TÉCNICO LISBOA

Planning, Learning and Decision Making

Lecture 13. Learning from examples - Probabilistic approaches

Learning from examples

- Set of possible situations, $\mathcal{X}$ ⟶ Input space
- Set of possible actions, $\mathcal{A}$ ⟶ Output space

Learning from examples

TASK:

- Decision rule (policy):
  - Deterministic $\pi : \mathcal{X} \rightarrow \mathcal{A}$ ⟶ Discriminant function
  - Stochastic $\pi : \mathcal{X} \rightarrow \Delta(\mathcal{A})$ ⟶ Discriminative model

Learning from examples

PERFORMANCE:

- We select a policy π* to minimize some cost functional *J*, i.e.,

$$\pi^* = \underset{\pi}{\operatorname{argmin}}\, L(\pi)$$

- *L*(π) usually cannot be computed exactly
- We select π* to minimize empirical estimate of *L*

Learning from examples

PERFORMANCE:

- Examples:
  - 0-1 loss

$$\hat{L}_N(\pi) = \frac{1}{N}\sum_{n=1}^{N}(1 - \pi(a_n \mid x_n))$$

- Negative log-likelihood

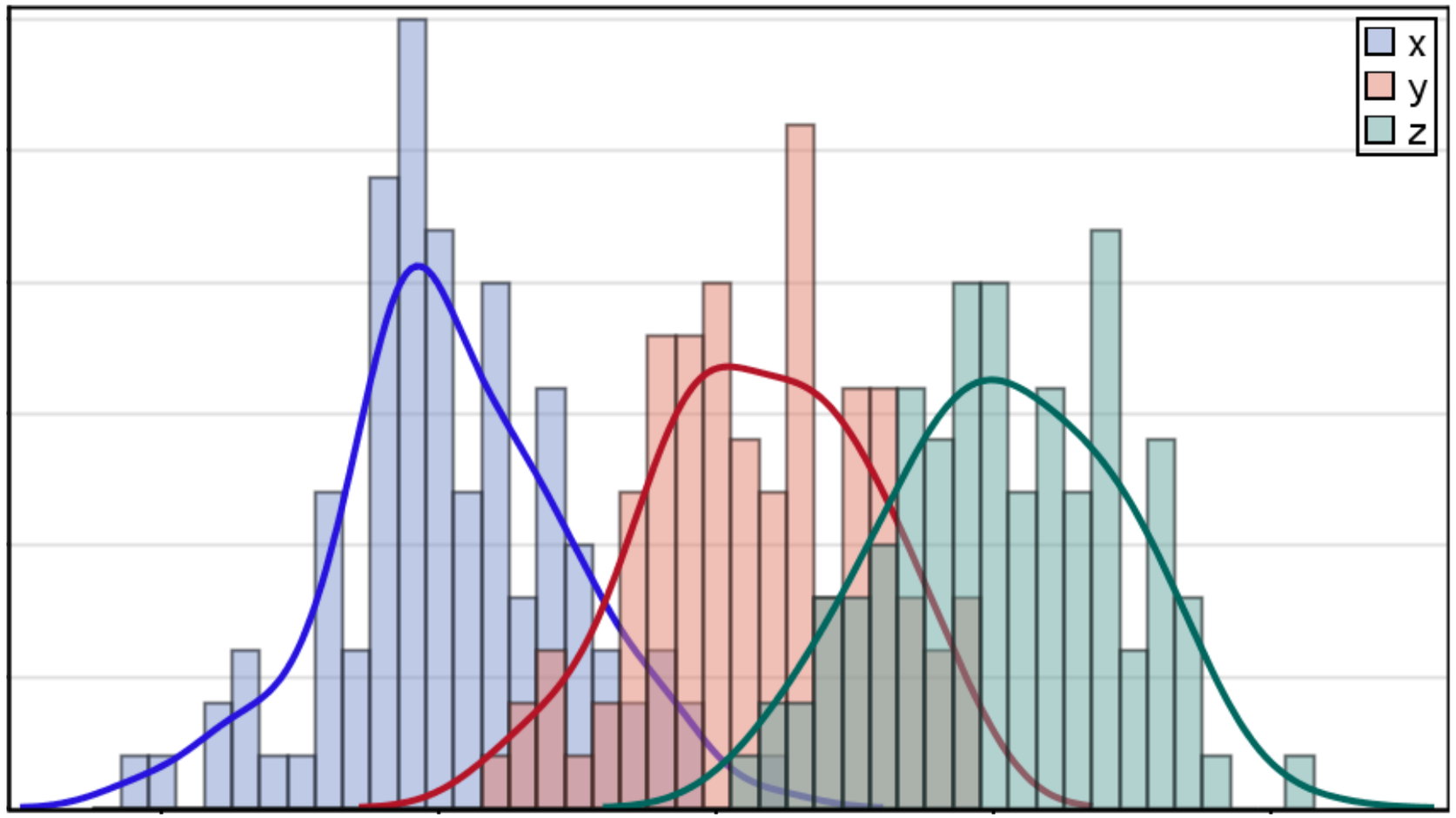
$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N}\log \pi(a_n \mid x_n)$$

Learning from examples

EXPERIENCE:

- Dataset of examples $\mathscr{D} = \{(x_0, a_0), (x_1, a_1), \ldots, (x_N, a_N)\}$

- Pairs (x, a) generated from an unknown distribution μ_D

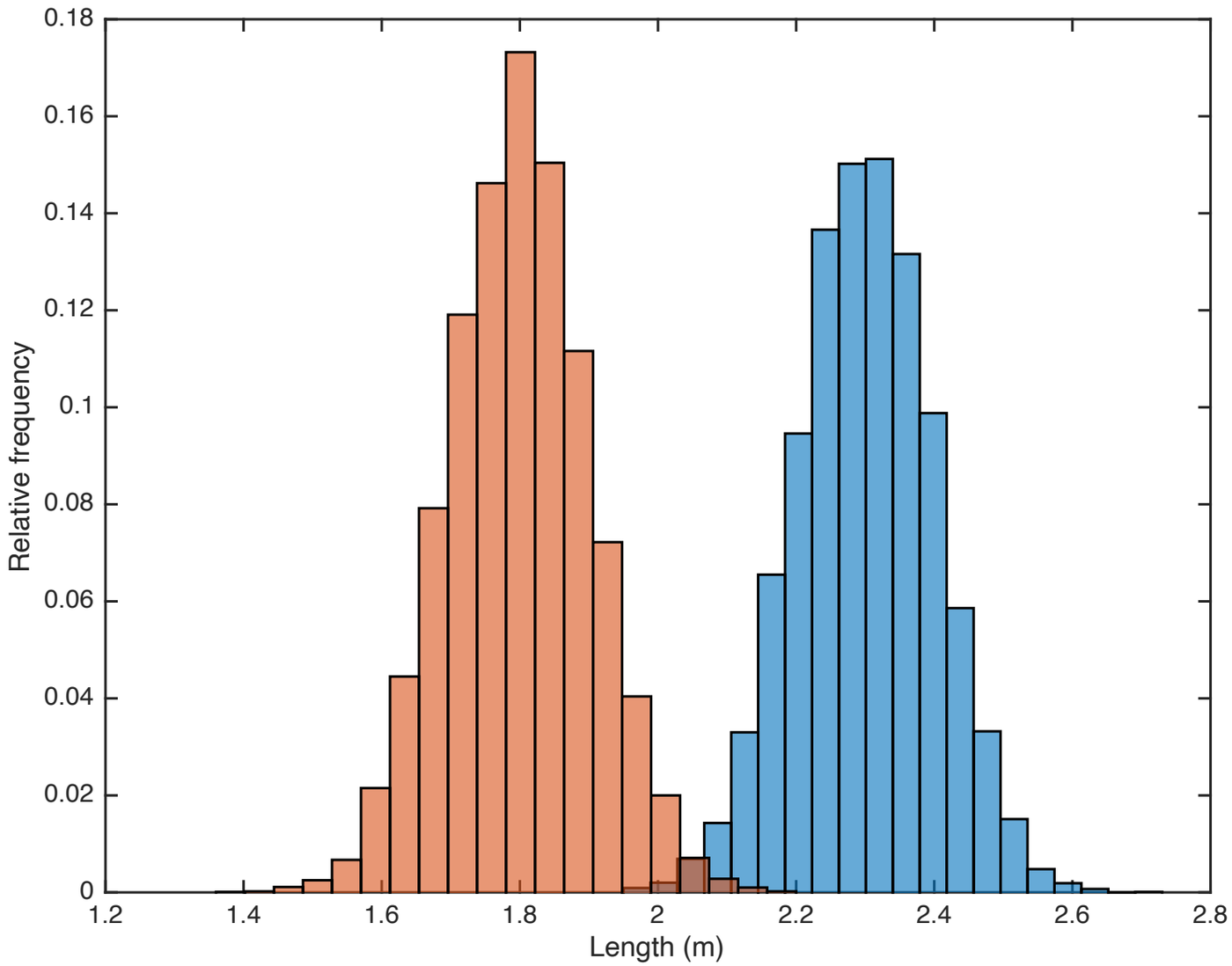


Probabilistic approaches

Example

- Sea lions are **sexually dimorphic** (males and females look different)

Example

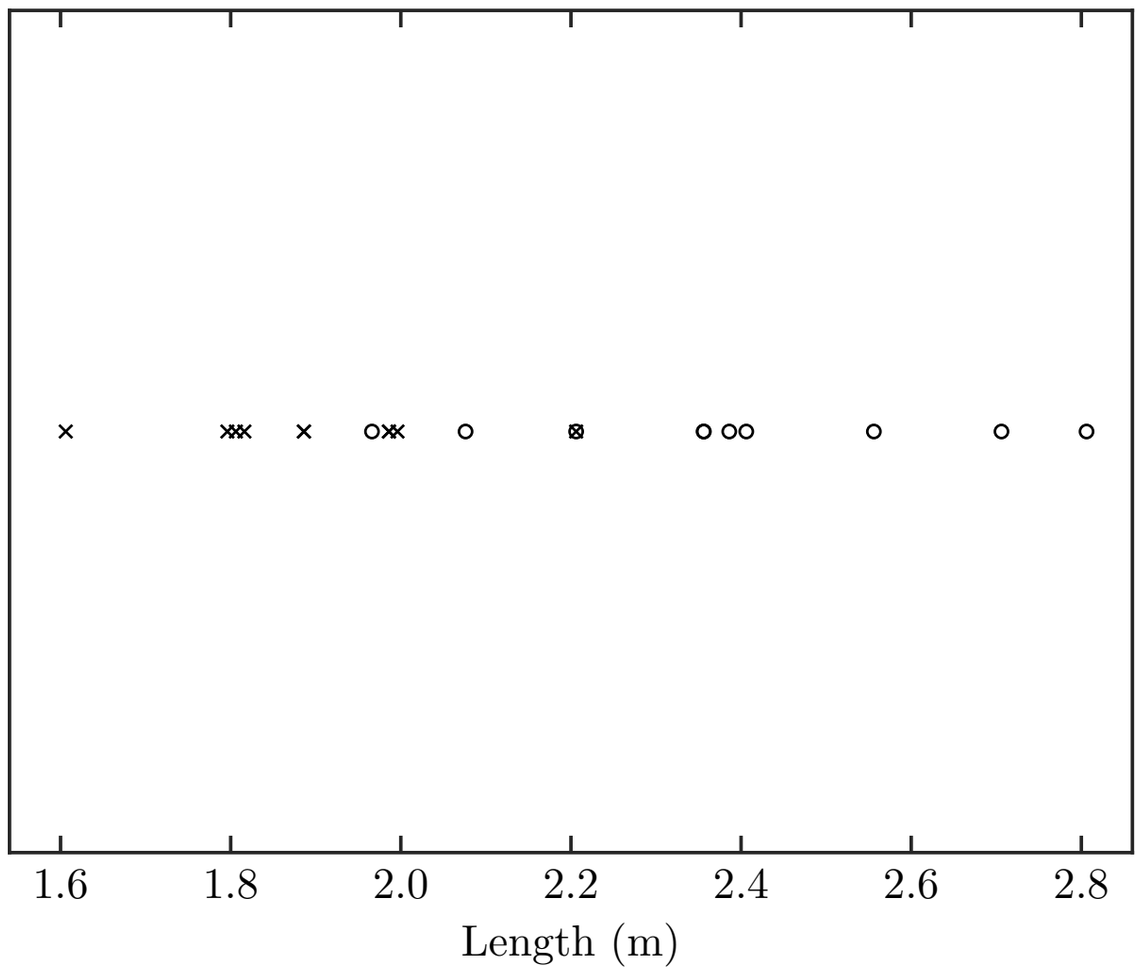

0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Relative frequency
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
Length (m)

Example

- Our agent must discriminate male and female sea lions
- We give the agent a set of examples

Length (m)	Gender
2.20	F
1.99	F
2.35	M
2.40	M
1.88	F
2.20	M
1.79	F
1.60	F
1.96	M
2.55	M
2.35	M
1.80	F
1.98	F
1.98	F
2.70	M
1.81	F
2.38	M
1.88	F
2.80	M
2.07	M

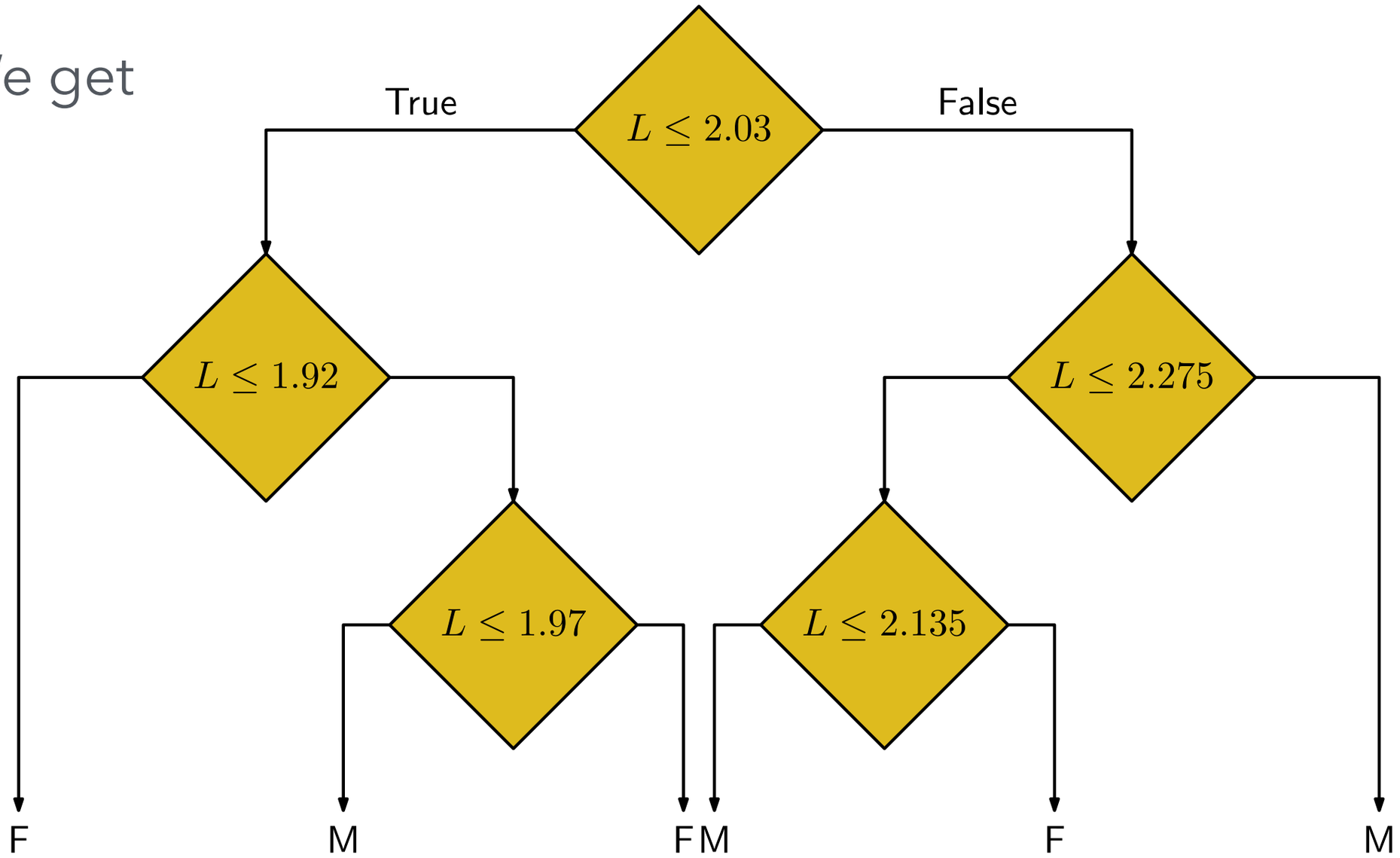
Example



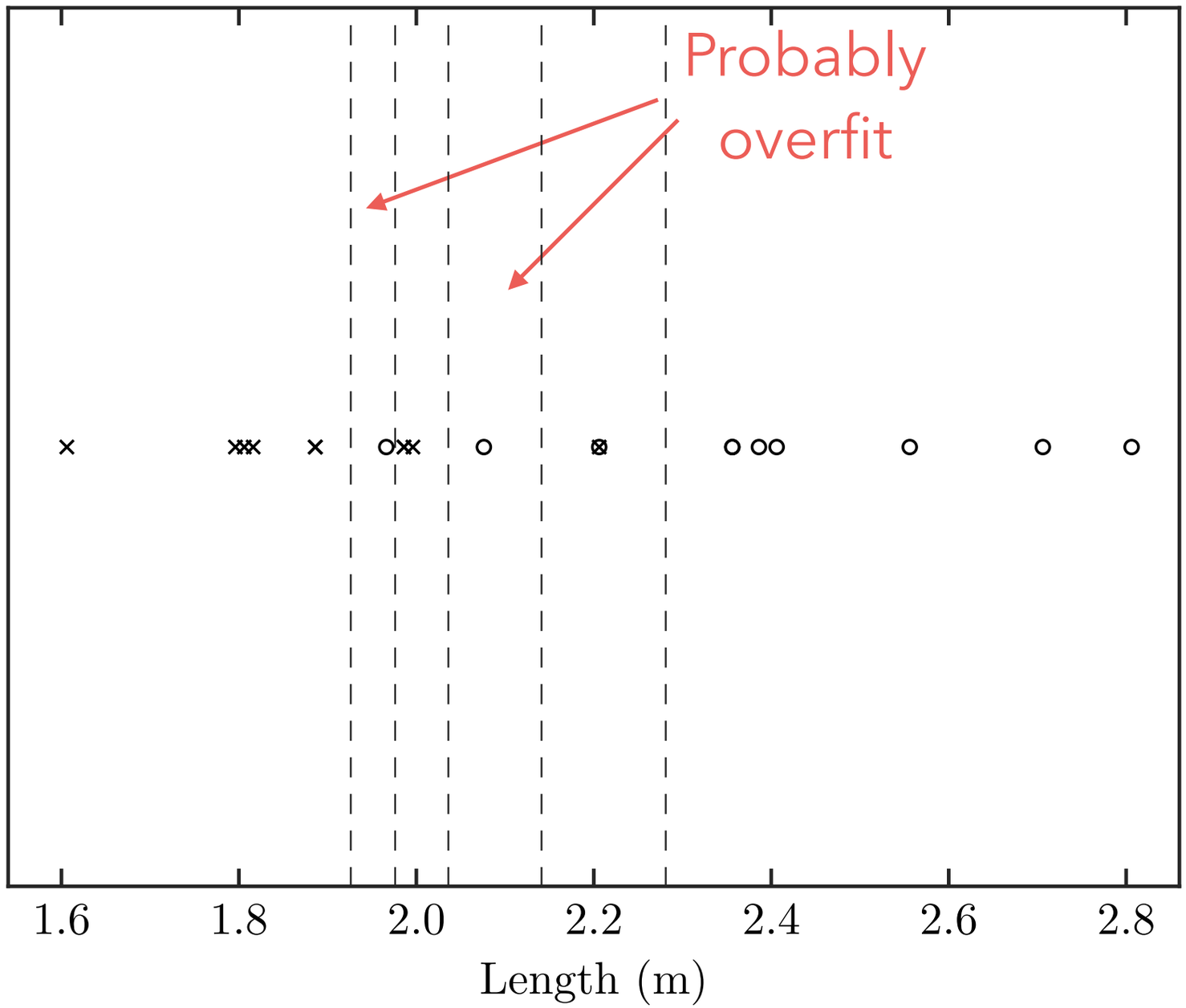
Length (m)	Gender
2.20	F
1.99	F
2.35	M
2.40	M
1.88	F
2.20	M
1.79	F
1.60	F
1.96	M
2.55	M
2.35	M
1.80	F
1.98	F
1.98	F
2.70	M
1.81	F
2.38	M
1.88	F
2.80	M
2.07	M

Decision tree

- We can use a decision tree
- We get



Decision tree

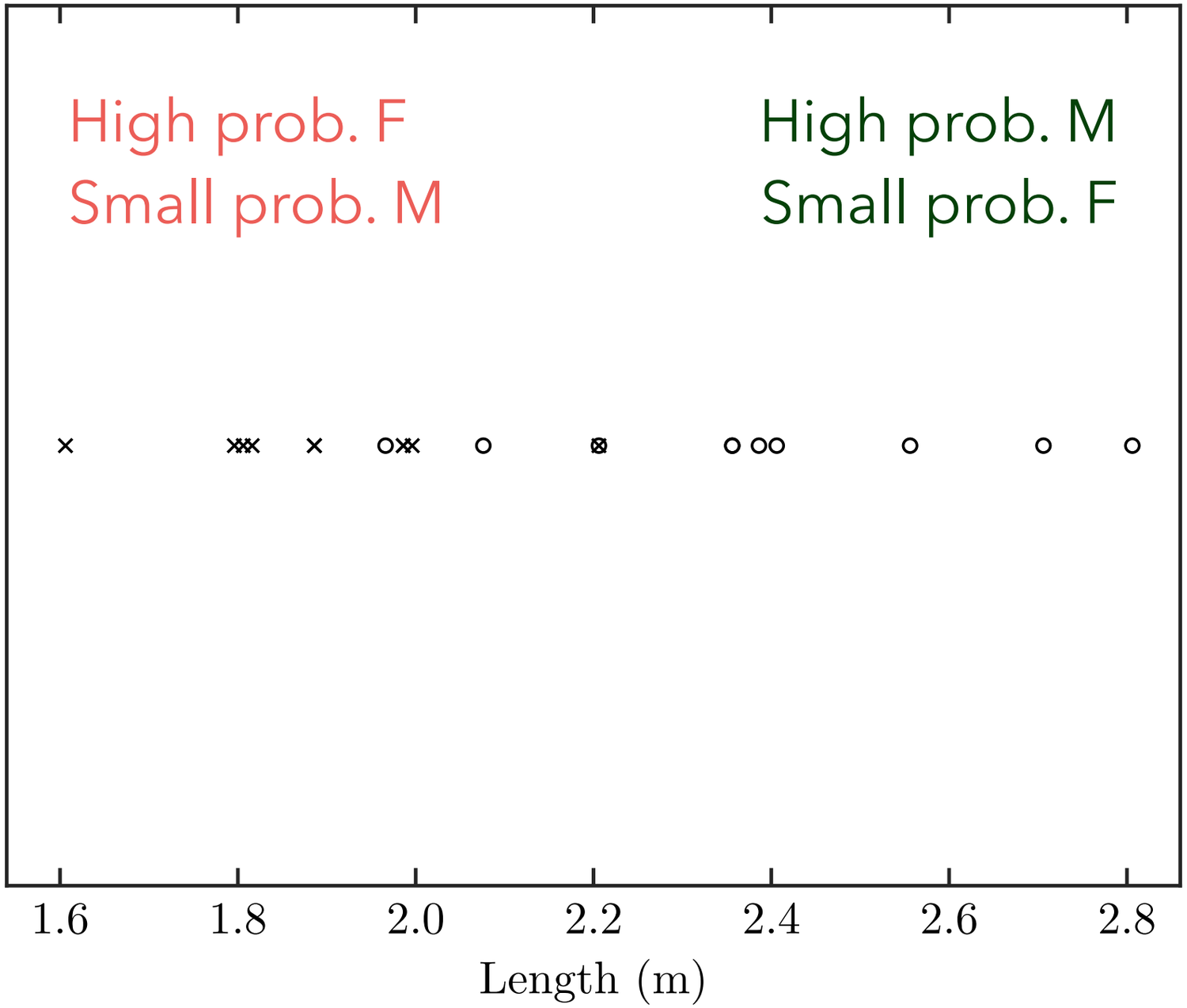


Discriminative model

- We want to compute a **discriminative model**
- A discriminative mode is a policy $\pi : \mathcal{X} \rightarrow \Delta(\mathcal{A})$
- $\pi(a \mid x)$ is the probability of action/label a given x

Length (m)	Gender
2.20	F
1.99	F
2.35	M
2.40	M
1.88	F
2.20	M
1.79	F
1.60	F
1.96	M
2.55	M
2.35	M
1.80	F
1.98	F
1.98	F
2.70	M
1.81	F
2.38	M
1.88	F
2.80	M
2.07	M

Discriminative model

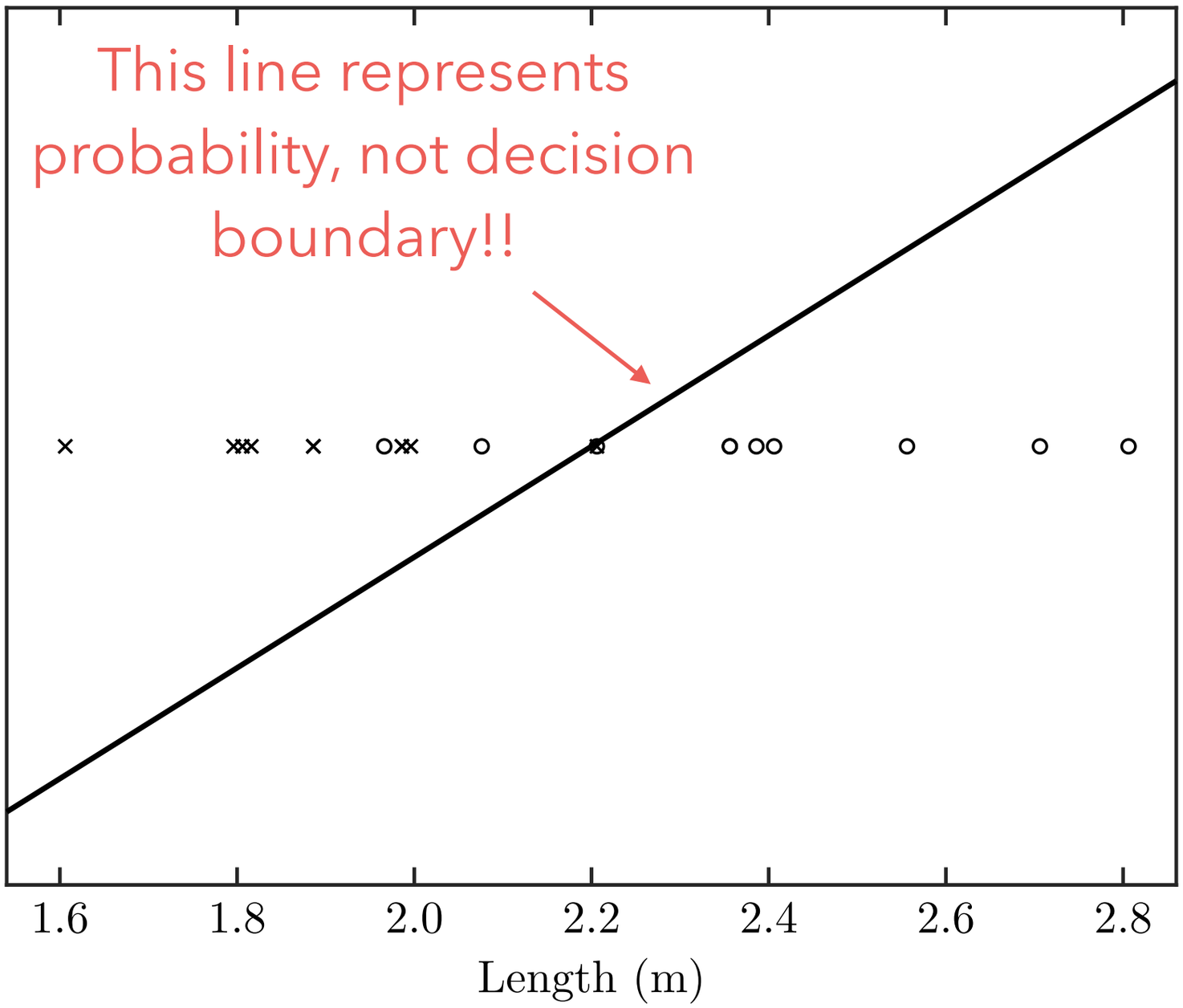


Discriminative model

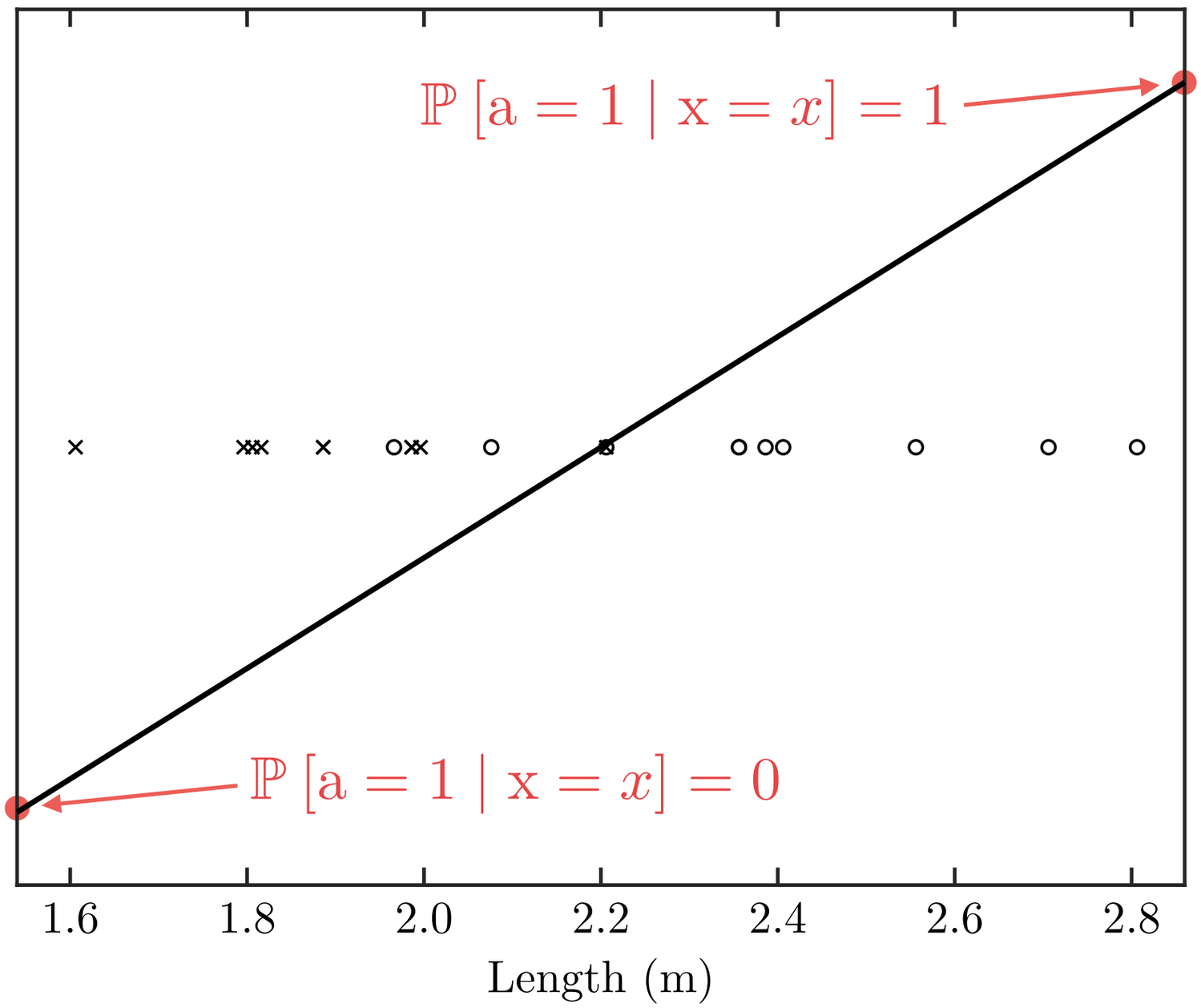
- We can focus on one probability alone
  - E.g., π(1 | *x*)
- We use the simplest approximation
  - π(1 | *x*) is linear in *x*:

$$\pi(1 \mid x) = \mathbb{P}\left[\mathrm{a} = 1 \mid \mathrm{x} = x\right] = w_0 + w_1\phi(x)$$

Discriminative model

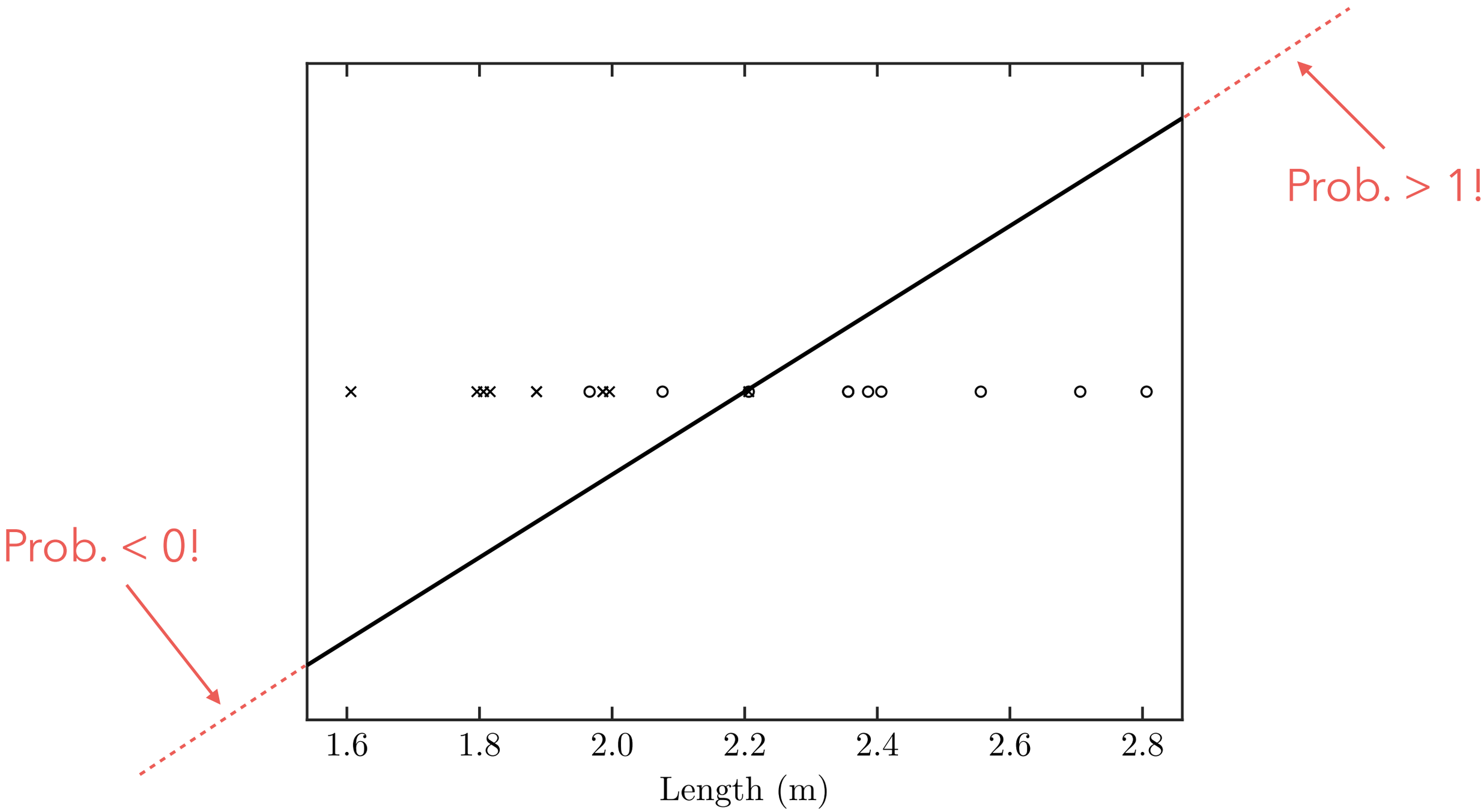


Discriminative model



Problem?

Problem



Discriminative model

- Let's make it more complicated:
  - Ratio between π(1 | *x*) and π(0 | *x*) is linear in *x*:

$$\frac{\pi(1 \mid x)}{\pi(0 \mid x)} = w_0 + w_1 \phi(x)$$

Goes from
0 to ∞

Discriminative model

- Let's make it more complicated:
  - Ratio between π(1 | *x*) and π(0 | *x*) is linear in *x*:

$$\frac{\pi(1 \mid x)}{1 - \pi(1 \mid x)} = w_0 + w_1 \phi(x)$$

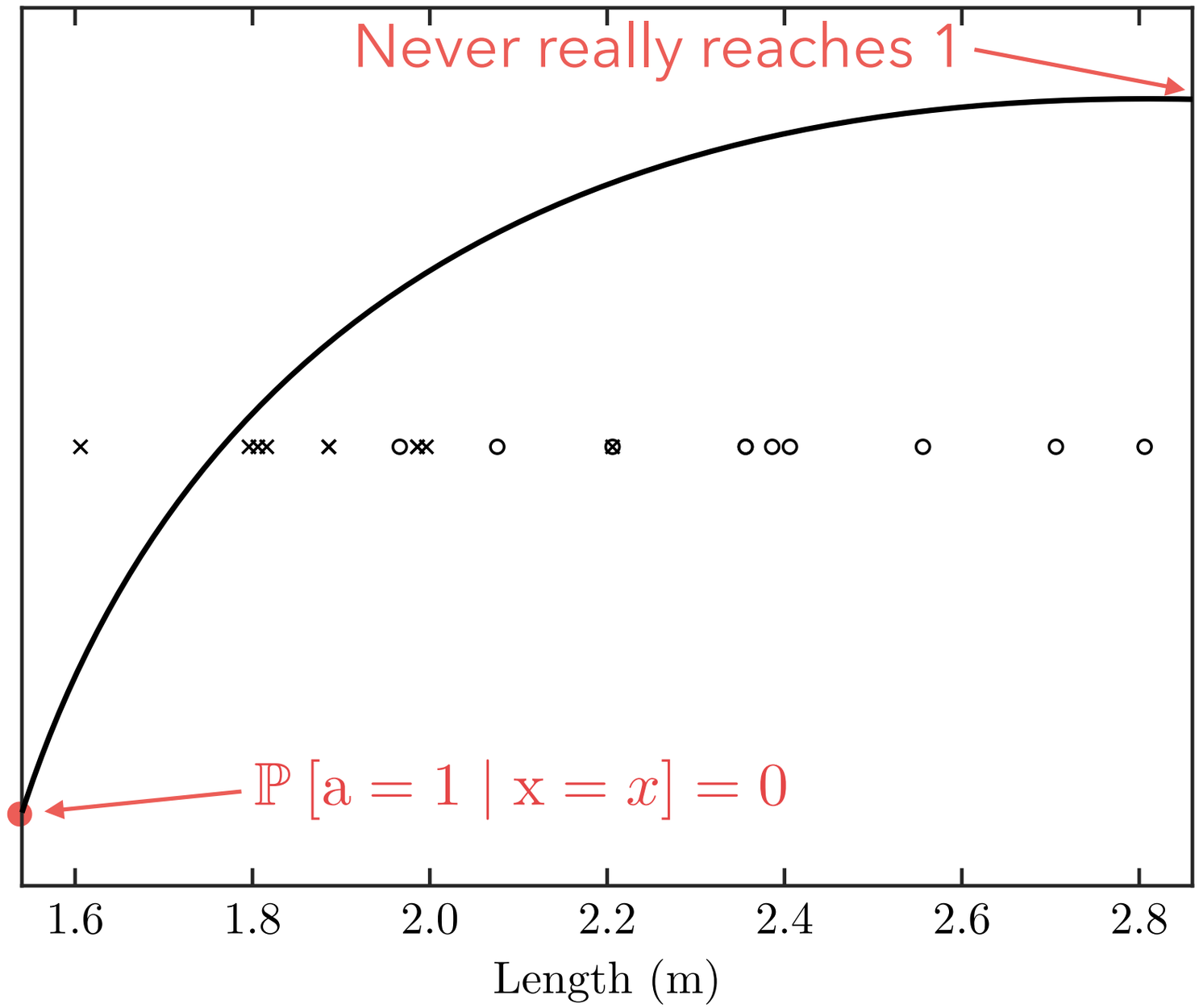
Discriminative model

- Let's make it more complicated:
  - Ratio between π(1 | *x*) and π(0 | *x*) is linear in *x*:

$$\frac{\pi(1 \mid x)}{1 - \pi(1 \mid x)} = w_0 + w_1 \phi(x)$$

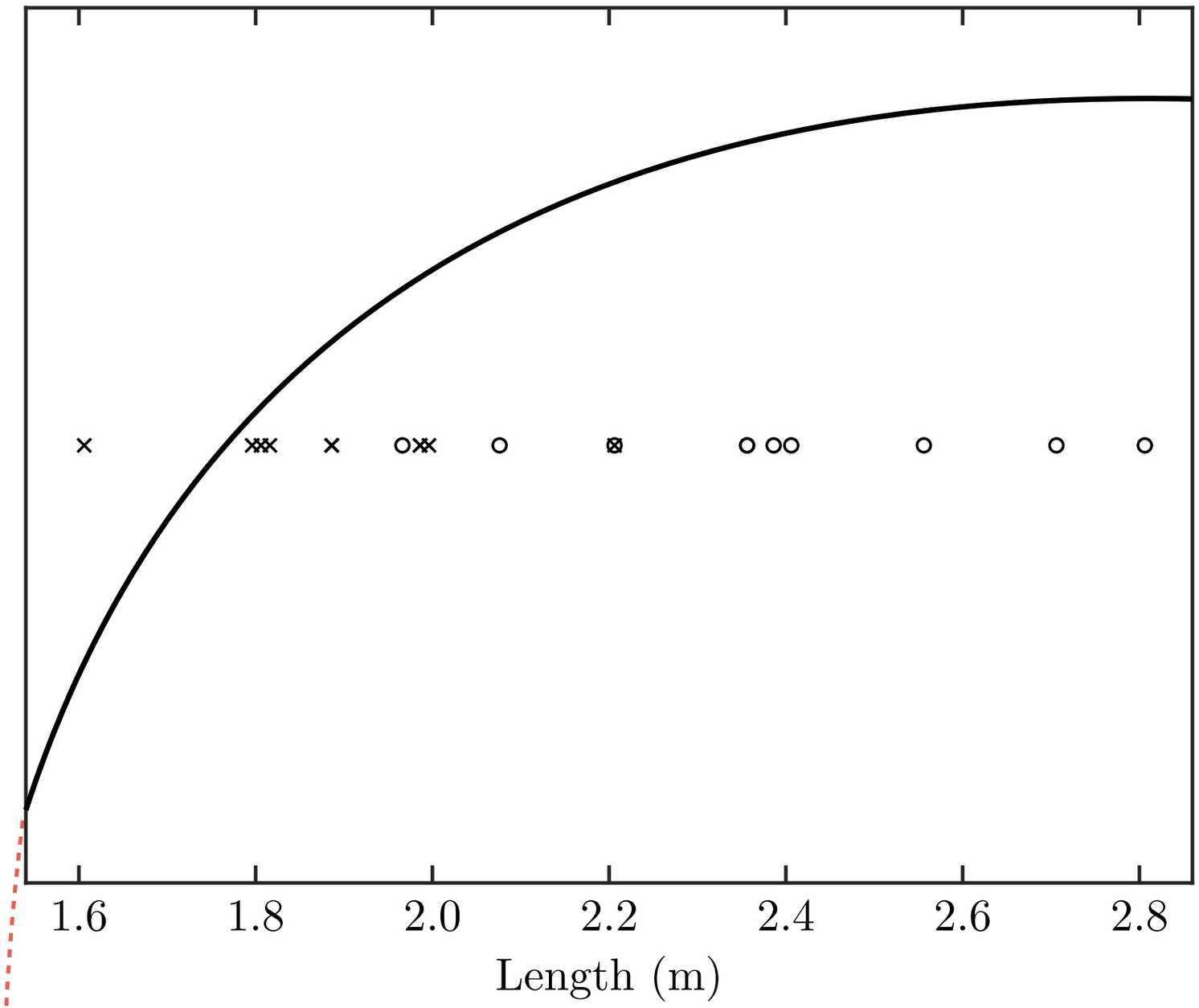
$$\pi(1 \mid x) = \frac{w_0 + w_1 \phi(x)}{1 + w_0 + w_1 \phi(x)}$$

Discriminative model



Problem?

Problem



Prob. < 0!

Discriminative model

- Let's make it more complicated:
  - Log-ratio between π(1 | *x*) and π(0 | *x*) is linear in *x*:

$$\log \frac{\pi(1 \mid x)}{\pi(0 \mid x)} = w_0 + w_1 \phi(x)$$

Goes from
-∞ to +∞

Discriminative model

- Let's make it more complicated:
  - Log-ratio between π(1 | *x*) and π(0 | *x*) is linear in *x*:

$$\log \frac{\pi(1 \mid x)}{1 - \pi(1 \mid x)} = w_0 + w_1 \phi(x)$$

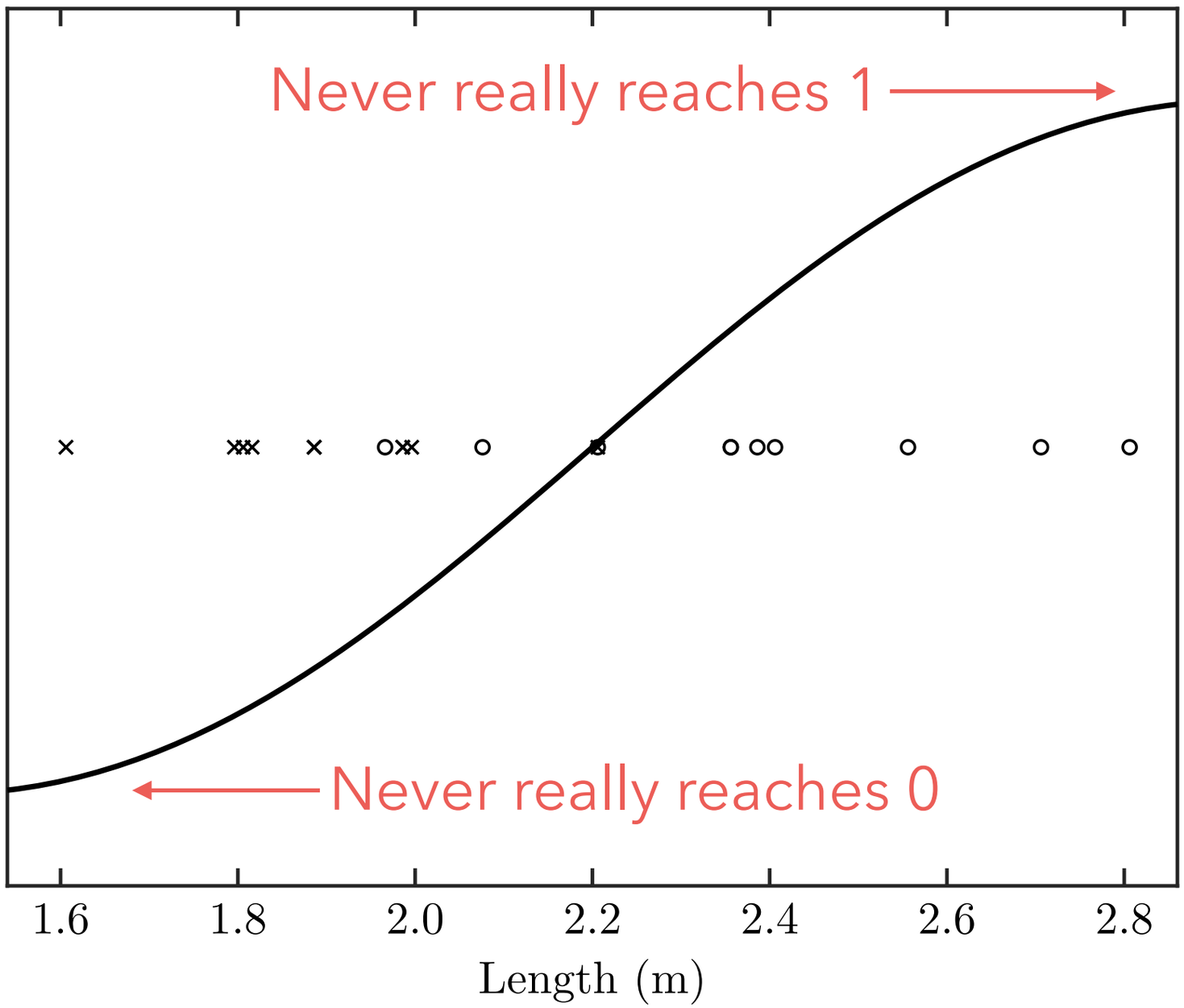
Discriminative model

- Let's make it more complicated:
  - Log-ratio between π(1 | *x*) and π(0 | *x*) is linear in *x*:

$$\log \frac{\pi(1 \mid x)}{1 - \pi(1 \mid x)} = w_0 + w_1 \phi(x)$$

$$\pi(1 \mid x) = \frac{1}{1 + e^{-(w_0 + w_1 \phi(x))}}$$

Discriminative model



Logistic regression

- Training the classifier corresponds to computing w_0 and w_1

$$\pi(1 \mid x) = \frac{1}{1 + e^{-(w_0 + w_1 \phi(x))}}$$

- How can we evaluate the quality of the classifier?
  - Negative log-likelihood!

$$\ell(x, a; \pi) = -\log \pi(a \mid x)$$

$$\hat{L}_N(\pi) = -\frac{1}{N} \sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

Motivation

- If all samples are independent, likelihood of data is

$$\text{likelihood}(\mathcal{D}) = \prod_{n=1}^{N} \pi(a_n \mid x_n)$$

- Negative log-likelihood comes:

$$\text{negative log-likelihood}(\mathcal{D}) = -\sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

$$\hat{L}_N(\pi) = -\frac{1}{N} \sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

Empirical risk

- Departing from the empirical risk

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

- If $a_n = 1$,

$$\pi(1 \mid x_n)^{a_n} = \pi(a_n \mid x_n)$$

$$\pi(0 \mid x_n)^{1-a_n} = 1$$

$$\downarrow$$

$$\pi(a_n \mid x_n) = \pi(1 \mid x_n)^{a_n} \pi(0 \mid x_n)^{1-a_n}$$

Empirical risk

- Departing from the empirical risk

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

- If $a_n = 0$,

$$\pi(1 \mid x_n)^{a_n} = 1$$

$$\pi(0 \mid x_n)^{1-a_n} = \pi(a_n \mid x_n)$$

$$\pi(a_n \mid x_n) = \pi(1 \mid x_n)^{a_n} \pi(0 \mid x_n)^{1-a_n}$$

Empirical risk

- Departing from the empirical risk

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

- Putting everything together…

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log \left(\pi(1 \mid x_n)^{a_n} \pi(0 \mid x_n)^{1-a_n}\right)$$

Empirical risk

- Finally,

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} a_n \log \pi(1 \mid x_n) + (1 - a_n) \log(1 - \pi(1 \mid x_n))$$

Learning with LR

- Training the classifier corresponds to computing w_0 and w_1

$$\pi(1 \mid x) = \frac{1}{1 + e^{-(w_0 + w_1 \phi(x))}}$$

- How can we train the classifier?
  - Select w_0 and w_1 to minimize negative log-likelihood

Learning with LR

- For example, using the gradient:

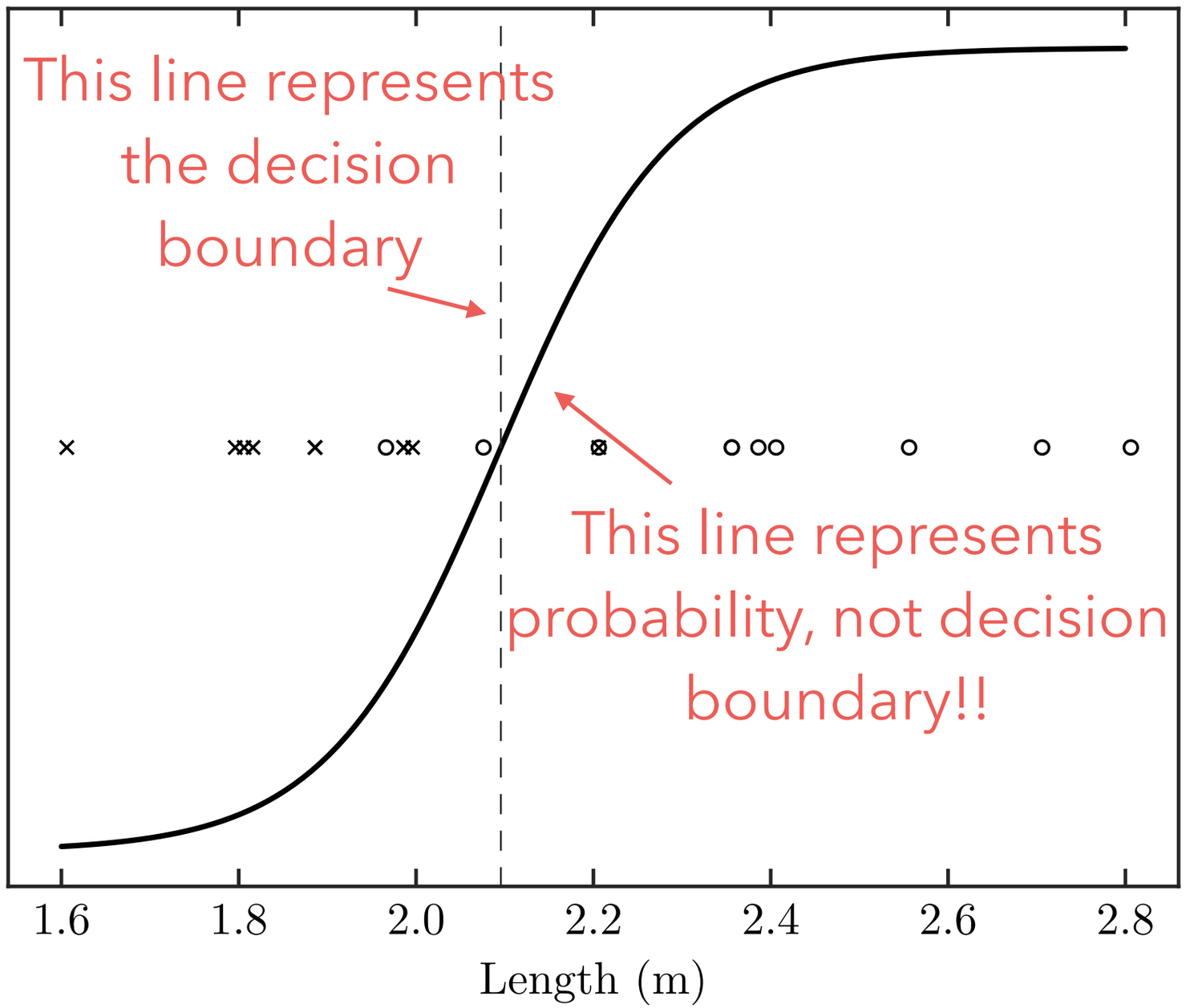
$$\frac{\partial \hat{L}_N(\pi)}{\partial w_0} = \frac{1}{N}\sum_{n=1}^{N}(\pi(1 \mid x_n) - a_n)$$

$$\frac{\partial \hat{L}_N(\pi)}{\partial w_1} = \frac{1}{N}\sum_{n=1}^{N}\phi(x_n)(\pi(1 \mid x_n) - a_n)$$

we can compute w_0 and w_1 using gradient descent:

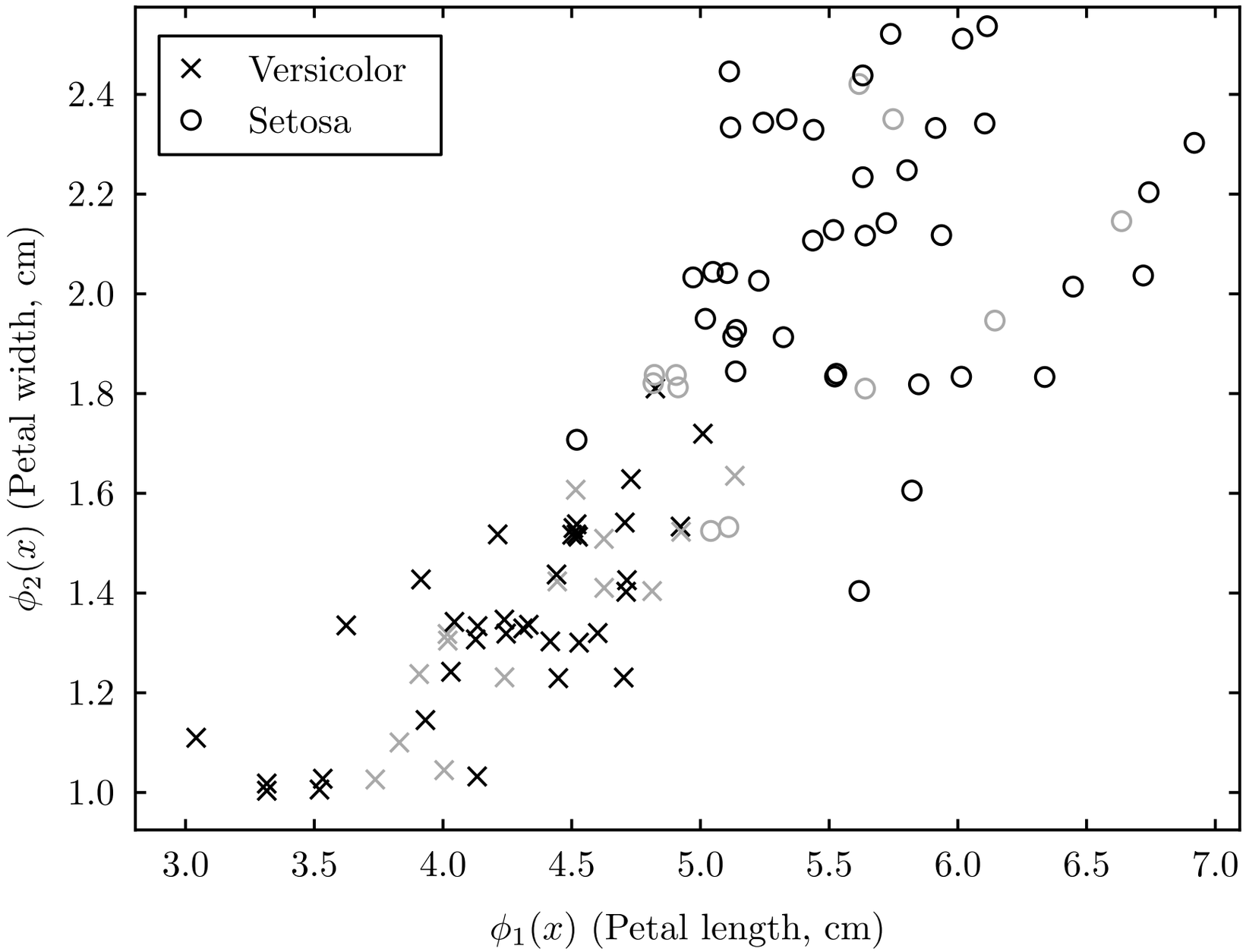
$$\boldsymbol{w} = \boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} \hat{L}_N(\pi)$$

Example


This line represents the decision boundary
This line represents probability, not decision boundary!!
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Length (m)

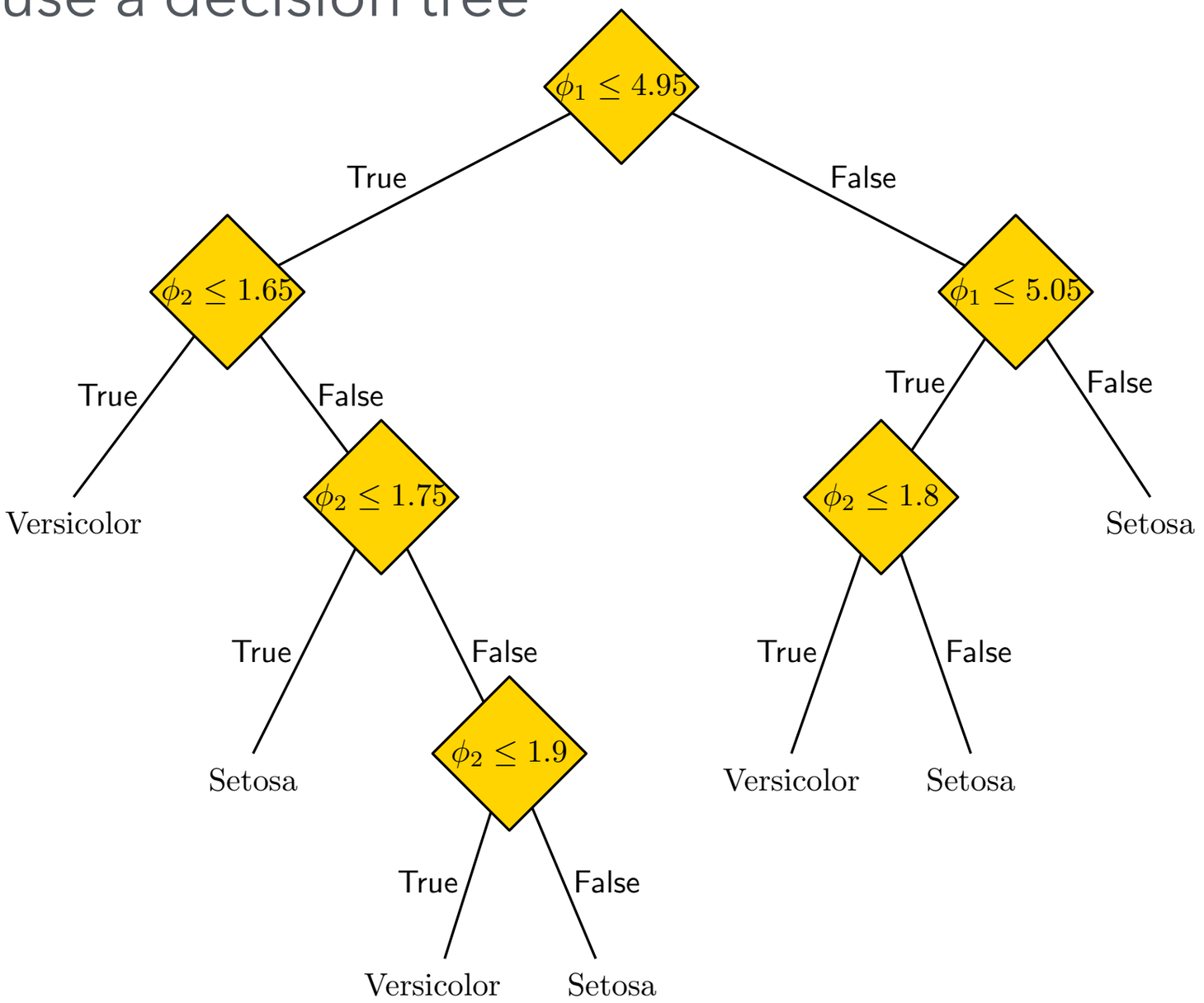
Example

Example

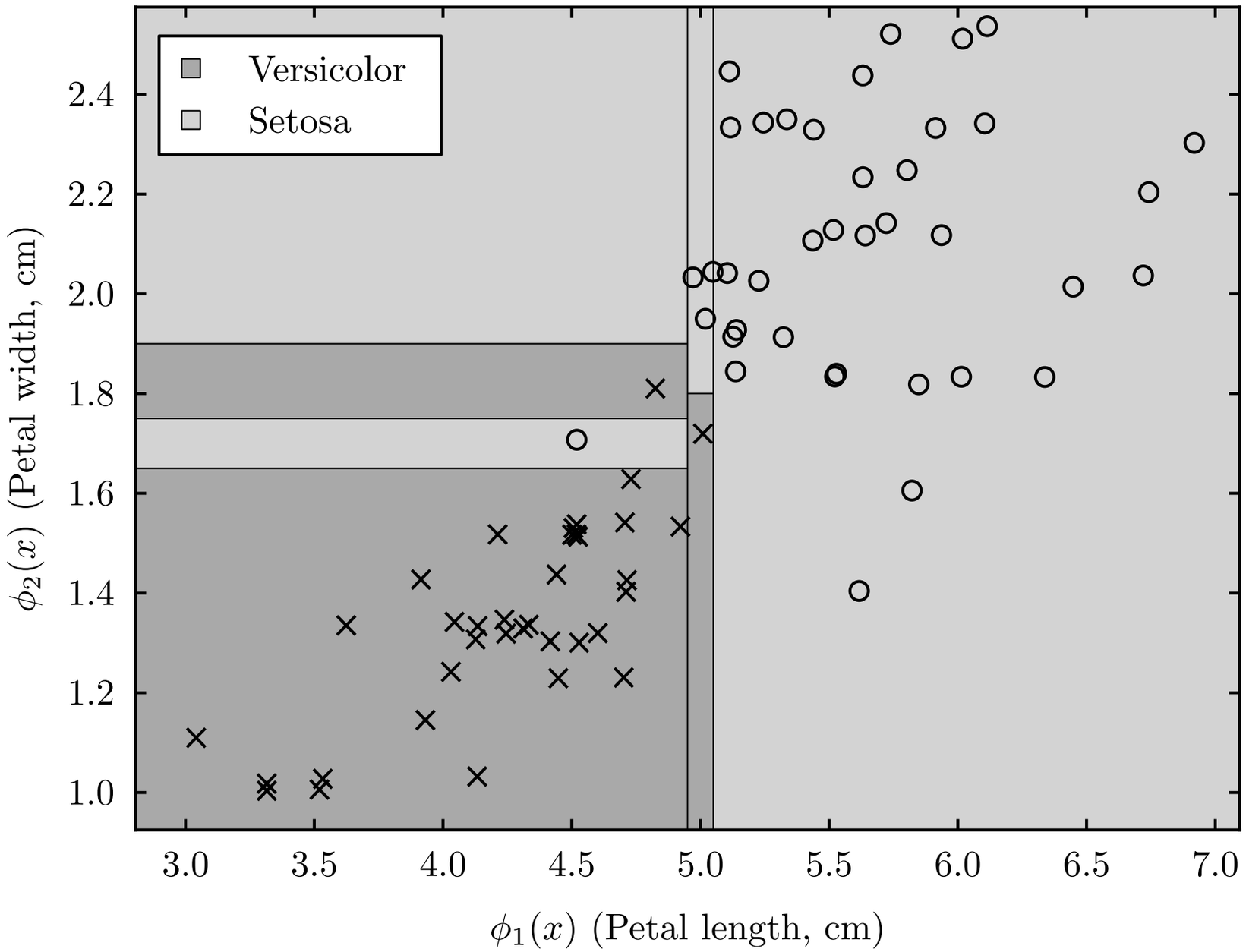

Versicolor
Setosa
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
$\phi_2(x)$ (Petal width, cm)
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
$\phi_1(x)$ (Petal length, cm)

Decision tree

- We can use a decision tree



Decision tree



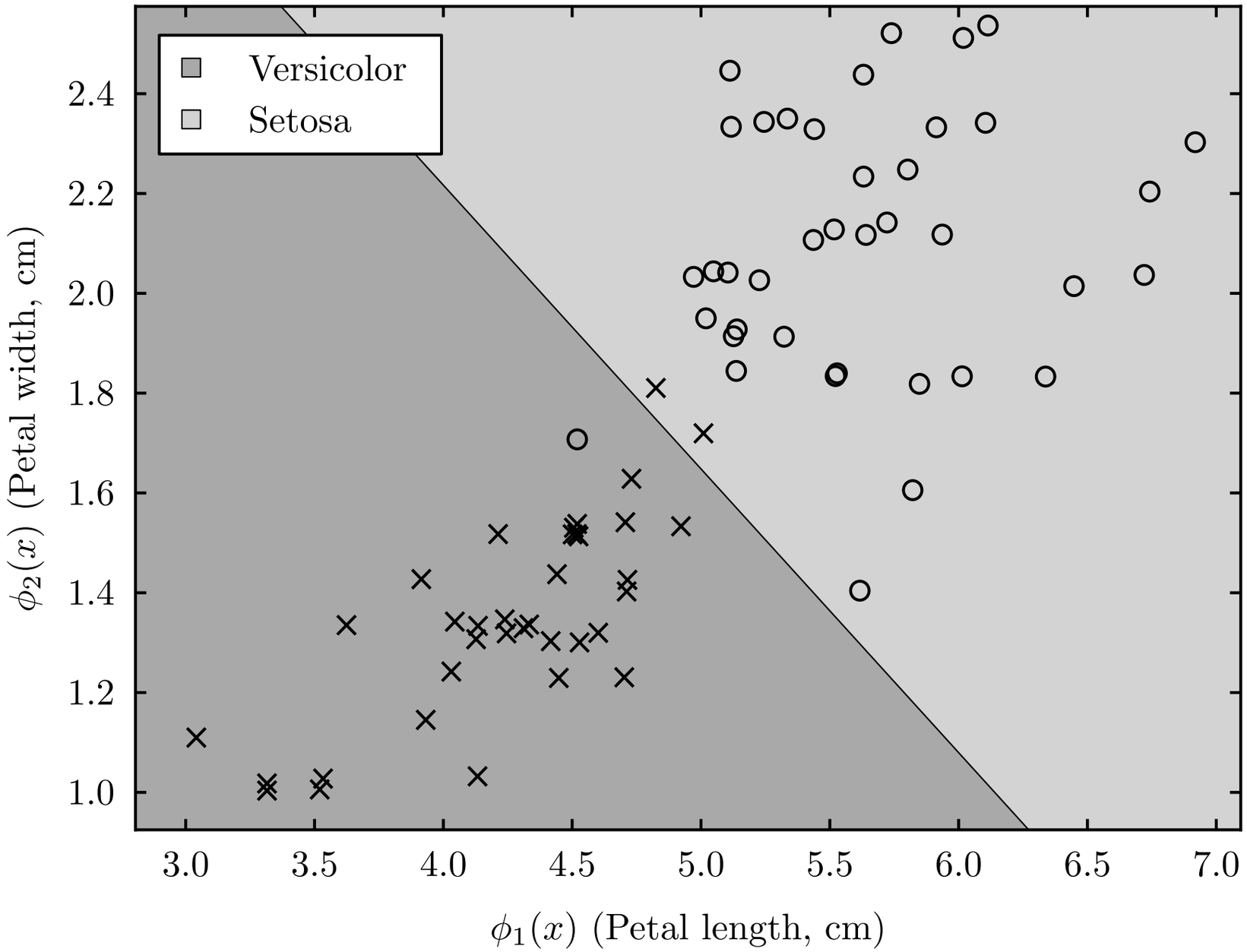
Using logistic regression

- The LR classifier now becomes

$$\pi(1 \mid x) = \frac{1}{1 + e^{-(w_0 + w_1\phi_1(x) + w_2\phi_2(x))}}$$
$$= \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{\phi}(x)}}$$

- Note that there is one extra parameter (data-points have 2 coordinates)
- Select w_0, w_1 and w_2 to minimize negative log-likelihood

Logistic regression



Discriminative model

- Logistic regression provides a discriminative model, computing the probability

$$\pi(1 \mid x) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{\phi}(x)}}$$

- However, it is possible to compute that same probability using Bayes rule:

$$\begin{aligned}\pi(1 \mid x) &= \mathbb{P}[\mathrm{a} = 1 \mid \mathrm{x} = x] \\ &= \frac{\mathbb{P}[\mathrm{x} = x \mid \mathrm{a} = 1]\mathbb{P}[\mathrm{a} = 1]}{\mathbb{P}[\mathrm{x} = x]}\end{aligned}$$

Discriminative model

- Logistic regression provides a discriminative model, computing the probability

$$\pi(1 \mid x) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{\phi}(x)}}$$

- However, it is possible to compute that same probability using Bayes rule:

$$\begin{aligned}\pi(1 \mid x) &= \mathbb{P}[\mathrm{a} = 1 \mid \mathrm{x} = x] \\ &= \frac{\mathbb{P}[\mathrm{x} = x, \mathrm{a} = 1]}{\mathbb{P}[\mathrm{x} = x]}\end{aligned}$$

Discriminative model

- Logistic regression provides a discriminative model, computing the probability

$$\pi(1 \mid x) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{\phi}(x)}}$$

- However, it is possible to compute that same probability using Bayes rule:

$$\begin{aligned}\pi(1 \mid x) &= \mathbb{P}[\mathrm{a} = 1 \mid \mathrm{x} = x] \\ &= \frac{\mathbb{P}[\mathrm{x} = x, \mathrm{a} = 1]}{\sum_{a' \in \{0,1\}} \mathbb{P}[\mathrm{x} = x, \mathrm{a} = a']}\end{aligned}$$

Generative model

- In general, the use of Bayes rule involves knowing

$$\mathbb{P}\left[\mathrm{x} = x \mid \mathrm{a} = a\right] \mathbb{P}\left[\mathrm{a} = a\right] = \mathbb{P}\left[\mathrm{x} = x, \mathrm{a} = a\right]$$

Joint distribution

Generative model

- A model of the joint distribution $\mathbb{P}\left[\mathrm{x} = x, \mathrm{a} = a\right]$ is called a **generative model**
- It allows the generation of pairs (x, a) – not just actions a from x

Generative model

- The joint distribution $\mathbb{P}\left[\mathrm{x} = x, \mathrm{a} = a\right]$ is usually estimated by estimating both:
  - The prior, $\mathbb{P}\left[\mathrm{a} = a\right]$
  - The likelihood, $\mathbb{P}\left[\mathrm{x} = x \mid \mathrm{a} = a\right]$

Generative model

- The prior can usually be computed easily
  - Simply count relative frequency of each class
- The likelihood is often unpractical to compute
  - … particularly if the state x is multidimensional

Naive Bayes

- Naive Bayes is designed to alleviate the computation of the likelihood

The Naive Bayes assumption

Naive Bayes assumes that the different features describing the state are independent given the class/action.

Naive Bayes

- If each datapoint x is described by attributes/features ϕ_1, …, ϕ_K, the likelihood is given by

$$\mathbb{P}\left[\boldsymbol{\phi}(x) = \mathbf{u} \mid a\right] = \mathbb{P}\left[\phi_1(x) = u_1, \ldots, \phi_K(x) = u_K \mid a\right]$$

Large K-dimensional table

Naive Bayes

- If each datapoint x is described by attributes/features ϕ_1, …, ϕ_K, the likelihood is given by

$$\mathbb{P}\left[\boldsymbol{\phi}(x) = \mathbf{u} \mid a\right] = \mathbb{P}\left[\phi_1(x) = u_1, \ldots, \phi_K(x) = u_K \mid a\right]$$

- Naive Bayes assumes that

$$\mathbb{P}\left[\boldsymbol{\phi}(x) = \mathbf{u} \mid a\right] = \prod_{k=1}^{K} \mathbb{P}\left[\phi_k(x) = u_k \mid a\right]$$

K

1-dimensional tables

Naive Bayes

- For discrete attributes,

$$\mathbb{P}\left[\phi_k(x) = u_k \mid a\right] = \frac{N_{u_k,a}}{N_a}$$

Examples where action *a* occurs and attr. *k* takes value u_k

Examples where action *a* occurs

Naive Bayes

- For continuous attributes, we represent likelihood

$$\mathbb{P}\left[\mathrm{x} = x \mid \mathrm{a} = a\right]$$

  as a Gaussian

- Gaussian parameters:
  - Mean:

$$\mu_{a,k} = \frac{1}{N_a} \sum_{n=1}^{N} \phi_k(x_n) \mathbb{I}(a_n = a)$$

  - Variance:

$$\sigma^2_{a,k} = \frac{1}{N_a - 1} \sum_{n=1}^{N} (\phi_k(x_n) - \mu_{a,k})^2 \mathbb{I}(a_n = a)$$

NB & LR

- For continuous attributes, under the Gaussian assumption, the likelihood of NB and the probabilities computed by LR are equivalent
  - The two approaches compute the parameters of the likelihood in different ways
- … may lead to different results

Example

Gun possession	Authority contacts	Movement	Appearance	Criminal
?	Yes	Slowly	Suited	No
No	Yes	Average	Suited	No
?	No	Average	Suited	Yes
Yes	No	Fast	Casual	Yes
?	No	Fast	Rough	Yes
No	No	Slowly	Casual	No
No	No	Fast	Casual	No
?	Yes	Average	Rough	No

$$\mathbb{P}\left[a = 1\right] = \frac{3}{8}$$

$$\mathbb{P}\left[a = 0\right] = \frac{5}{8}$$

Example

Gun possession	Authority contacts	Movement	Appearance	Criminal
?	Yes	Slowly	Suited	No
No	Yes	Average	Suited	No
?	No	Average	Suited	Yes
Yes	No	Fast	Casual	Yes
?	No	Fast	Rough	Yes
No	No	Slowly	Casual	No
No	No	Fast	Casual	No
?	Yes	Average	Rough	No

$$\mathbb{P}\left[\text{Gun pos.} \mid a=1\right] = \frac{1}{3}$$
$$\mathbb{P}\left[\neg\text{Gun pos.} \mid a=1\right] = 0$$
$$\mathbb{P}\left[\text{Unk. gun pos.} \mid a=1\right] = \frac{2}{3}$$

$$\mathbb{P}\left[\text{Gun pos.} \mid a=0\right] = 0$$
$$\mathbb{P}\left[\neg\text{Gun pos.} \mid a=0\right] = \frac{3}{5}$$
$$\mathbb{P}\left[\text{Unk. gun pos.} \mid a=0\right] = \frac{2}{5}$$

Example

Gun possession	Authority contacts	Movement	Appearance	Criminal
?	Yes	Slowly	Suited	No
No	Yes	Average	Suited	No
?	No	Average	Suited	Yes
Yes	No	Fast	Casual	Yes
?	No	Fast	Rough	Yes
No	No	Slowly	Casual	No
No	No	Fast	Casual	No
?	Yes	Average	Rough	No

$$\mathbb{P}\left[\text{Contact auth.} \mid a = 1\right] = 0$$
$$\mathbb{P}\left[\neg\text{Contact auth.} \mid a = 1\right] = 1$$

$$\mathbb{P}\left[\text{Contact auth.} \mid a = 0\right] = \frac{3}{5}$$
$$\mathbb{P}\left[\neg\text{Contact auth.} \mid a = 0\right] = \frac{2}{5}$$

Example

Gun possession	Authority contacts	Movement	Appearance	Criminal
?	Yes	Slowly	Suited	No
No	Yes	Average	Suited	No
?	No	Average	Suited	Yes
Yes	No	Fast	Casual	Yes
?	No	Fast	Rough	Yes
No	No	Slowly	Casual	No
No	No	Fast	Casual	No
?	Yes	Average	Rough	No

$$\mathbb{P}\left[\text{Mov. slow} \mid a=1\right] = 0$$

$$\mathbb{P}\left[\text{Mov. aver.} \mid a=1\right] = \frac{1}{3}$$

$$\mathbb{P}\left[\text{Mov. fast} \mid a=1\right] = \frac{2}{3}$$

$$\mathbb{P}\left[\text{Mov. slow} \mid a=0\right] = \frac{2}{5}$$

$$\mathbb{P}\left[\text{Mov. aver.} \mid a=0\right] = \frac{2}{5}$$

$$\mathbb{P}\left[\text{Mov. fast} \mid a=0\right] = \frac{1}{5}$$

Example

Gun possession	Authority contacts	Movement	Appearance	Criminal
?	Yes	Slowly	Suited	No
No	Yes	Average	Suited	No
?	No	Average	Suited	Yes
Yes	No	Fast	Casual	Yes
?	No	Fast	Rough	Yes
No	No	Slowly	Casual	No
No	No	Fast	Casual	No
?	Yes	Average	Rough	No

$$\mathbb{P}\left[\text{Suited} \mid a=1\right] = \frac{1}{3} \qquad \mathbb{P}\left[\text{Suited} \mid a=0\right] = \frac{2}{5}$$

$$\mathbb{P}\left[\text{Casual} \mid a=1\right] = \frac{1}{3} \qquad \mathbb{P}\left[\text{Casual} \mid a=0\right] = \frac{2}{5}$$

$$\mathbb{P}\left[\text{Rough} \mid a=1\right] = \frac{1}{3} \qquad \mathbb{P}\left[\text{Rough} \mid a=0\right] = \frac{1}{5}$$

Example

- Let's classify an individual:

(?, No, Fast, Suited)

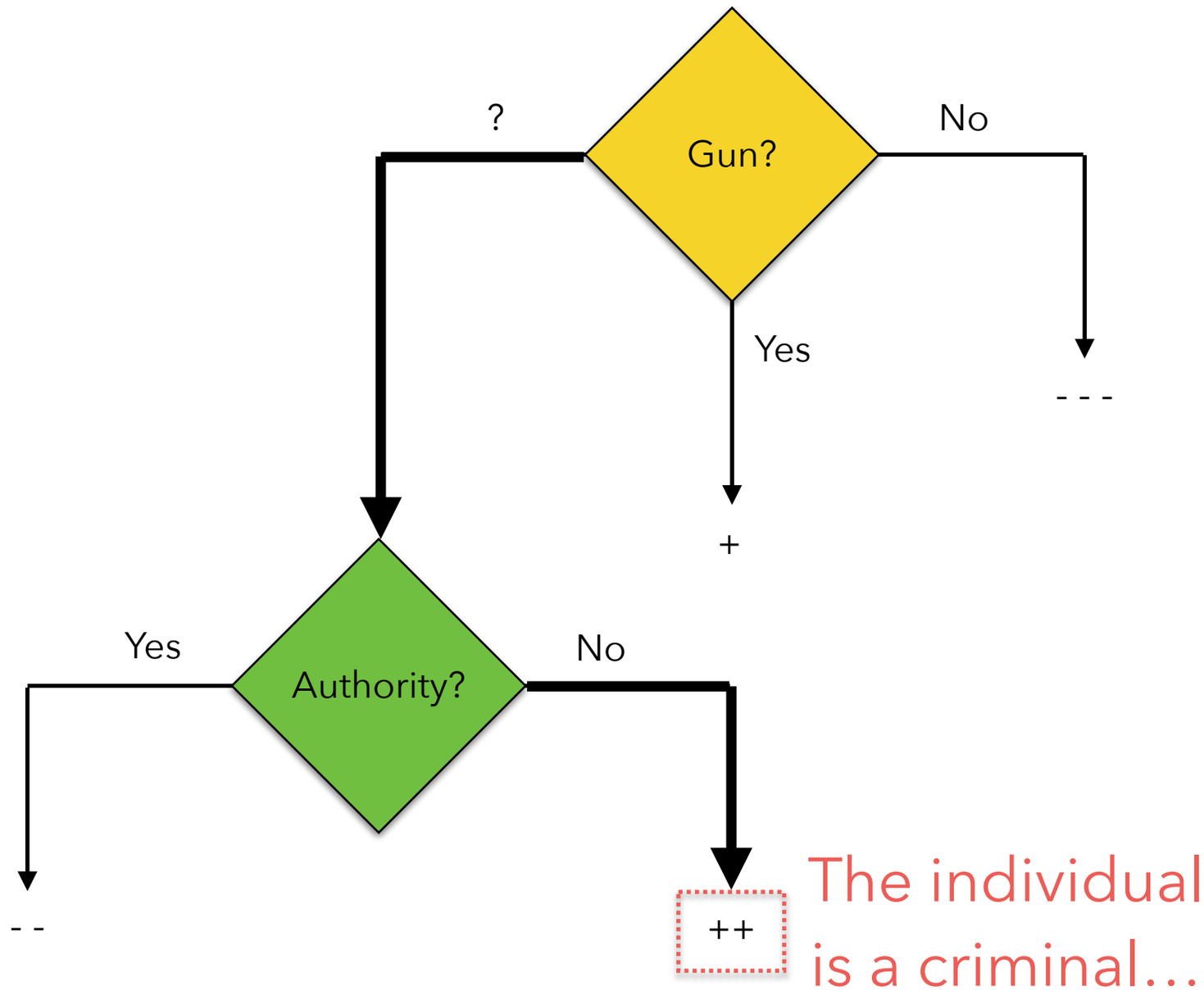
- Using NB, we get:

$$\mathbb{P}\left[a = 1 \mid \mathrm{x} = (?, \text{No}, \text{Fast}, \text{Suited})\right] \propto \frac{3}{8} \times \frac{2}{3} \times 1 \times \frac{2}{3} \times \frac{1}{3} = 0.0556$$

$$\mathbb{P}\left[a = 0 \mid \mathrm{x} = (?, \text{No}, \text{Fast}, \text{Suited})\right] \propto \frac{5}{8} \times \frac{2}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} = 0.008$$

- It is more likely that the individual is a criminal…

Recall the DT…



The individual is a criminal…

Example

Length (m)	Gender
2.38	M
1.85	F
2.4	M
1.67	F
1.73	F
2.34	M
1.69	F
1.82	F
1.88	F
1.41	F
2.32	M
2.28	M
2.2	M
2.21	M
2.16	M
1.94	M
1.74	F
1.8	F
1.95	F
2.2	M

- Class probabilities:

$$\mathbb{P}\left[a=1\right]=\frac{1}{2}$$

$$\mathbb{P}\left[a=0\right]=\frac{1}{2}$$

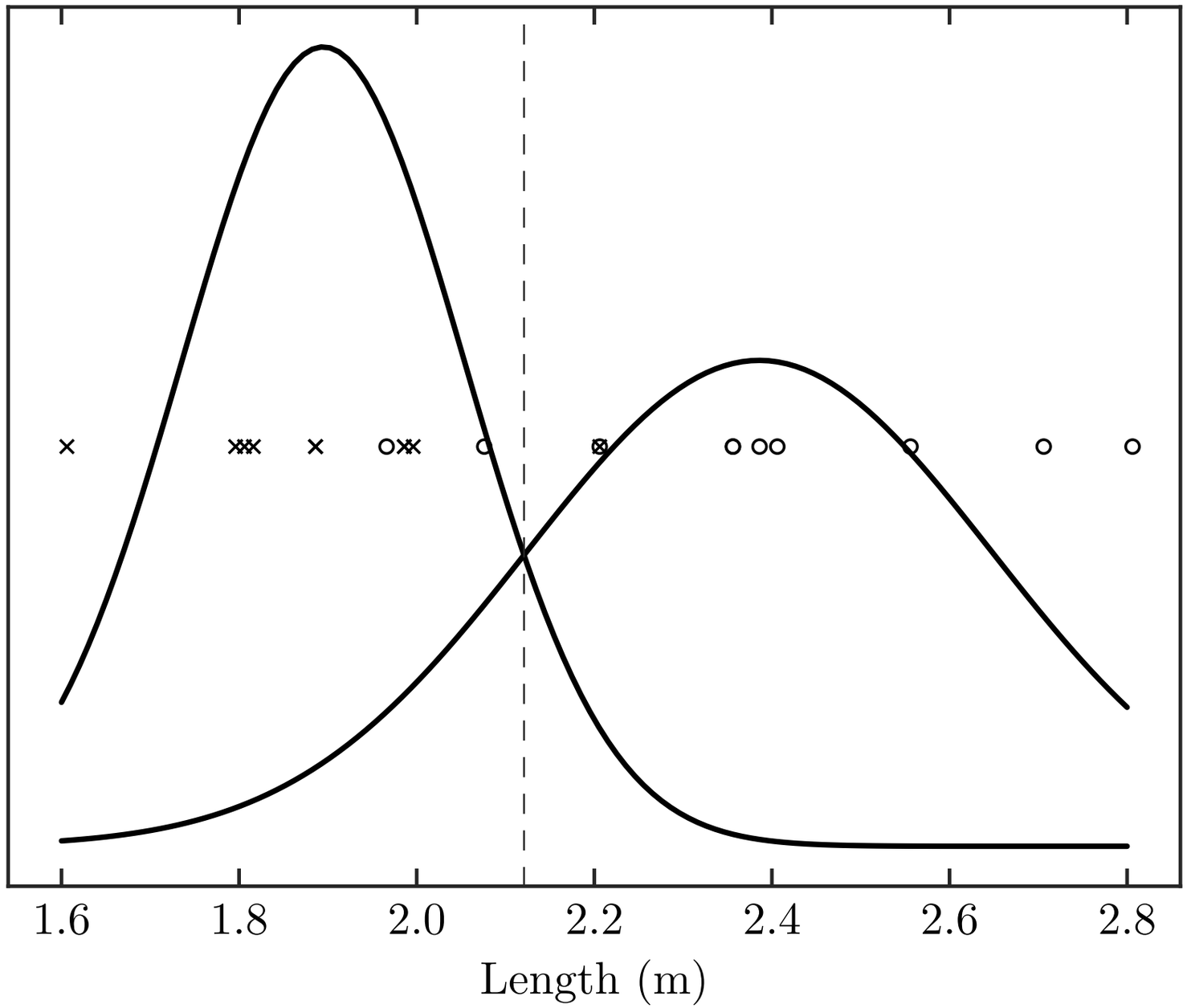
- Attribute modeling ($a = 1$):

$$\mu_1 = 2.3792 \qquad \sigma_1^2 = 0.0674$$

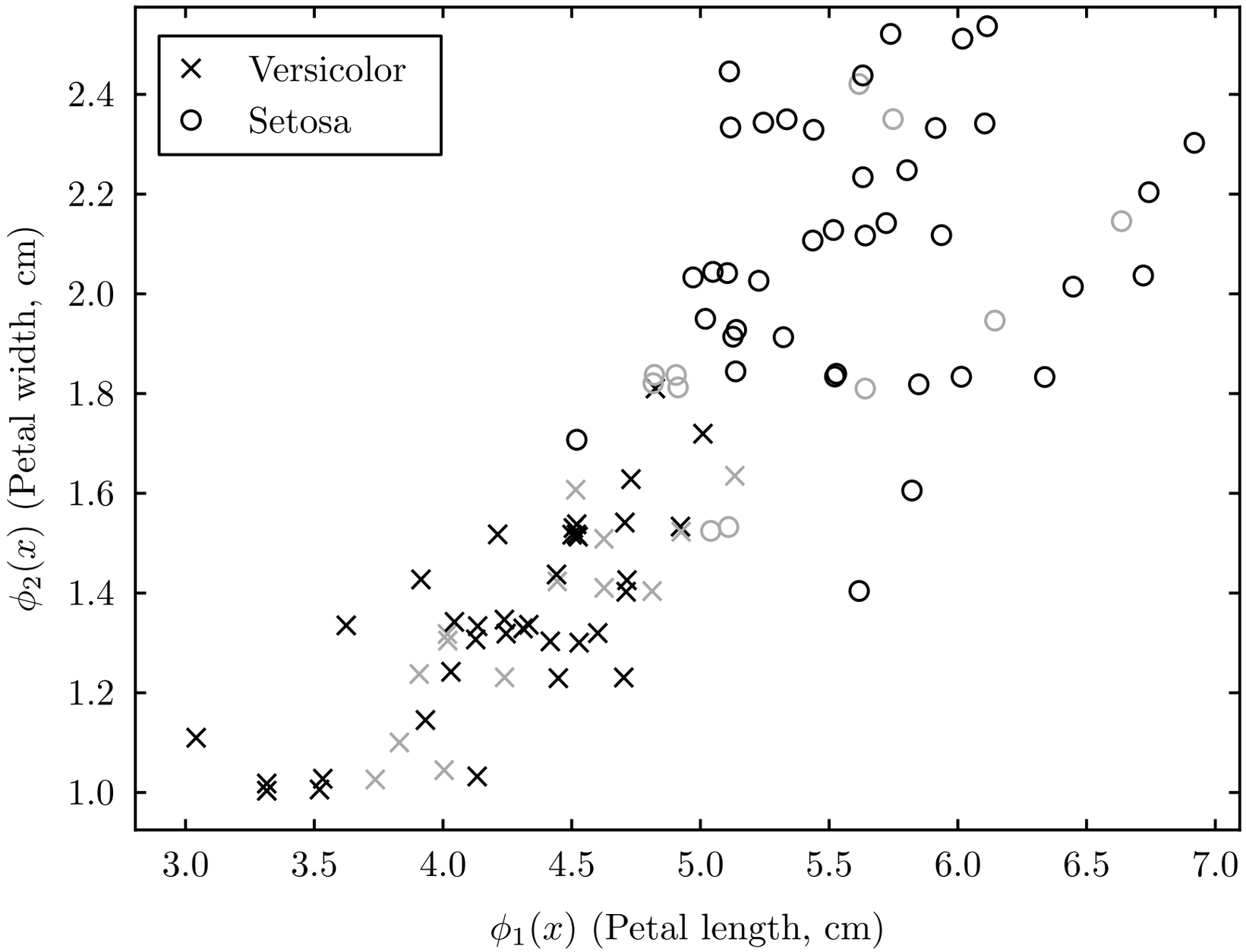
- Attribute modeling ($a = 0$):

$$\mu_0 = 1.8922 \qquad \sigma_0^2 = 0.0249$$

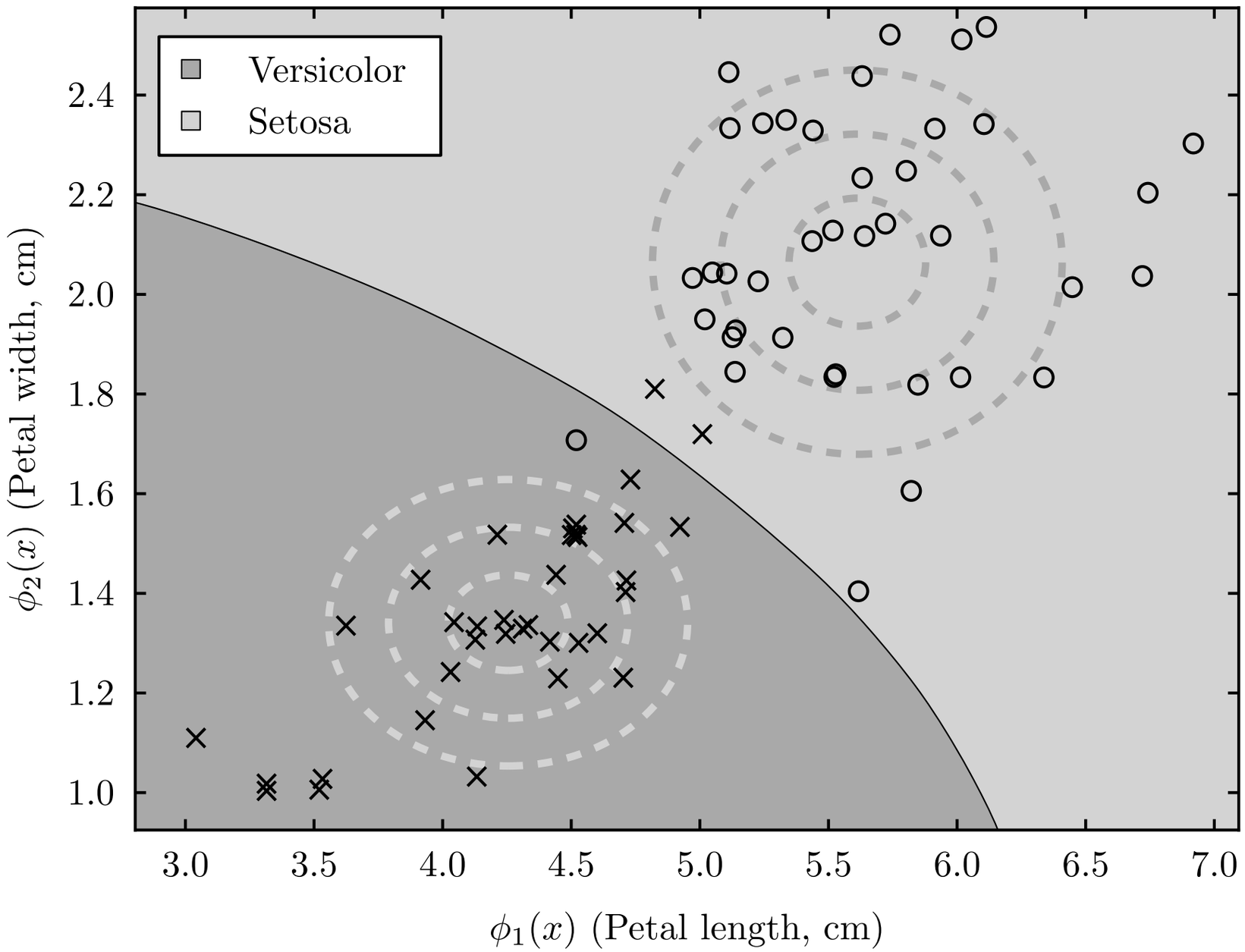
Example



Example


Versicolor
Setosa
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
$\phi_2(x)$ (Petal width, cm)
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
$\phi_1(x)$ (Petal length, cm)

Example


Versicolor
Setosa
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
$\phi_2(x)$ (Petal width, cm)
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
$\phi_1(x)$ (Petal length, cm)

To conclude…

- LR and NB are two “cousin” classifiers
  - LR is a discriminative classifier
  - NB is a generative classifier

To conclude…

- NB assumes independence of attributes given class
  - Good when there is little data
  - Non-independent attributes counted “twice”

To conclude…

- LR does not assume independence
- We should watch out for overfitting