CSC4007 Advanced Machine Learning

Lesson 04: Classification-Discriminant Functions

by Vien Ngo
EEECS / ECIT / DSSC

Outline

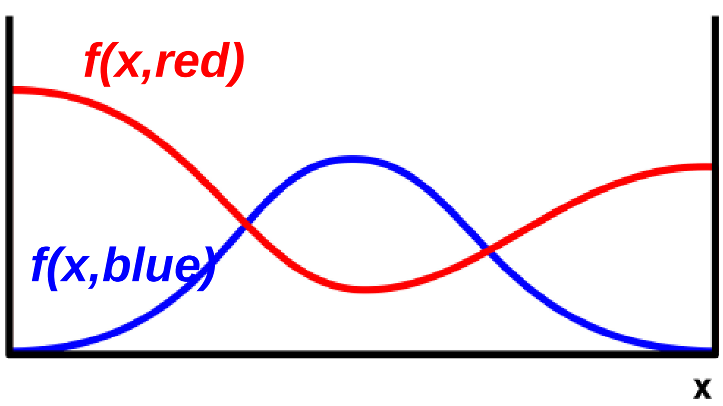
- The simplest approach: k-nearest neighbour
- Discriminate function
- Logistic regression for binary classification
- Multi-class classification

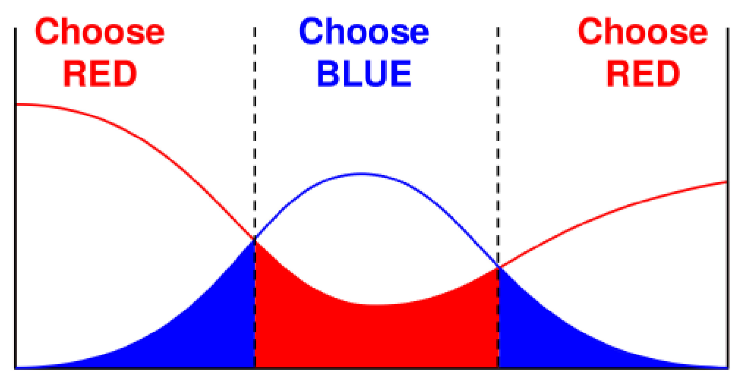
Outline

- The simplest approach: k-nearest neighbour
- **Discriminant function**
- Logistic regression for binary classification
- Multi-class classification

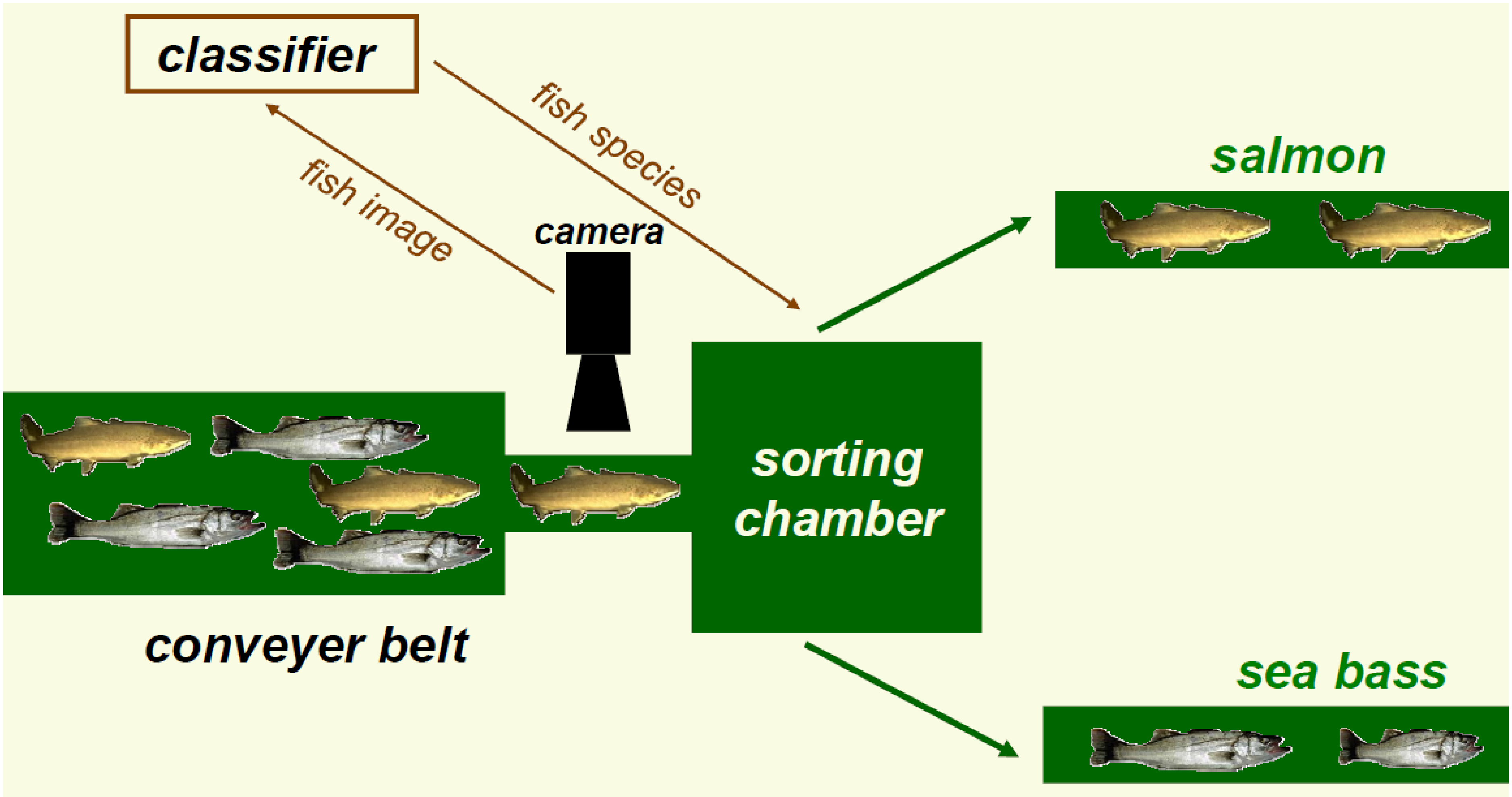
Discriminant function

- **How to represent hypothesis?**
  - e.g. classifer or classification functions F(x)
  - M-class classifier can be viewed as a set of M functions which computes M discriminant functions and selects category corresponding to the largest discriminant
  - Works for both the binary and multi-way classification
- **Ideas**:
  - For every class m, define a function *f(x,m)*
  - When making the decision on input x, choose the class with the highest value f(x,m)





Classification Example: Fish Sorting

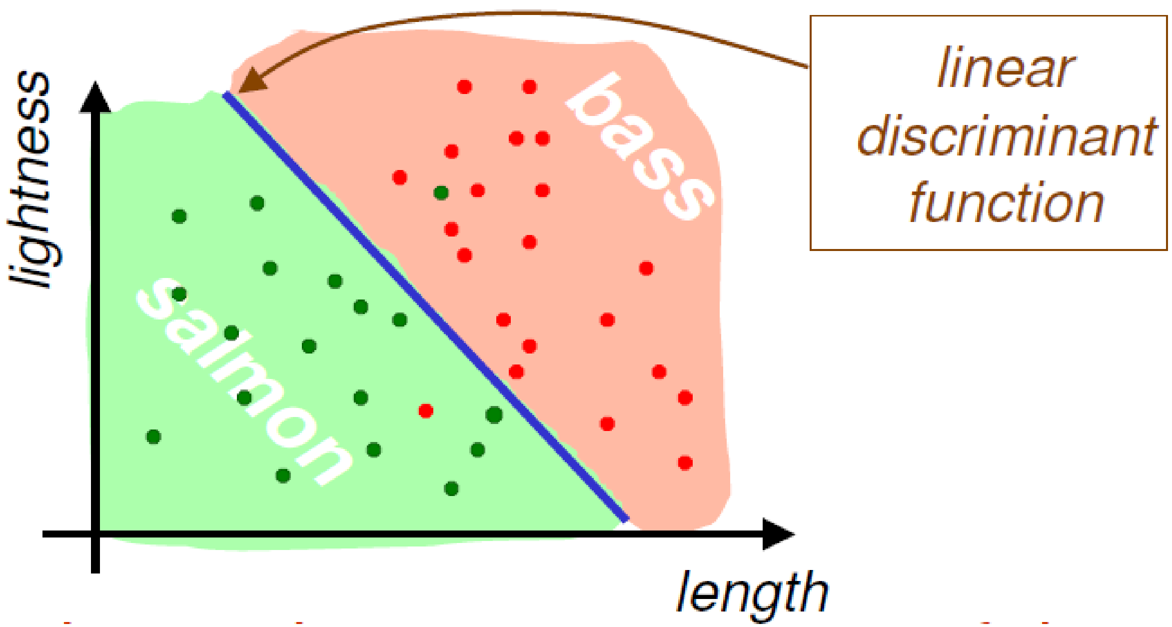


Olga Veksler

Classification Example: Fish Sorting

- **Given labeled data (binary class)**

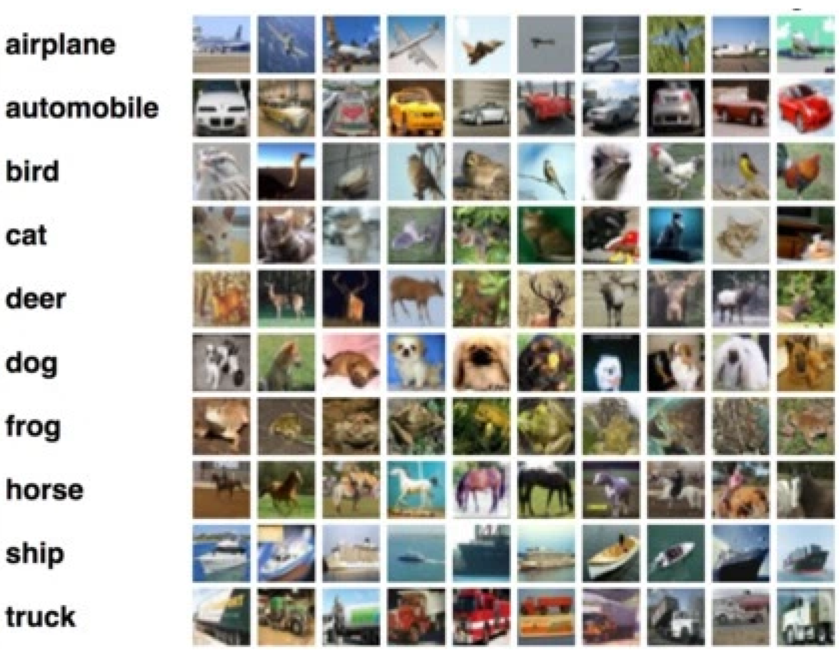
- **The shape of the discriminant function is known (linear)**



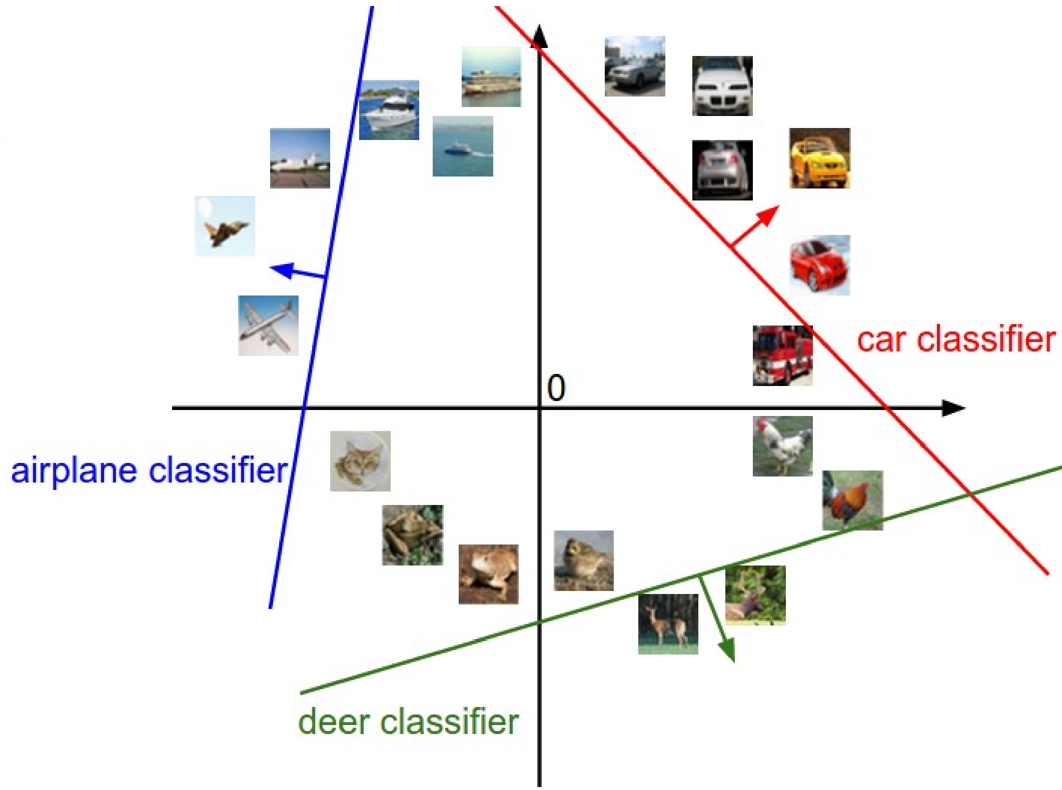
Olga Veksler

- **Need to estimate the parameters of the classifier?**

Classification Example: Multi-class classification



Given labeled data (CIFAR100)



Linear discriminant function (one vs. all classifiers)

Stanford lecture

Need to estimate the parameters of the classifiers.

Example: movie recommendation system

- A data-driven system to recommend a new movie: e.g. should I like *Gravity* if I know its rating and my own likes on some other movies?

Movie name	Mary's rating	John's rating	I like?
Lord of the Rings II	1	5	No
...	...	...	...
Star Wars I	4.5	4	Yes
Gravity	3	3	?

Le V. Quoc's tutorial

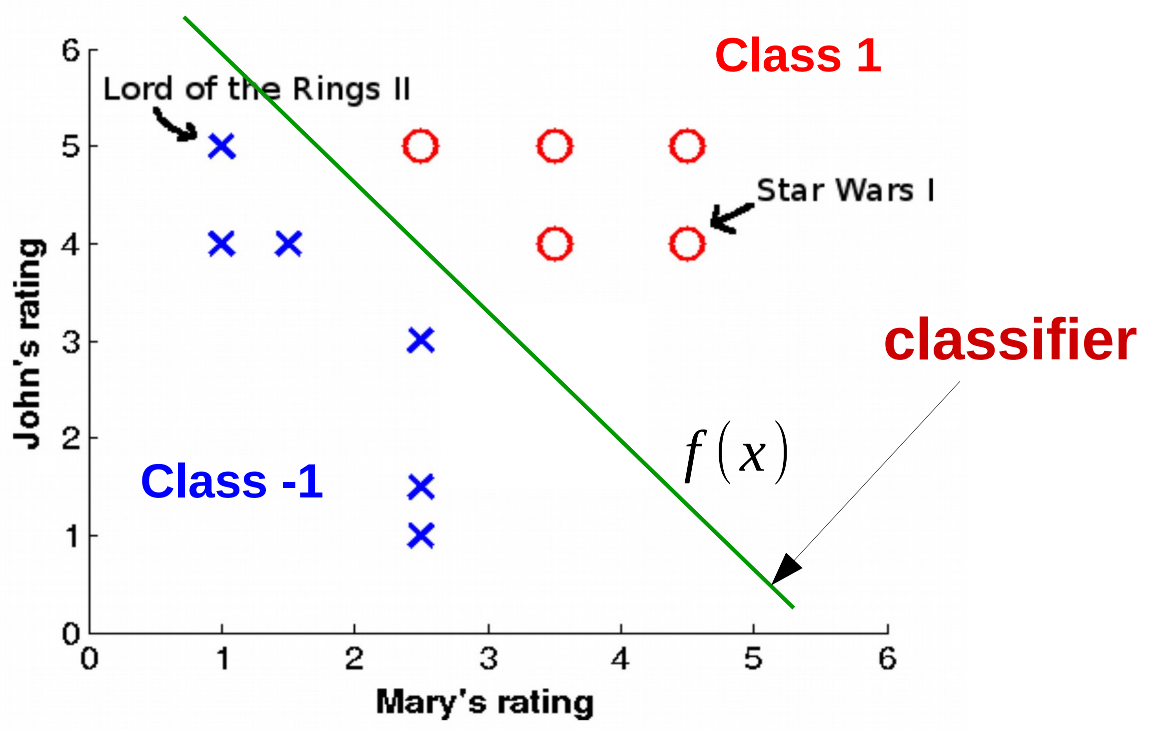
- **inputs** x_i: 2-dimensional ($x_{i1} =$ Mary's rating, $x_{i2} =$John's rating)
  – $x_1 = [1, 5] \quad x_2 = [4.5, 4]$
- **outputs** y: Yes or No
  – $y_1 =$ No $\quad y_2 =$ Yes
- **predictions**: given a new x, predict the label of y
  – $x = [3, 3]$, $y =$? (I would like the movie Gravity or not?)

Discriminant function: binary classification

- **How to represent hypothesis?**
  - e.g. classifer or classification functions
  - Output is {-1,1} (binary output)
  - Binary case: don’t need to maintain two functions for two classes (*f(x,1)* and *f(x,-1)* are replaced by *f(x)*)

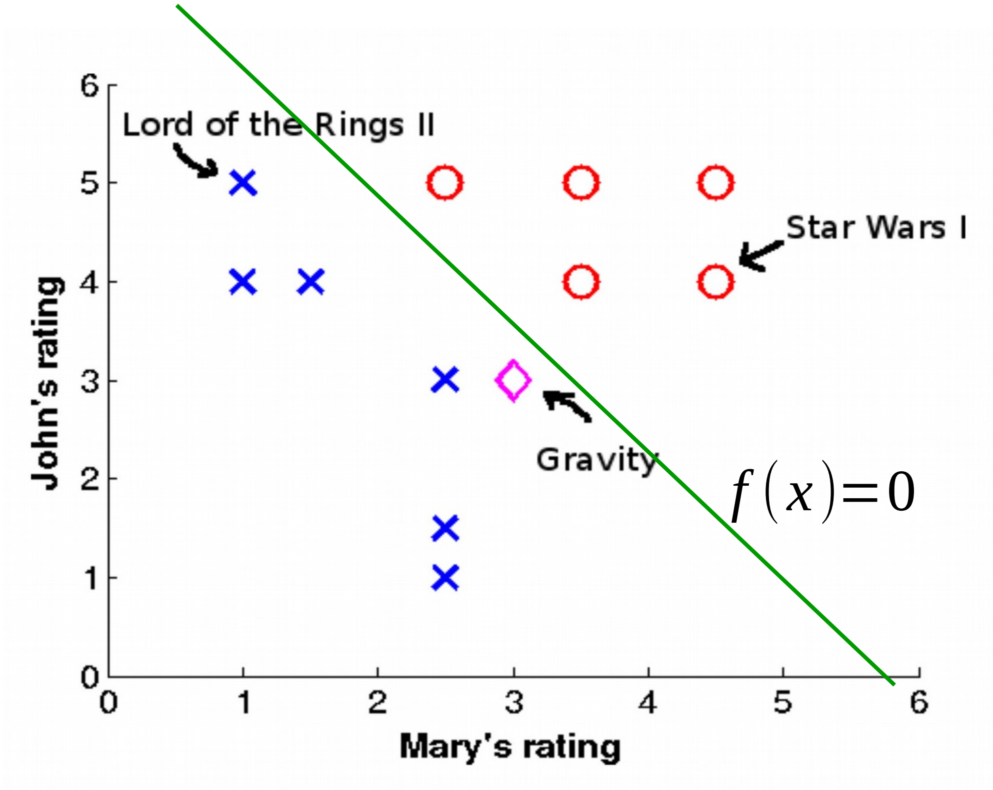
e.g.: f(x) = f(x,1) – f(x,-1)

Classifier: {**Yes**=1, **No**=-1}



Discriminant function: binary classification

- **Continuous function f(x) is used as a classifier**
  - If the value of *f(x)* is negative, then predict No (-1)
  - If the value of *f(x)* is positive, then predict Yes (1)
  - Points on classifier have zero values.

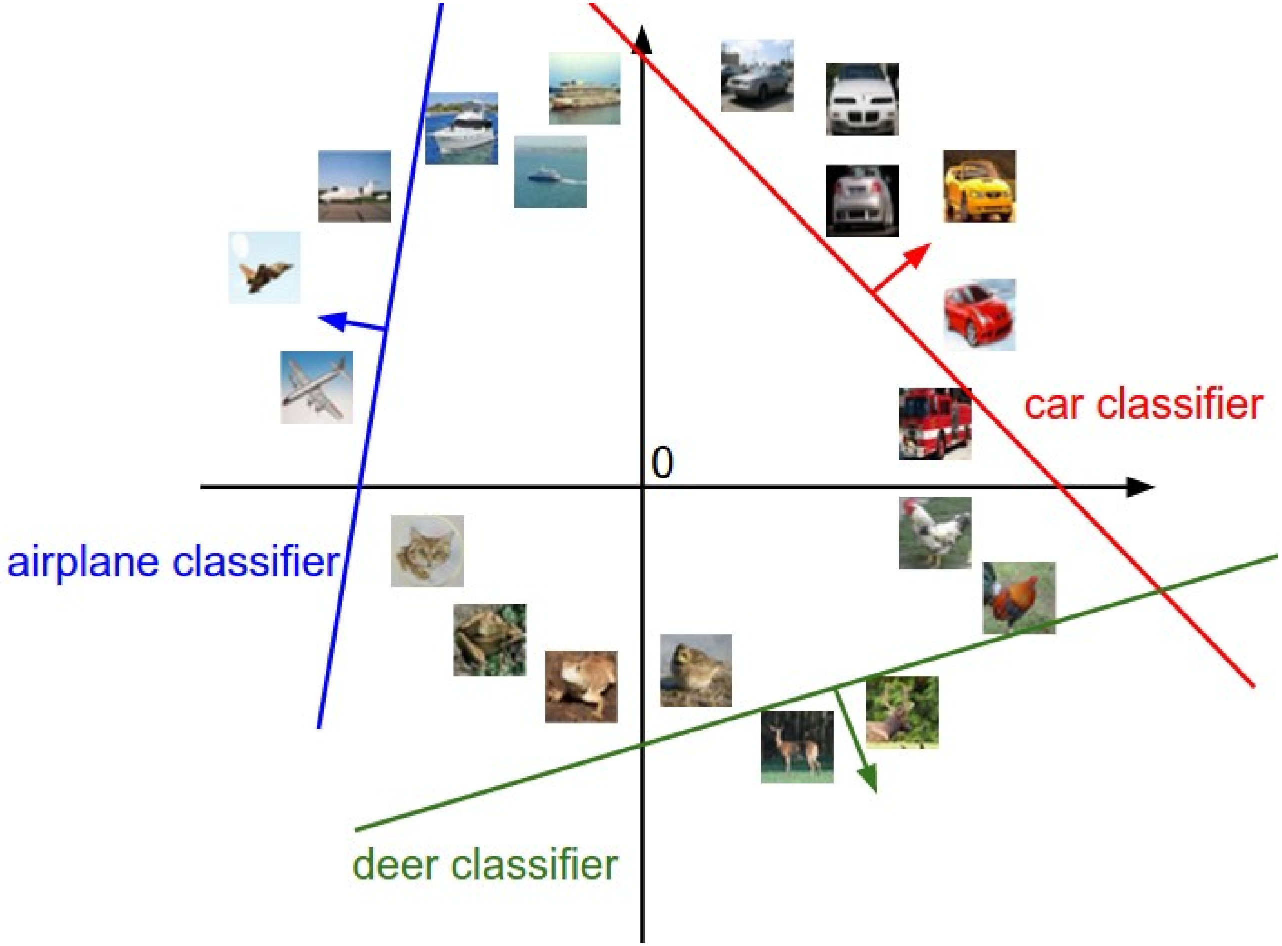


Regression function: $f(x)=0$ → **classifier**

Regression function ← $f(x)>0$ **predict** → **Class 1** → classification function

Regression function ← $f(x)<0$ **predict** → **Class -1** → classification function

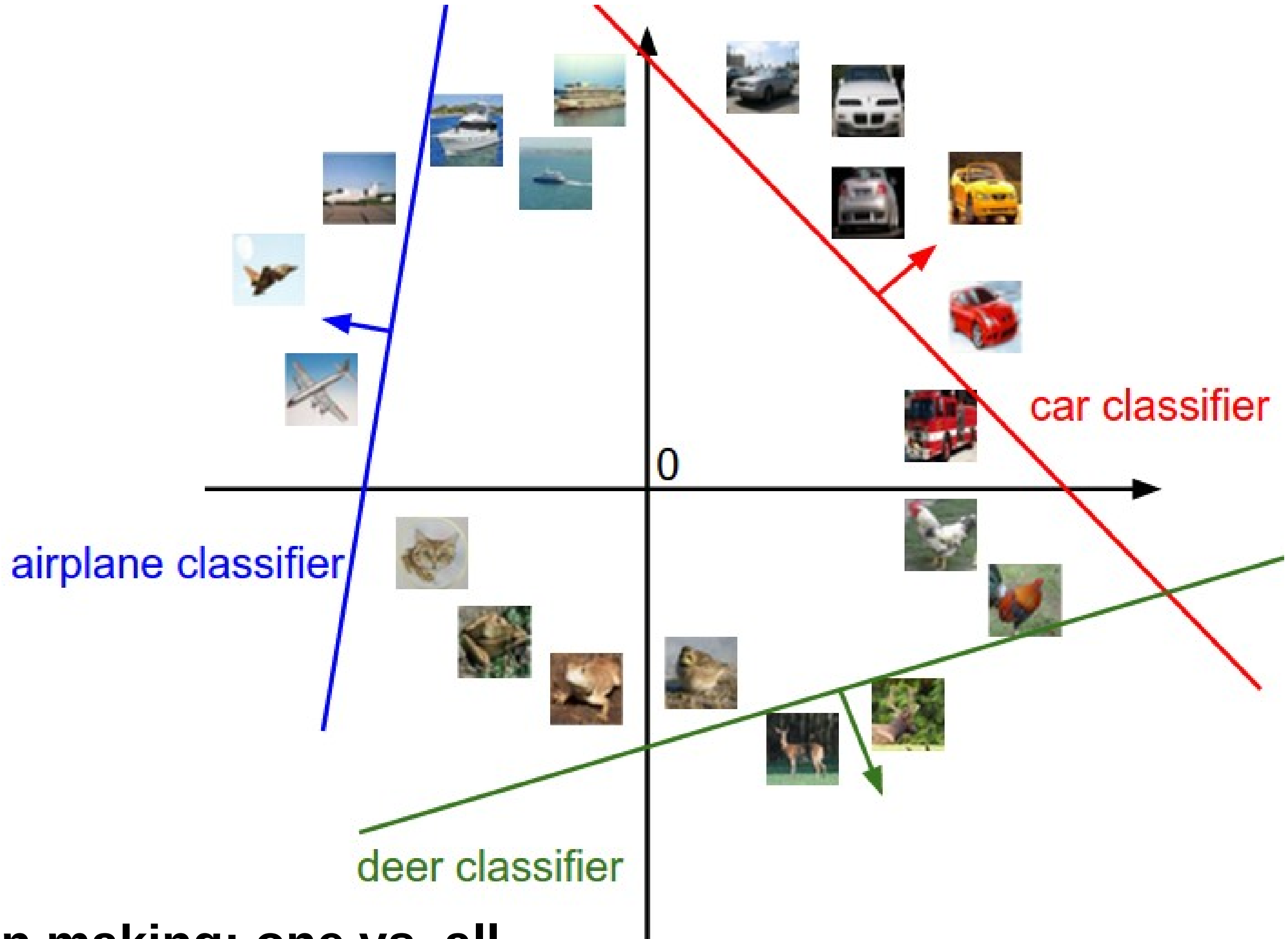
Discriminant function: multi-class classification



- **For each class (car, deer, airplane, etc.), construct one discriminant function**

$$f(x, car), f(x, deer), f(x, airplane), \ldots$$

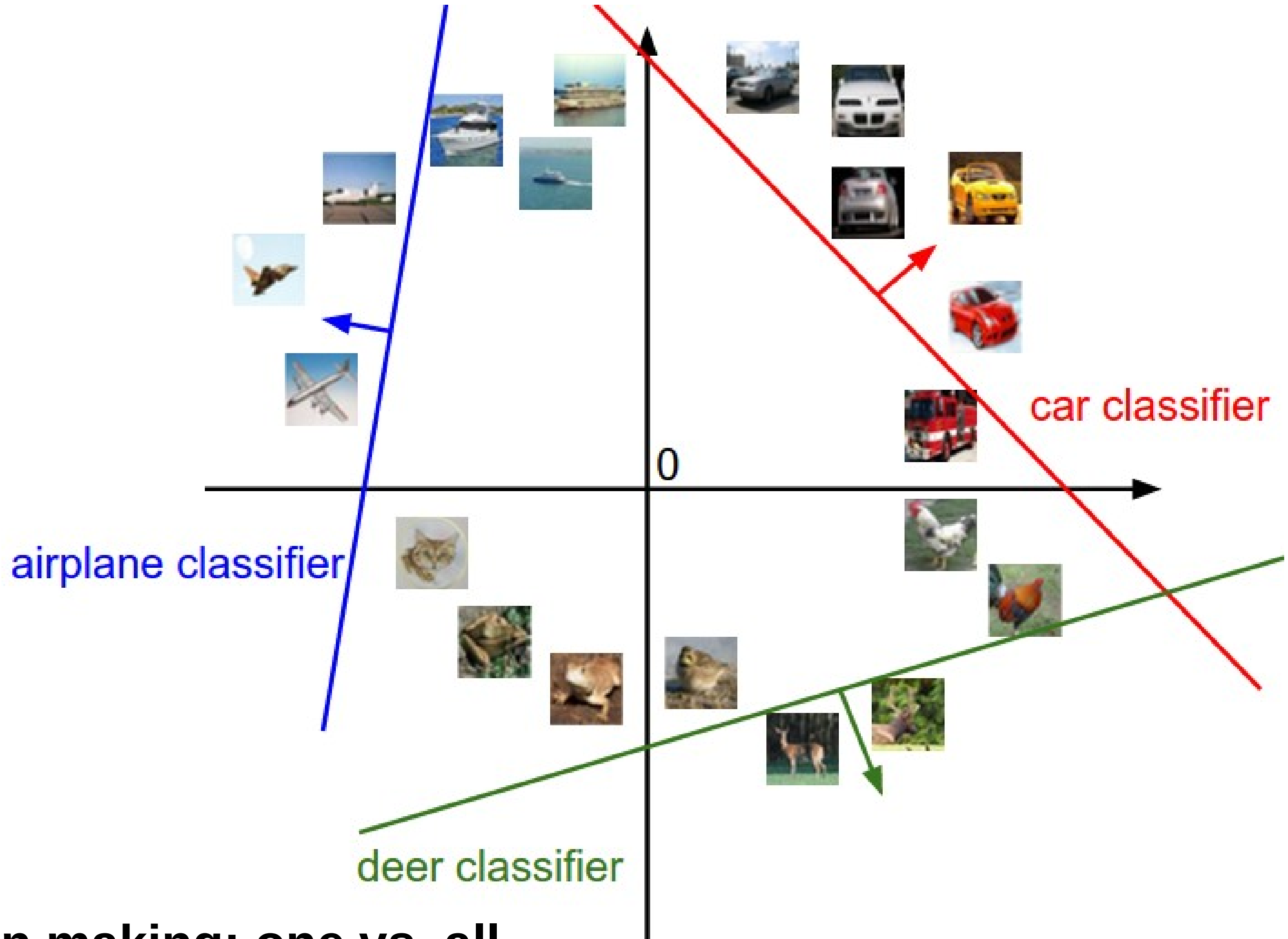
Discriminative function: multi-class classification



- **Decision making: one vs. all**

$$\left.\begin{array}{l} f(x, car) > f(x, deer) \\ f(x, car) > f(x, airplane) \\ \vdots \end{array}\right\} \Rightarrow$$ **y=car**

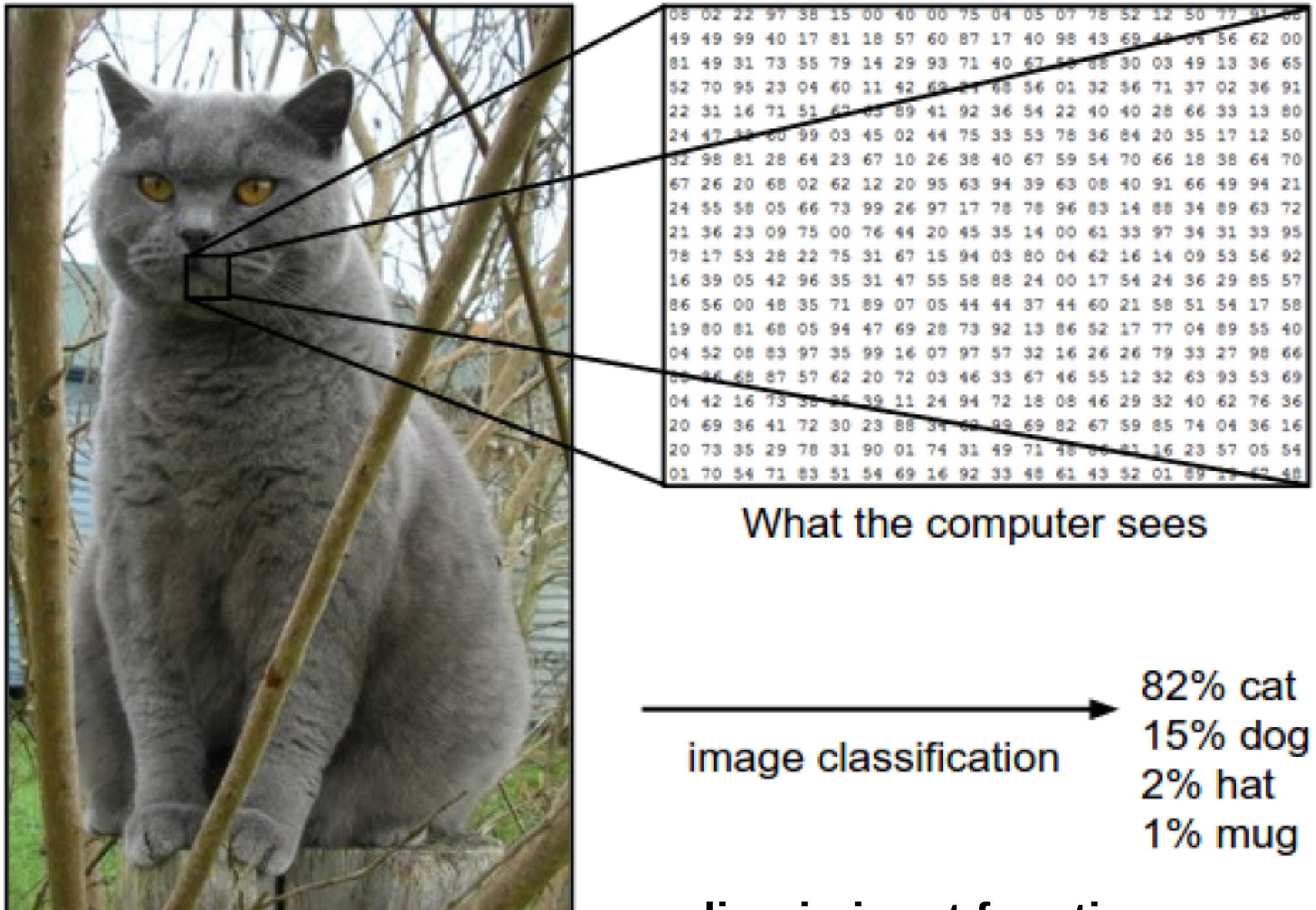
Discriminative function: multi-class classification



- **Decision making: one vs. all**

$$\left.\begin{array}{l} f(x, deer) > f(x, car) \\ f(x, deer) > f(x, airplane) \\ \vdots \end{array}\right\} \Rightarrow$$ **y=deer**

- example of image classification as probability function



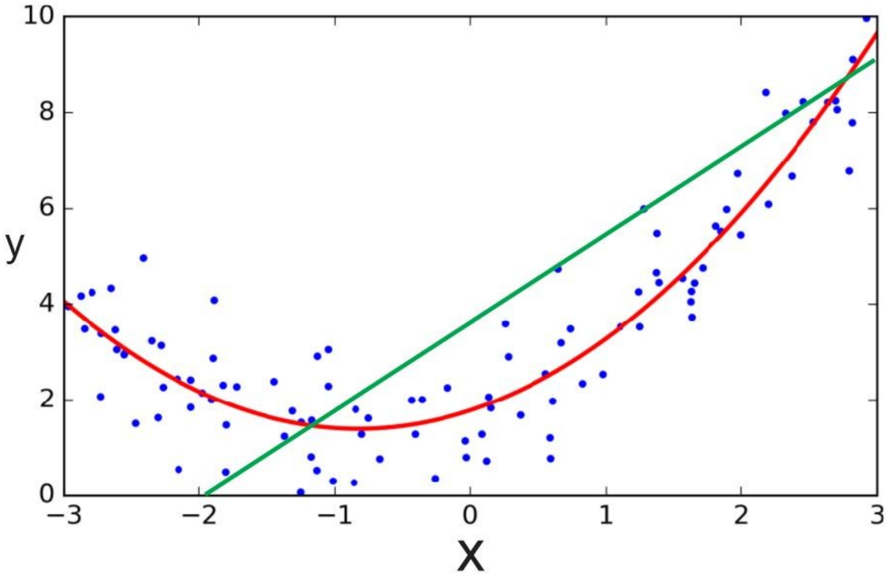
- **discriminant function**

– $p(y=\text{cat}|x)=0.82,\ p(y=\text{dog}|x)=0.15,\ p(y=\text{hat}|x)=0.02,\ p(y=\text{mug}|x)=0.01$

$f(x, cat)$ $f(x, dog)$ $f(x,hat)$ $f(x, mug)$

Discriminant function: representation

- Discriminant function f(x,y) is function of two arguments x (input data) and y (output data)
- Let’s approximate it similar to the regression function f(x)



Linear regression with polynomial features $f(x) = \beta^T \phi(x)$

Discriminant function: representation

- linear in features!

$$f(x, y) = \sum_{j=1}^{k} \phi_j(x, y)\beta_j = \phi(x, y)^\top \beta$$

- **example** (linear feature): Let $x \in \mathbb{R}$ and $y \in \{1, 2, 3\}$. Typical features might be

Regression Linear feature:

$$\phi(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$\phi(x, y) = \begin{pmatrix} 1 & [y = 1] \\ x & [y = 1] \\ 1 & [y = 2] \\ x & [y = 2] \\ 1 & [y = 3] \\ x & [y = 3] \end{pmatrix}$$

– where we denote $[y = k]$ means: $[y = k] = 1$ if (y==k), $= 0$ otherwise

– linear features rewritten: $\phi(x, y) = \begin{pmatrix} \phi(x)[y = 0] \\ \phi(x)[y = 1] \\ \phi(x)[y = 2] \end{pmatrix}$

where $\phi(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}$

Discriminant function: representation

- linear in features!

$$f(x,y) = \sum_{j=1}^{k} \phi_j(x,y)\beta_j = \phi(x,y)^\top \beta$$

Linear Discriminant Function

Linear Discriminant Function: Binary Classification

Movie recommendation system example

- an example of two data points $(x_1, y_1), (x_2, y_2)$

  x_1 (lord of the ring) $= \{1, 5\}, \quad y_1 = -1$ (No)

  x_2 (star wars I) $= \{4.5, 4\}, \quad y_2 = 1$ (Yes)

- Discriminant function (binary case) is

$$f(x_i) = \beta^T \phi(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- linear feature

$$\phi(x_1) = \begin{pmatrix} 1 \\ x_{11} \\ x_{12} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 5 \end{pmatrix} \quad \phi(x_2) = \begin{pmatrix} 1 \\ x_{21} \\ x_{22} \end{pmatrix} = \begin{pmatrix} 1 \\ 4.5 \\ 4 \end{pmatrix}$$

Linear Discriminant Function: Binary Classification

Movie recommendation system example

- an example of two data points $(x_1, y_1), (x_2, y_2)$
  x_1 (lord of the ring) $= \{1, 5\}$, $\quad y_1 = -1$ (No)
  x_2 (star wars I) $= \{4.5, 4\}$, $\quad y_2 = 1$ (Yes)
- so parameters $\beta \in \mathbb{R}^3$, for example

$$f(x_1) = \phi(x_1)^\top \beta = \beta_0 + \beta_1 + 5\beta_2 \quad f(x_2) = \phi(x_2)^\top \beta = \beta_0 + 4.5\beta_1 + 4\beta_2$$

Prediction for binary classification

$$f(x) > 0 \xrightarrow{\text{predict}} \text{Class 1}$$

$$f(x) < 0 \xrightarrow{\text{predict}} \text{Class -1}$$

Linear Discriminant Function: Binary Classification

Movie recommendation system example

- an example of two data points $(x_1, y_1), (x_2, y_2)$

  x_1 (lord of the ring) $= \{1, 5\}, \quad y_1 = -1$ (No)

  x_2 (star wars I) $= \{4.5, 4\}, \quad y_2 = 1$ (Yes)

- classifying?
  – for example: if current parameters are $\beta = [1, 1, 2]$, so

$$f(x_1) = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}^T \begin{pmatrix} 1 \\ 1 \\ 5 \end{pmatrix} = 1 + 1 + 5 \cdot 2 = 12$$

decision: $f(x_1) > 0$, then the system predicts y = 1 (Yes)

$y_1 = -1$

Different from the ground truth. Let's find optimum β

Linear Discriminant Function: Binary Classification

Movie recommendation system example

- an example of two data points $(x_1, y_1), (x_2, y_2)$

  x_1 (lord of the ring) $= \{1, 5\}, \quad y_1 = -1$ (No)

  x_2 (star wars I) $= \{4.5, 4\}, \quad y_2 = 1$ (Yes)

- **Exercise**: Predict if I like to watch the movie *"Star Wars I"* or not, if our model's current parameter is at $\beta = [1, 1, 2]$

Our linear discriminant function is:

$$f(x_i) = \beta^T \phi(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Linear Discriminant Function: Binary Classification

Movie recommendation system example

- an example of two data points $(x_1, y_1), (x_2, y_2)$
  x_1 (lord of the ring) $= \{1, 5\}, \quad y_1 = -1$ (No)
  x_2 (star wars I) $= \{4.5, 4\}, \quad y_2 = 1$ (Yes)

- **Exercise**: Predict if I like to watch the movie Star Wars I or not, if our model's current parameter is at $\beta = [1, 1, 2]$

Our linear discriminant function is:

$$f(x_i) = \beta^T \phi(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

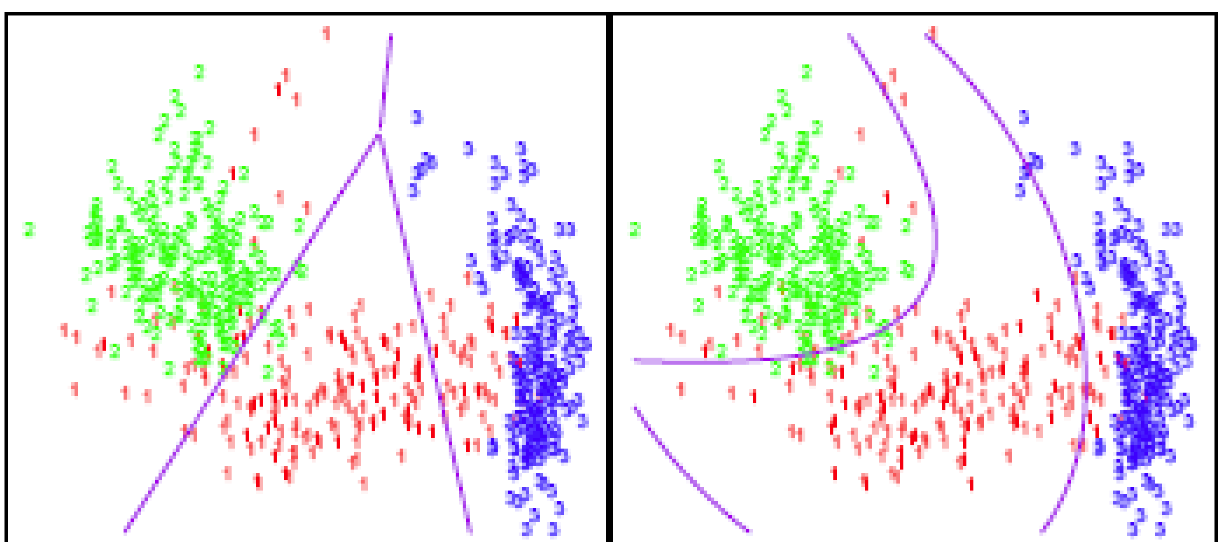
- **Solution**: $f(x_2) = 1 + 4.5 + 2 \times 4 = 13.5 \longrightarrow$ Predict y = 1 (Yes)

Example: quadratic feature, multi-class cases

- Example (**quadratic feature**): Let $x \in \mathbb{R}$ and $y \in \{1, 2, 3\}$. Typical features might be

$$\phi(x, y) = \begin{pmatrix} 1 & [y = 1] \\ x & [y = 1] \\ x^2 & [y = 1] \\ 1 & [y = 2] \\ x & [y = 2] \\ x^2 & [y = 2] \\ 1 & [y = 3] \\ x & [y = 3] \\ x^2 & [y = 3] \end{pmatrix}$$

linear features | quadratic features



Example: quadratic feature, multi-class cases

- Example (**quadratic feature**): Let $x \in \mathbb{R}$ and $y \in \{1, 2, 3\}$. Typical features might be

$$\phi(x, y) = \begin{pmatrix} 1 & [y=1] \\ x & [y=1] \\ x^2 & [y=1] \\ 1 & [y=2] \\ x & [y=2] \\ x^2 & [y=2] \\ 1 & [y=3] \\ x & [y=3] \\ x^2 & [y=3] \end{pmatrix}$$

Given input data: x

Prediction (one vs. all): If $f(x,1) > f(x,2)$ and $f(x,1) > f(x,3)$

$\longrightarrow$ Predict y = 1

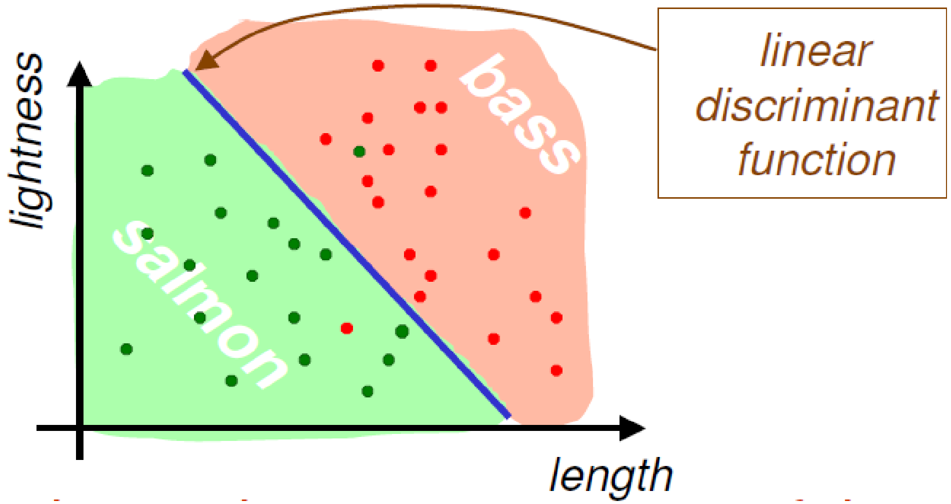
Prediction (one vs. all): If $f(x,2) > f(x,1)$ and $f(x,2) > f(x,3)$

$\longrightarrow$ Predict y = 2

Prediction (one vs. all): If $f(x,3) > f(x,1)$ and $f(x,3) > f(x,2)$

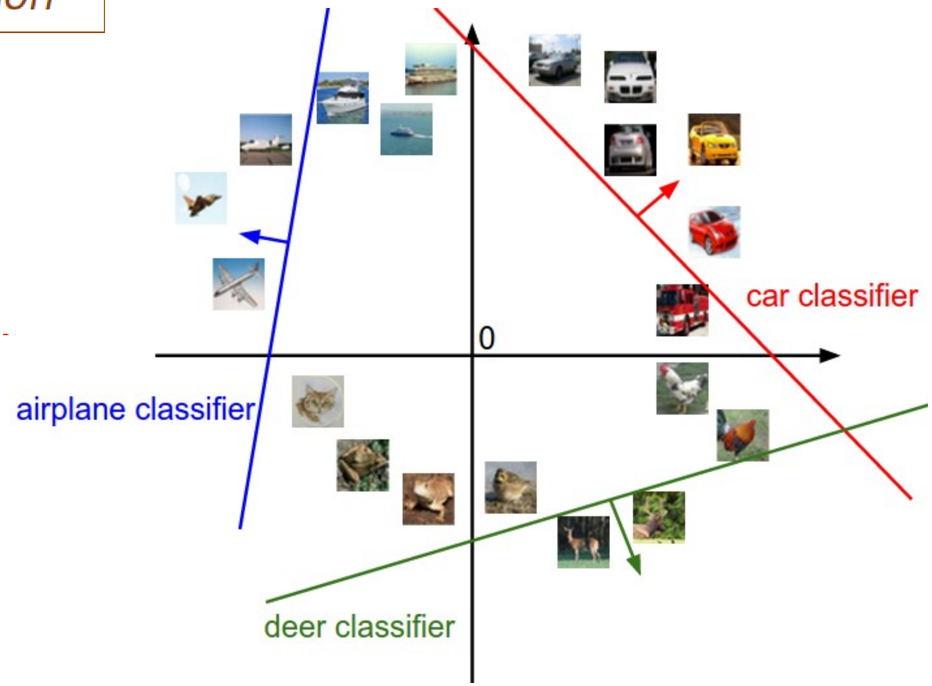
$\longrightarrow$ Predict y = 3

Discriminant Functions: Optimization?



Binary classification

$$f(x) = \beta^T \phi(x)$$



Multi-class classification

$$f(x, y) = \beta^T \phi(x, y)$$

Finding optimum parameters for classification:

Logistic Regression