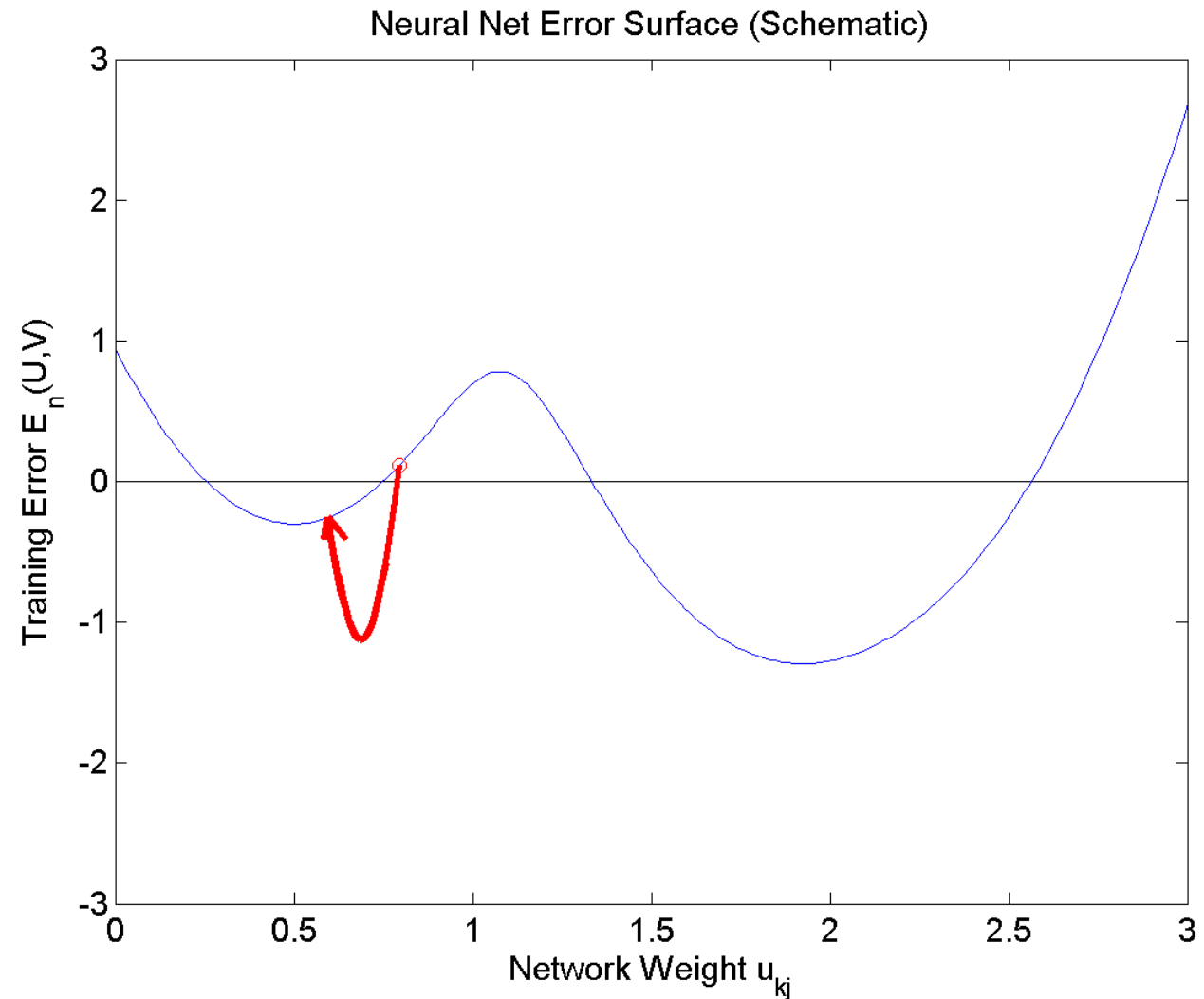
- Answer: because if we want to improve it, we can adjust the weight vector in order to reduce the error:

$$\vec{w} = \vec{w} - \eta \nabla_{\vec{w}} L$$

- This is called “gradient descent.” We move  $\vec{w}$  “downhill,” i.e., in the direction that reduces the value of the loss  $L$ .

The weights get updated according  
to

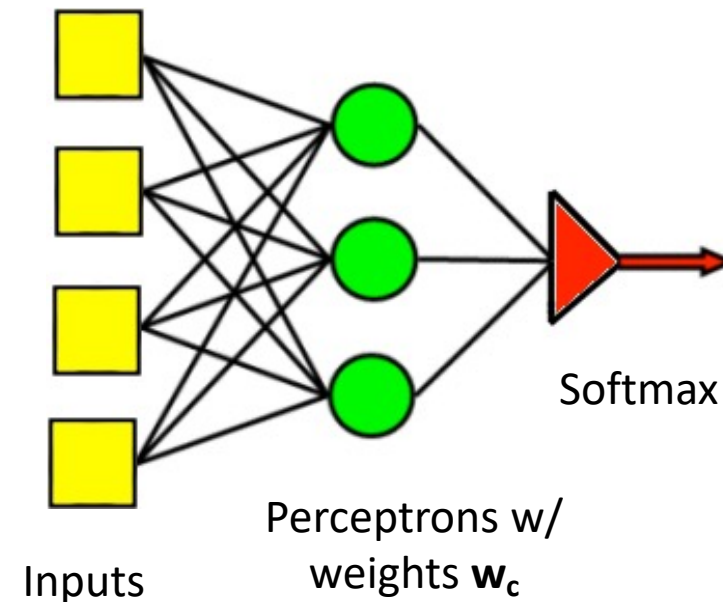
$$\vec{w} = \vec{w} - \eta \nabla_{\vec{w}} L$$



Same idea works for multi-class perceptrons. We replace the non-differentiable decision rule  $c = \operatorname{argmax}_c \mathbf{w}_c \cdot \mathbf{f}$  with the differentiable decision rule  $c = \operatorname{softmax}_c \mathbf{w}_c \cdot \mathbf{f}$ , where the softmax function is defined as

Softmax:

$$p(c|\vec{f}) = \frac{e^{\vec{w}_c \cdot \vec{f}}}{\sum_{k=1}^{\# \text{ classes}} e^{\vec{w}_k \cdot \vec{f}}}$$



- Then we can define the loss to be:

$$L(y_1, \dots, y_n, \vec{f}_1, \dots, \vec{f}_n) = - \sum_{i=1}^n \ln p(c = y_i | \vec{f}_i)$$

- And because the probability term on the inside is differentiable, we can reduce the loss using gradient descent:

$$\vec{w} = \vec{w} - \eta \nabla_{\vec{w}} L$$



You now know SEVEN!! different types of linear classifiers. These 5 types are things you should completely understand already now:

- One bit per document Naïve Bayes
- One bit per word token Naïve Bayes
- Linear classifier can implement some logical functions, like AND and OR, but not others, like XOR
- Perceptron
- Multi-class Perceptron

These 2 types of linear classifiers have been introduced today, and you should know the general idea, but you don't need to understand the equations yet. We will spend lots more time talking about those equations later in the semester.

- Differentiable Perceptron a.k.a. Logistic Regression
- Differentiable Multi-class perceptron