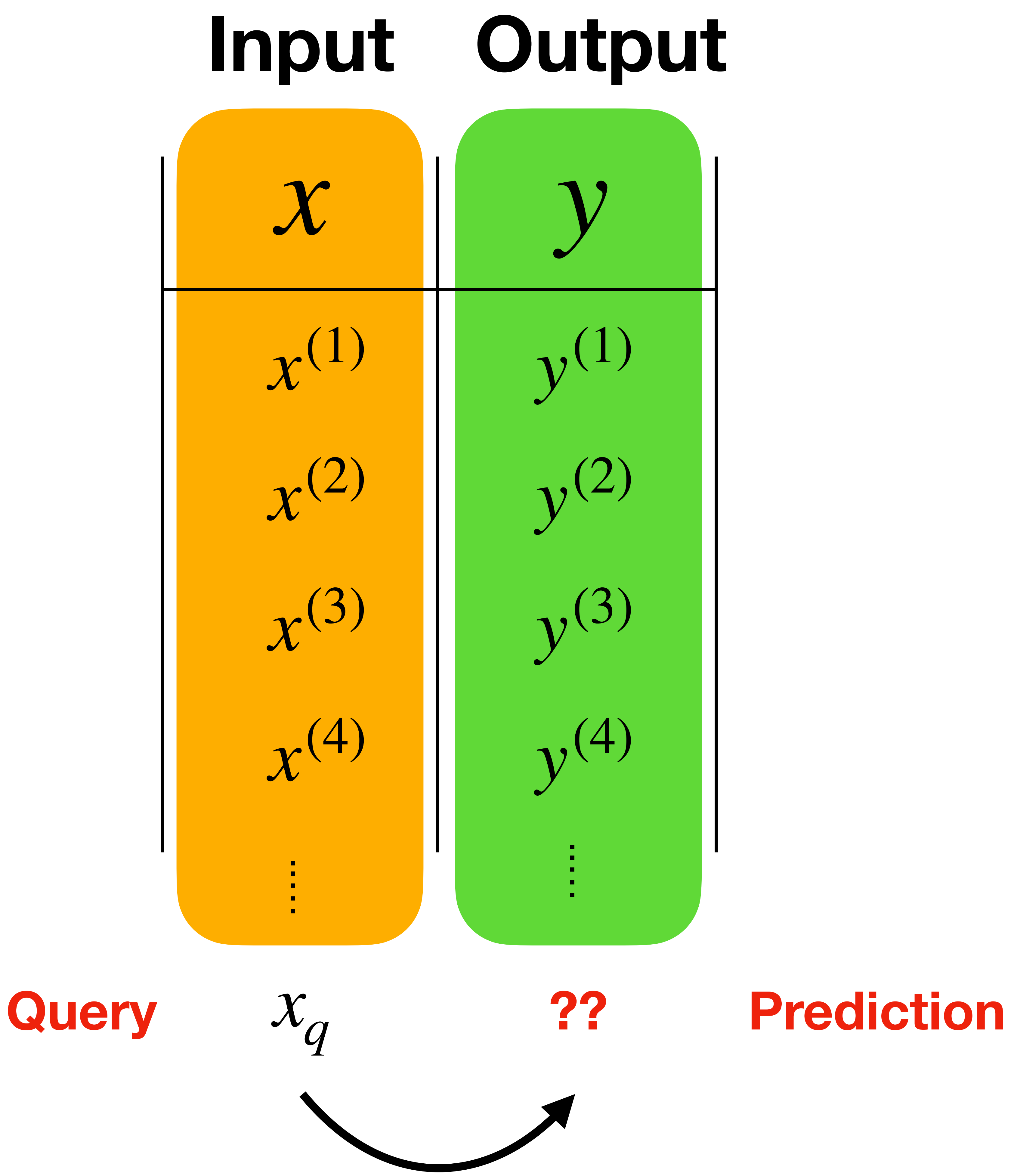
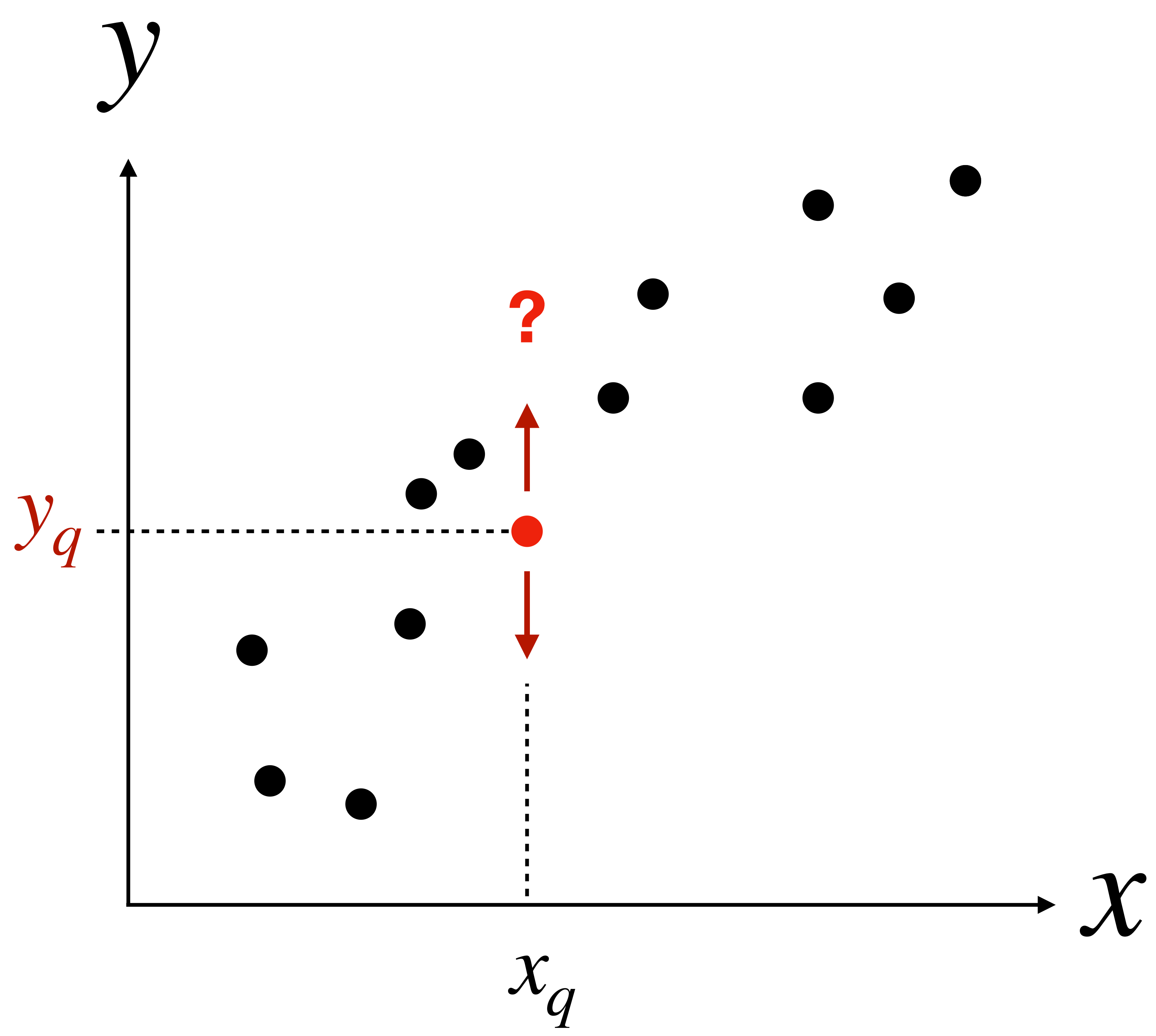
Logistic Regression

Prepared by: Joseph Bakarji

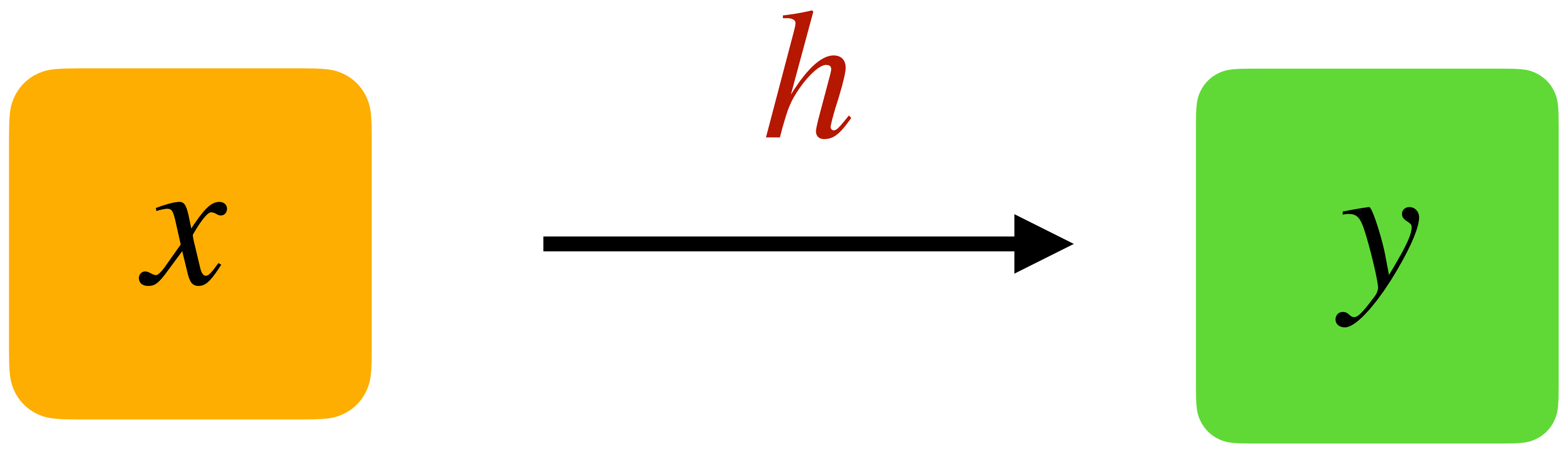
Given new input, what's the output?





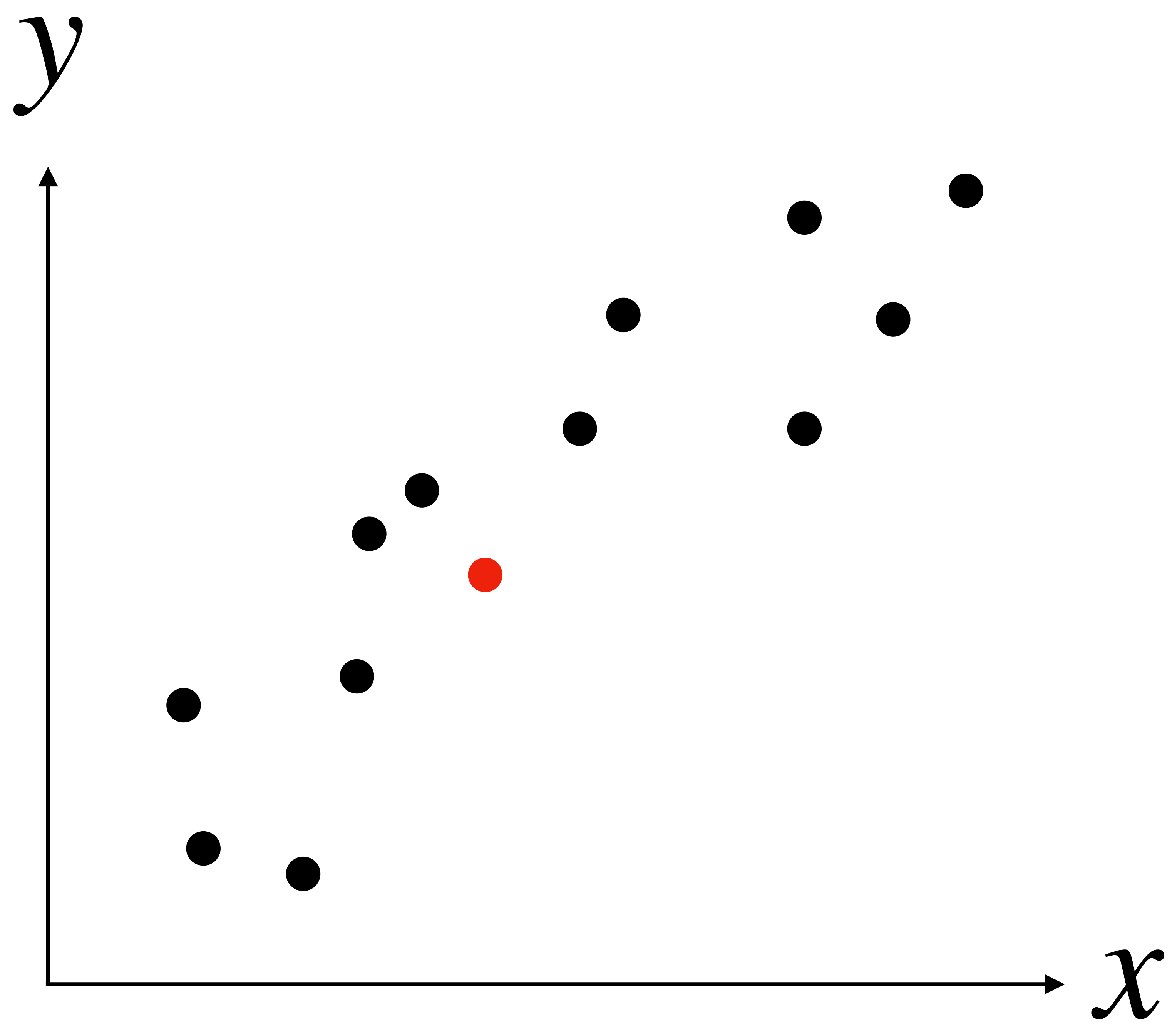
Given new input, what's the output?

Assuming $y \in \mathbb{R}$

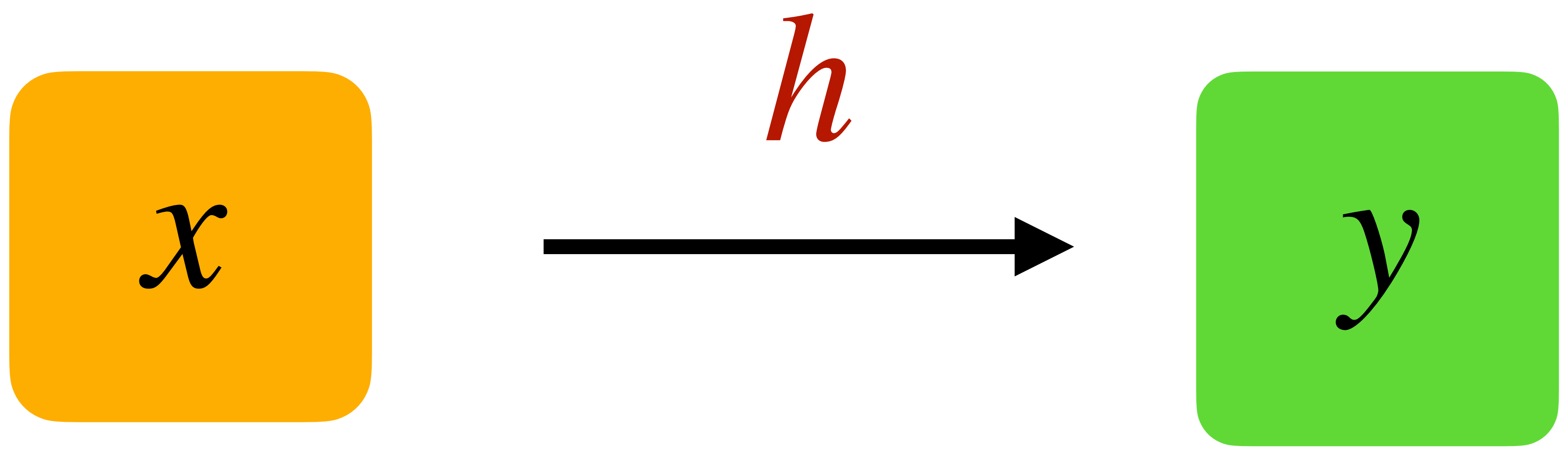


Given the data,
find a **function** h,
that predicts y, given x

$$\mathbf{y} = h(\mathbf{x})$$



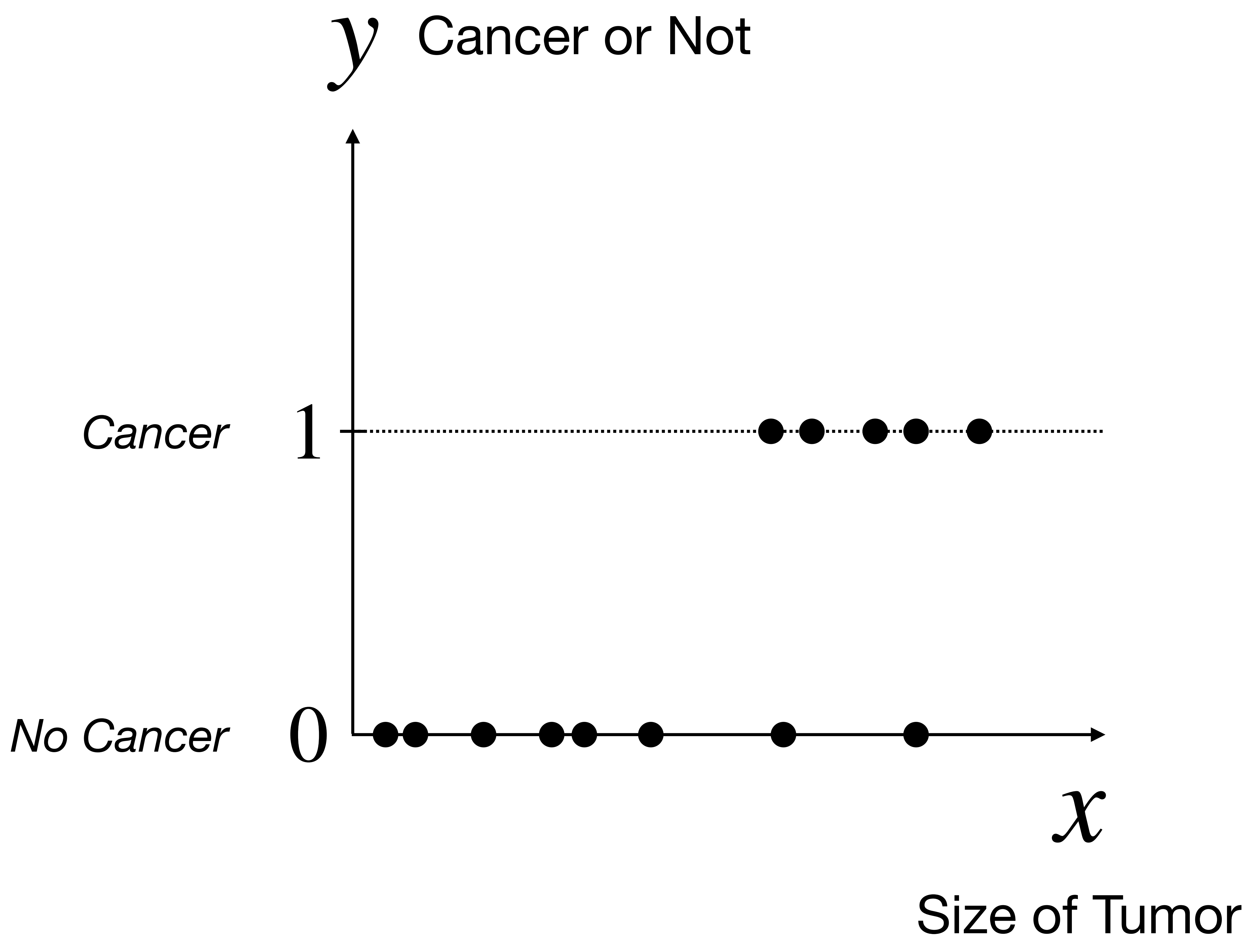
What if y is a label?



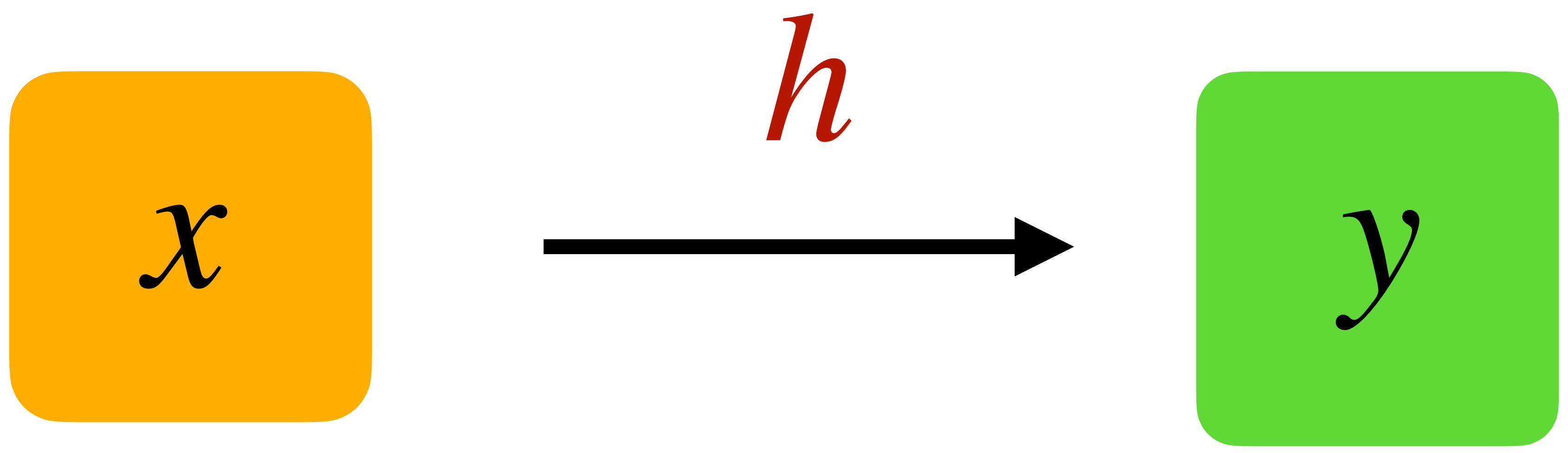
Given the data,
find a **function** h,
that predicts y, given x

$$\mathbf{y} = h(\mathbf{x})$$

$$y \in [0,1]$$



What if y is a label?

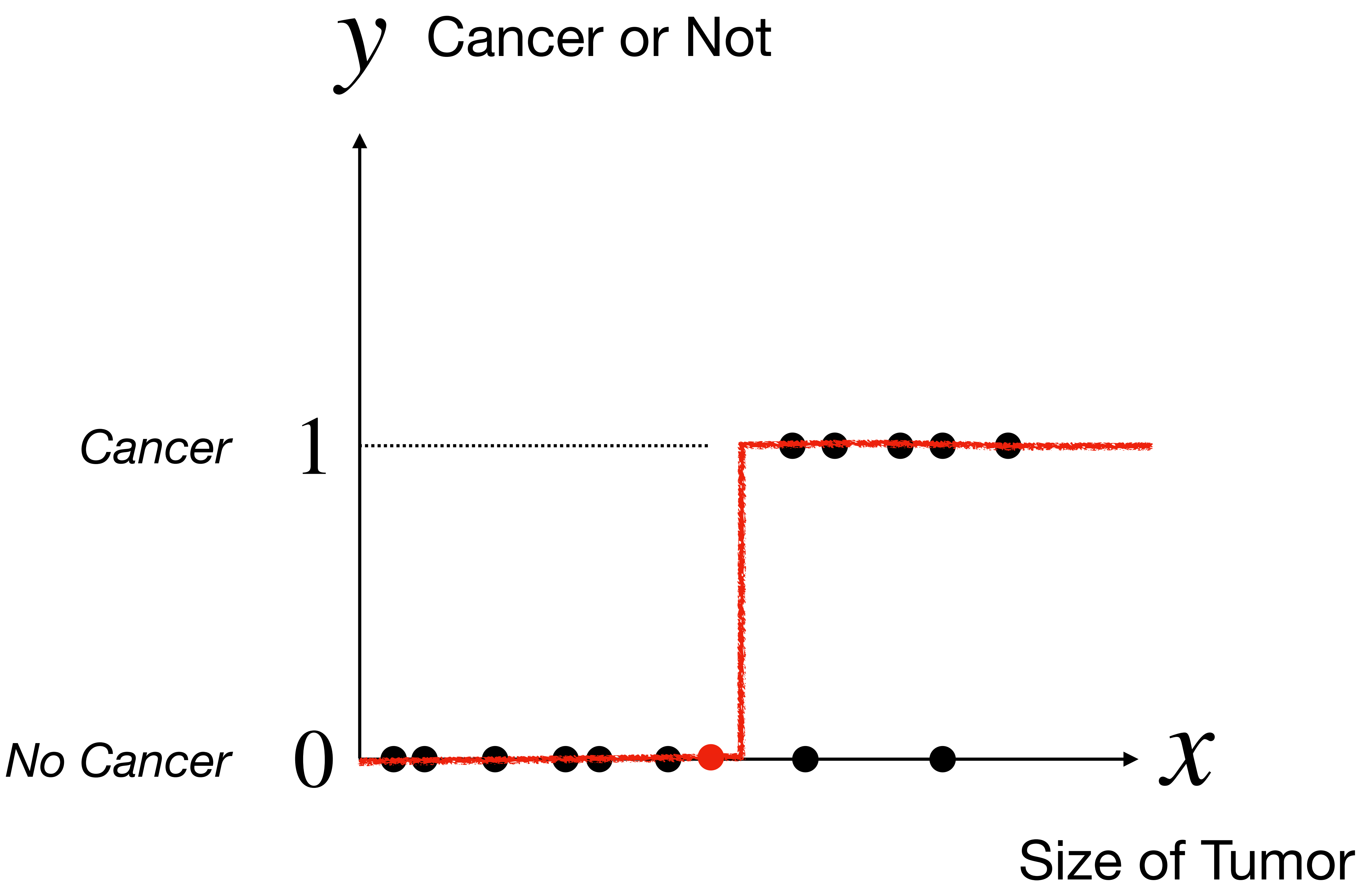


Given the data,
find a **function** h,
that predicts y, given x

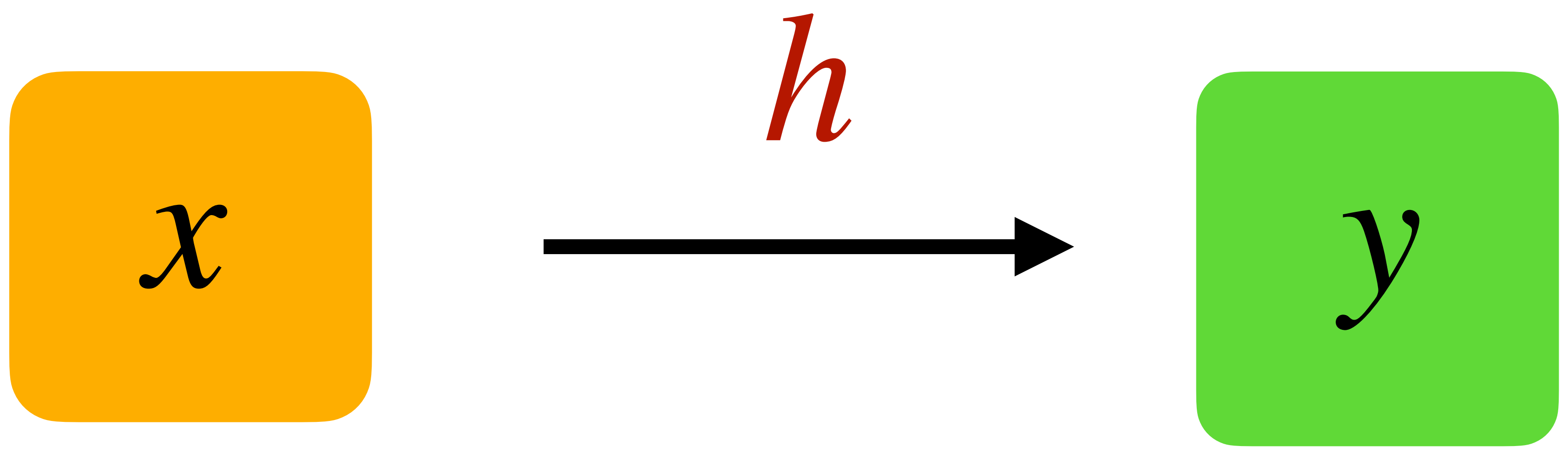
$$\mathbf{y} = h(\mathbf{x})$$

$$y \in [0,1]$$

A step function, or threshold



What if y is a label?

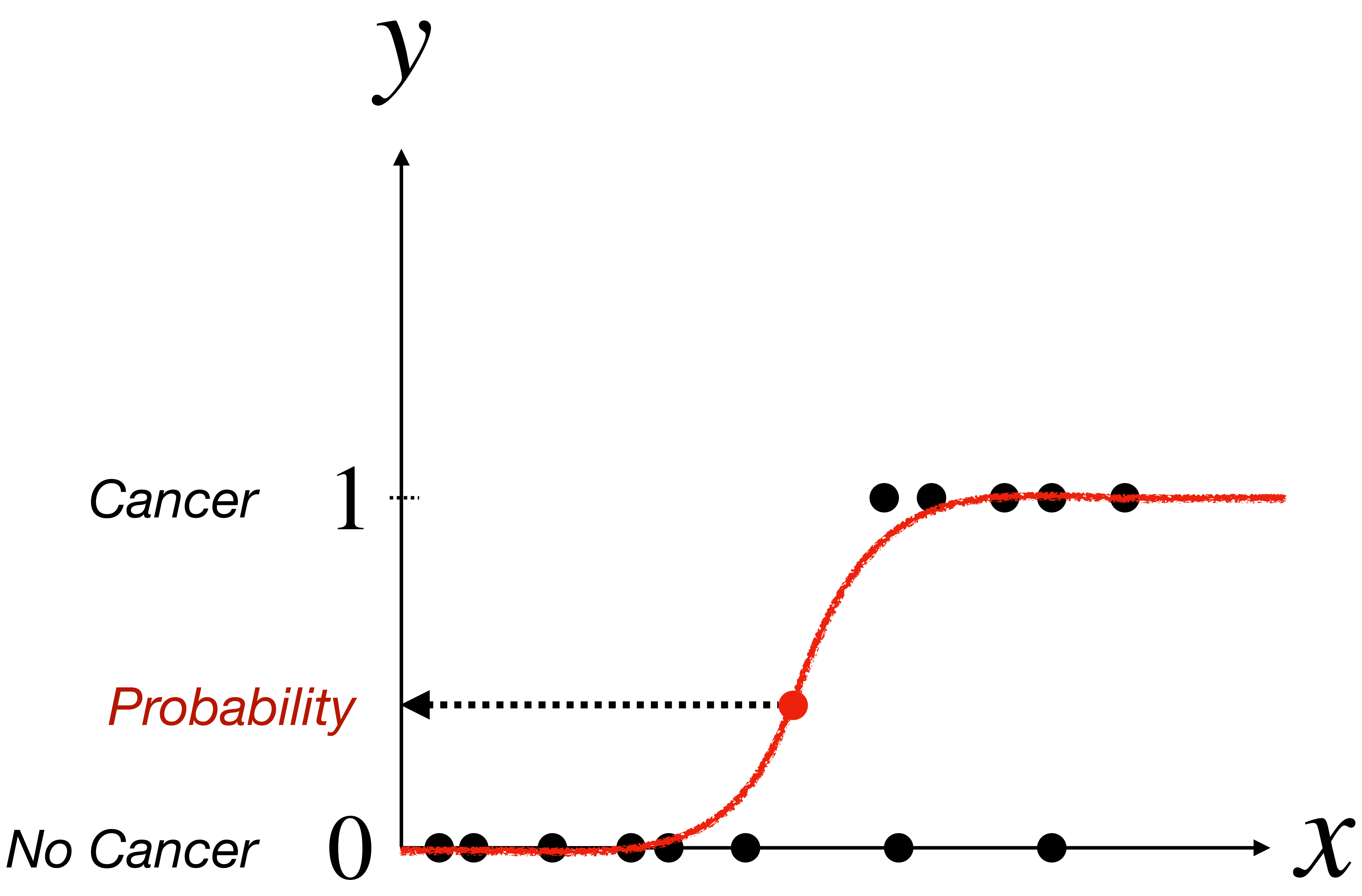


Given the data,
find a **function** h,
that predicts y, given x

$$\mathbf{y} = h(\mathbf{x})$$

$$y \in [0,1]$$

A smooth function that returns
probability of occurrence



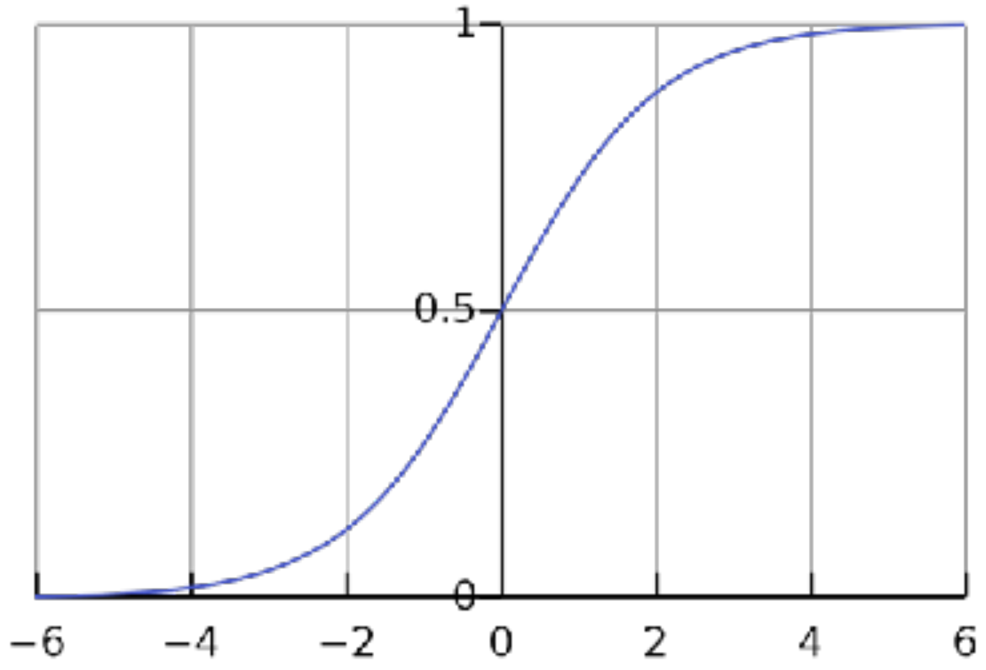
What if y is a label?

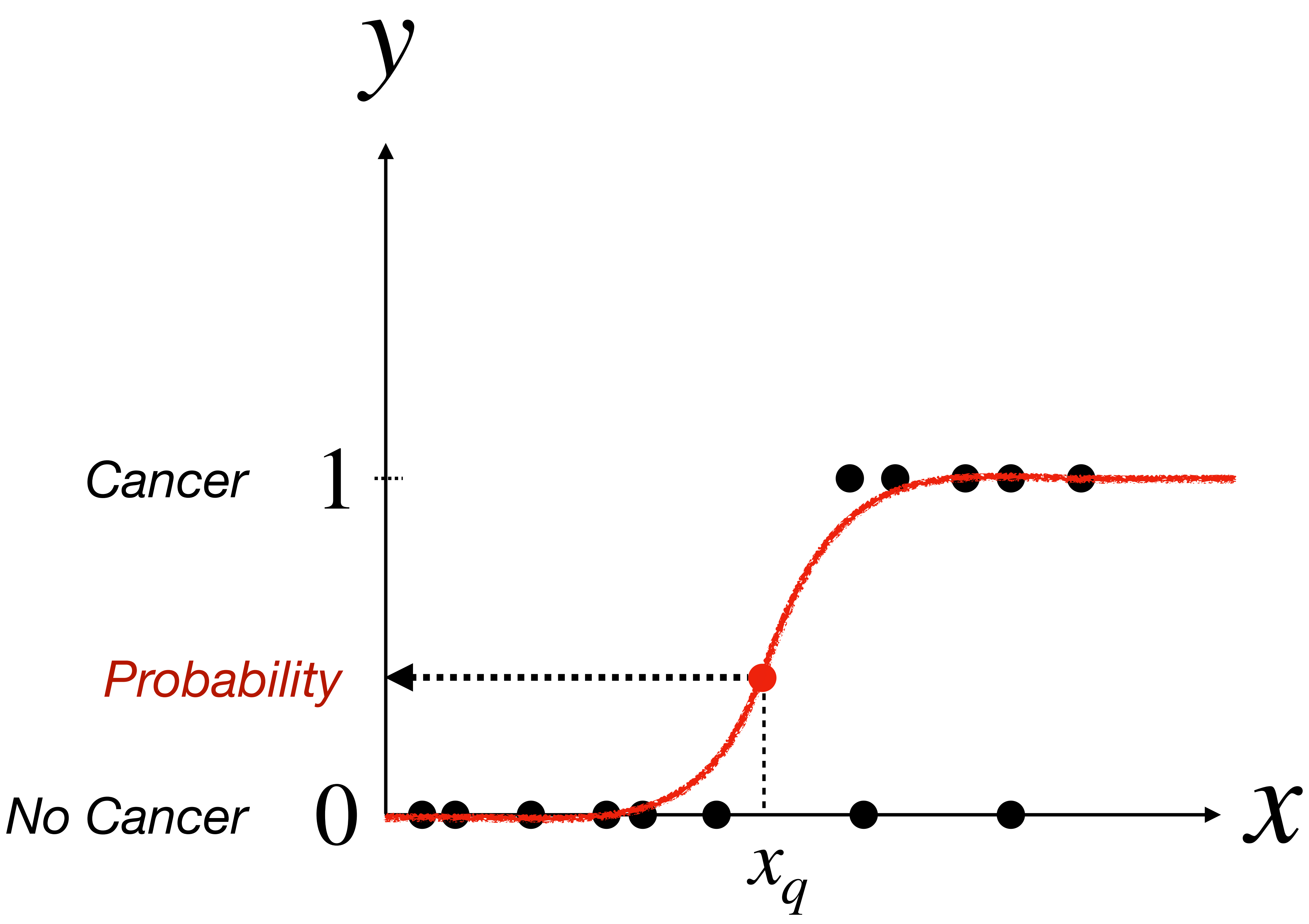
$$y = h_\theta(x) \quad \& \quad y \in [0,1]$$

A smooth function that returns probability of occurrence

$$y = \frac{1}{1 + e^{-x}}$$

Logistic Function





$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

What if y is a label?

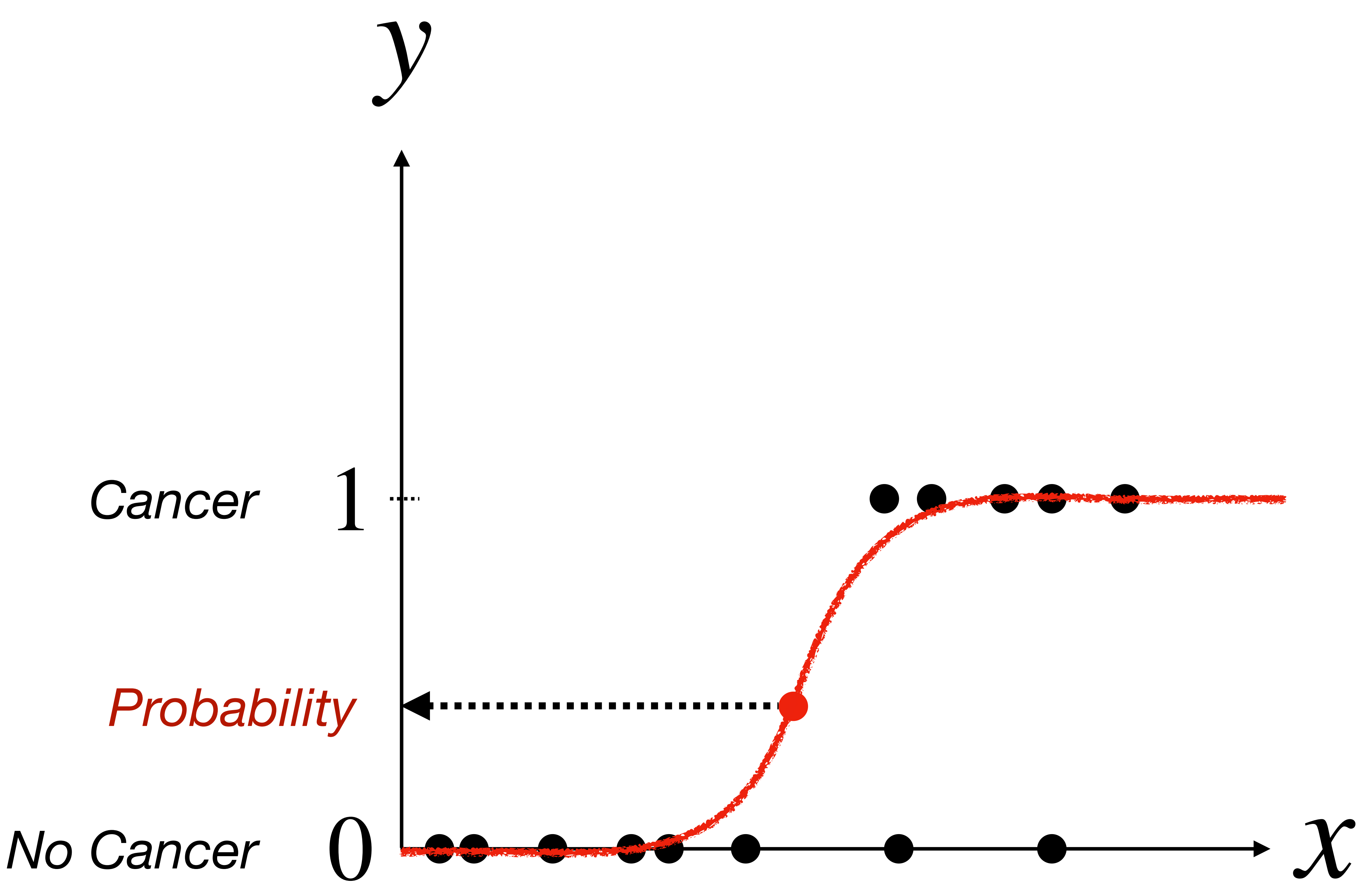
$$y = h_\theta(x) \quad \& \quad y \in [0,1]$$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta^\top x)}}$$

Where $\theta^\top x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$

$$\theta = [\theta_0, \theta_1, \ldots]$$
$$x = [x_0, x_1, \ldots]$$

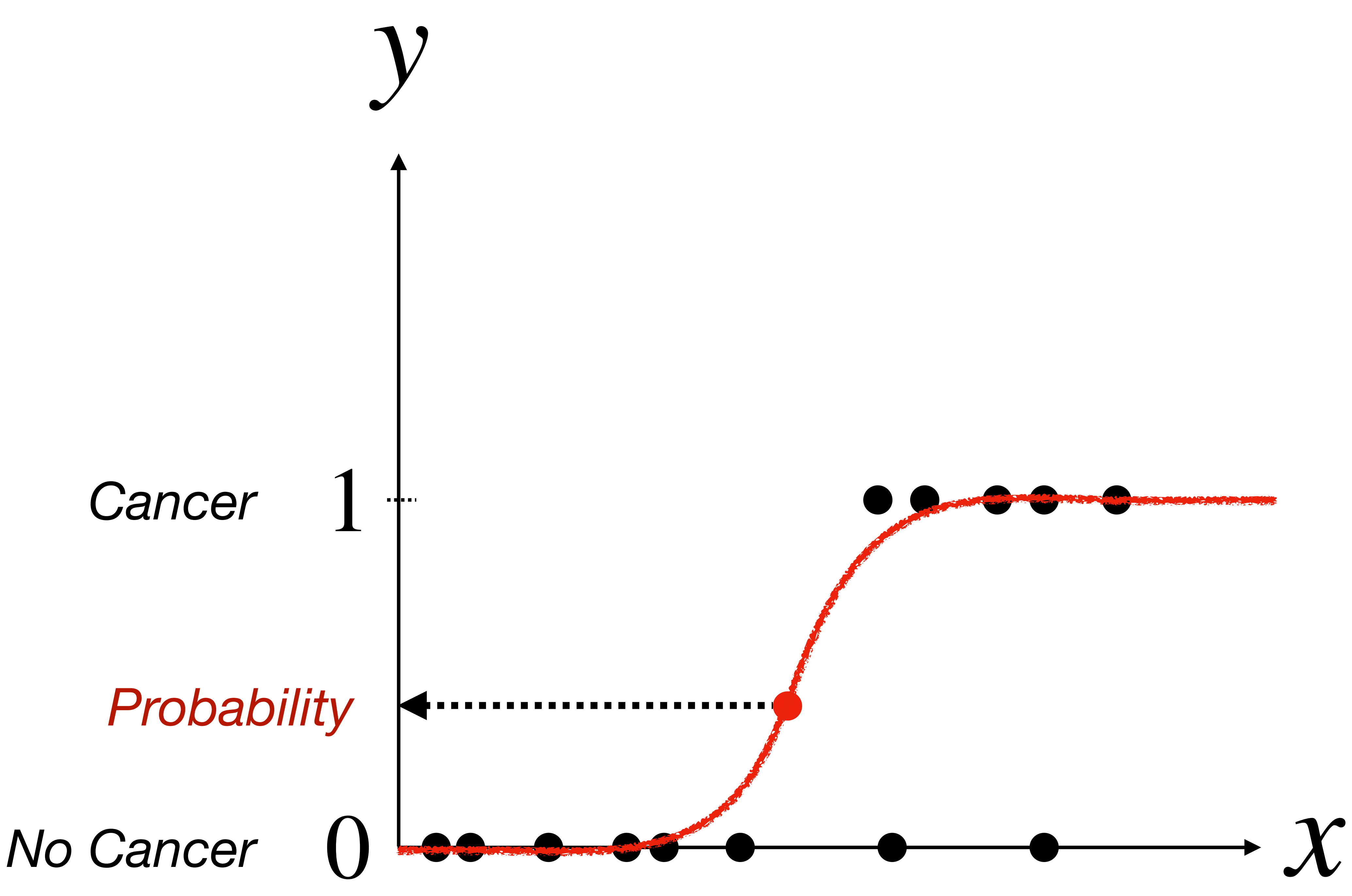
A **smooth** **function that returns probability of occurrence**



What if y is a label?

$$y = h_\theta(x) \quad \& \quad y \in [0,1]$$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta^\top x)}}$$



1. **Define a predictor:** the logistic function ✅
2. **Define a loss:** distance between function and data ❓
3. **Optimize loss**
4. **Test model**

How do we pick the best parameters θ ?

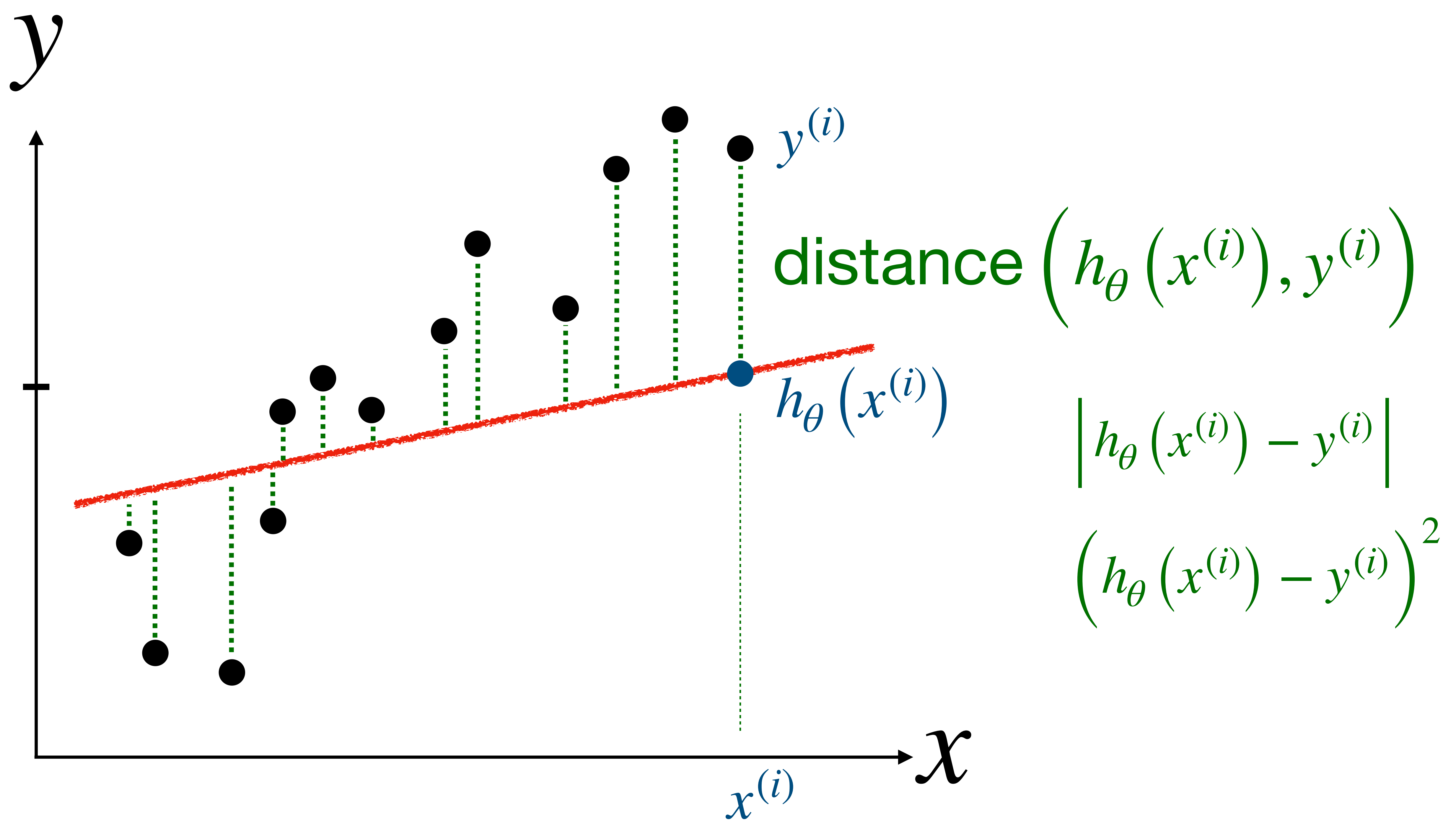
$$h_\theta(\mathbf{x}) = \theta^\top \mathbf{x} = \sum_{i=0}^{d} \theta_i x_i$$

Cost function

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{d} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$$

$$= \frac{1}{2} \sum_{i=1}^{d} \left(\theta^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2$$

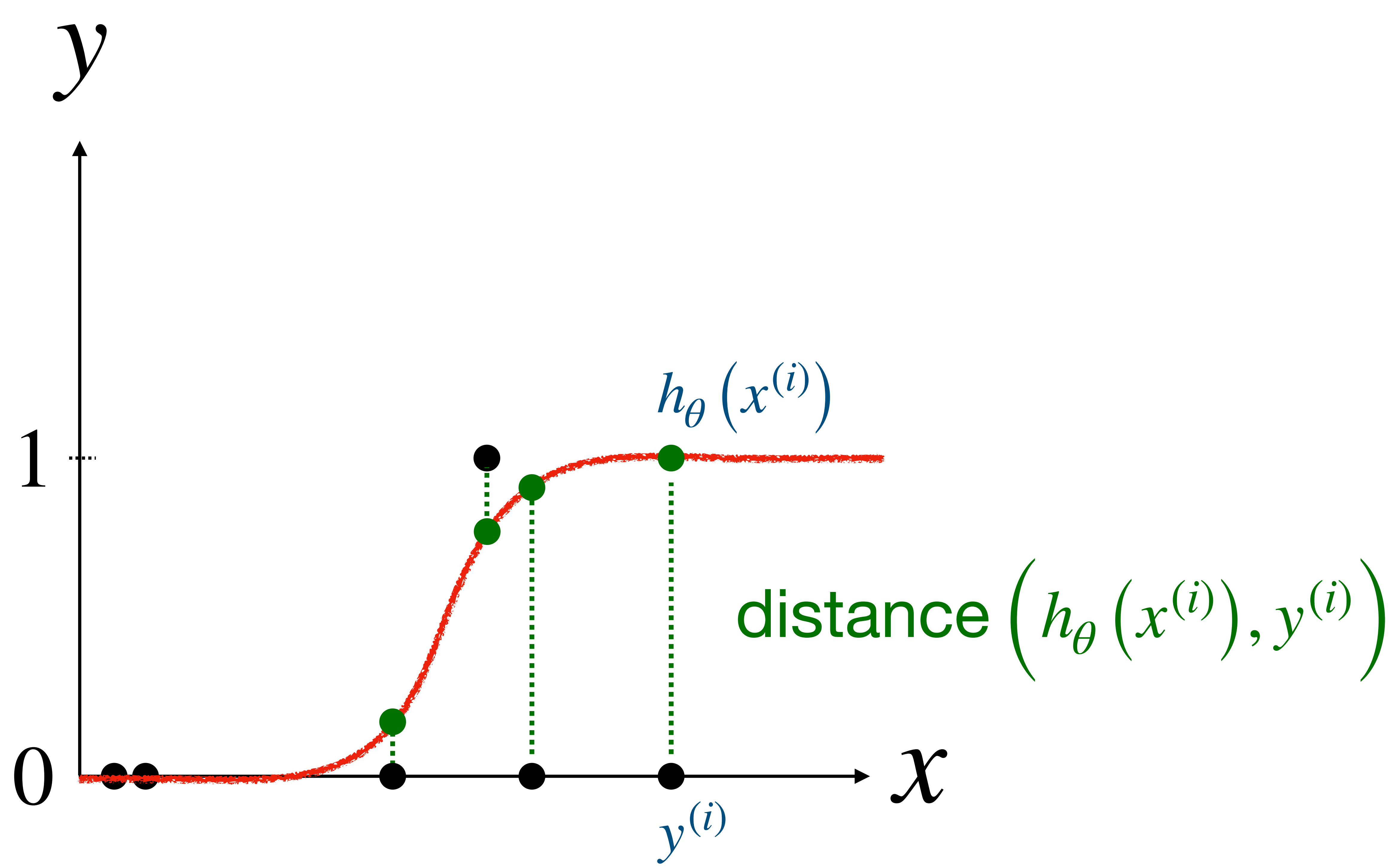
Ordinary least squares



Logistic Regression

$$y = h_\theta(x)$$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta^\top x)}} = g\left(\theta^\top x\right)$$



Linear predictor

negative log-likelihood or OLS

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{d}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Logistic predictor

Binary-cross entropy loss

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

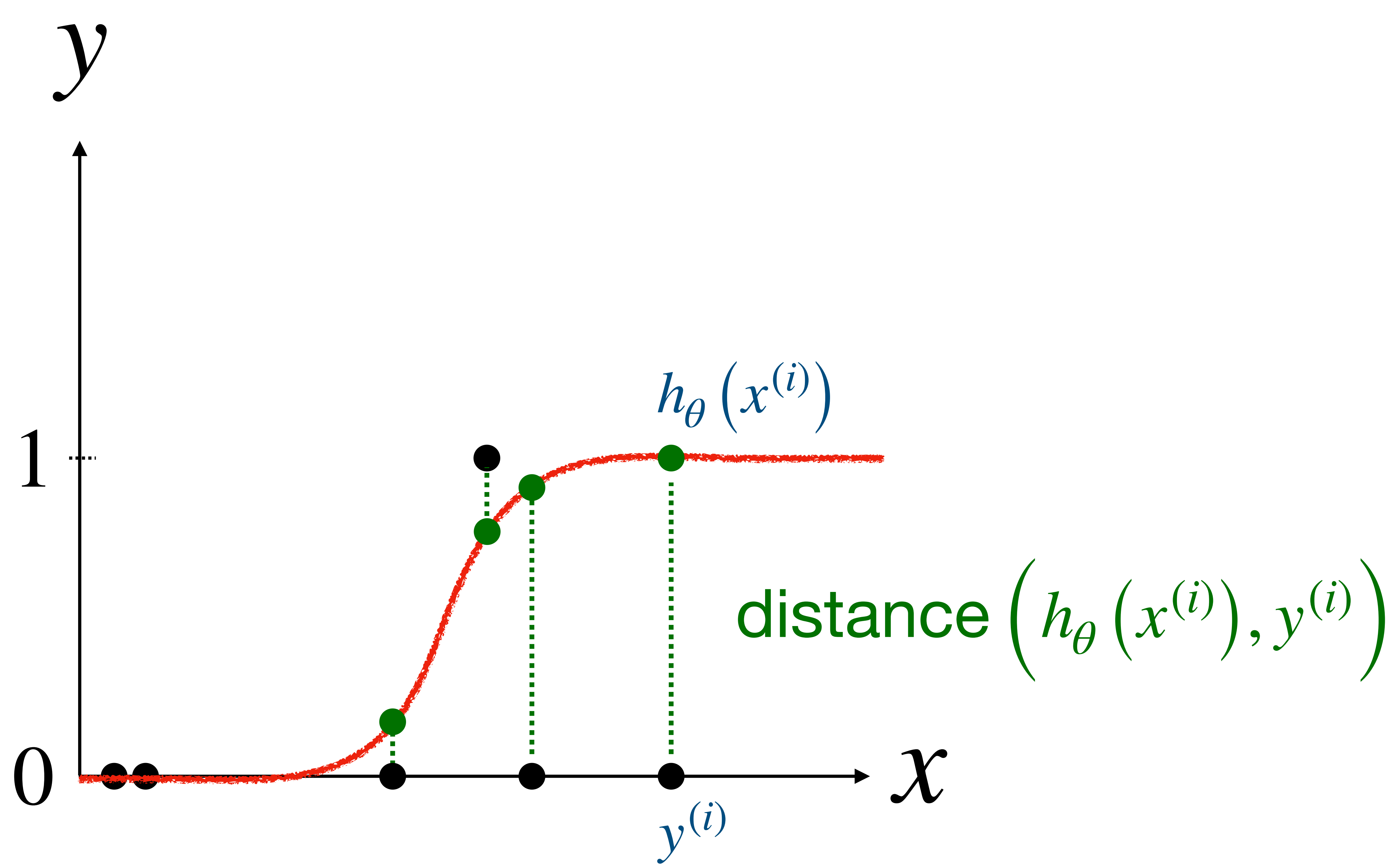
Compute gradient $\nabla \mathscr{L}(\theta)$

Gradient descent → Done!

Logistic Regression

$$y = h_\theta(x)$$

$$h_\theta(x) = \frac{1}{1 + e^{-\left(\theta^\top x\right)}} = g\left(\theta^\top x\right)$$



Linear predictor

negative log-likelihood or OLS

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{d}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Why not use an ordinary least squares loss?

Why not use an ordinary least squares loss?

Using ordinary least squares (OLS) with the logistic function for logistic regression is generally not appropriate due to several key reasons:

1. **Non-Linearity**: The logistic function is non-linear, mapping a linear combination of inputs to a probability between 0 and 1. OLS is designed to minimize the sum of squared differences between the observed values and a linear model's predictions. However, in logistic regression, the relationship between the input variables and the probability of the outcome is non-linear. OLS would not appropriately handle this non-linear relationship.
2. **Non-Gaussian Residuals**: OLS assumes that the residuals (errors between the observed and predicted values) are normally distributed. In logistic regression, the residuals follow a binomial distribution, not a normal distribution. Therefore, applying OLS would violate the assumptions of the method, leading to biased and inefficient estimates.
3. **Prediction Outside (0, 1) Interval**: OLS does not inherently restrict predictions to the interval [0, 1]. Since probabilities must lie within this range, OLS could produce predicted values that are less than 0 or greater than 1, which is not meaningful in the context of probabilities.
4. **Inefficient Estimation:** The estimates obtained using OLS in the context of a logistic regression model would not be the most efficient (i.e., they would not have the lowest variance among unbiased estimators). Maximum likelihood estimation (MLE), used in logistic regression, provides more efficient and reliable parameter estimates in this setting.
5. **Interpretation of Results**: Logistic regression models the log-odds of the outcome as a linear combination of the predictors. The OLS approach does not provide a straightforward interpretation in terms of odds or probabilities, which are the natural scales for binary outcomes.

Probabilistic Interpretation of Linear Regression

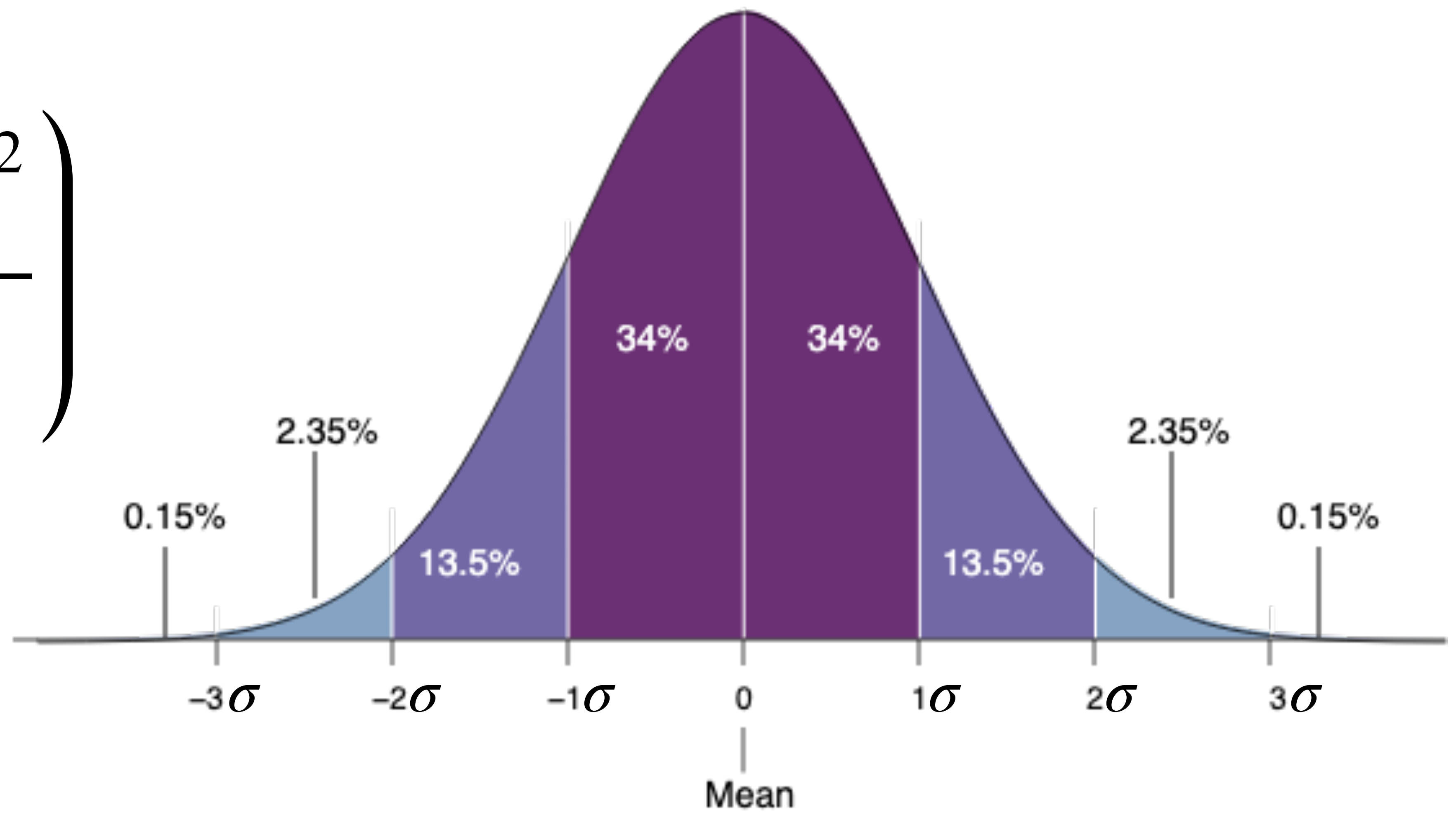
Assume noise is normally distributed around model

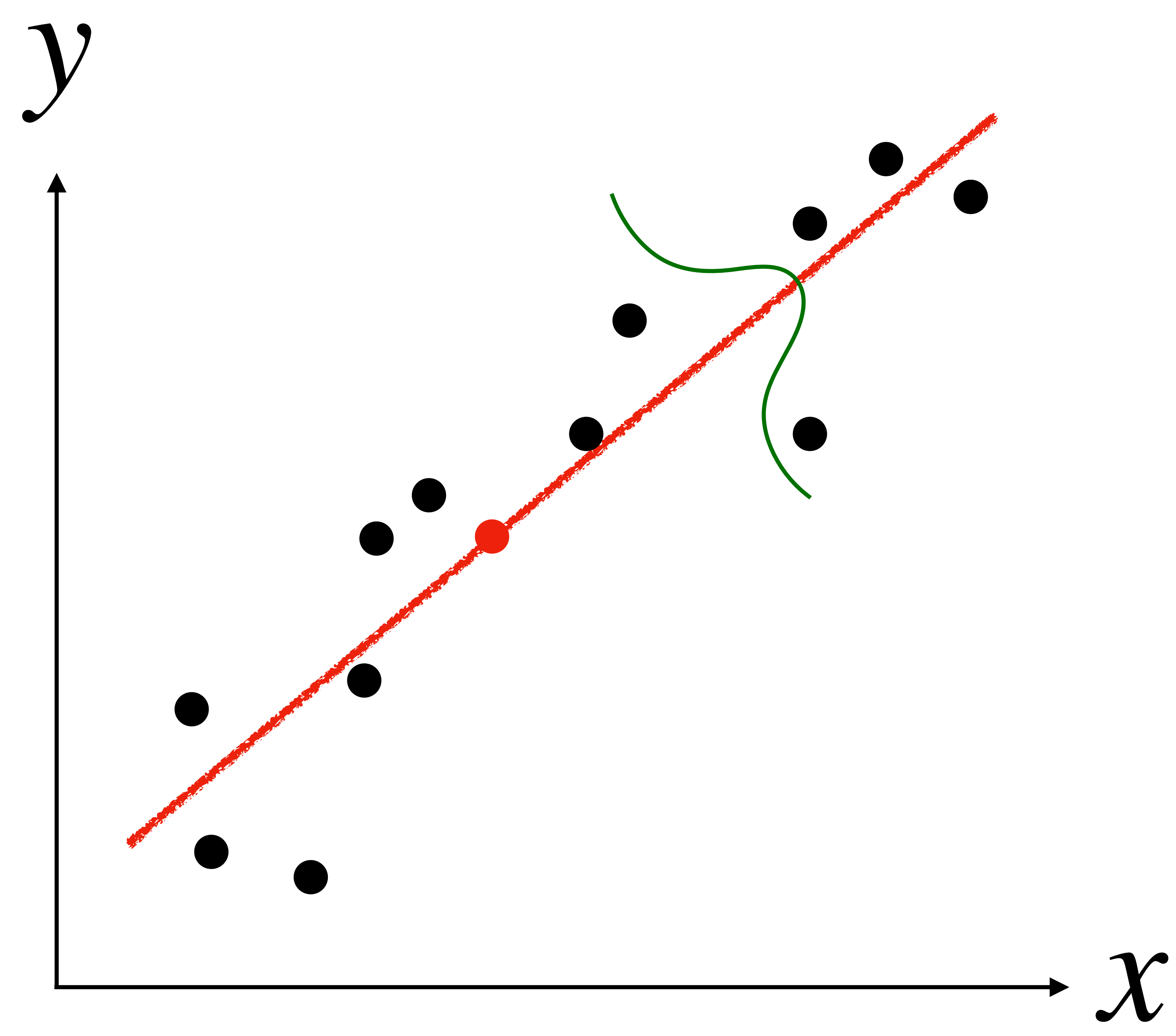
$$y^{(i)} = \theta^\top x^{(i)} + \varepsilon^{(i)}$$

Normally distributed

$$\mathcal{N}\left(0, \sigma^2\right)$$

$$p\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\varepsilon^{(i)}\right)^2}{2\sigma^2}\right)$$



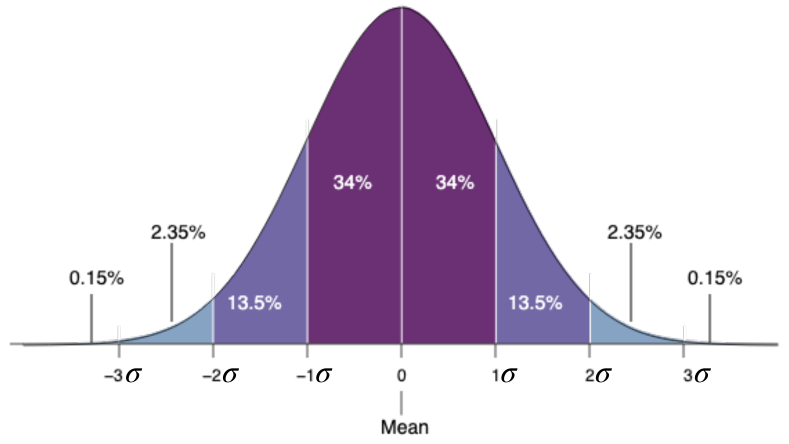


Probabilistic Interpretation

Assume noise is normally distributed around model

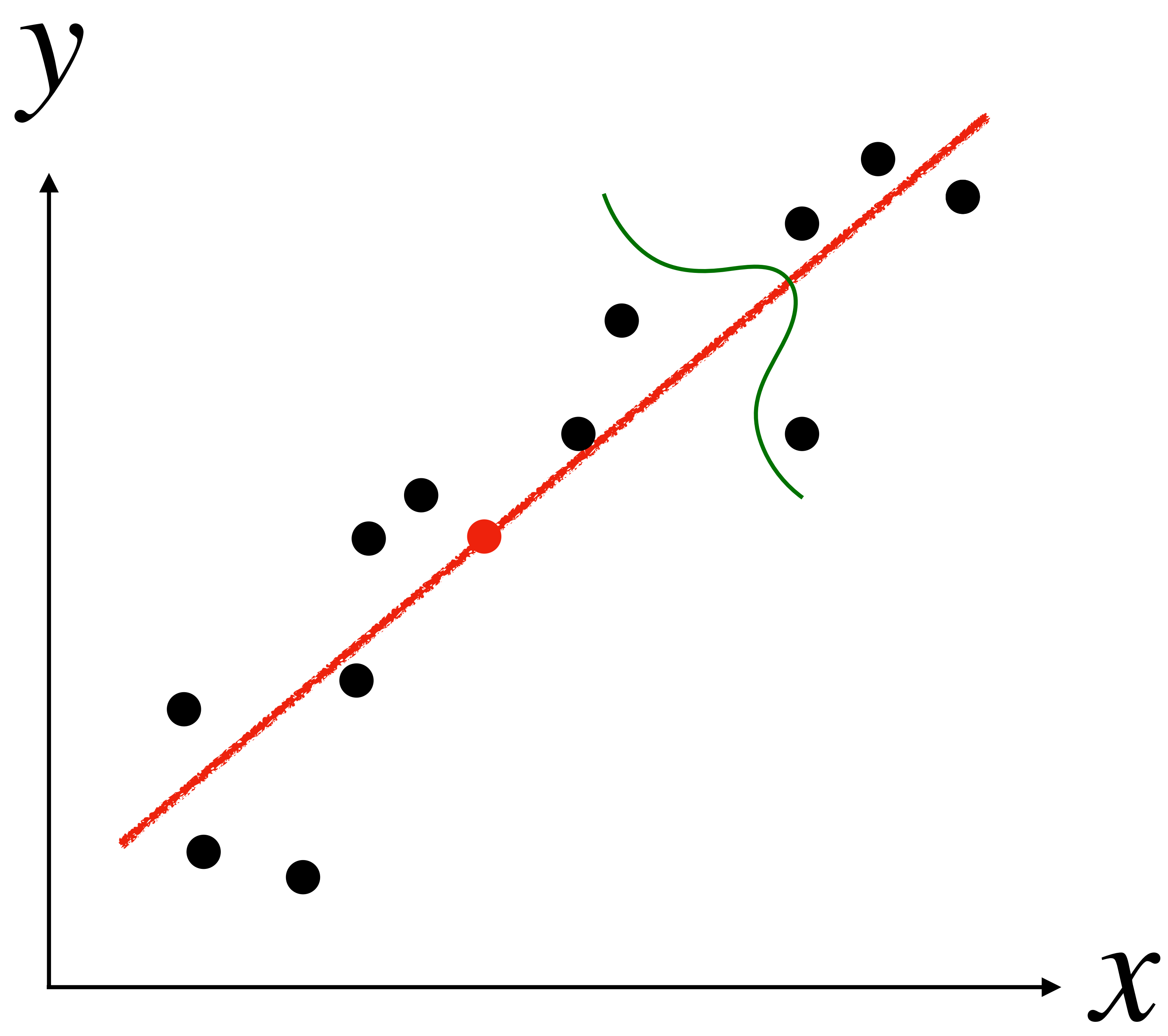
$$y^{(i)} = \theta^\top x^{(i)} + \varepsilon^{(i)}$$

$$\mathcal{N}\left(0, \sigma^2\right)$$



$$p\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\varepsilon^{(i)}\right)^2}{2\sigma^2}\right)$$

$$p\left(y^{(i)} \mid x^{(i)}; \theta\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^\top x^{(i)}\right)^2}{2\sigma^2}\right)$$

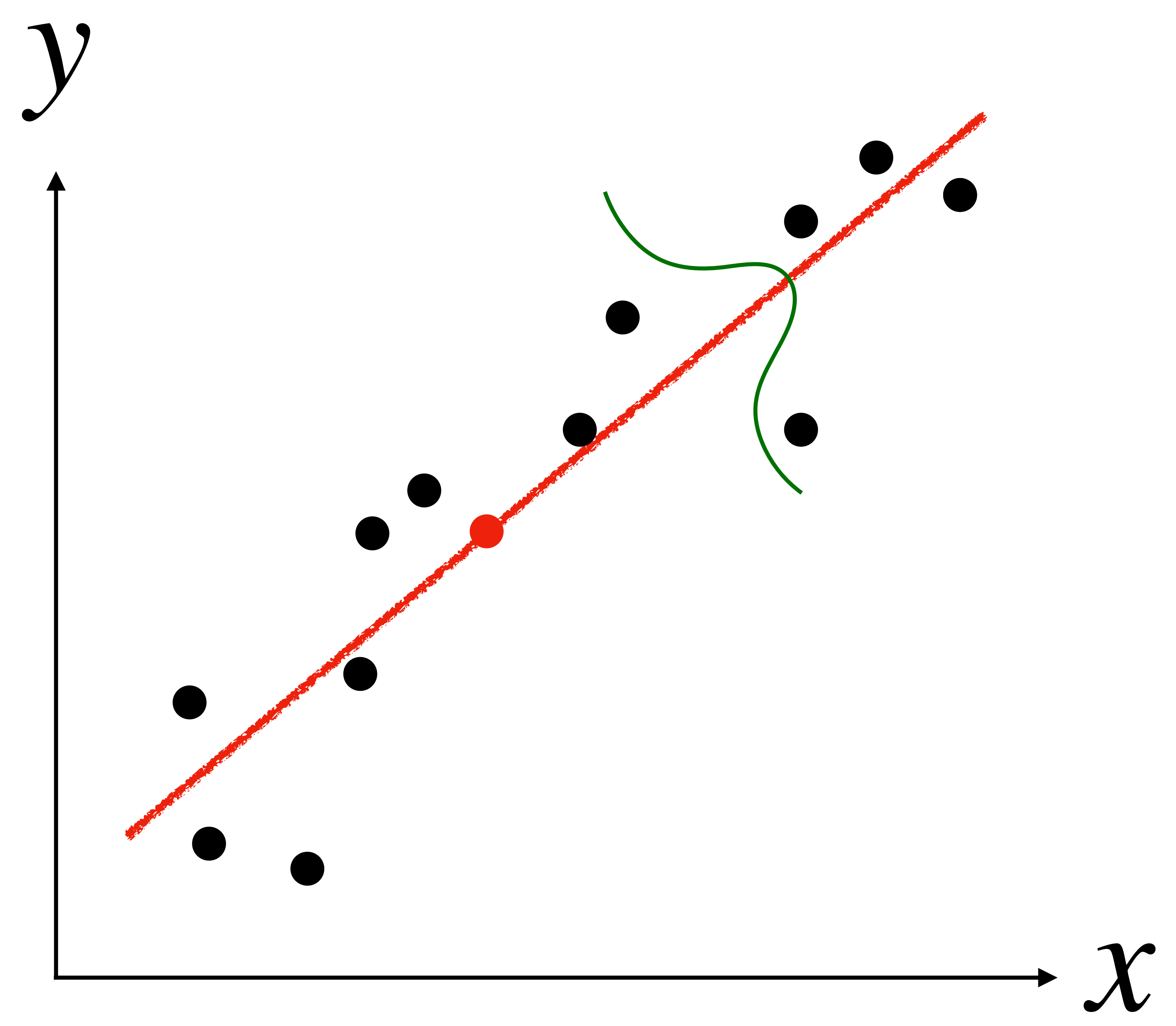


Likelihood of output given input

$$L(\theta) = \prod_{i=1}^{n} p\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

Independent and Identically Distributed (IID)

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2}{2\sigma^2}\right)$$



Log-likelihood

$$\mathscr{L}(\theta) = \log L(\theta)$$

$$= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2}{2\sigma^2}\right)$$

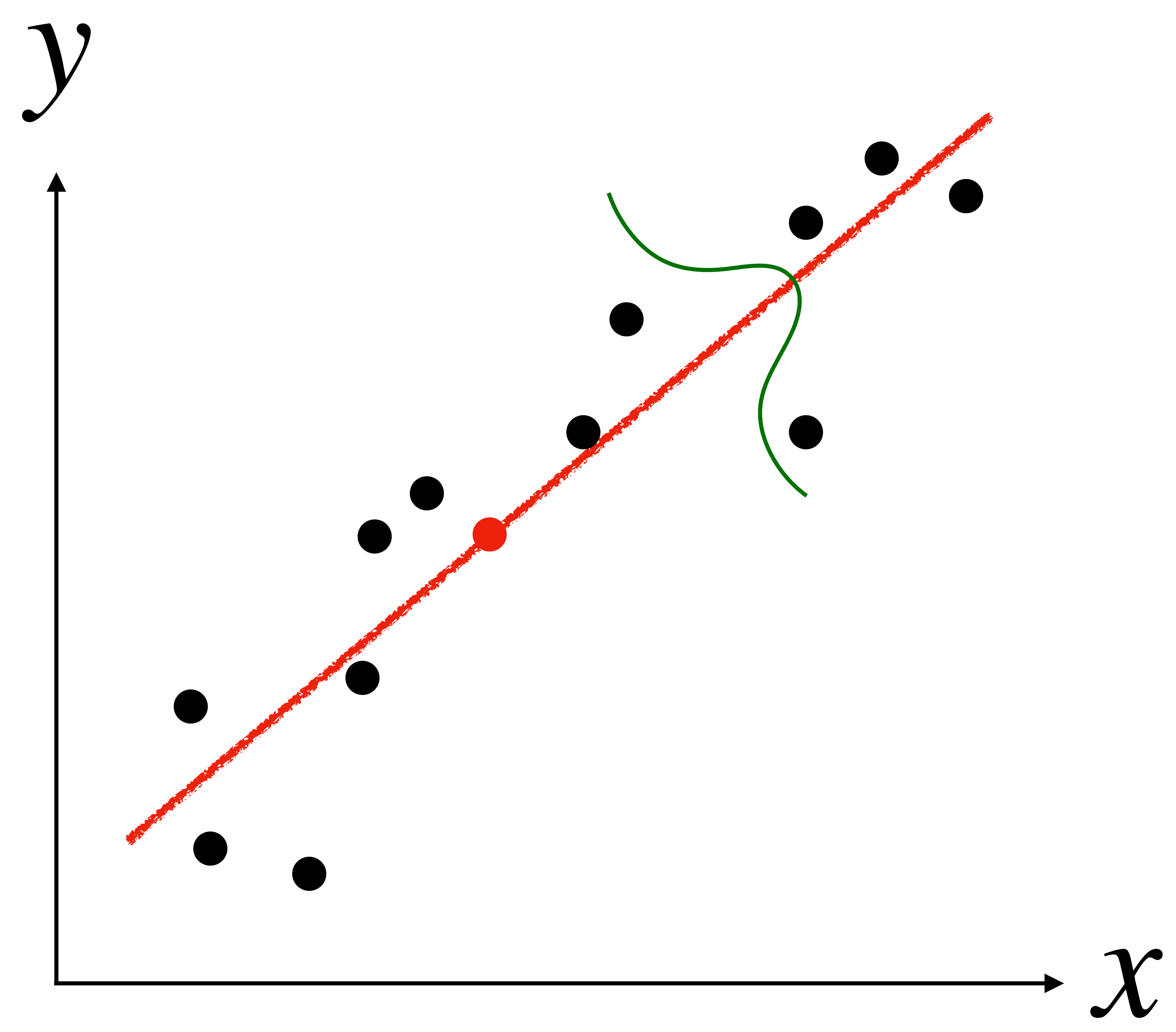
$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2}{2\sigma^2}\right) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^{\top} x^{(i)}\right)^2$$

Maximize Log-likelihood

$$\mathcal{L}(\theta) = \log L(\theta)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \theta^\top x^{(i)}\right)^2}{2\sigma^2}\right)$$

$$= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^\top x^{(i)}\right)^2$$



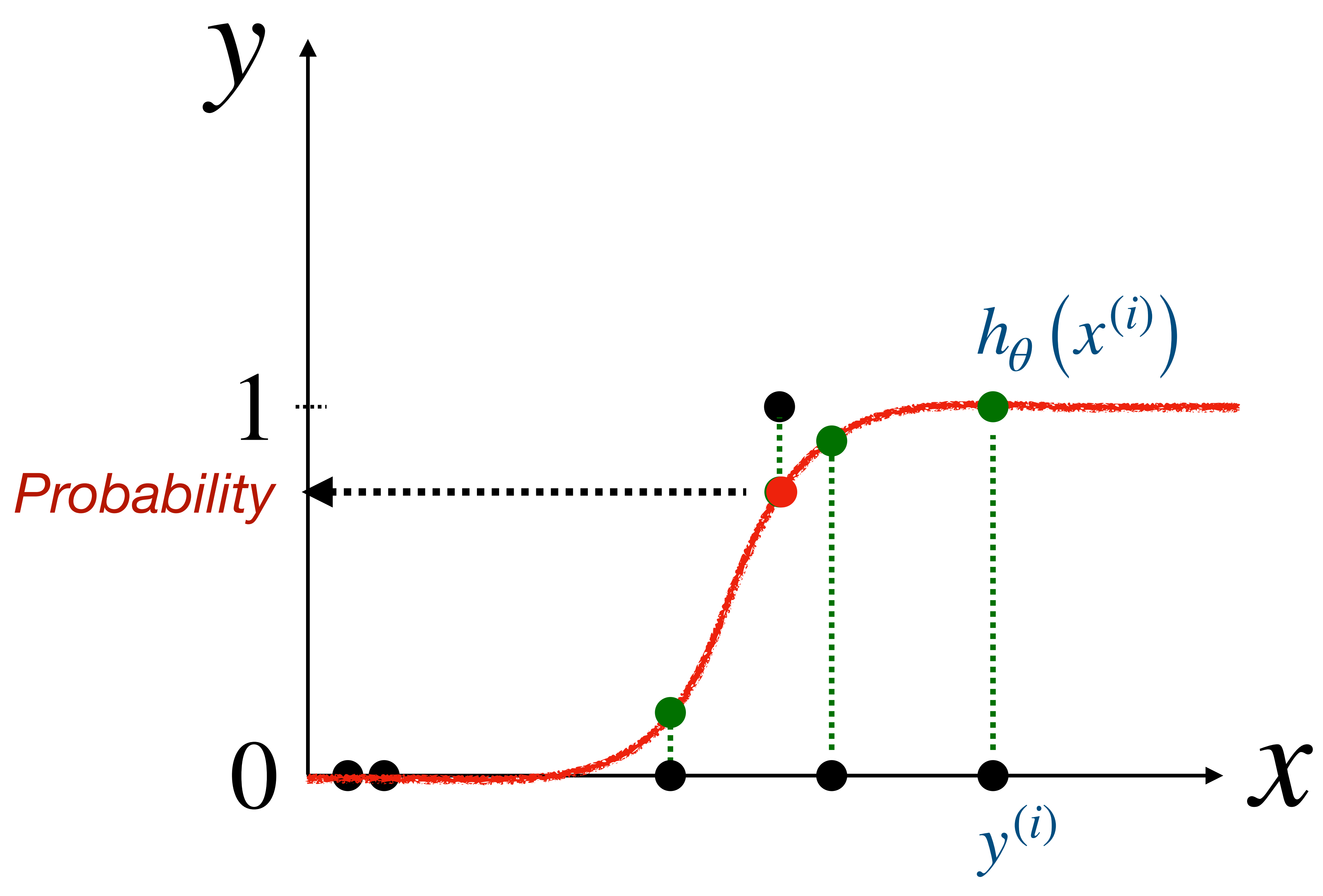
Maximize $\mathcal{L}(\theta)$ $\longrightarrow$ **Minimize** $\frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^\top x^{(i)}\right)^2$

What if the noise is not Gaussian?

Why not Least Squares?

$$y = h_\theta(x)$$

$$h_\theta(x) = \frac{1}{1 + e^{-\left(\theta^\top x\right)}} = g\left(\theta^\top x\right)$$



Probability of output given input

$$P\left(y = 1 \,\middle|\, x; \theta\right) = h_\theta(x)$$
$$P\left(y = 0 \,\middle|\, x; \theta\right) = 1 - h_\theta(x)$$

→

True label

$$p(y \,|\, x; \theta) = \left(h_\theta(x)\right)^{y} \left(1 - h_\theta(x)\right)^{1-y}$$

Likelihood!

For Bernoulli Distributed Noise

Bernoulli Distribution

Properties [edit]

If X is a random variable with a Bernoulli distribution, then:

$$\Pr(X=1)=p=1-\Pr(X=0)=1-q.$$

The probability mass function f of this distribution, over possible outcomes *k*, is

$$f(k;p)=\begin{cases}p & \text{if } k=1,\\ q=1-p & \text{if } k=0.\end{cases}$$ [3]

This can also be expressed as

$$f(k;p)=p^k(1-p)^{1-k}\quad \text{for } k\in\{0,1\}$$

or as

$$f(k;p)=pk+(1-p)(1-k)\quad \text{for } k\in\{0,1\}.$$

The Bernoulli distribution is a special case of the binomial distribution with $n=1$.[4]

Define Log-likelihood

Likelihood

$$p(y\,|\,x;\theta) = \left(h_\theta(x)\right)^y \left(1 - h_\theta(x)\right)^{1-y}$$

for all (x, y) pair

$$L(\theta) = \prod_{i=1}^{n} p\left(y^{(i)}\,|\,x^{(i)};\theta\right)$$

$$= \prod_{i=1}^{n} h_\theta\left(x^{(i)}\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}$$

log

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Maximize Log-likelihood

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

Update rule

while not converged:

$$\theta_j := \theta_j + \alpha \frac{\partial \mathcal{L}(\theta)}{\partial \theta_j}$$

Derive →

Gradient Descent

for t = 1…T:

for all parameters j:

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

$$\frac{\mathcal{L}(\theta)}{\partial \theta_j} = \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

Base Code Snippet

Scikit-Learn Code Snippet

Example