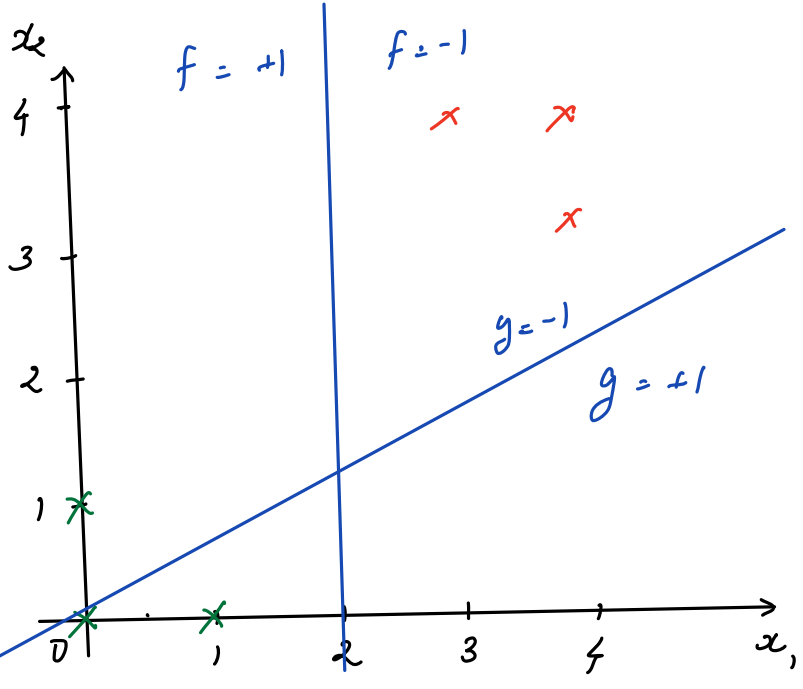
Classification

- E.g. Predict if rooms>3 from area and price.
- Training data: $\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^n, y^n)\}$
- $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \{+1, -1\}$
- Algorithm outputs a model $f : \mathbb{R}^d \rightarrow \{+1, -1\}$
- Loss[f] $= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(f(\mathbf{x}^i) \neq y^i)$ = Fraction of training data classified wrongly by f
- $f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$

Linear separator

Classification Illustration 1

x	y	f	g
[0, 0]	+1	+1	+1
[0, 1]	+1	+1	-1
[1, 0]	+1	+1	+1
[4, 4]	-1	-1	-1
[3, 4]	-1	-1	-1
[4, 3]	-1	-1	-1



$f(x) = \text{sign}(2 - x_1)$

$g(x) = \text{sign}(x_1 - 2x_2)$

$\text{Loss}[f] = \frac{1}{6}(0) = 0$

$\text{Loss}[g] = \frac{1}{6}(1) = 1/6$

Classification Illustration 2

Area	Price	Rooms
9	5.0	-1
7	3.1	-1
12	6.9	+1
16	9.7	+1
15	8.5	+1
11	7.1	+1

Rooms=1 or 2 or 3 is encoded as -1.
Rooms > 3 is encoded as +1.

f	g	h
-1	-1	-1
-1	-1	-1
+1	+1	-1
+1	+1	+1
+1	+1	-1
+1	+1	-1

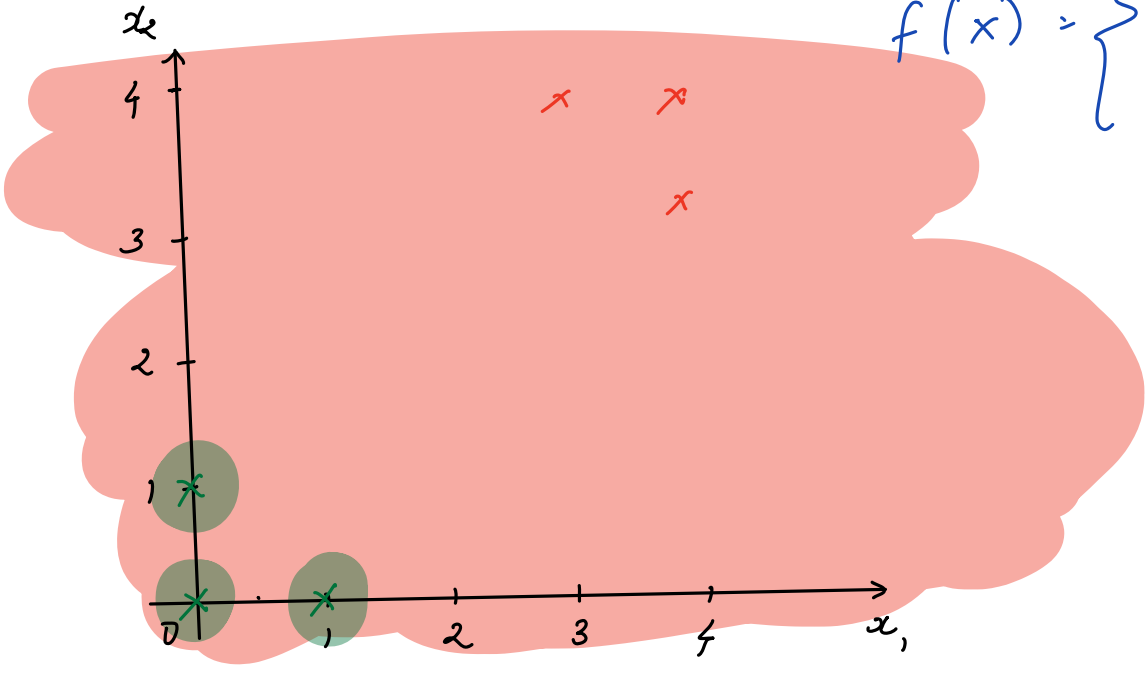
$f(x)$: sign (area - 10)

$g(x)$: sign (Price - 6)

$h(x)$: sign (Price - 9)

Evaluating Learned Models : Test Data

- Learning algorithm uses training data $(x^1, y^1), \ldots, (x^n, y^n)$ to get model f.
- But evaluating the learned model must **not** be done on the training data itself.
- Use test data that is **not** in the training data for model evaluation.



$$f(x) = \begin{cases} +1 & \text{if } x = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ +1 & \text{if } x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ +1 & \text{if } x = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ -1 & \text{otherwise} \end{cases}$$

Model Selection : Validation Data

- Learning algorithms just find the “best” model in the collection of models given by the human.
- How to find the right collection of models?
- This is called model selection, and it is done by using another subset of data called **validation data** that is distinct from train and test data.

$$Price = w_1 * (\#rooms) + w_2 (area) + w_3 (distance) + b$$