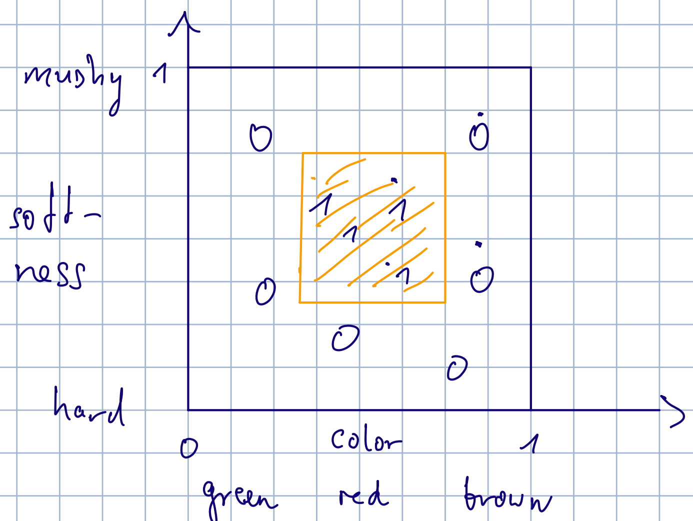
Suppose we want to predict when a papaya is tasty.

Based on experience with other fruits, we assume that it depends on the color and on the softness of the papaya. Assume we measure both parameters on an interval $[0,1]$.



We try a couple of fruits and estimate color and softness and label them with a 1 for tasty or 0 for not tasty.

Goal: Find a function $h: X = [0,1]^2 \to \{0,1\} = Y$ s.t.

$$h(a,b) = \begin{cases} 1 & \text{if papayas with color value } a \text{ and softness value } b \text{ are (usually) tasty} \\ 0 & \text{" " " " not tasty.} \end{cases}$$

Unknown distribution $\mathcal{D}$ on $X \times \{0,1\}$ (or on X) represents environment.

Given our labeled data $(x_1, y_1) = (\underbrace{(a_1, b_1)}_{x_1}, \underset{\{0,1\}}{y_1})$,

(x_2, y_2), ..., (x_m, y_m)

how should we choose h?

- h should satisfy $h(x_i) = y_i$ for (at least most of) our data points (x_i, y_i).

But there may be many such functions...

E.g. we could choose:

$$h(x) = \begin{cases} 1 & \text{if there is a training example } (x_i, y_i) = (x, 1) \\ 0 & \text{otherwise} \end{cases}$$

Then $h(x_i) = y_i \quad \forall\, i = 1, \dots, m$

(if all x_i are distinct)

Does not seem to be realistic ($\to$ overfitting)

Other possibility (hopefully more realistic)

assuming there are intervals $I_c \subset [0, 1]$ and $I_s \subset [0, 1]$ such that the tasty papayes are those with color value in I_c and softness value in I_s. Then h would be the characteristic function of $I_c \times I_s$ i.e.

$$h(a, b) = \begin{cases} 1 & \text{if } (a, b) \in I_c \times I_s \\ 0 & \text{otherwise} \end{cases}$$

$\rightsquigarrow$ We want to define a class H of functions $h: [0, 1]^2 \to \{0, 1\}$ from which we choose h.

E.g. $\mathcal{H}$ could be the set of characteristic functions of axis aligned rectangles.

Choosing a suitable $\mathcal{H}$ requires preknowledge about the problem.

$\mathcal{H}$ should <u>not</u> be the class of all functions

$$h: \underbrace{[0,1]^2}_{X} \longrightarrow \{0,1\}$$

Will see later rigorously and more generally why.