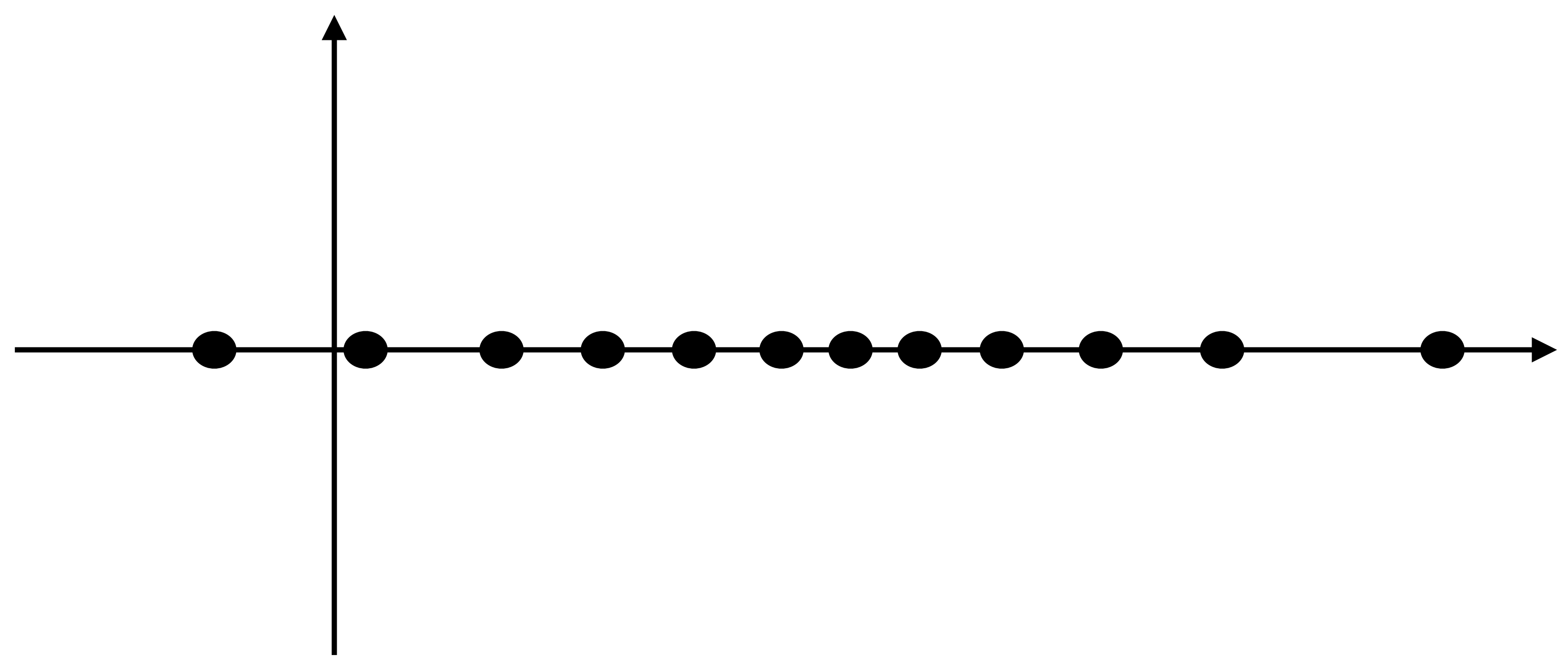
Bayes Classifier and Naive Bayes

Announcements

HW 2 is out — start early

Recap on MLE



$$\mathcal{D} = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}$$

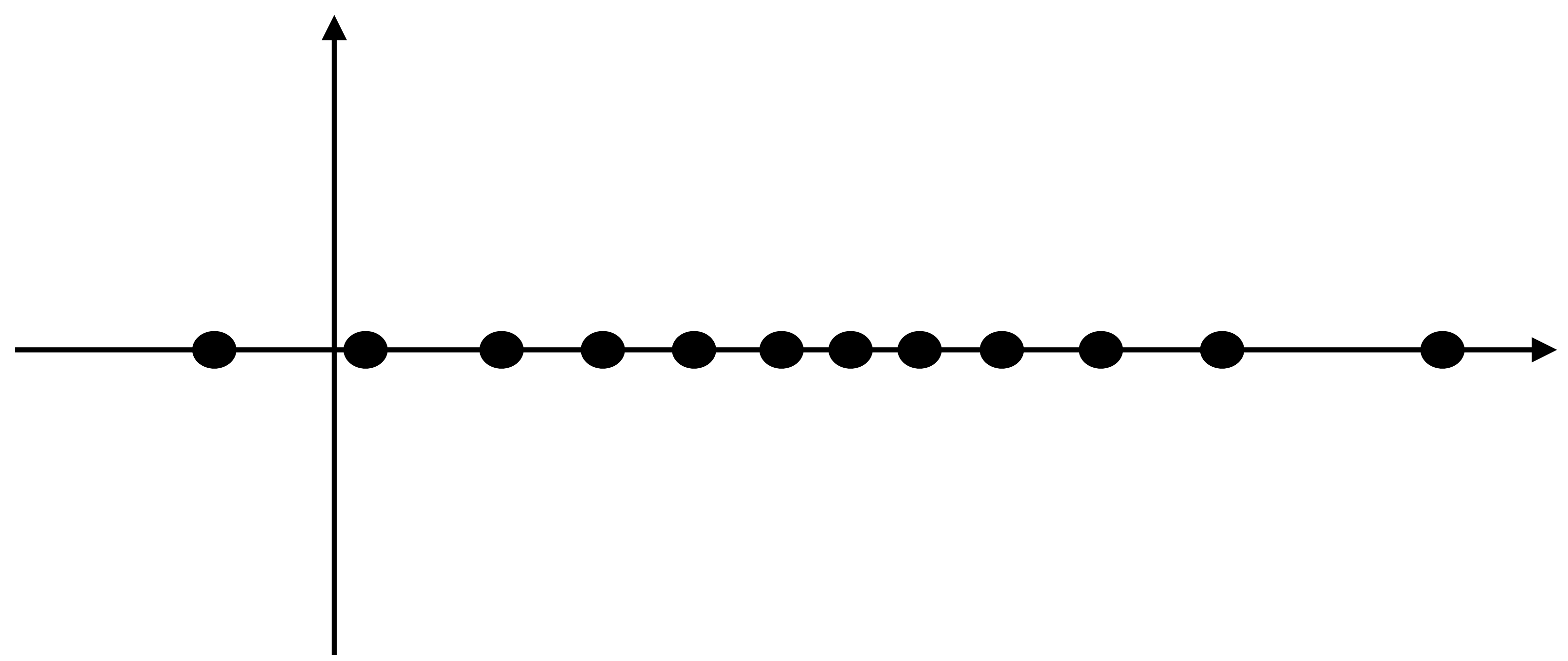
Assume data is from $\mathcal{N}(\mu^\star, \sigma^2)$, want to estimate $\mu^\star, \sigma$ from the data $\mathcal{D}$ MLE

$$P(\mathcal{D} \mid \mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$$

The solution that maximizes the log-likelihood:

$$\hat{\mu} = \sum_{i=1}^n x_i/n, \;\; \hat{\sigma}^2 = \sum_{i=1}^n (x_i - \hat{\mu})^2/n$$

Recap on MAP



$$\mathcal{D} = \{x_i\}_{i=1}^{n}, x_i \in \mathbb{R}$$

Now if we want to use **MAP:**

$P(\mu)P(\sigma)$

1. Pick a prior: $P(\mu, \sigma)$

2. Write down data-likelihood: $P(\mathcal{D} \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2/\sigma^2\right)$

3. Form posterior $P(\mu, \sigma \mid \mathcal{D}) \propto P(\mu, \sigma)P(\mathcal{D} \mid \mu, \sigma)$

Today

Objective: learn our second classification algorithm—Naive Bayes (derived via MLE)

Outline

1. General formulation of Naive Bayes

2. Example

3. Connection to linear classifier

Generative modeling

Setting: binary classification w/ dataset $\{x_i, y_i\}_{i=1}^n, (x_i, y_i) \sim P$, where $x \in \mathbb{R}^d, y \in \{-1,1\}$

Goal: estimate $P(y \mid x)$

We take a **generative modeling** approach here:

$$P(y \mid x) \propto P(x \mid y)P(y)$$

Estimate $P(x \mid y)$ & $P(y)$ from data
(hence generative modeling)

(Discriminative modeling: directly estimate $P(y \mid x)$)

Naive Bayes

Estimate $P(y)$ from data:

$$P(y \mid x) \propto P(x \mid y)P(y)$$

Estimate $P(y)$ is easy:

$$P(y = 1) \approx \frac{\sum_{i=1}^{n} \mathbf{1}(y_i = 1)}{n}$$

Naive Bayes

Estimate $P(x \mid y)$ from data:

$$P(y \mid x) \propto P(x \mid y) P(y)$$

Estimate $P(x \mid y)$ is not easy:

x can be high-dimensional, e.g., d is large!

There may not be repetitions in $\{x_i\}_{i=1}^n$!

The key assumption in Naive Bayes

The Naive Bayes assumption:

$$P(x \mid y) = \prod_{\alpha=1}^{d} P(x[\alpha] \mid y)$$

Conditioned on label y, feature values are **independent**!

About the independence assumption

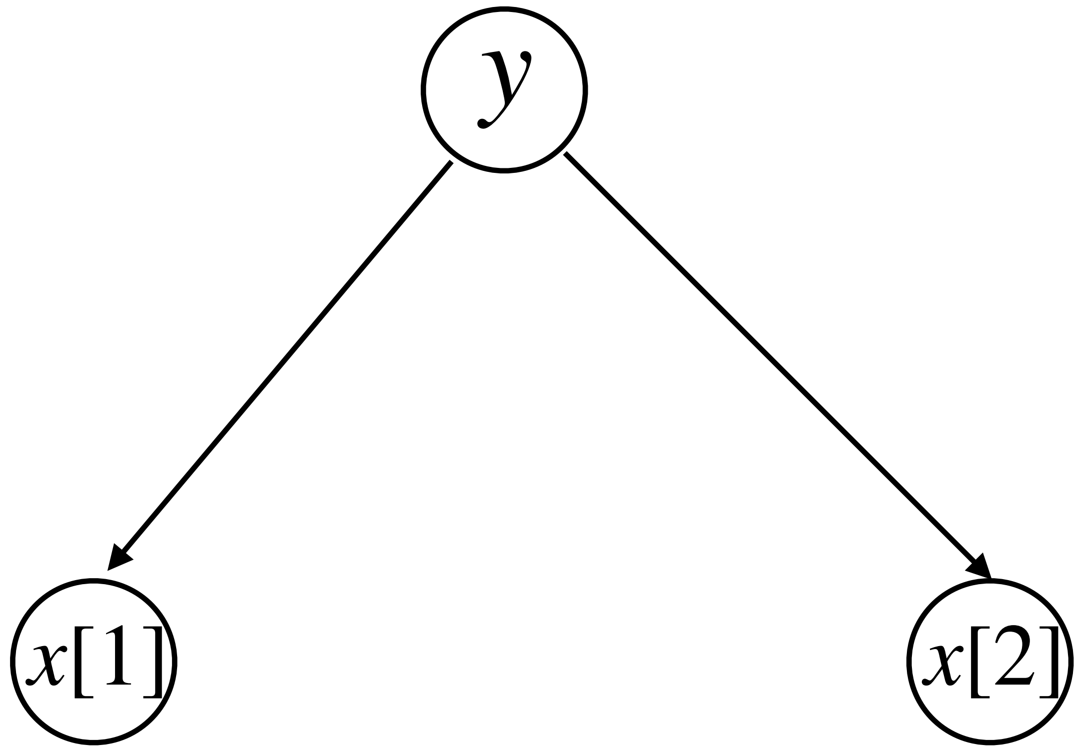
The Naive Bayes assumption:

$$P(x \mid y) = \prod_{\alpha=1}^{d} P(x[\alpha] \mid y)$$

Conditioned on label y, feature values are **independent**!

Q: does conditional independence imply global independence?

Age of a child



Height

# of words the child knowns

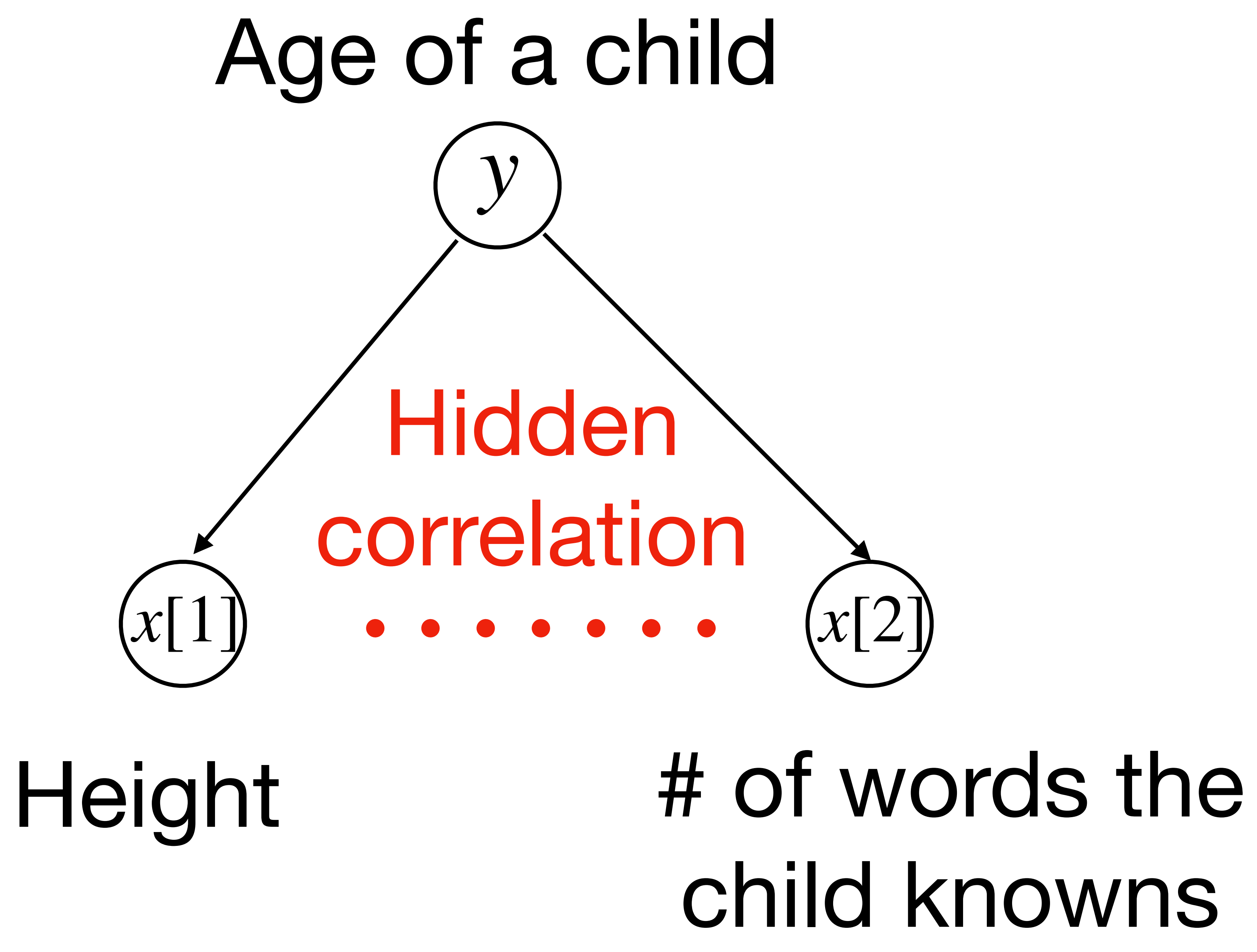
About the independence assumption

The Naive Bayes assumption:

$$P(x \mid y) = \prod_{\alpha=1}^{d} P(x[\alpha] \mid y)$$

Conditioned on label y, feature values are **independent**!

Q: why it is also a naive assumption?



Naive Bayes

Estimate $P(x \mid y)$ from data:

W/ the NB assumption $P(x \mid y) = \prod_{\alpha=1}^{d} P(x[\alpha] \mid y)$

Now we can estimate $P(x[\alpha] \mid y)$ for each α

1-dim problem!

Naive Bayes

Once estimated $P(y)$ and $P(x \mid y)$, we can make prediction:

$$P(y \mid x) \propto P(x \mid y)P(y)$$

In test time, given x:

$$\hat{y} = \arg\max_y P(y \mid x)$$

$$= \arg\max_y P(x \mid y)P(y) = \arg\max_y \left(\prod_{\alpha=1}^{d} P(x[\alpha] \mid y)\right)P(y)$$

Outline

1. General formulation of Naive Bayes ✓

2. Example

3. Connection to linear classifier

Case study

Continuous features

$x[\alpha] \in \mathbb{R}$, for all $\alpha \in \{1,2,\ldots d\}$

We model each $P(x[\alpha] \mid y)$ using a 1-dim Gaussian distribution:

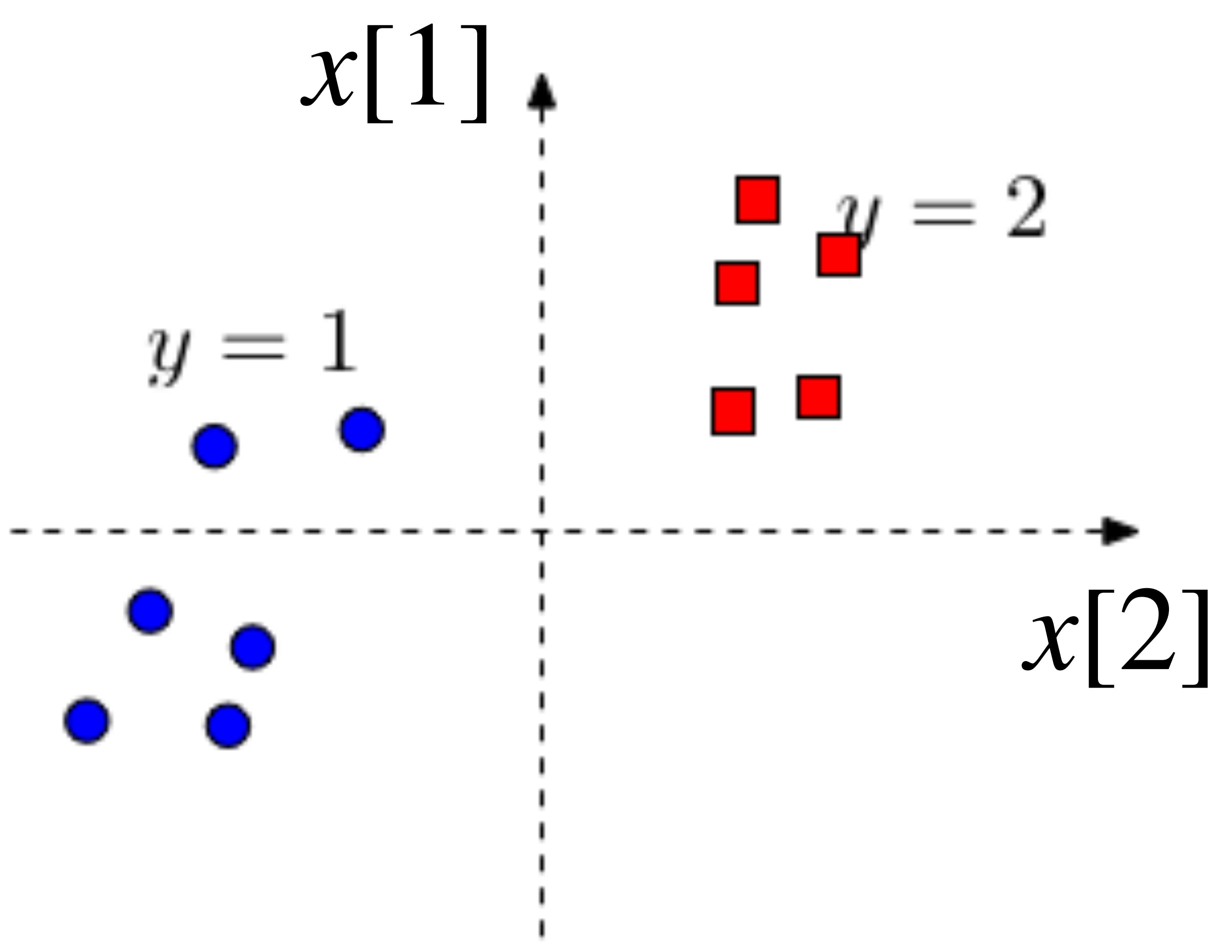
$$P(x[\alpha] \mid y) = \mathcal{N}(\mu_{\alpha,y} , \sigma^2_{\alpha,y})$$

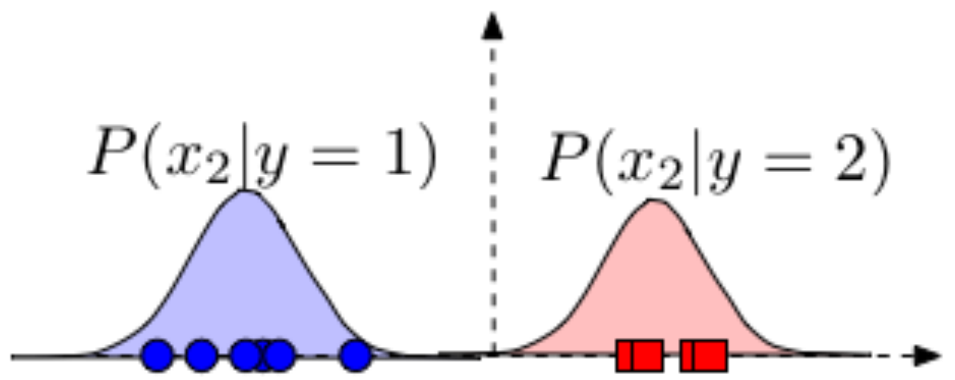
Case study

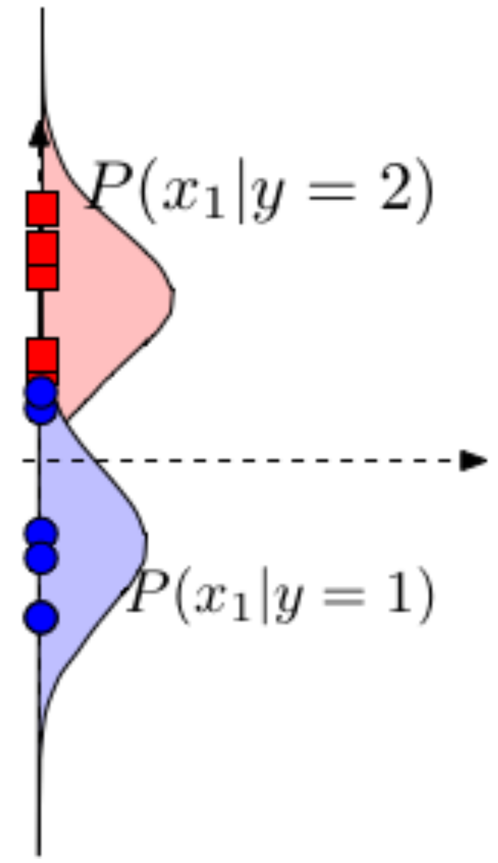
Estimate the mean/std parameter $\mu_{\alpha,y}, \sigma_{\alpha,y}$:

$$P(x[\alpha]\,|\,y) = \mathcal{N}(\mu_{\alpha,y}\,, \sigma^2_{\alpha,y})$$

Original data



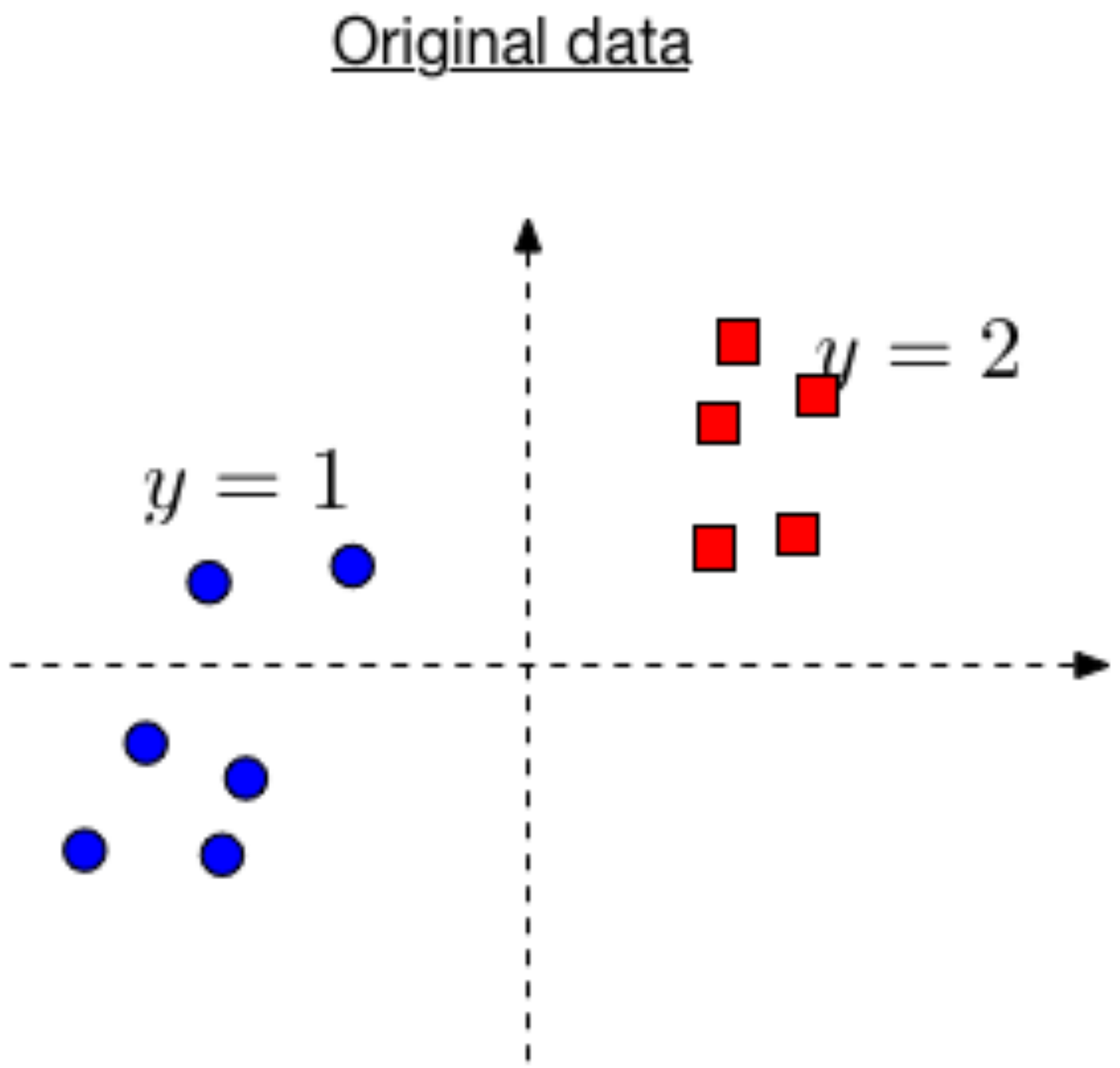


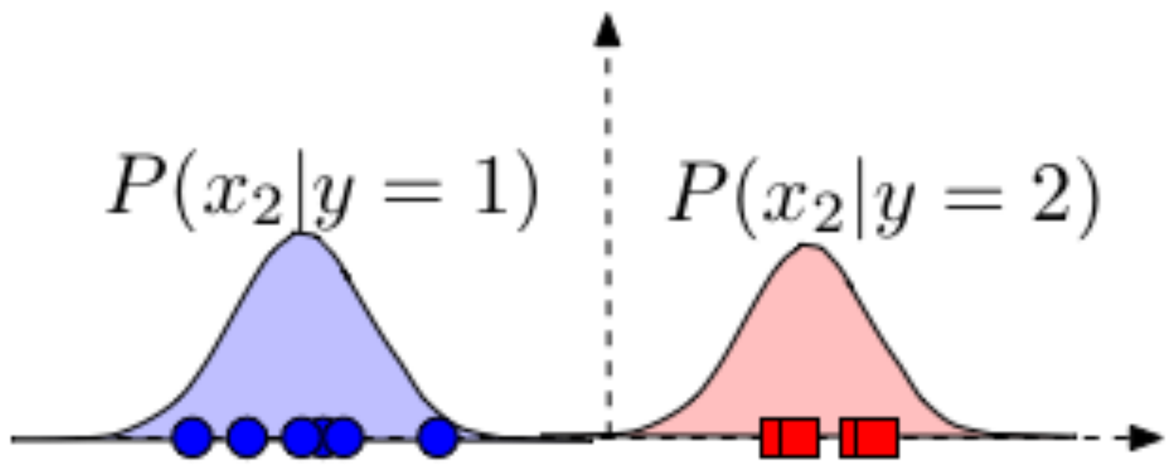


Case study

Estimate the mean/std parameter $\mu_{\alpha,y}, \sigma_{\alpha,y}$ via **MLE**:

$$P(x[\alpha] \,|\, y) = \mathcal{N}(\mu_{\alpha,y}\,, \sigma^2_{\sigma,\alpha})$$





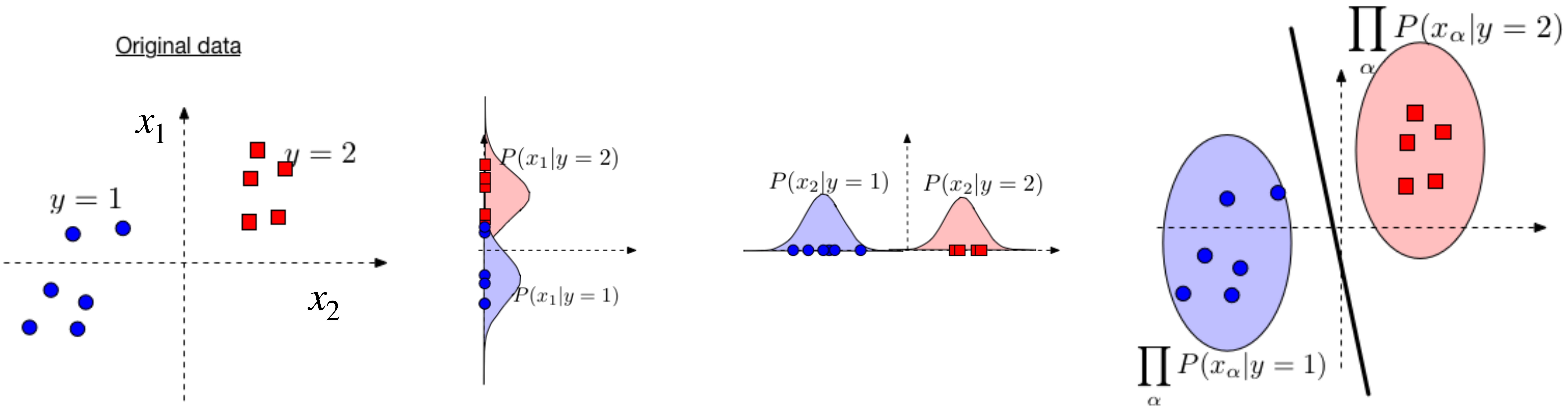
$$\mu_{\alpha,y} = \frac{\sum_{i=1}^{n} x_i[\alpha]\mathbf{1}(y_i = y)}{\sum_{i=1}^{n} \mathbf{1}(y_i = y)}$$

$$\sigma^2_{\alpha,y} = \frac{\sum_{i=1}^{n} (x_i[\alpha] - \mu_{\alpha,y})^2\mathbf{1}(y_i = y)}{\sum_{i=1}^{n} \mathbf{1}(y_i = y)}$$

Case study

Formulate the joint conditional distribution

$$P(x \mid y) = \prod_{\alpha=1}^{d} P(x[\alpha] \mid y)$$



Outline

1. General formulation of Naive Bayes ✓

2. Example ✓

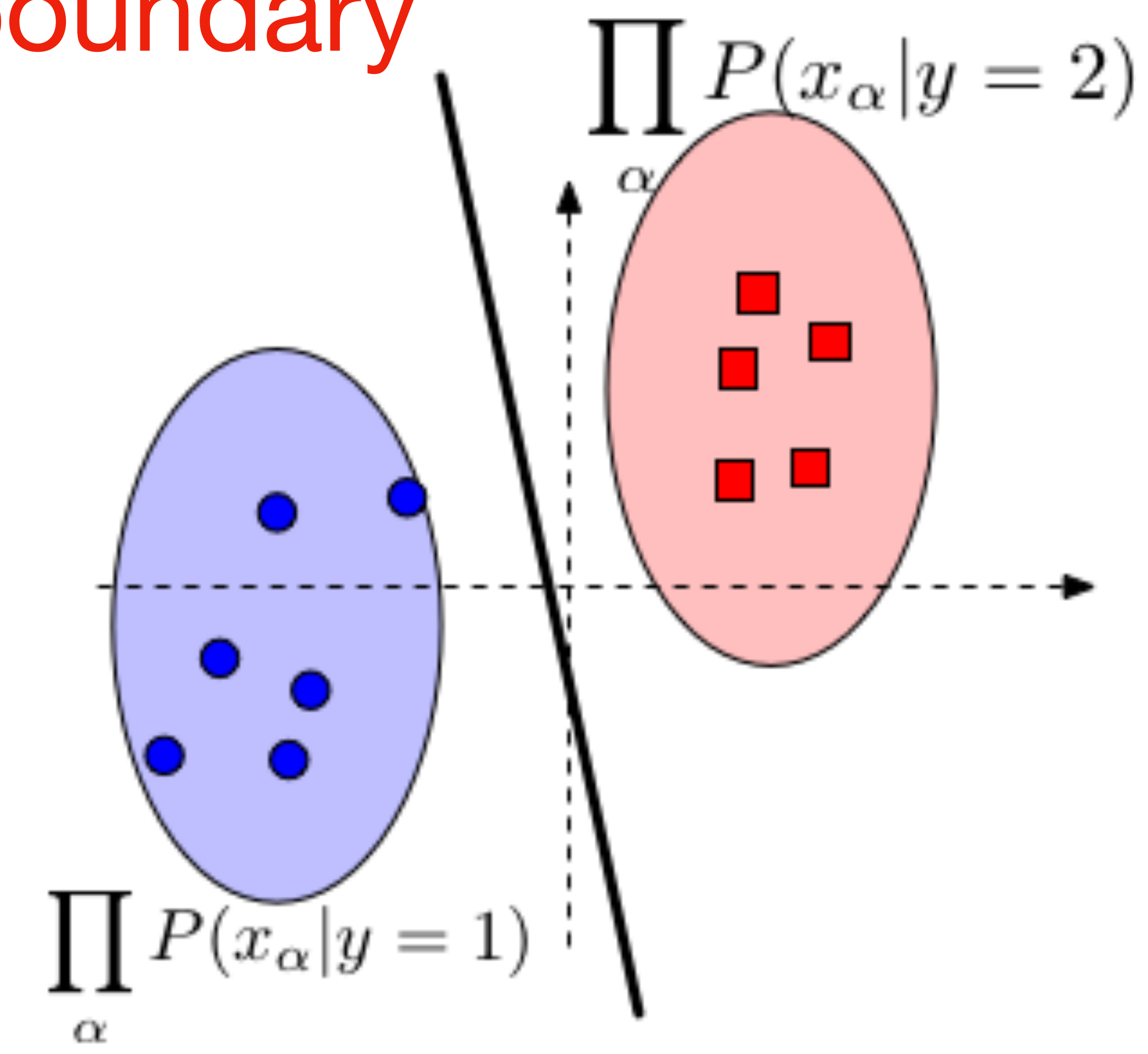
3. Connection to linear classifier

Gaussian Naive Bayes induces a linear classifier

When $P(x[\alpha] \mid y) = \mathcal{N}(\mu_{\alpha,y}, \sigma_\alpha^2)$

i.e., give α, STD σ_α is the same across all labels,

Linear decision boundary



$\exists w, b$ (i.e., a hyperplane), such that:

$$\arg\max_y P(y \mid x) = 1 \iff w^\top x + b > 0$$

(Try this out in HW3)

Summary for today

We start from Bayes rules:

$$P(y \mid x) \propto P(x \mid y)P(y)$$

The Naive Bayes assumption

$$P(x \mid y) = \prod_{\alpha=1}^{d} P(x[\alpha] \mid y)$$

Estimate each $P(x[\alpha] \mid y)$ via MLE (or MAP)

Easy to estimate via MLE

NB classifier: $\arg\max_y P(y \mid x)$

Take-home Q: Perceptron VS NB classifier