The University Of Sheffield.

# Linear Models

Tony Dodd

# Overview

- Linear models.
- Parameter estimation.
- Linear in the parameters.
- Classification.
- The nonlinear bits.

# Linear models

- Linear model has general form

$$\hat{f}(x) = \sum_{i=0}^{m} w_i x_i$$

where $x_i$ is the $i$th component of input $x$ .

- Assume $x_0 = 1$ and therefore $w_0$ is the bias.
- Can represent lines and planes.
- Should ALWAYS try simplest model first!

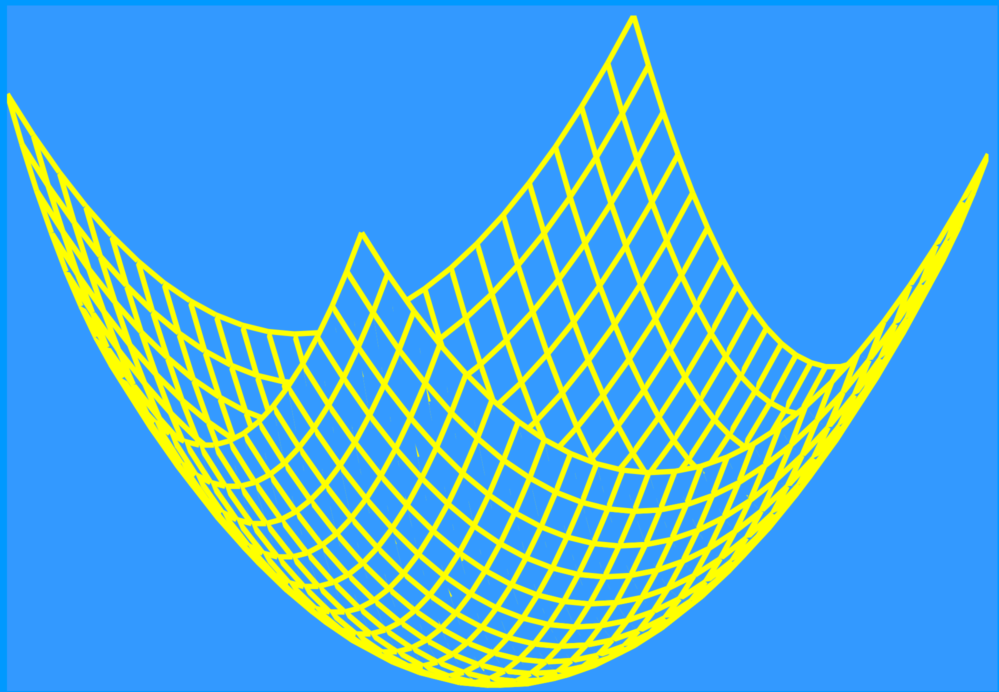
# Parameter estimation

- Least squares estimation.
- Choose parameters that minimise, SSE

$$\sum_{i=1}^{N}\left[\hat{f}(x_i)-z_i\right]^2$$

- Unique minimum…
- Optimum when noise is Gaussian.
- Corresponds to maximum-likelihood estimate.

# Least squares cost function



# Least squares parameters

Define the design matrix

$$\Phi = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,m} \end{bmatrix}$$

Then

$$z = \Phi w + e$$

The University Of Sheffield.

# A bit of maths

$$SSE = \left(z - \Phi w\right)^T \left(z - \Phi w\right)$$

$$= z^T z - z^T \Phi w - w^T \Phi^T z + w^T \Phi^T \Phi w$$

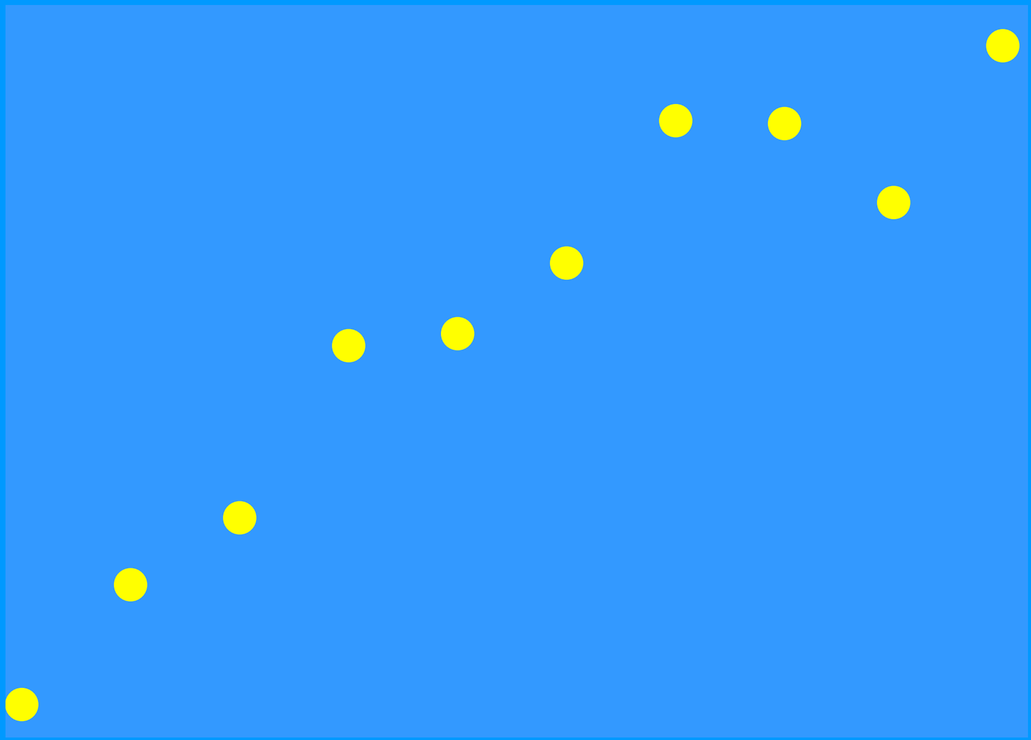
$$= z^T z - 2 z^T \Phi w + w^T \Phi^T \Phi w$$

$$\frac{\partial SSE}{\partial w} = -2 z^T \Phi + 2 w^T \Phi^T \Phi = 0$$

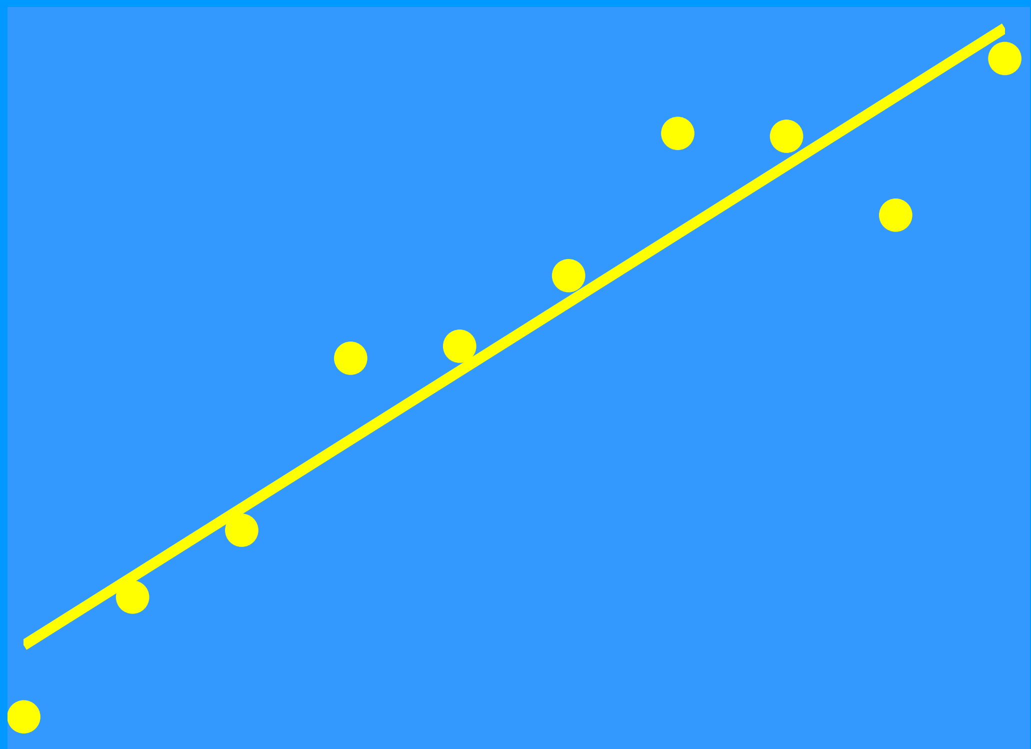
$$w^T \Phi^T \Phi = z^T \Phi$$

$$\hat{w} = \left(\Phi^T \Phi\right)^{-1} \Phi^T z$$

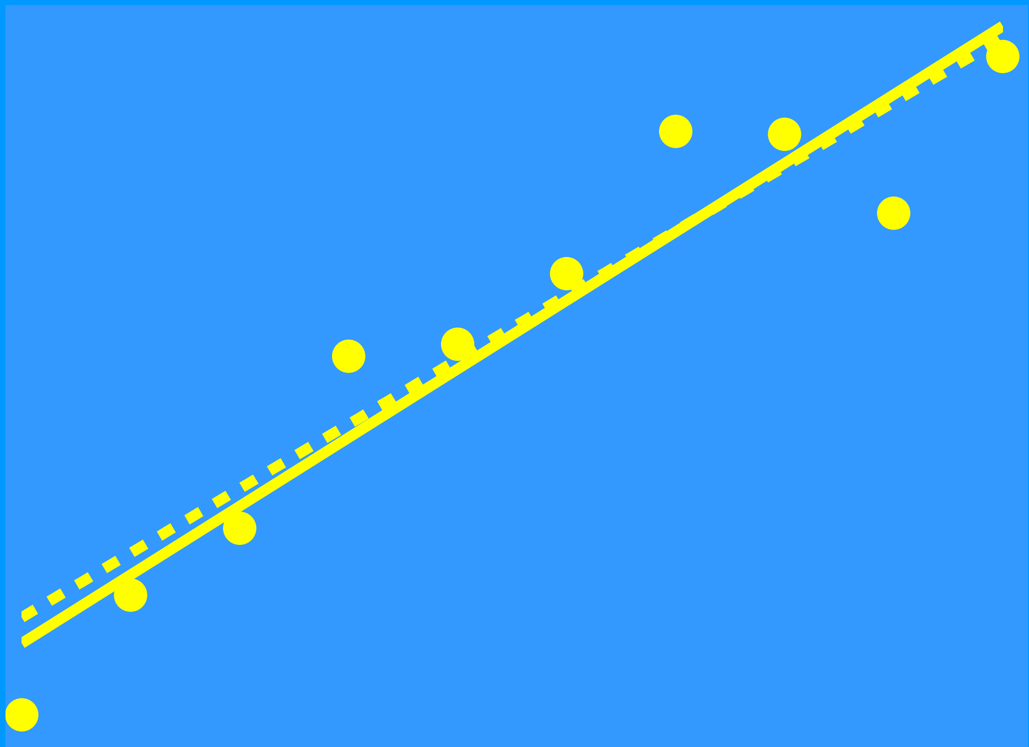
# Example



# Example



# Example



# How can we generalise this?

- Consider instead

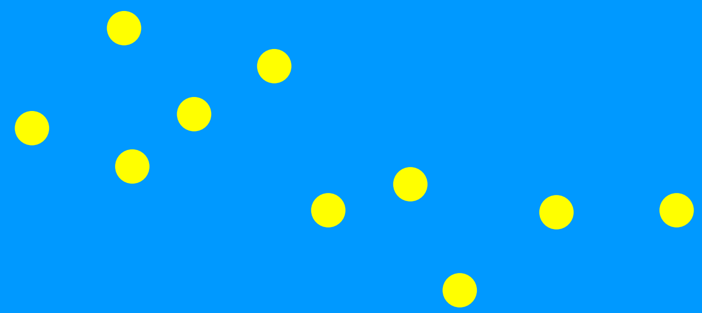
$$\hat{f}(x) = \sum_{i=1}^{m} w_i \phi_i(x)$$

- Where $\phi(x_i)$ is a nonlinear function of the inputs.
- Nonlinear transform of the inputs and then form a linear model.

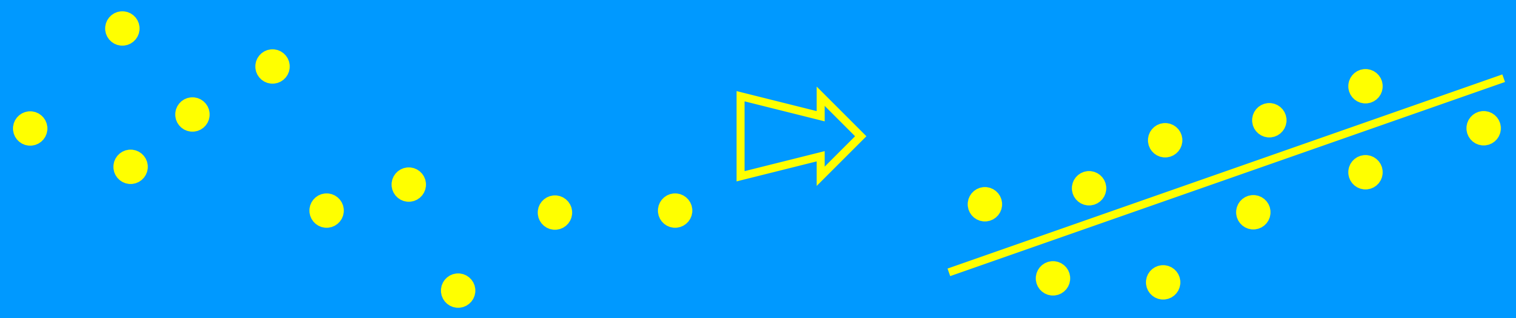
# Linear in the parameters

- A nonlinear model that is often called linear.
- Can apply simple estimation to the parameters.
- But… it is nonlinear in the basis functions.
- Linear in the parameters but nonlinear input-output relationship.

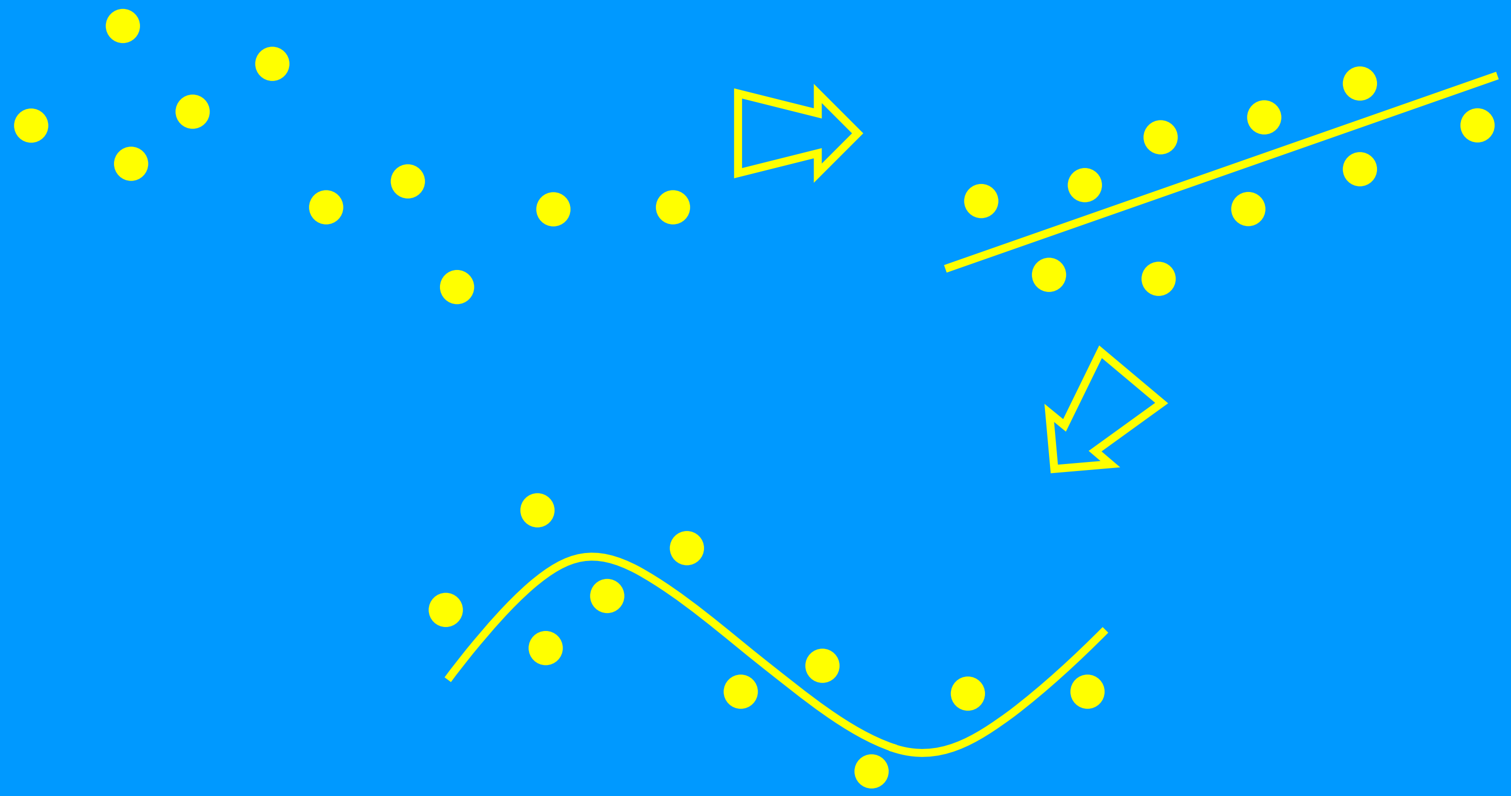
# Nonlinear mapping (regression)



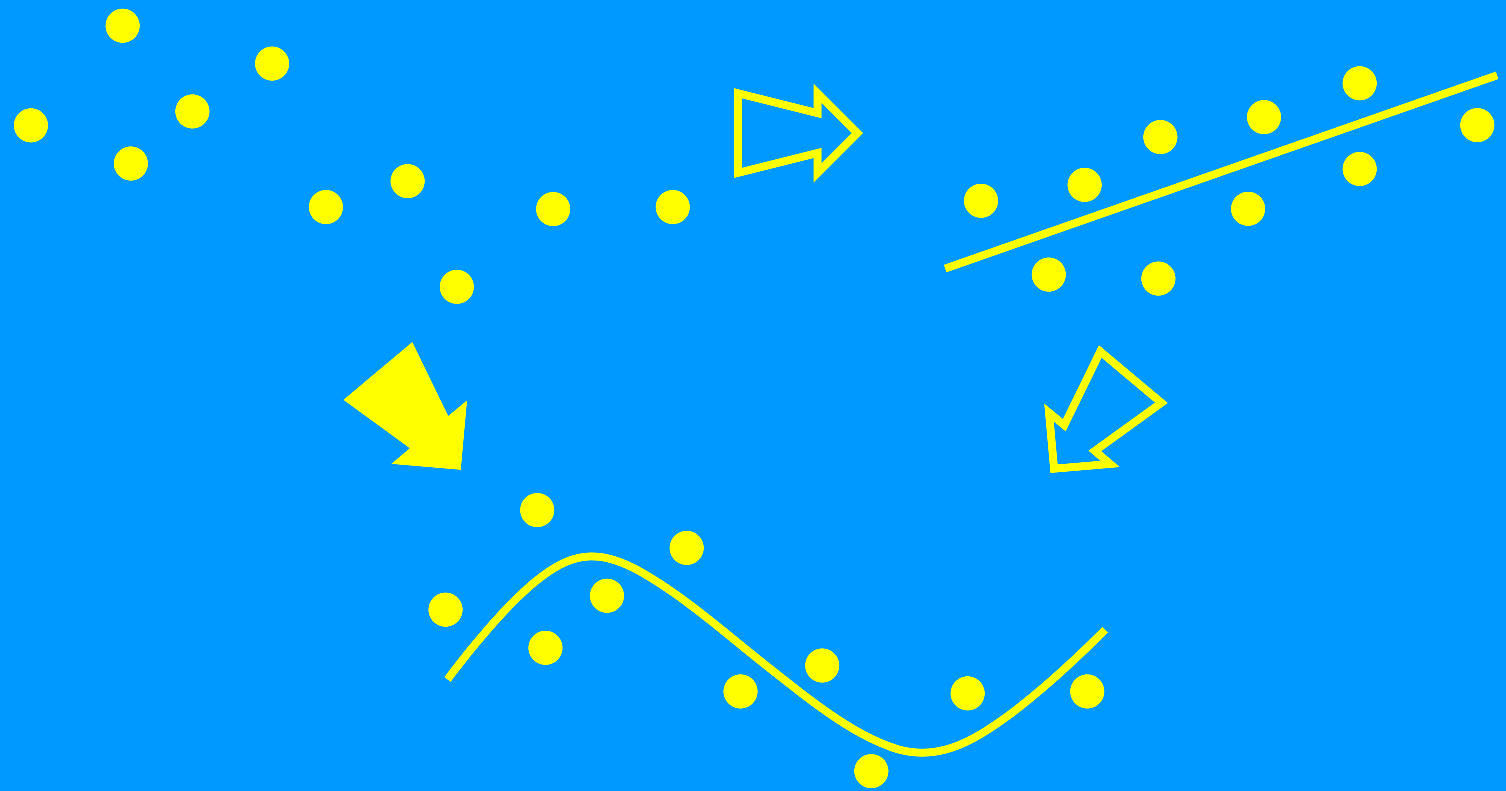
# Nonlinear mapping (regression)



# Nonlinear mapping (regression)



# Nonlinear mapping (regression)



# Parameter estimation

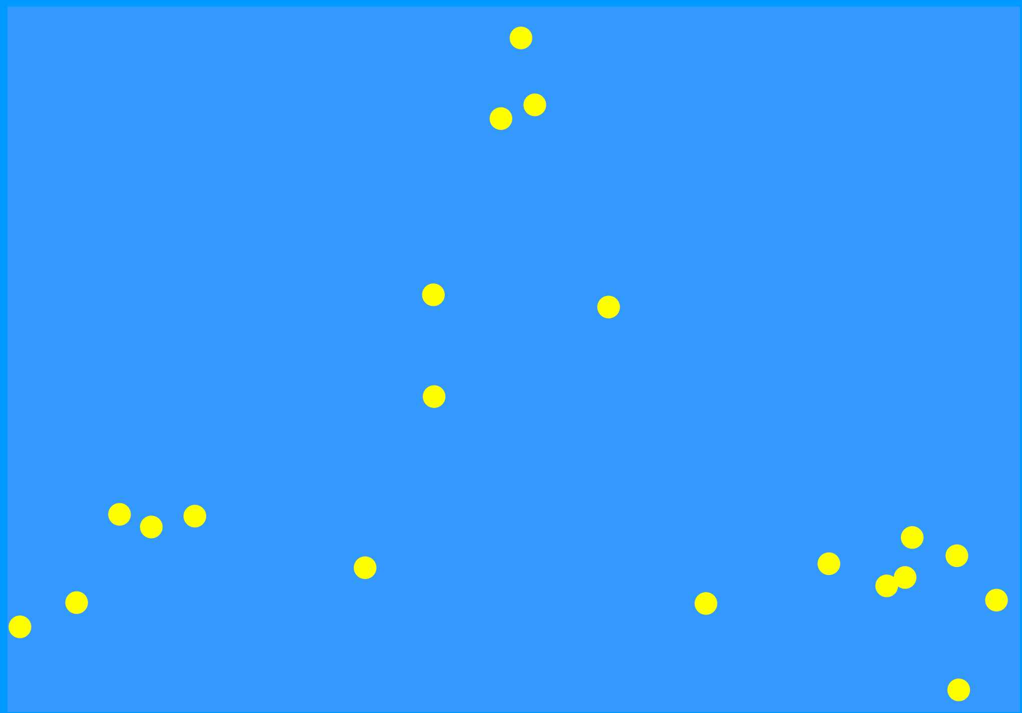
- Define the design matrix

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_m(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \cdots & \phi_m(x_N) \end{bmatrix}$$

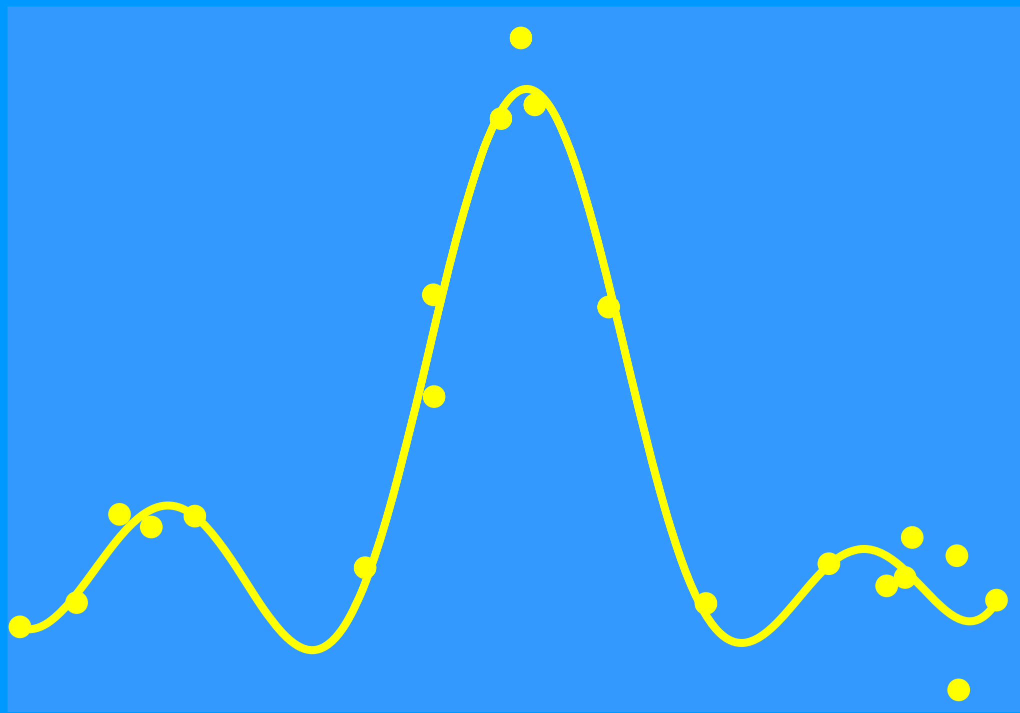
- Then the optimal parameters given by

$$\hat{w} = \left(\Phi^T \Phi\right)^{-1} \Phi^T z$$

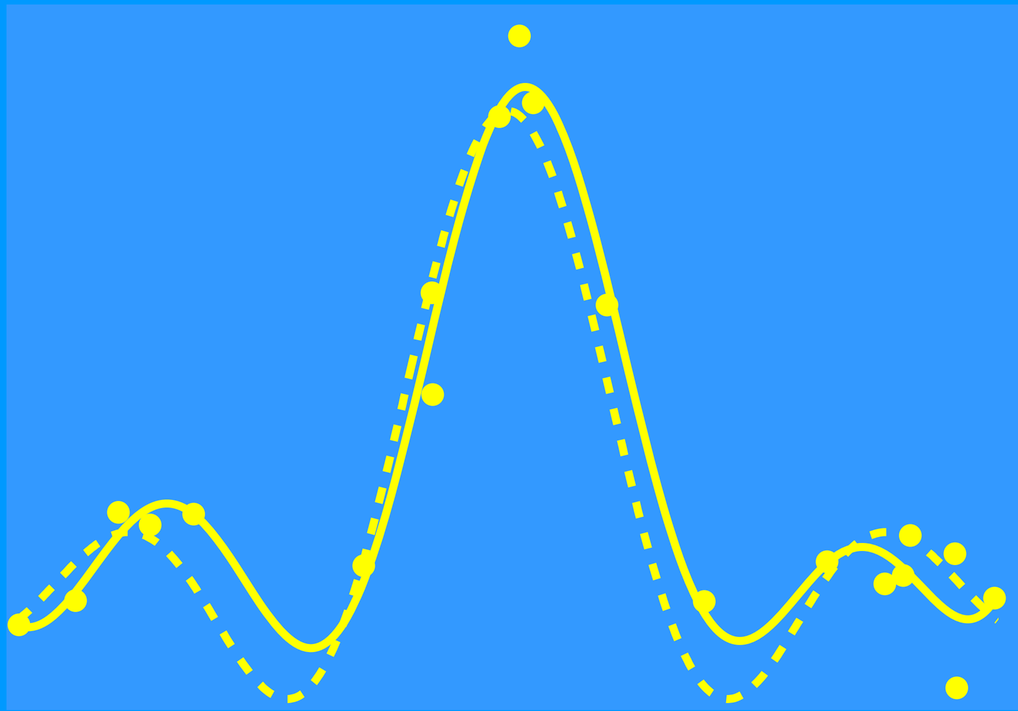
# Example



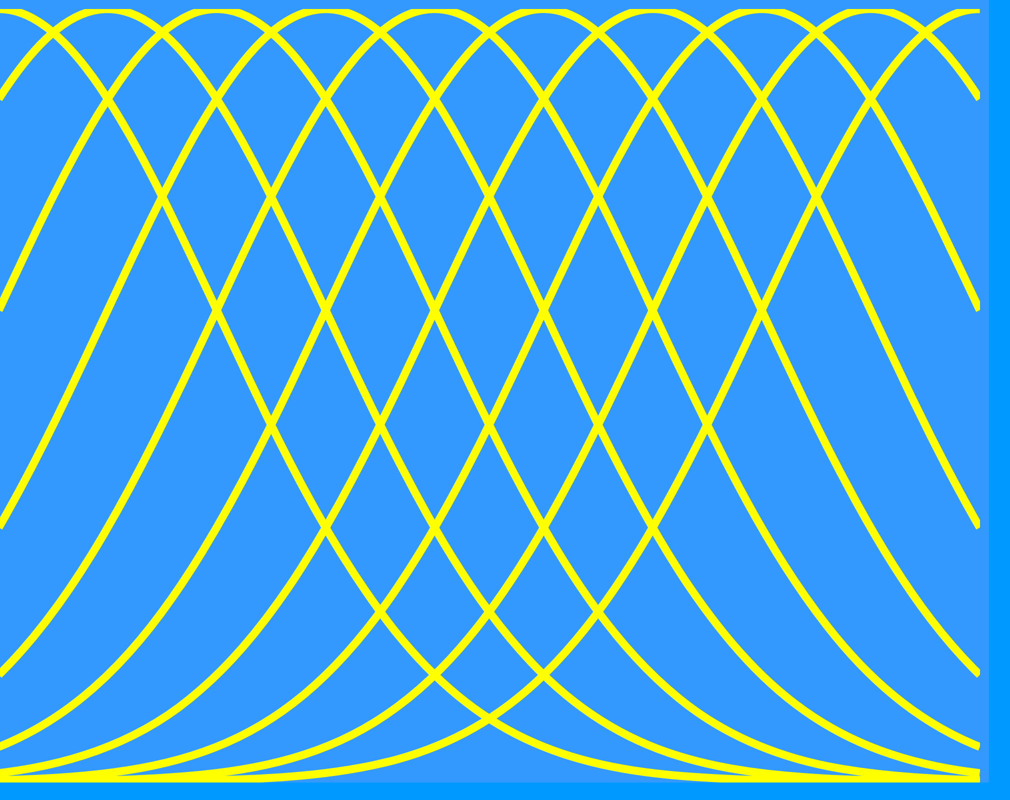
# Example



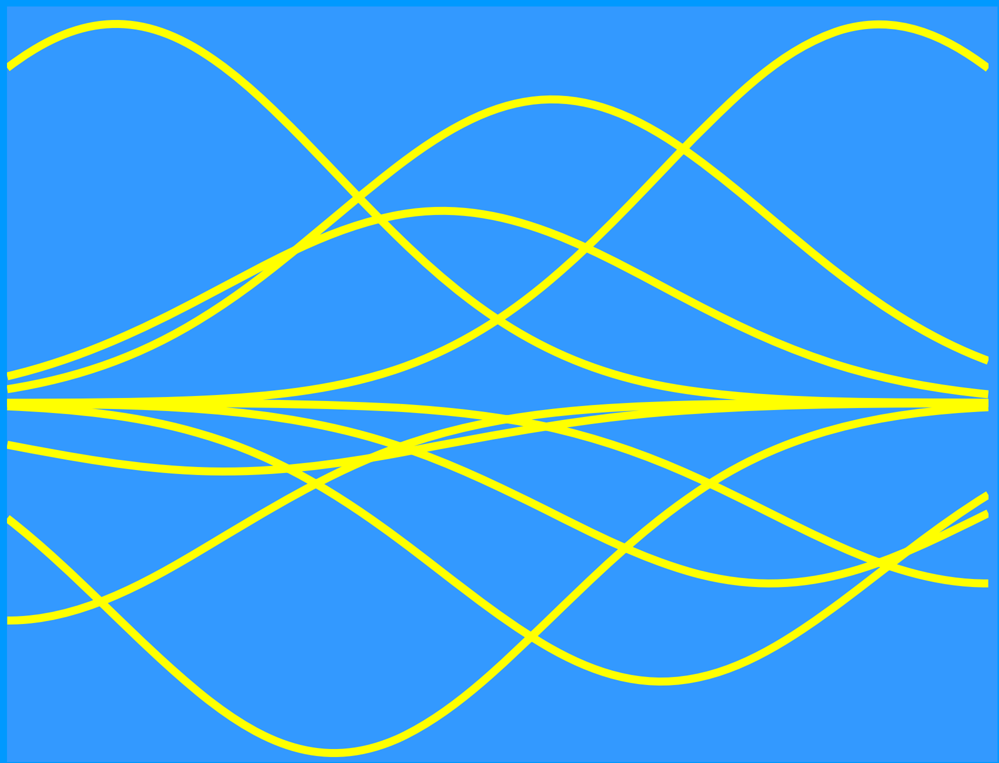
# Example



# Example – how does it work?

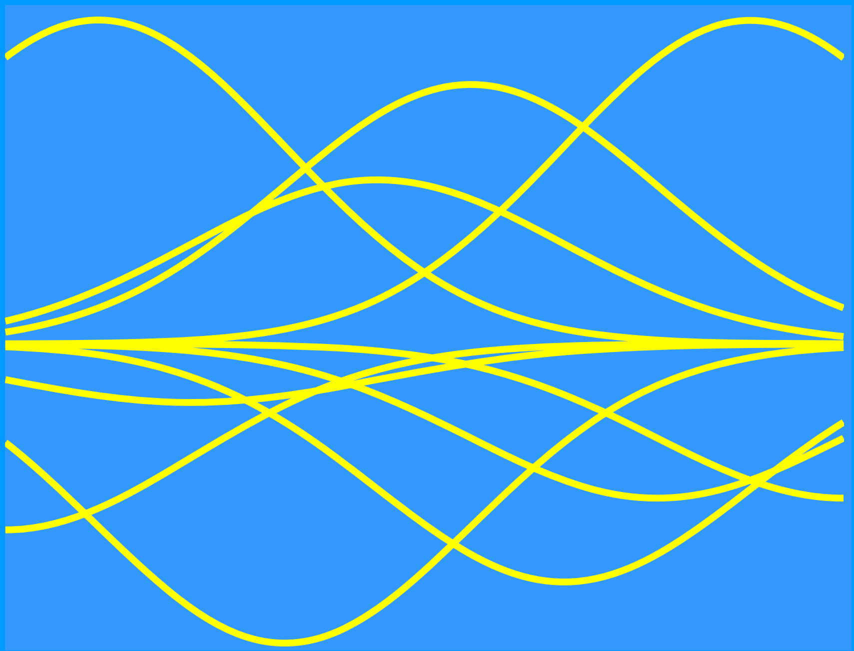


# Example – how does it work?

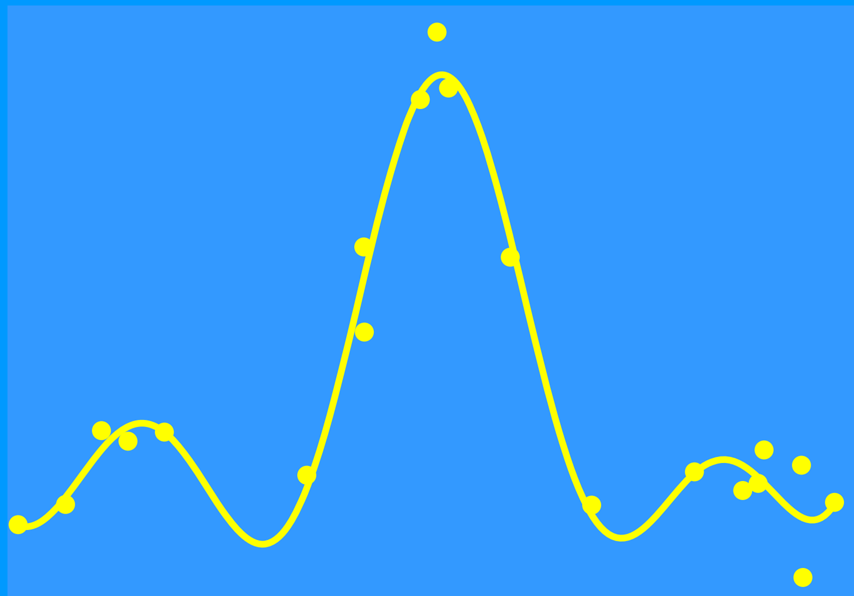


# Example – how does it work?

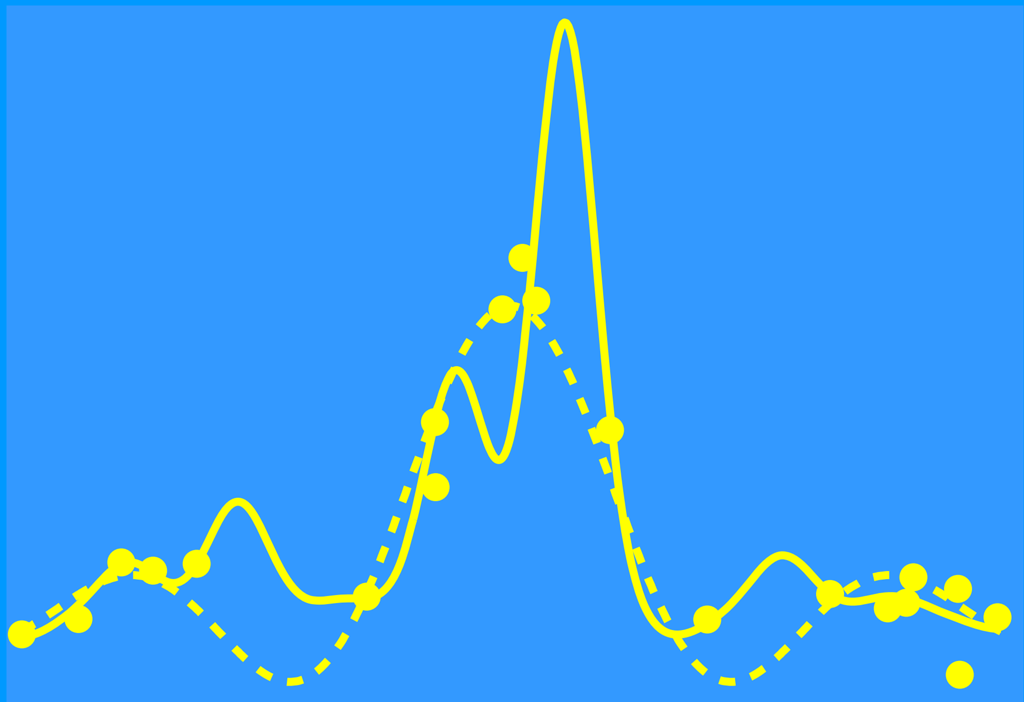
Add all these together



To get the function estimate



# Example – when it all goes wrong



# Linear classification

How do we apply linear models to classification – output is now categorical?

- Discriminant analysis.
- Probit analysis.
- Log-linear regression.
- Logistic regression.

Aim is to get a linear decision boundary.

# Logistic regression

- A regression model for Bernoulli-distributed targets.
- Form the linear model

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \sum_{i=0}^{m} w_i x_i$$

where $p = \Pr(y = 1 \mid x) = \dfrac{e^{w_0 + w_1 x_1 + \cdots}}{1 + e^{w_0 + w_1 x_1 + \cdots}}$.

# Can we generalise it?

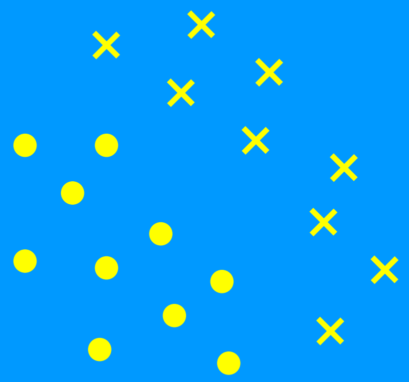
- Instead of

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \sum_{i=0}^{m} w_i x_i$$

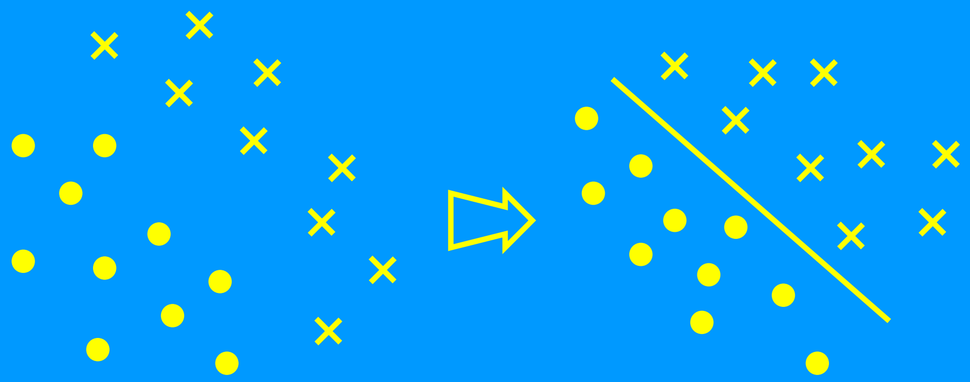
use a linear in the parameters model

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \sum_{i=1}^{m} w_i \phi_i(x)$$

# Nonlinear mapping (classification)

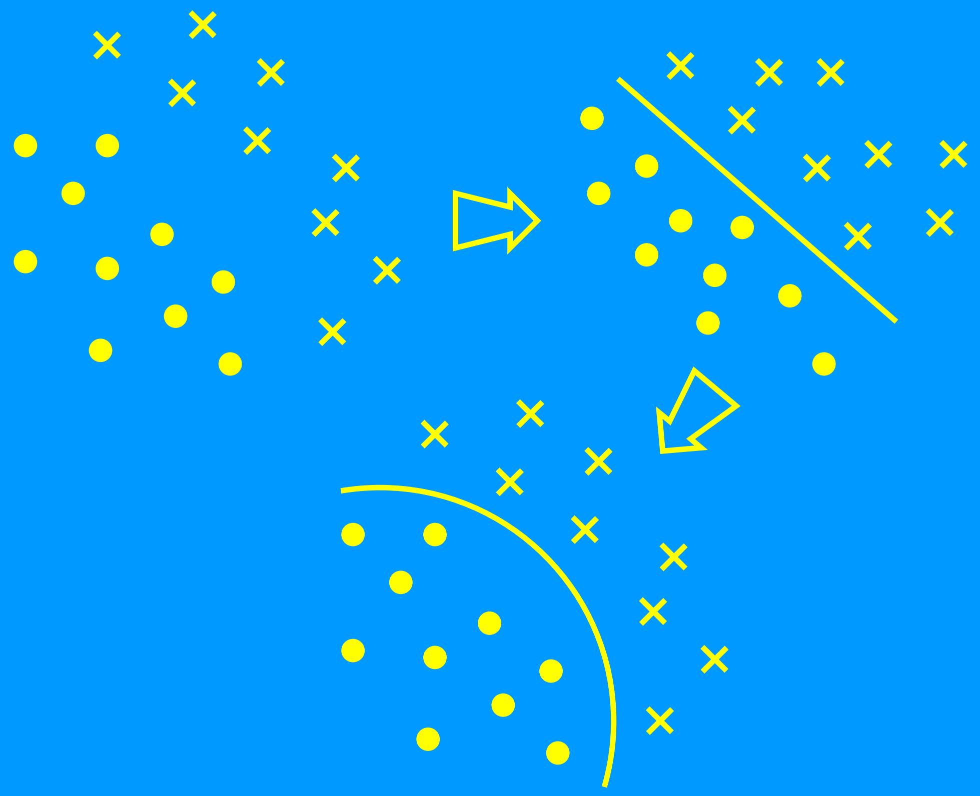


# Nonlinear mapping (classification)

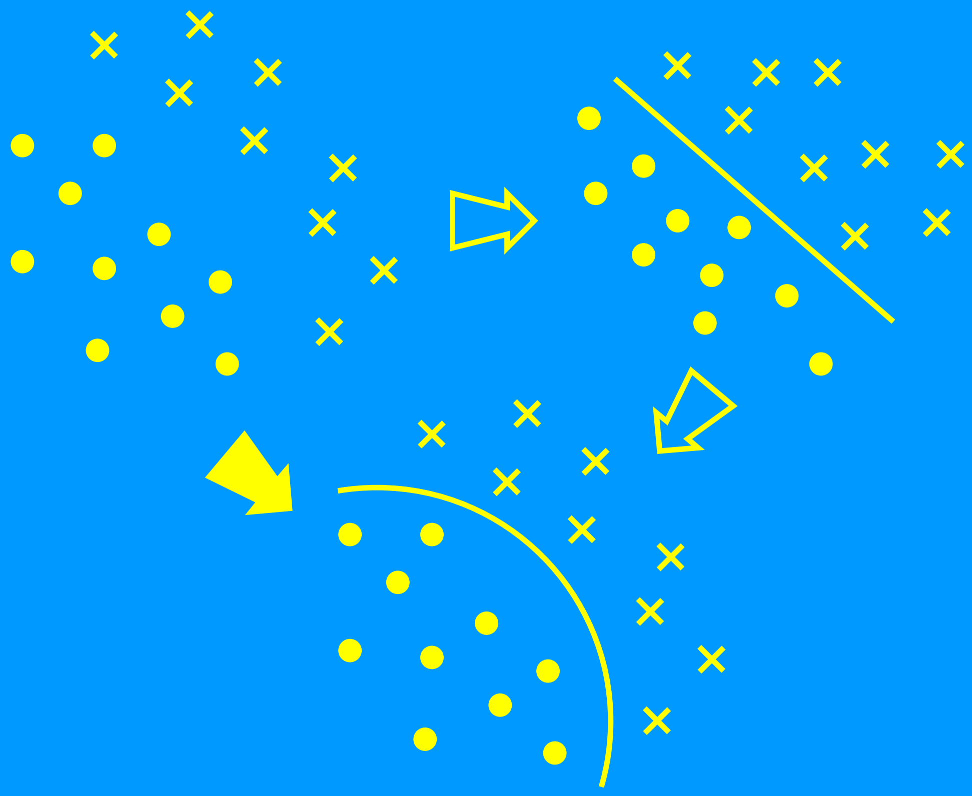


The University Of Sheffield.

# Nonlinear mapping (classification)



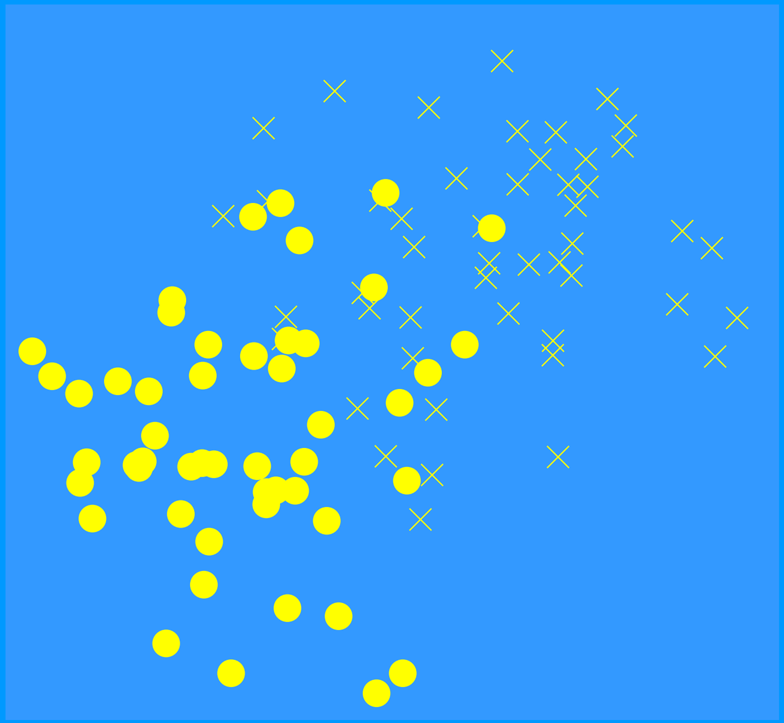
# Nonlinear mapping (classification)



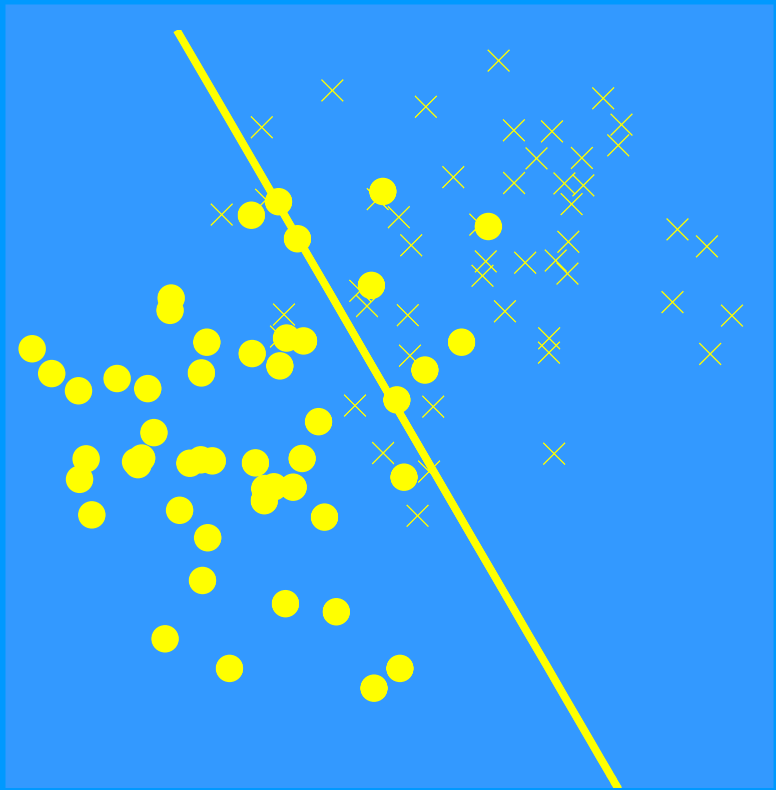
# Parameter estimation

- Maximum likelihood.
- Maximise the probability of getting the observed results given the parameters.
- Although unique minimum need to use iterative techniques (no closed form solution).

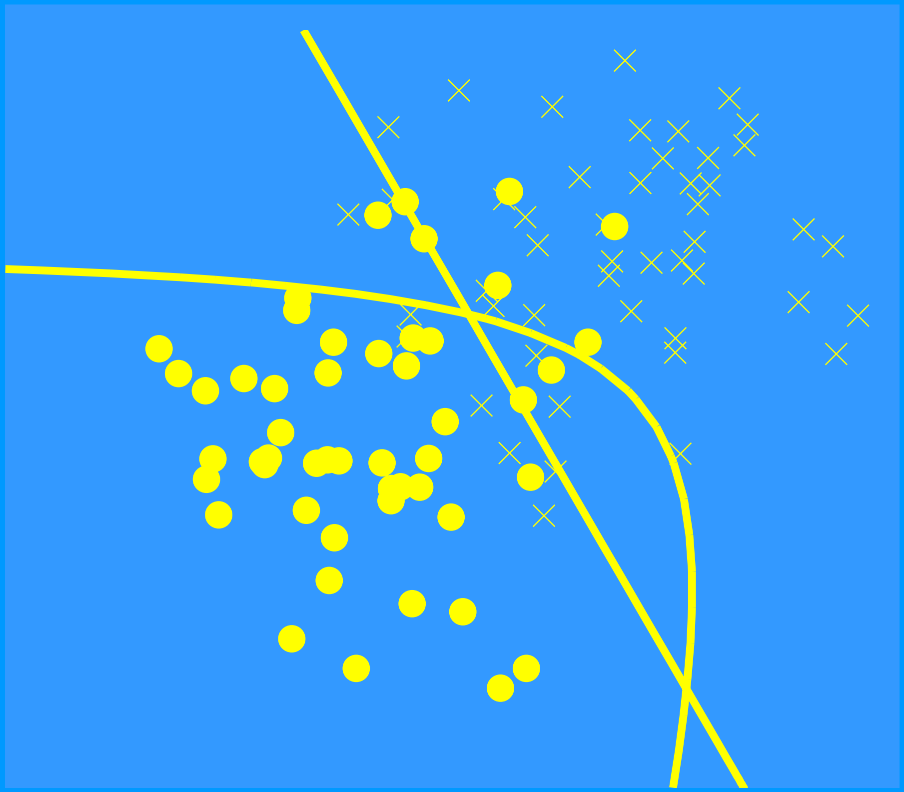
# Example



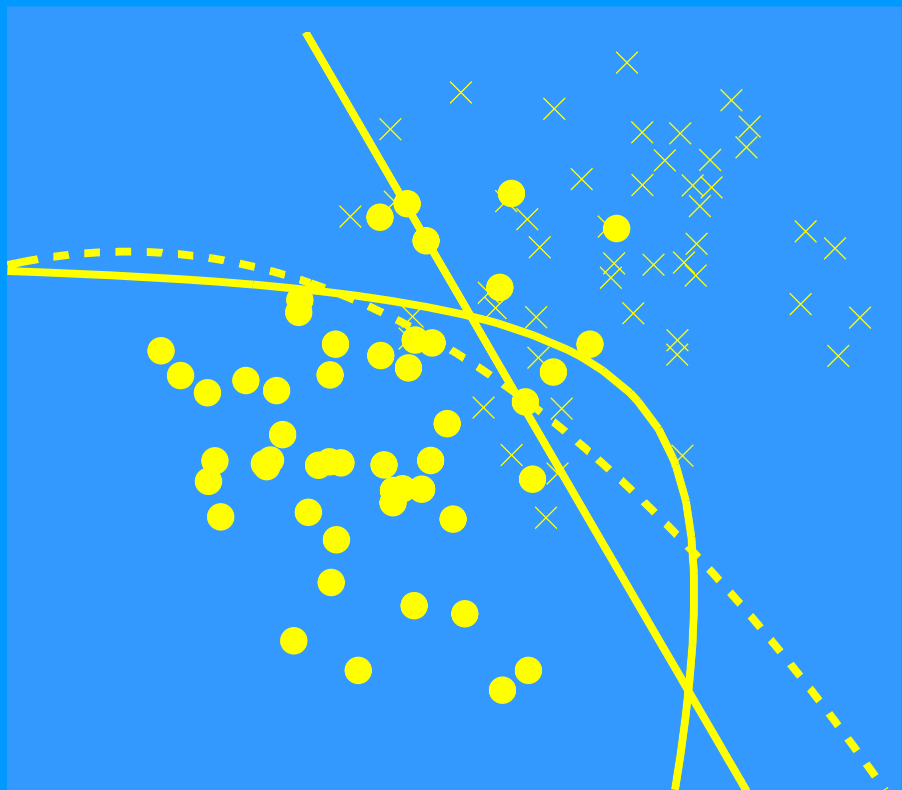
# Example



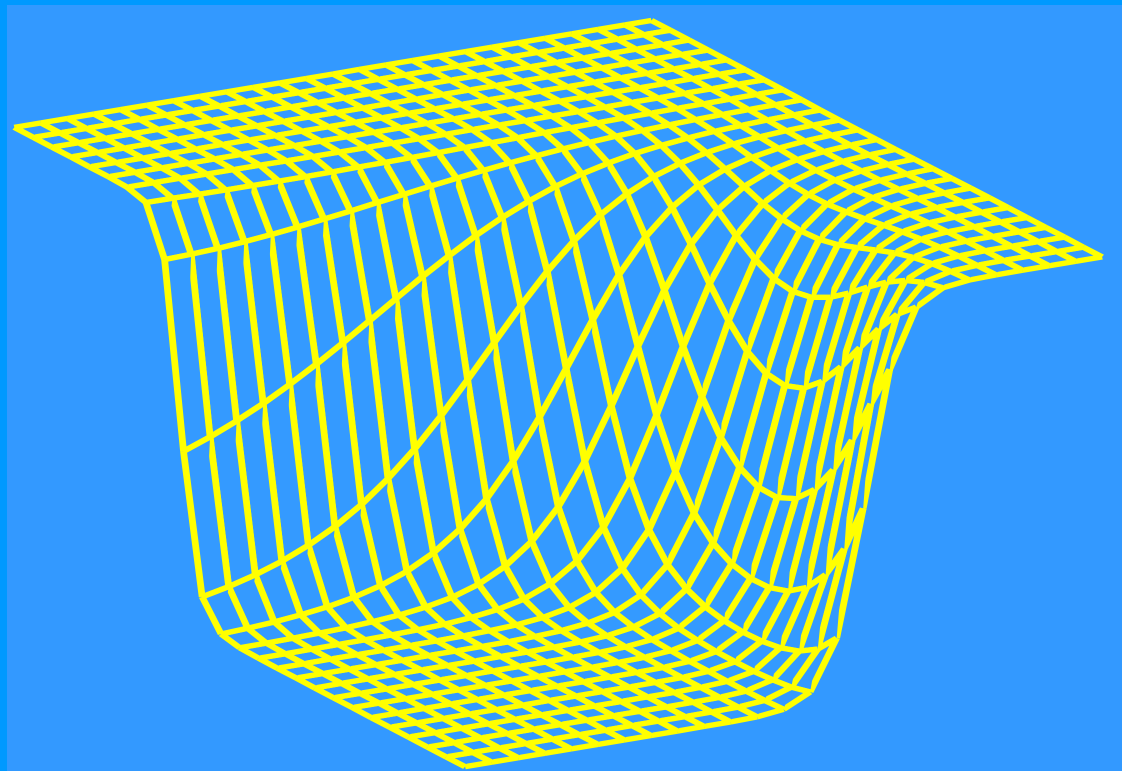
# Example



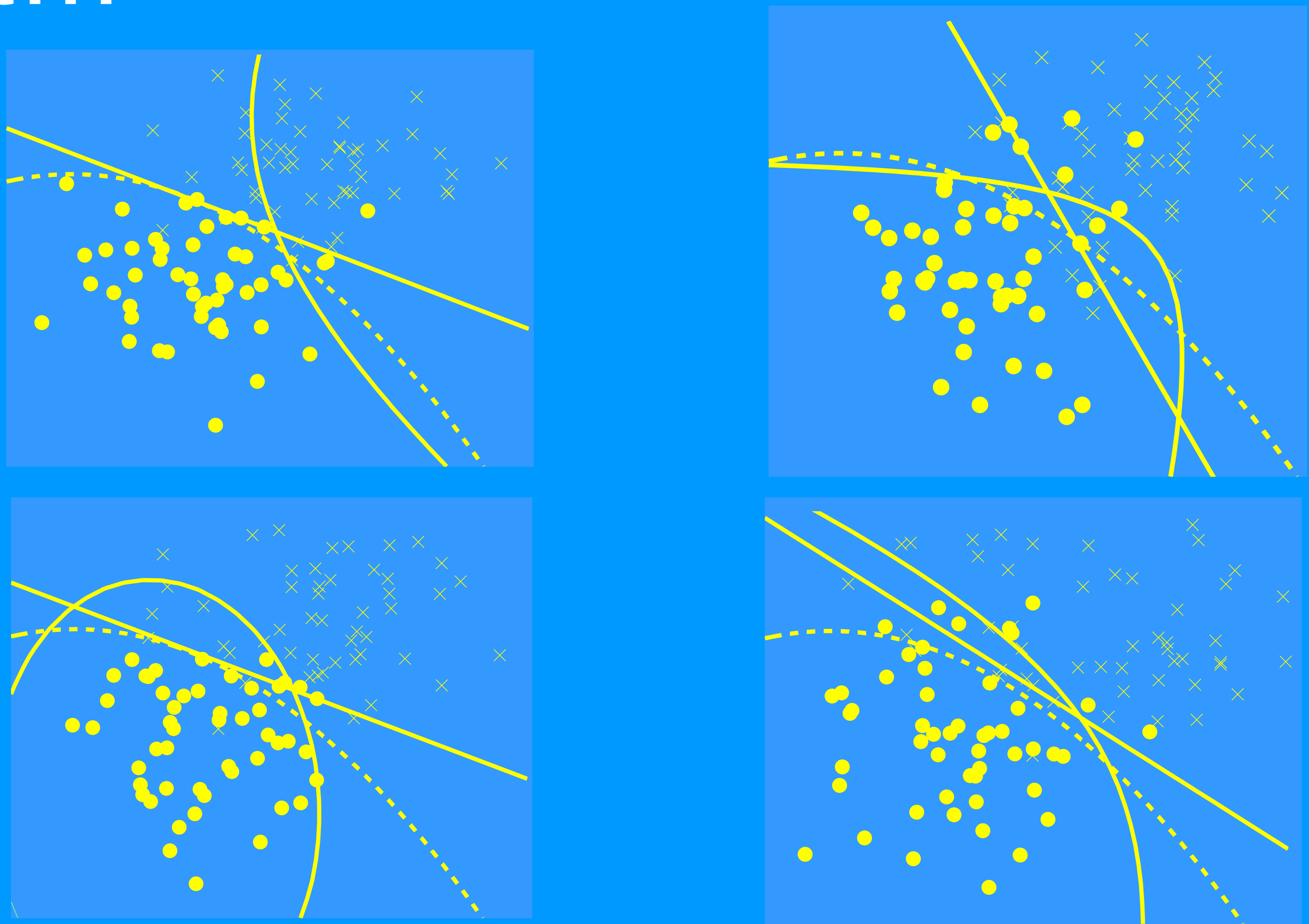
# Example



# Example – class probabilities



# But…



# Basis function optimisation

Need to estimate:

- Type of basis functions.
- Number of basis functions.
- Positions of basis functions.

These are nonlinear problems – difficult!

# Types of basis functions

- Usually choose a favourite!
- Examples include:

Polynomials: $\phi(x) = \{1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, \ldots\}$

Gaussians:
$$\phi_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma^2}\right)$$

Fourier: $\phi(x) = \sin nx, \cos nx \ldots$

…

# Number of basis functions

- How many basis functions?
- Slowly increase number until overfit data.
- Exploratory vs optimal.

# Positions of basis functions

- This is really difficult!
- One easy possibility is to put one basis function on each data point.
- Uniform grid (but curse of dimensionality).
- Advantage of global basis functions e.g. polynomials – don't need to optimise positions.

# Note on Data

How much data do we need?

- Enough to train the model?
- But how much is this?
- What about validating and testing the model?
- Need train, validate and test data!

# Concluding remarks

- Always try the simplest possible model first (e.g. linear).
- Can make nonlinear in the input but linear in the parameters.
- But becomes nonlinear optimisation.
- Is least squares/maximum likelihood the best way?