An Brief Introduction to Machine Learning

Zahra Dehghanighobadi

Who are we?

- Artifical Intelligence and Society Group @ RUB
- Research on human-centric and trustworthy AI/ML

Muhammad Bilal Zafar

Zahra Dehghanighobadi

Elisabeth Kirsten

Defining Machine Learning

A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

[Mitchell]

Defining Machine Learning

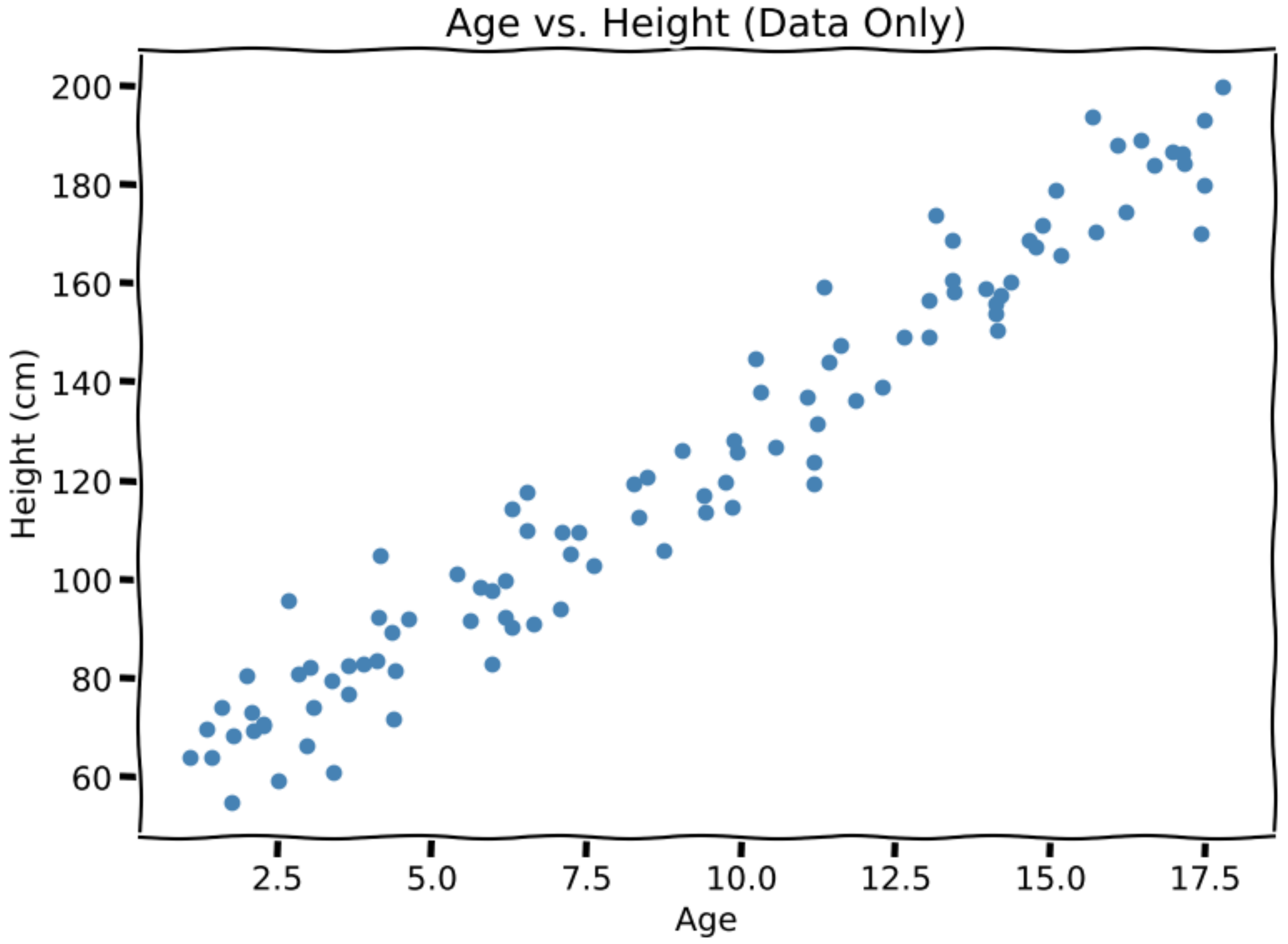
> *A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*
>
> [Mitchell]

Things to work out

- Task T
- Experience E
- Performance measure P

Task: Regression

- Task T: Given $\mathbf{x} \in \mathbb{R}^d$, predict $\hat{y} \in \mathbb{R}$
  - $\mathbf{x}$: Age of a child, y: Their height
  - $\mathbf{x}$: (time of day, month of year), y: Temperature



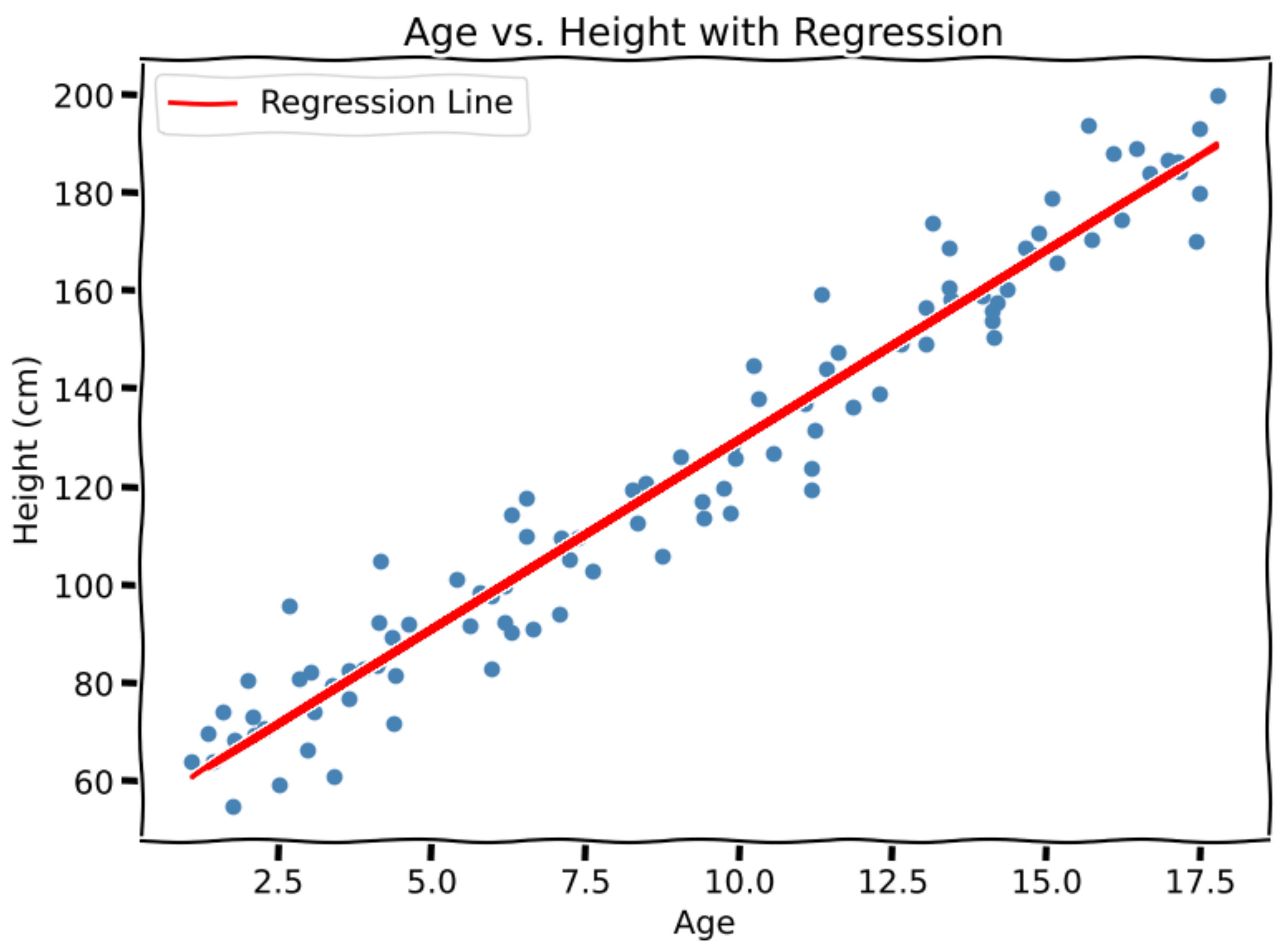
- Experience E: $(\mathbf{x}, y)$ pairs
  - Called **training data**

- Performance measure P
  - Mean squared error: $(y - \hat{y})^2$

* Fictitious dataset

Learning Linear Regression

- **Goal:** Learn $y = f(\mathbf{x}) \in \mathbb{R}$



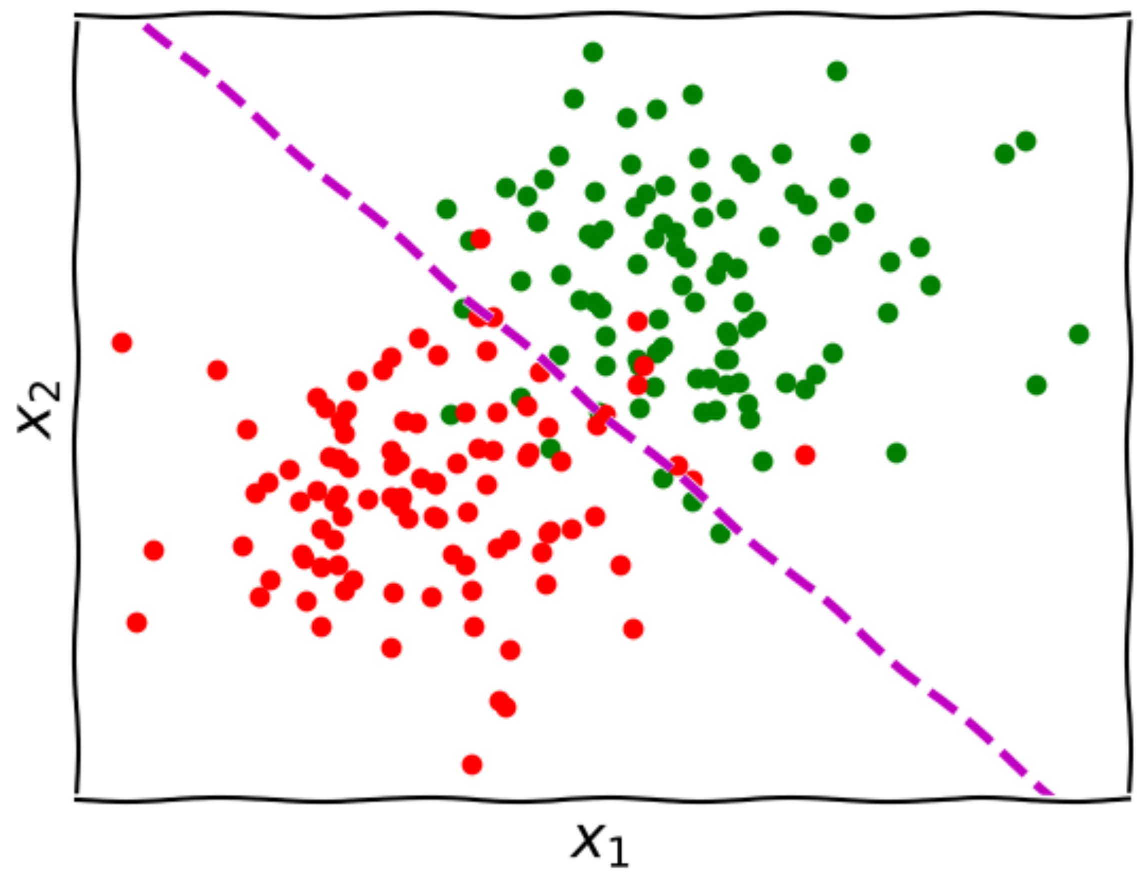
Steps

1. Specify the form of the function $wx + b$
2. Identify the parameters that you want to learn w, b
3. Compute the loss with the parameters $\left(y - (wx + b)\right)^2$
4. Change parameters such that the loss decreases

Next: Learning to classify

Regression to Classification

- **Goal:** Learn $y = f(\mathbf{x}) \in [0,1,\ldots,K-1]$

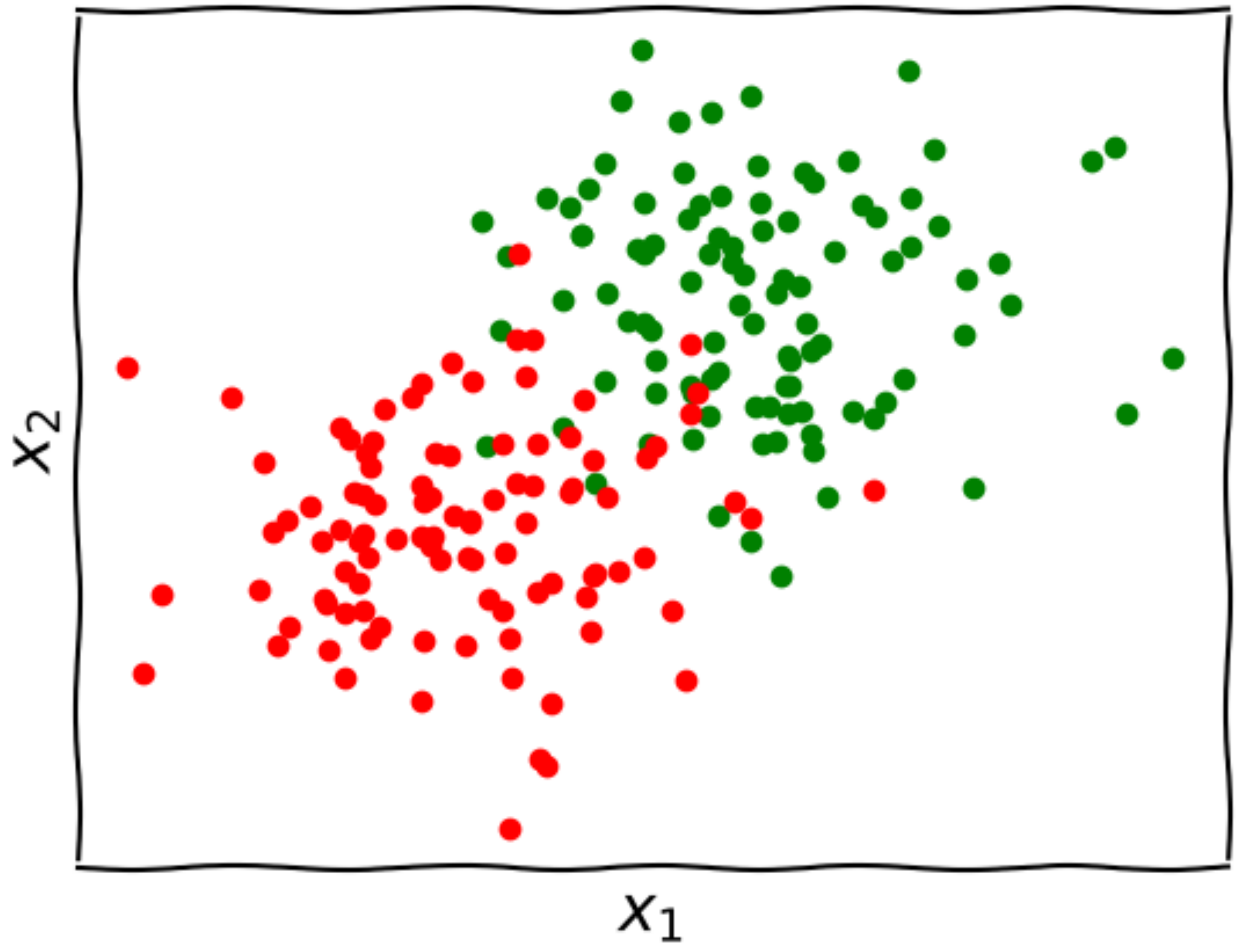


Steps

1. Specify the form of the function — $wx + b = 0$ specifies the decision boundary
2. Identify the parameters that you want to learn — w, b
3. Compute the loss with the parameters — Approximate 0-1 loss. Depends on the model
4. Change parameters such that the loss decreases

Task: Classification

- Task T: Given $\mathbf{x} \in \mathbb{R}^d$, predict $\hat{y} \in \{0,1,2,...\}$
  - $\mathbf{x}$: Income, education, y: Creditworthiness
  - $\mathbf{x}$: Image pixels, y: Cat or dog

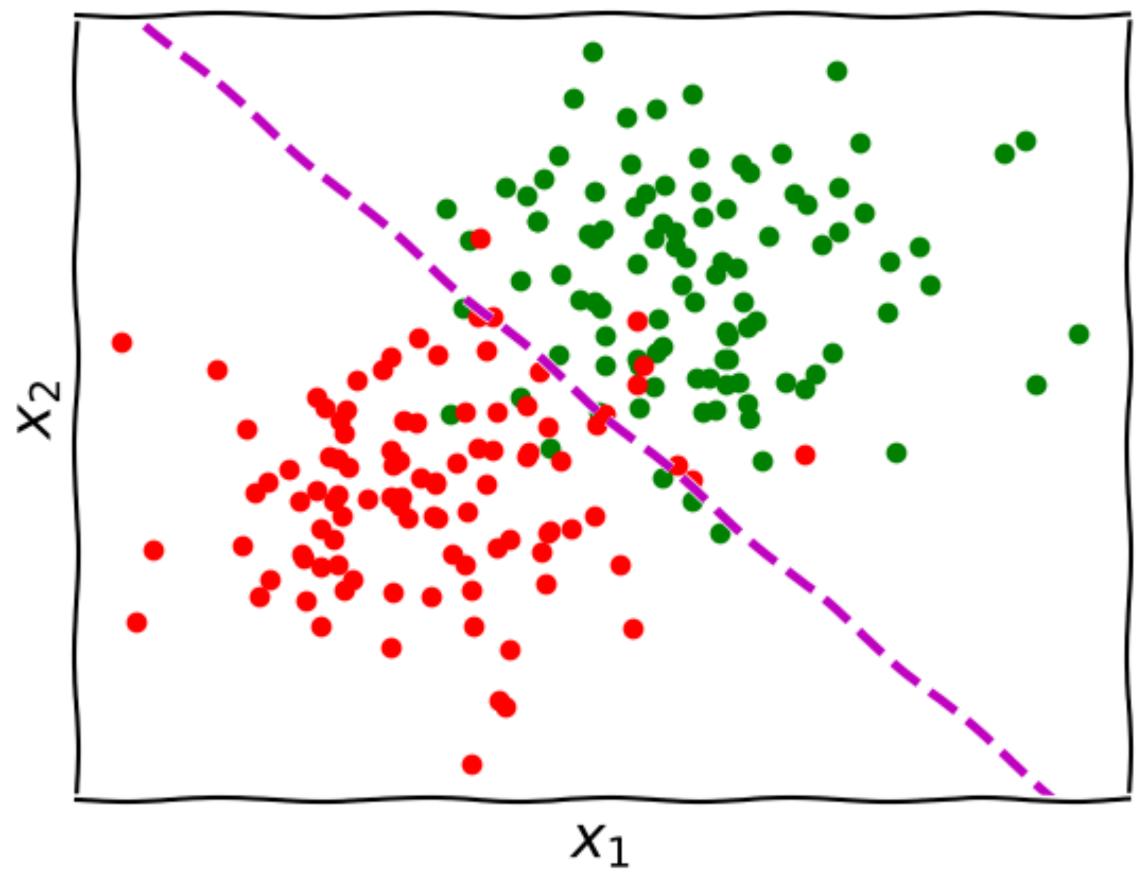


- Experience E: $(\mathbf{x}, y)$ pairs
  - Called **training data**

- Performance measure P
  - Binary loss: $y \neq \hat{y}$

The Logistic Regression Model

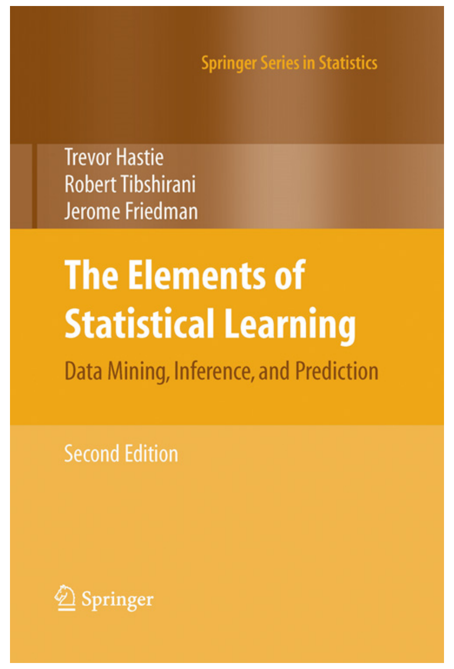
- Simplest form with two classes $y \in [0,1]$
- Distance from boundary: $d(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$
- $p(y = 1 \,|\, \mathbf{x}) = \dfrac{1}{1 - exp(-d(\mathbf{x}))}$
- $p(y = 0 \,|\, \mathbf{x}) = 1 - p(y = 1 \,|\, \mathbf{x})$
- Goal is to maximize negative log likelihood: $y \log p(y = 1 \,|\, \mathbf{x}) + (1 - y) \log p(y = 0 \,|\, \mathbf{x})$



Detailed derivation and K>1 case in class

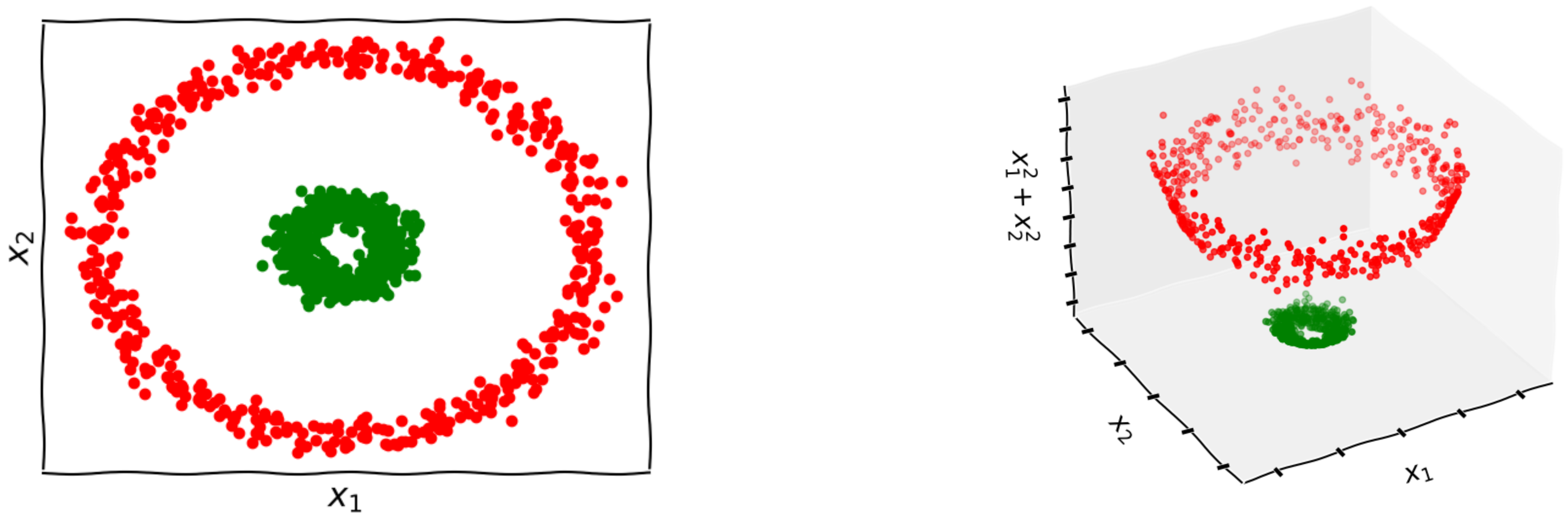
+

in ESLII Chapter 4.4

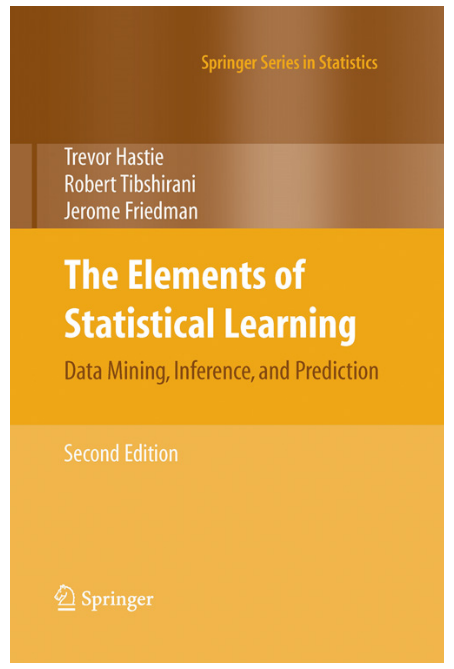


The case of non-linearly separable data

- Similar solution to regression
- Project the data to higher dimensional space
- Models like Kernel SVM

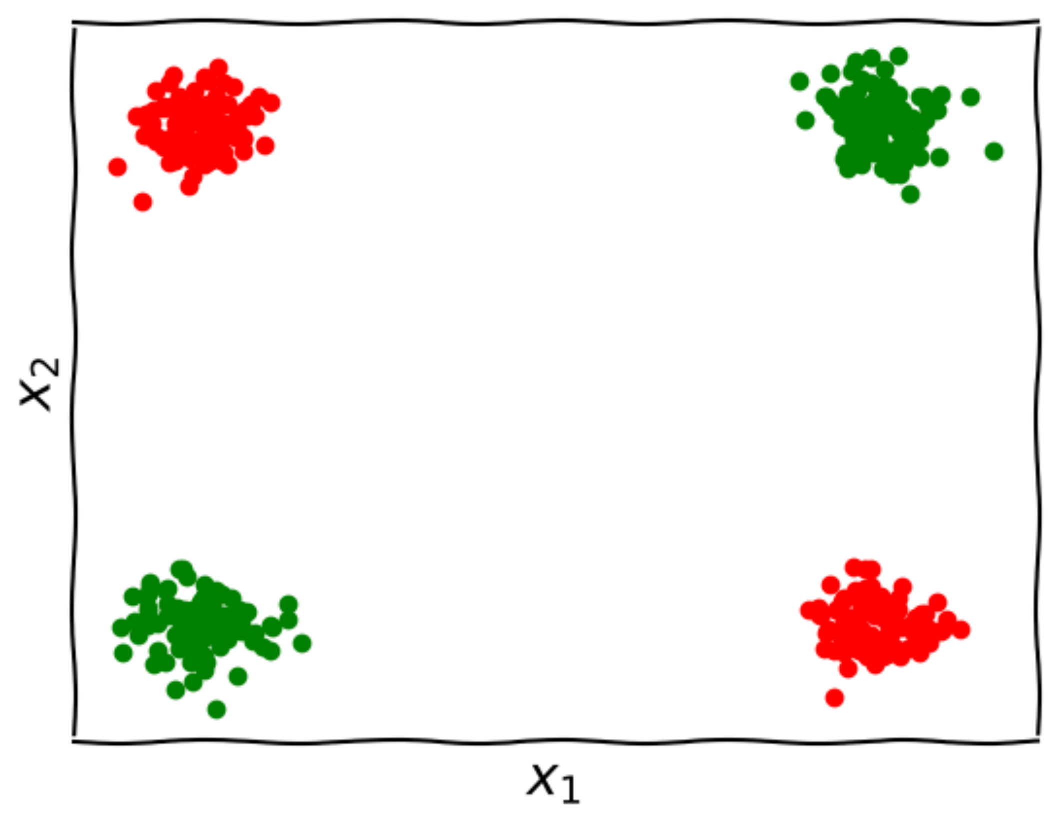


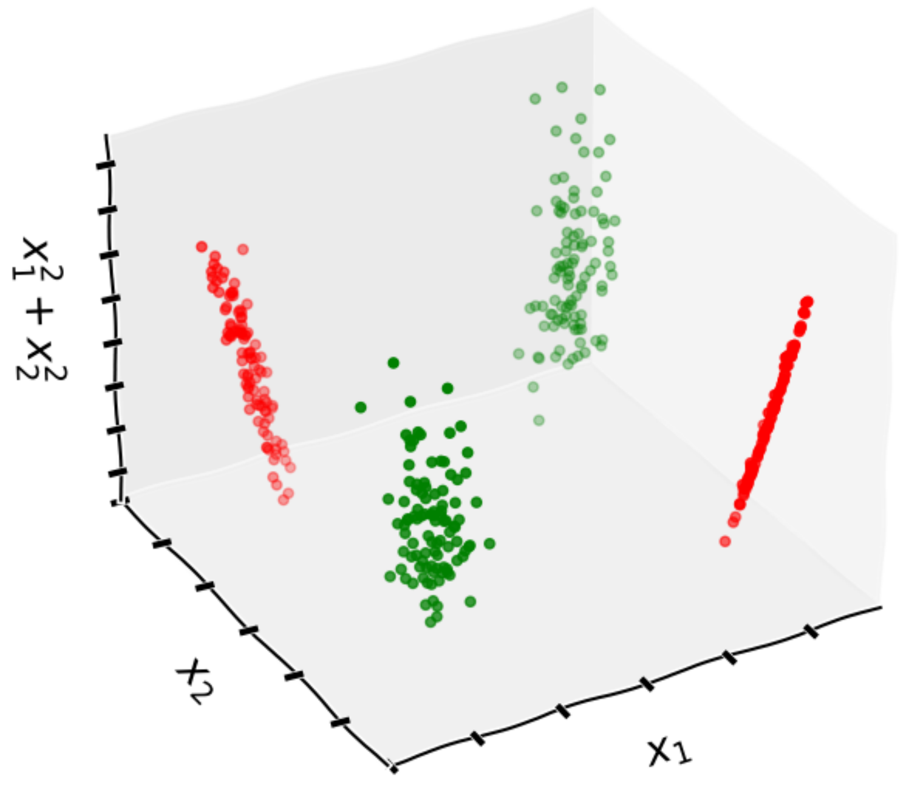
Kernel SVM description in <u>ESLII</u> Chapter 12



Datasets can get arbitrarily complex

- Higher dimensional projections may not work as expected

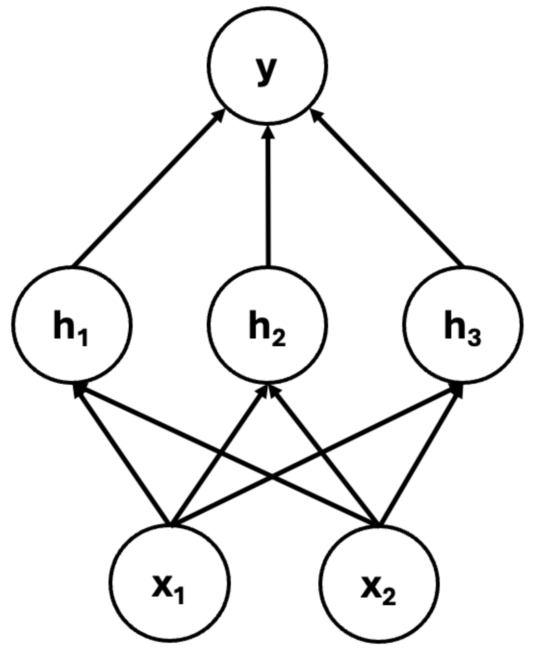


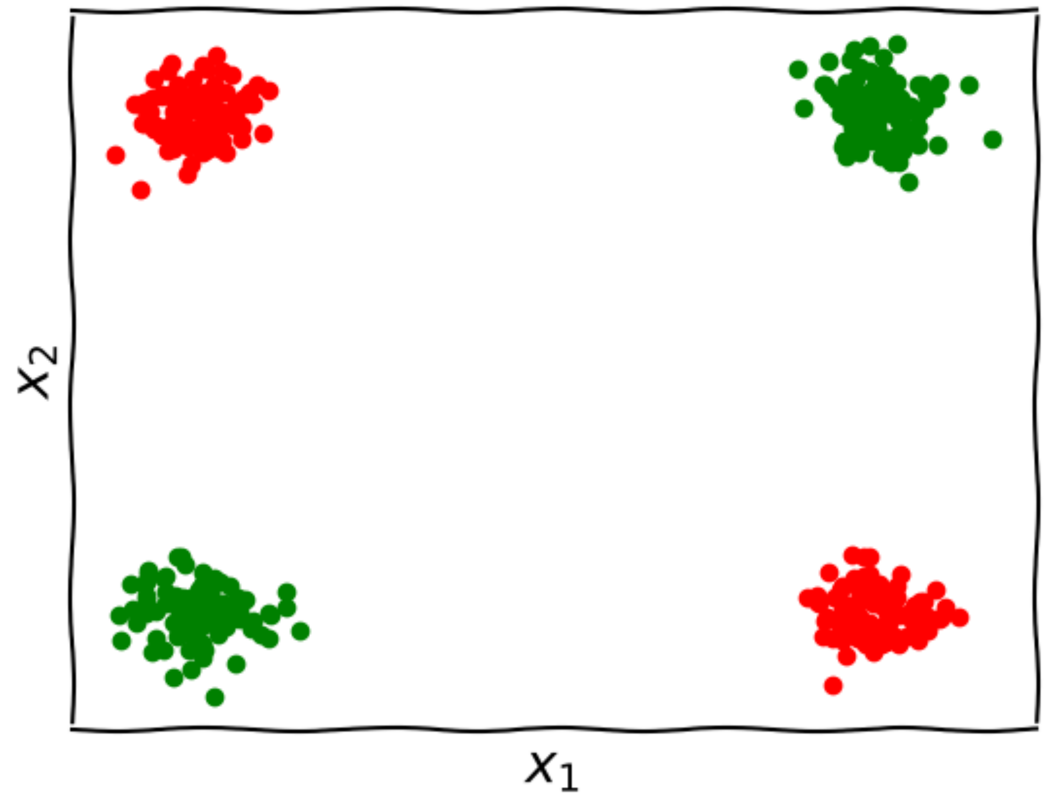


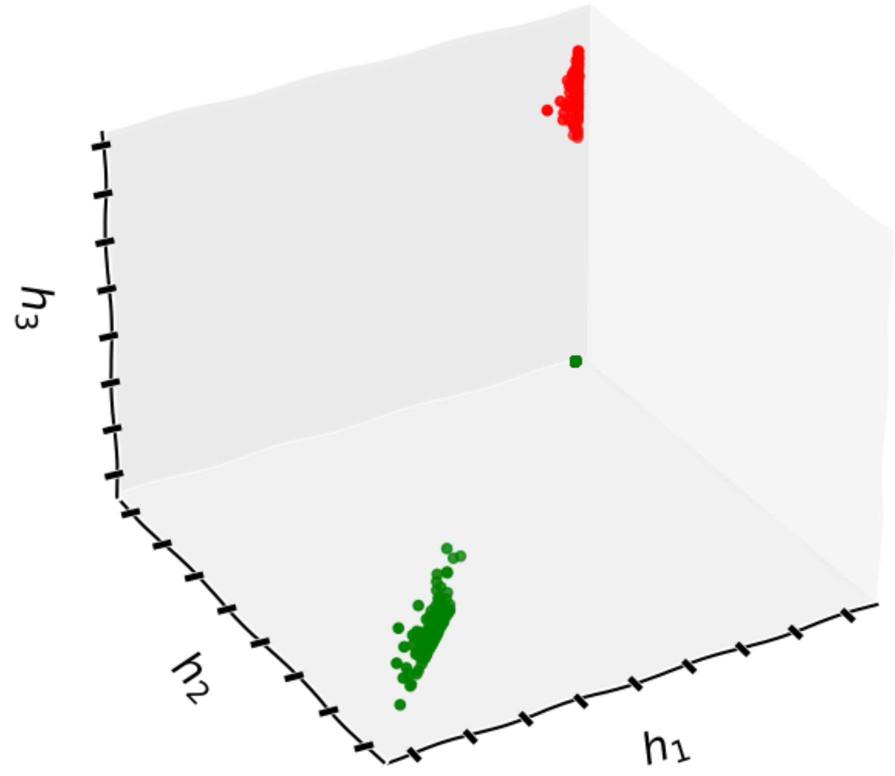
What about neural networks

Neural Networks for Classification

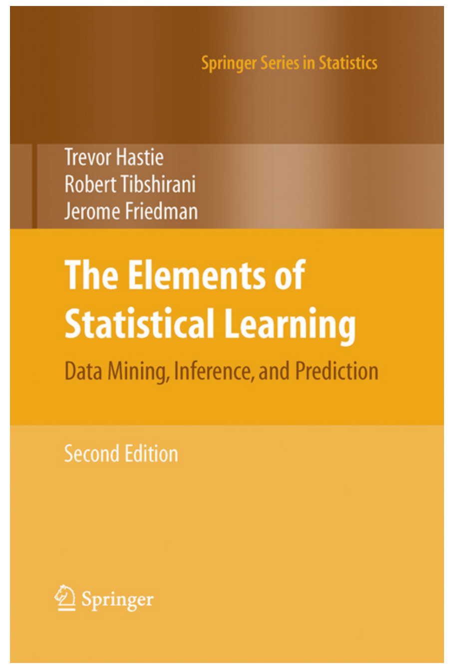
- Same as regression (last lectures)
- Project data repeatedly via hidden layers







Details in [ESLII](#) Chapter 11



Measuring performance

- 0/1 accuracy: Fraction of points that are correct
- Balanced accuracy
  - Assume two classes, class 0 comprising of 99% of the data and class 1 1%
  - A constant classifier always predicting class 0 gets 99% accuracy!
  - Balanced accuracy gives equal weight to each class
- Confusion matrices

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Same concept for multi-class accuracies & confusion matrices