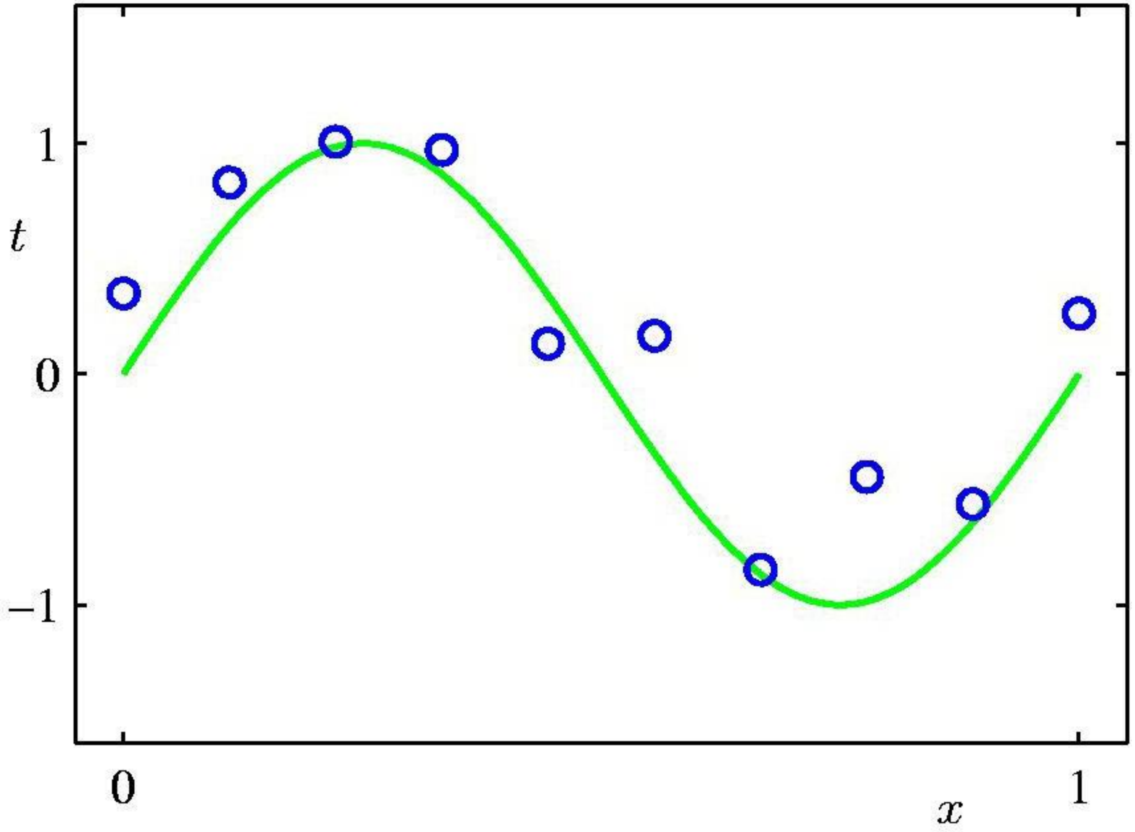
Learning from Examples (Chapter 19)

[Discussions on regression and classification]

Learning from Examples

- Linear Regression
- Feature Extraction
- Classification
- Discriminant functions
- Performance Measures

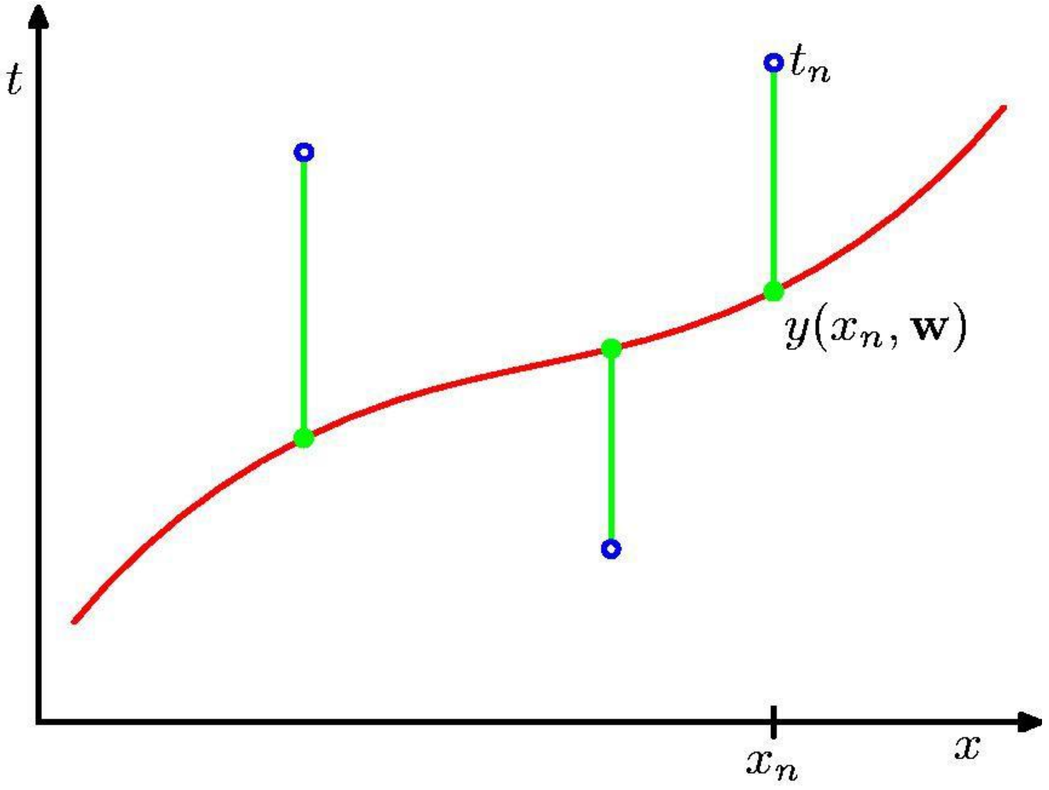
Polynomial curve fitting



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Machine Learning and Pattern Recognition (C. Bishop 2006)

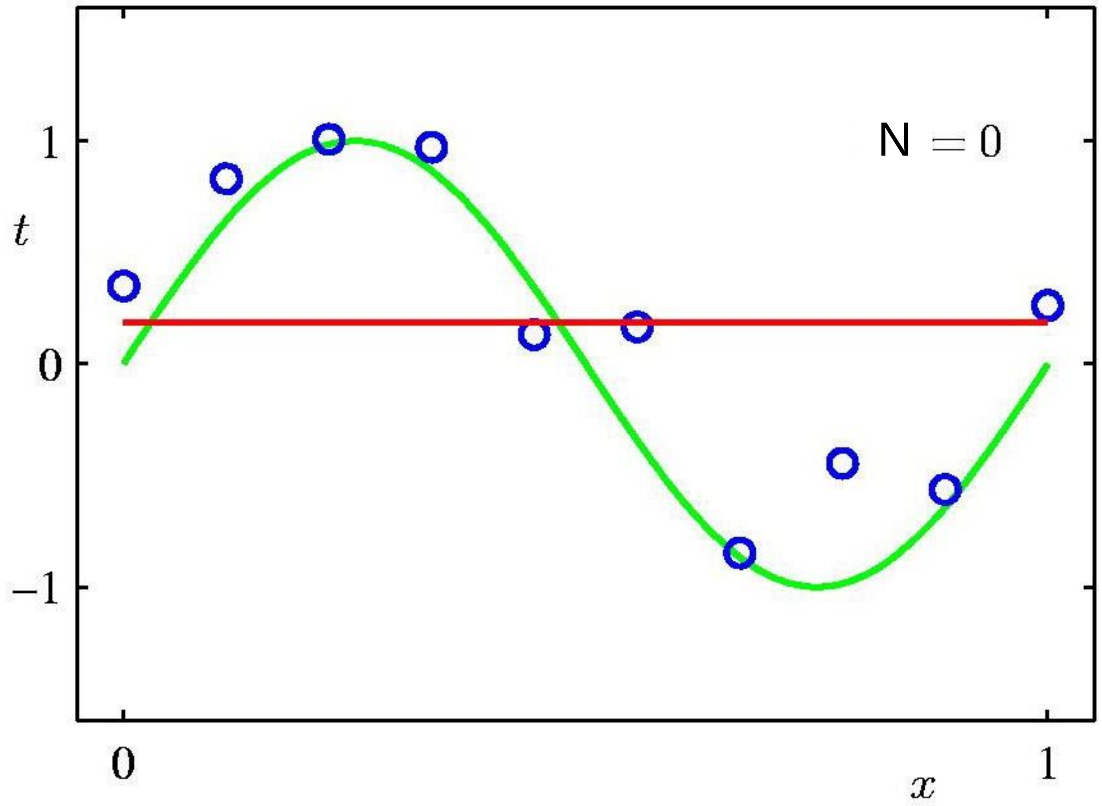
Sum of Squares Error Function



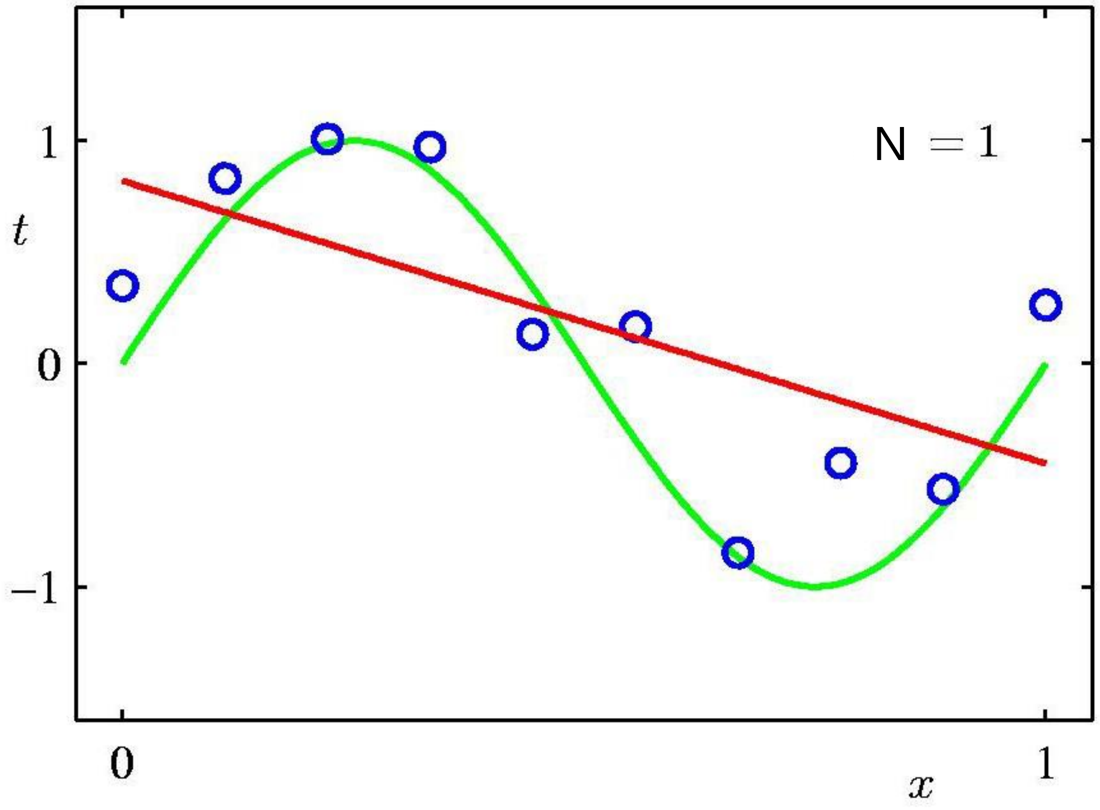
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Machine Learning and Pattern Recognition (C. Bishop 2006)

0th Order

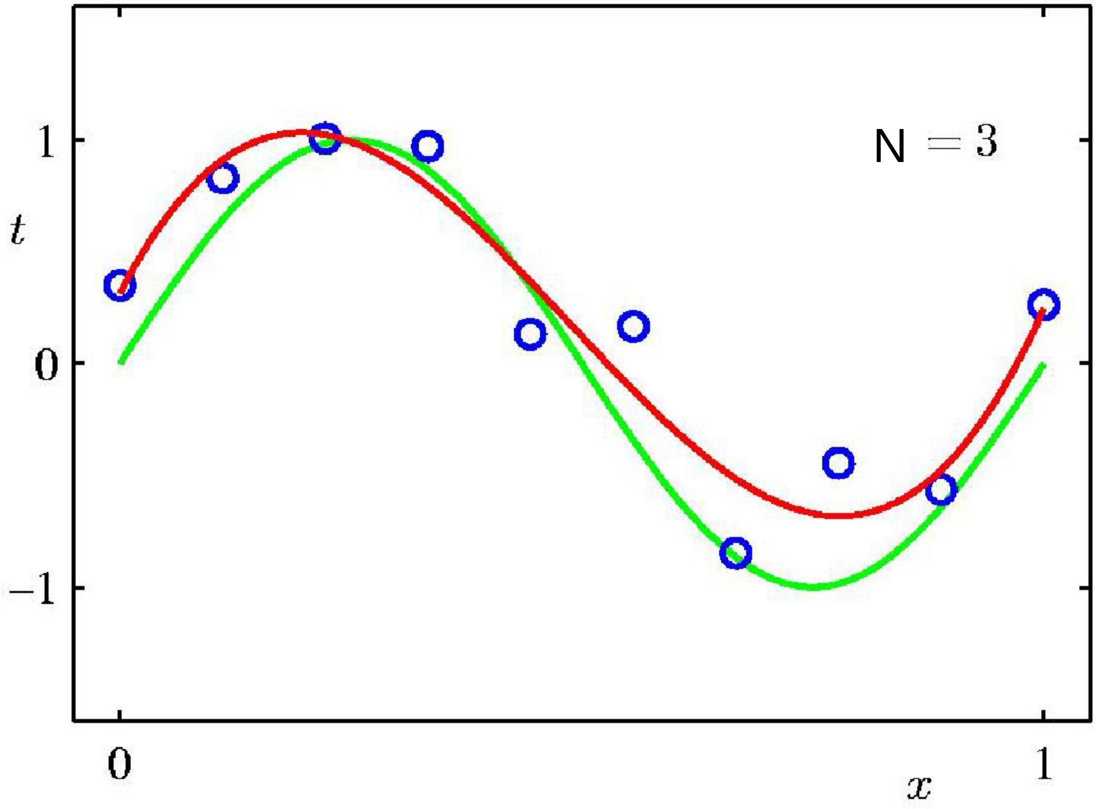


1^{st} Order



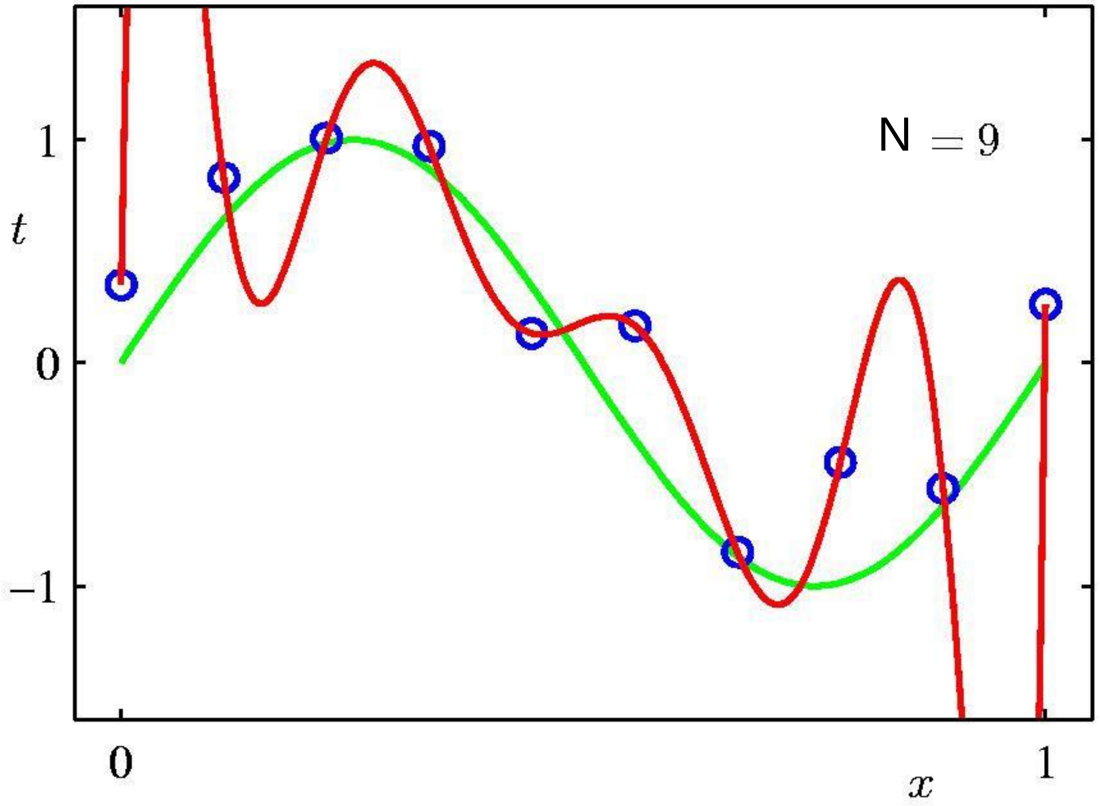
Machine Learning and Pattern Recognition (C. Bishop 2006)

3rd Order



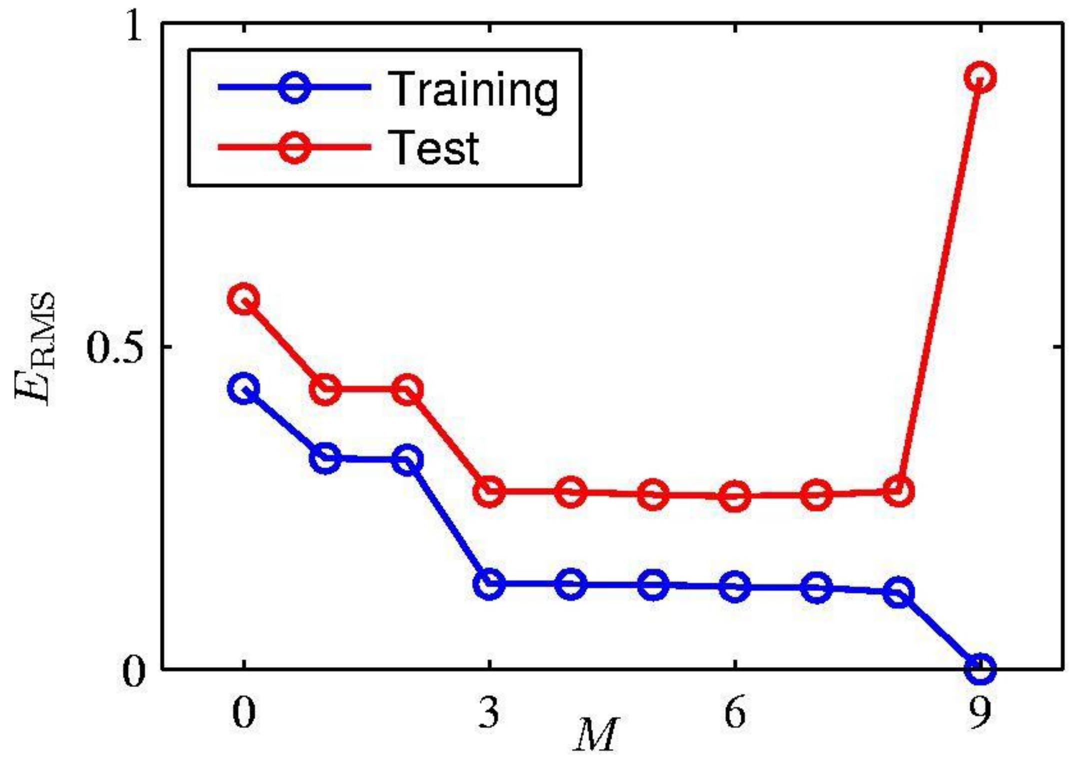
Machine Learning and Pattern Recognition (C. Bishop 2006)

9th Order



Machine Learning and Pattern Recognition (C. Bishop 2006)

Over-fitting

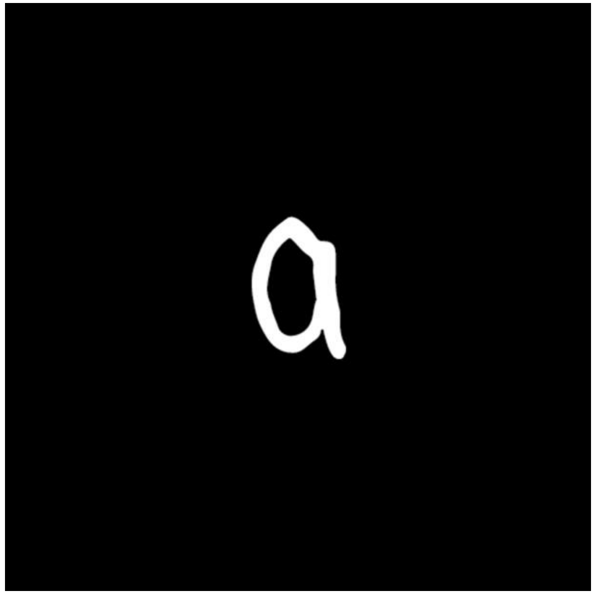
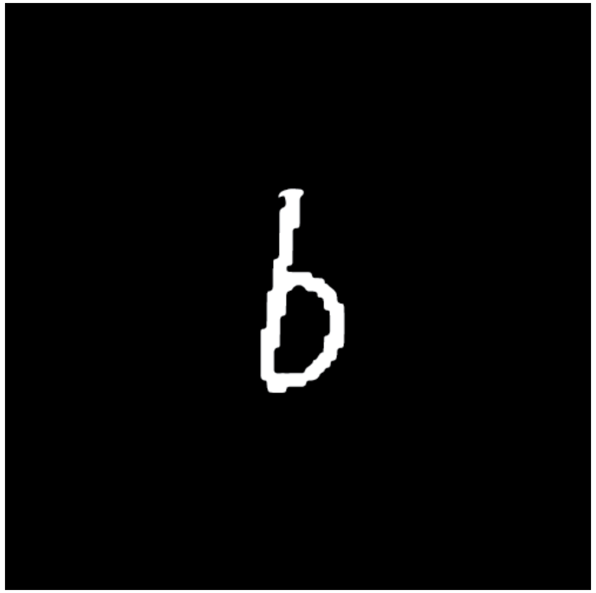


Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Machine Learning and Pattern Recognition (C. Bishop 2006)

Character recognition example

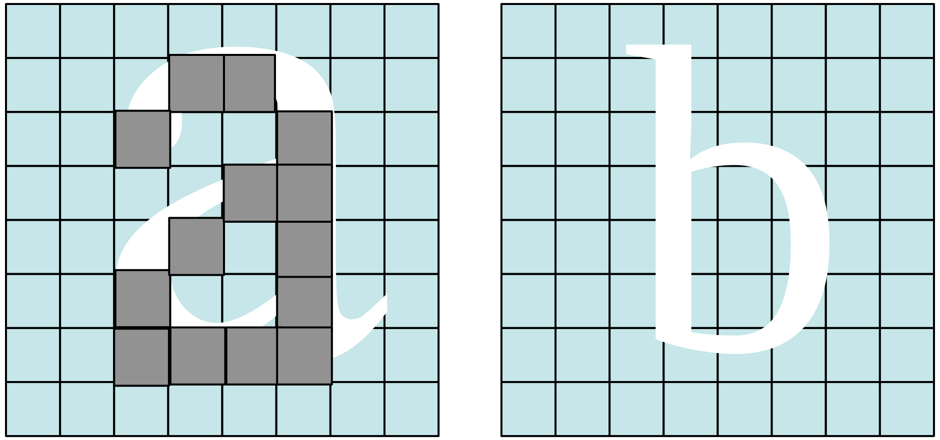
- We wish to distinguish the hand written characters "a" and "b" as reliably as possible.

Hand written letters by Emily Mahar
[https://emilymahar.com/The-Handwritten-A]

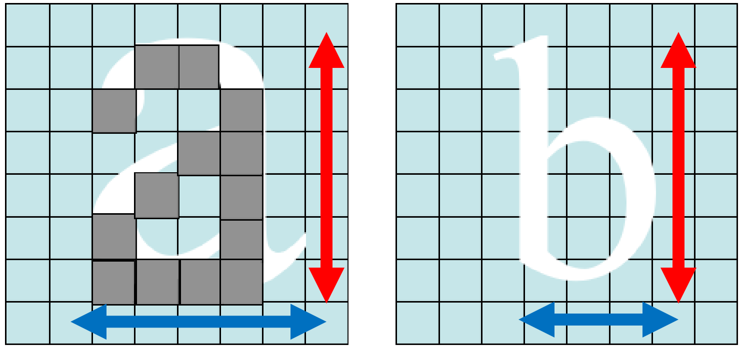
Character recognition example

- Images are captured by a scanner or camera and fed into the computer.
- Each character is represented by an array of pixels, each of a value x_i where i labels the individual pixels. x can be 0 for white and 1 for black.
- Our goal is to develop an algorithm that will assign an image to one of the classes.



Feature extraction

- We may only have a few thousand examples in our training set. The classifier must therefore be able to correctly classify an unseen image vector. This property is often referred to as generalization.
- A large number of input variables can create some problems for pattern recognition systems.
- Input variables can be combined together to make a smaller number of new variables called ***features (attributes)***
- An example for our problem might be to use the ratio of the height of the characters to the width.



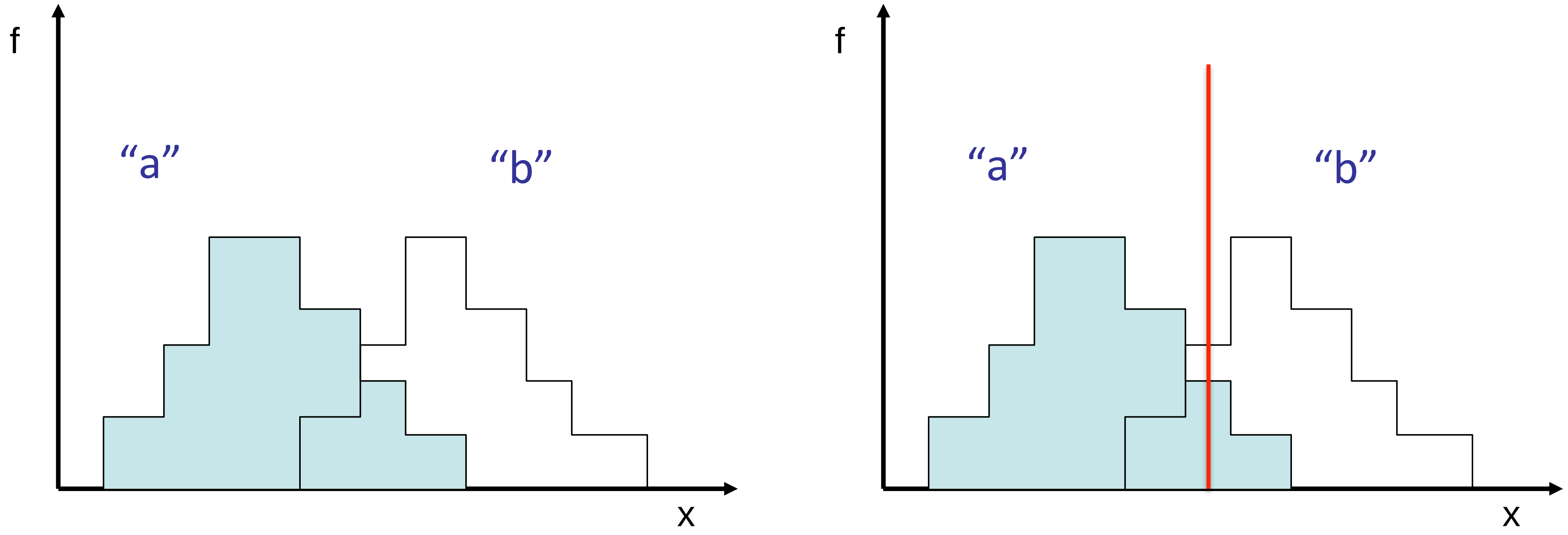
Feature extraction

- No classification can succeed if features are poorly chosen.
  - Can attributes/features be chosen that give unique descriptions of the patterns?
  - How easy are the attributes to measure?

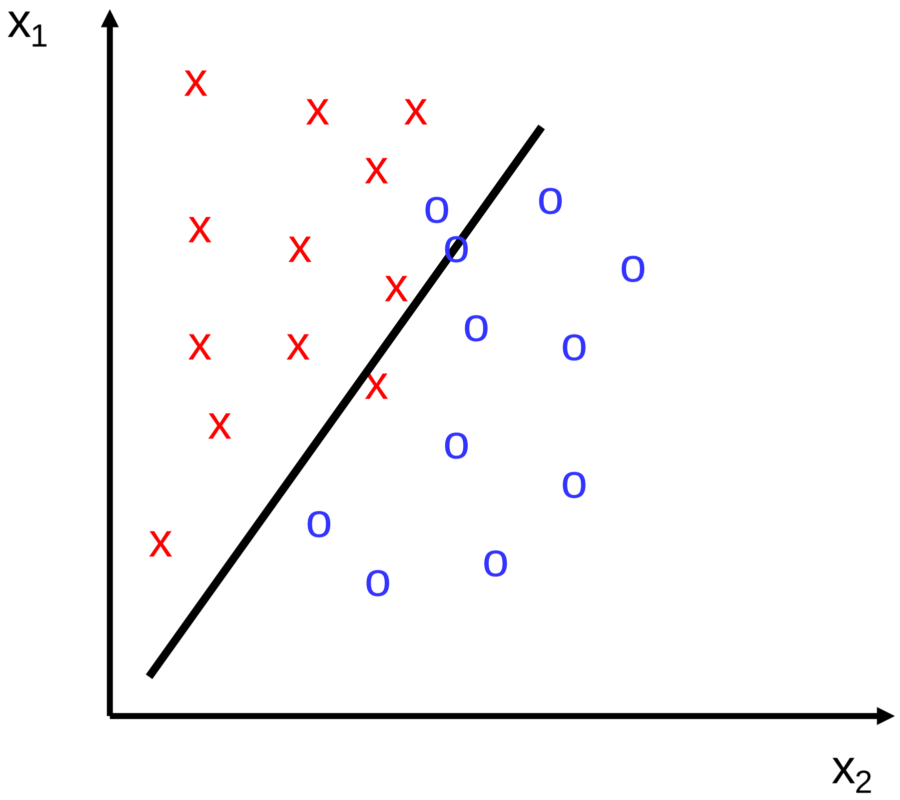
THE HAND-WRITTEN “A” by Emily Mahar
[https://emilymahar.com/The-Handwritten-A]

Classification based on a single feature

- Typically examples are higher for “b”rather than “a”, but there is overlap and therefore it is not a perfect classifier.



Multiple Features



- Further features can be added but eventually this could make things worse.
- It is inevitable that some overlap will exist highlighting the intrinsically probabilistic nature of pattern classification problems.

Have a look here to experiment with different decision boundaries: https://playground.tensorflow.org/

Feature Vectors

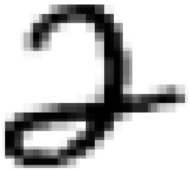
x | $f(x)$ | y

```
Hello,

Do you want free printr
cartriges?  Why pay more
when you can get them
ABSOLUTELY FREE!  Just
```

```
# free       : 2
YOUR_NAME    : 0
MISSPELLED   : 2
FROM_FRIEND  : 0
...
```

SPAM
or
+

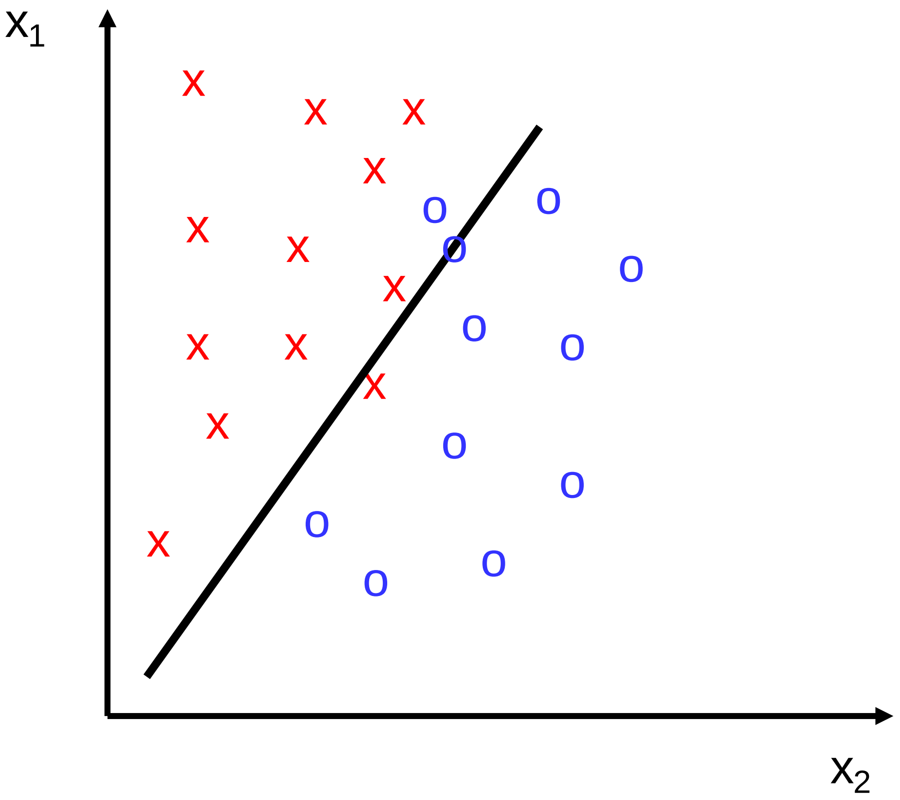


```
PIXEL-7,12   : 1
PIXEL-7,13   : 0
...
NUM_LOOPS    : 1
...
```

“2”

Discriminant Analysis

- In some cases we are able to separate clusters of data with some function



$$f(x) = \sum w_i x_i + \theta$$

f(x) < 0 for one cluster &

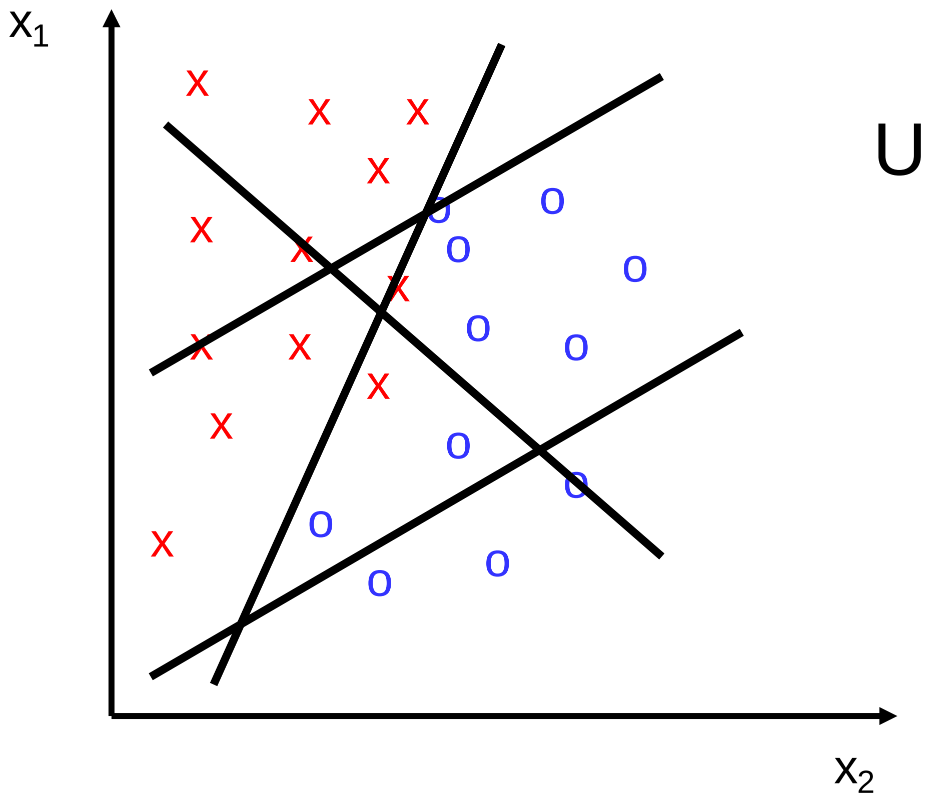
f(x) > 0 for another

f(x) is known as a *discriminant function*

Determining Discriminant Functions

- Discriminant functions are often determined through error minimisation
- Given a training set and desired output values. The error is measured as the difference between the actual and desired output
- Minimising the error to zero produces the best decision function.

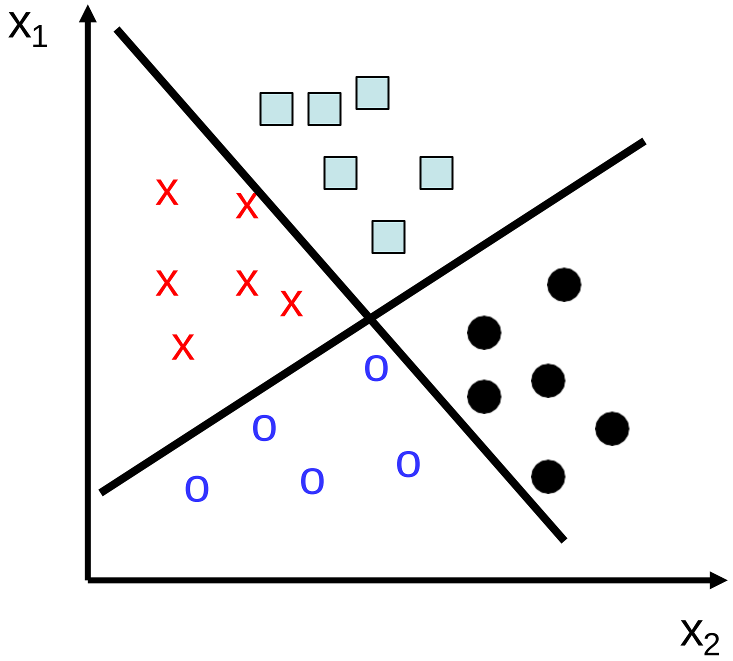
Determining Discriminant functions



Use error minimisation to adapt the weights

Multiple Discriminators

- You can have multiple discriminant functions
- Enables you to classify into multiple classes



$f_1(x) < 0$ and $f_2(x) < 0$ for one cluster

$f_1(x) > 0$ and $f_2(x) < 0$ for another

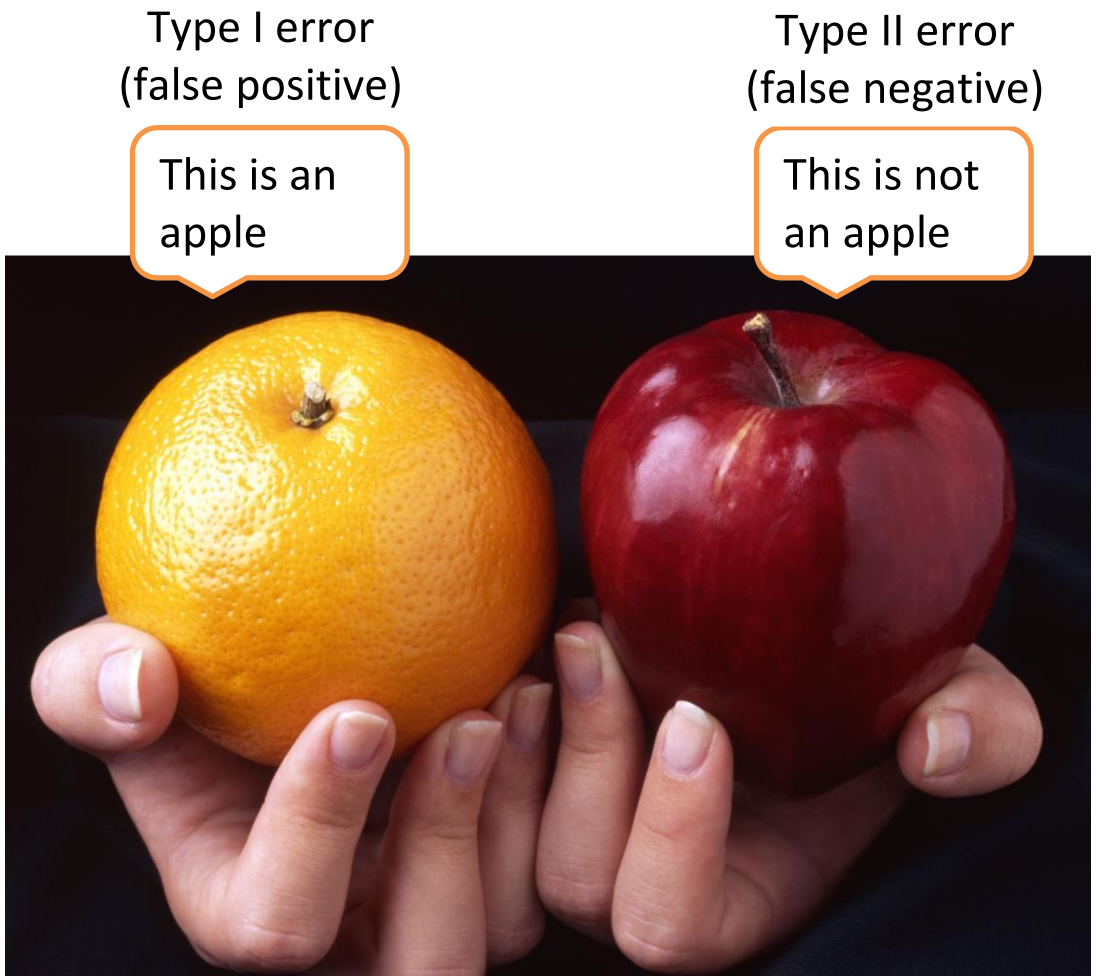
Have a look here to experiment with different machine learning models for a 2D classification task:

https://ml-playground.com/

Performance measures

- **Recall** = of those that exist, how many did you find (Sensitivity)
- **Precision** = of those you found, how many are correct, also known as positive predictive value
- **F-Score** (harmonic mean)

= 2P*R/R+P



Performance Measures for binary classifier

Confusion matrix or contingency table

Expected output

		1	0
Predicted output	1	**TP**	**FP**
	0	**FN**	**TN**

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

False Pos rate (type I error rate) $= \dfrac{FP}{TP + FN}$

False Neg rate (type II error rate) $= \dfrac{FN}{FP + TN}$

$$Specificity = \frac{TN}{FP + TN}$$

Performance Measures for binary classifier

		Predicted Class			
		Positive	**Negative**		
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$	(Recall)
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$	
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$	

From: https://manisha-sirsat.blogspot.com/

Be Careful!

A

Expected output

Predicted output	1	0
1	**0.9**	**0.1**
0	**0**	**0**

B

Expected output

Predicted output	1	0
1	**0.8**	**0.0**
0	**0.1**	**0.1**

C

Expected output

Predicted output	1	0
1	**0.78**	**0**
0	**0.12**	**0.1**

Metric	**A**	**B**	**C**
Accuracy	0.9	0.9	0.88
Precision	0.9	1.0	1.0
Recall	1.0	0.888	0.8667
F-score	0.947	0.941	0.9286

Which Model is Better?

Summary

- Linear Regression
- Error minimisation
- Feature Vectors
- Classification
- Discriminant functions
- Performance measures