Announcements

Quiz next week!

On the student feedback

On datasets and models

General questions

Attendance

INM431: Machine Learning

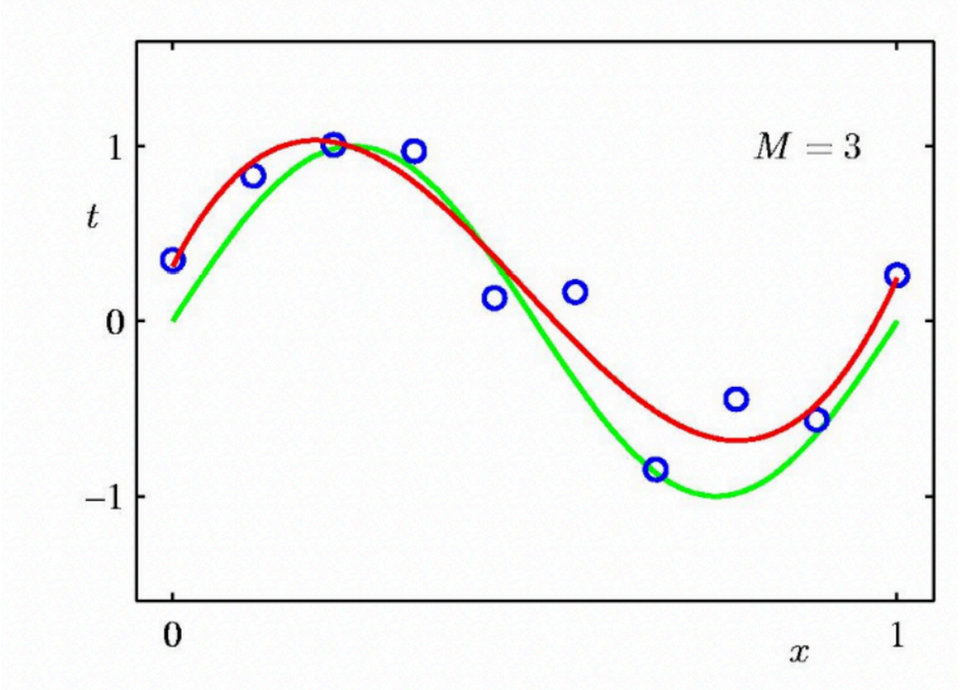
Linear regression and generalised linear models

Pranava Madhyastha (pranava.madhyastha@city.ac.uk)

Based on materials by Artur S. d'Avila Garcez + Bishop's book + Murphy's book

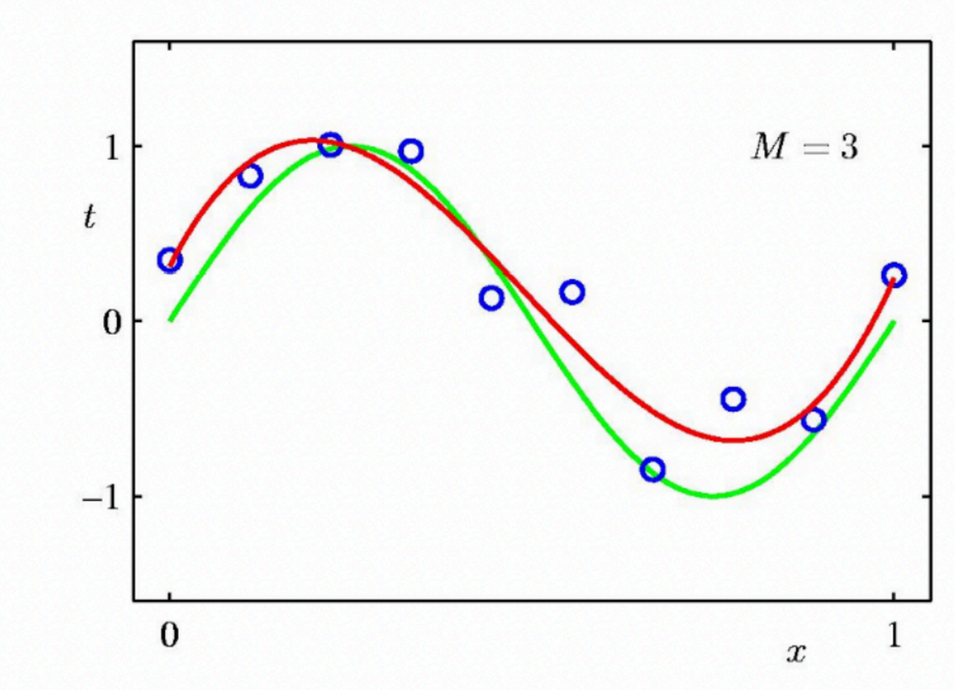
We have seen this before

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



We have seen this before

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



Why is this linear?

The setup for regression

$$\mathbf{y} = \sum_{j=0}^{m-1} w_j f_j(x)$$

Fit the best line fitting the data.

Goal: Write a rigorous model and estimate $\mathbf{w}$.

The setup for regression

$$\mathbf{y} = \sum_{j=0}^{m-1} w_j f_j(x)$$

In this case, let's assume $f(\mathbf{x})$ is linear (that is it has linear basis or is identity)

Also for notational convenience, here: $f(x_0) = 1$, thereby w_0 is the 'bias' or the intercept.

Linear regression

$$\mathbf{y} = \sum_{j=0}^{m-1} w_j x_j$$

Given a set of $n-$dimensional observations of m data points $\{x_1, \cdots, x_m\}$, a linear regression model learns a linear function of parameters $\mathbf{w}$

$\mathbf{x}$ and $\mathbf{w}$ are vectors: $\mathbf{y} = \mathbf{w}^\top \mathbf{x}$

We can also say in functional form as: $\mathbf{y}(\mathbf{w}, \mathbf{x})$

Linear regression: vocabulary

$\{x_1, \cdots, x_m\}$: Explanatory variables, independent variables, predictor variables

- a variable that can be used to observe changes

$\mathbf{y}$: Dependent variables, outcome variables, response variables

- a variable that responds to the explanatory variables
- the expected effect

Linear regression: vocabulary

$\{x_1, \cdots, x_m\}$: Explanatory variables, independent variables, predictor variables

- a variable that can be used to observe changes

$\mathbf{y}$: Dependent variables, outcome variables, response variables

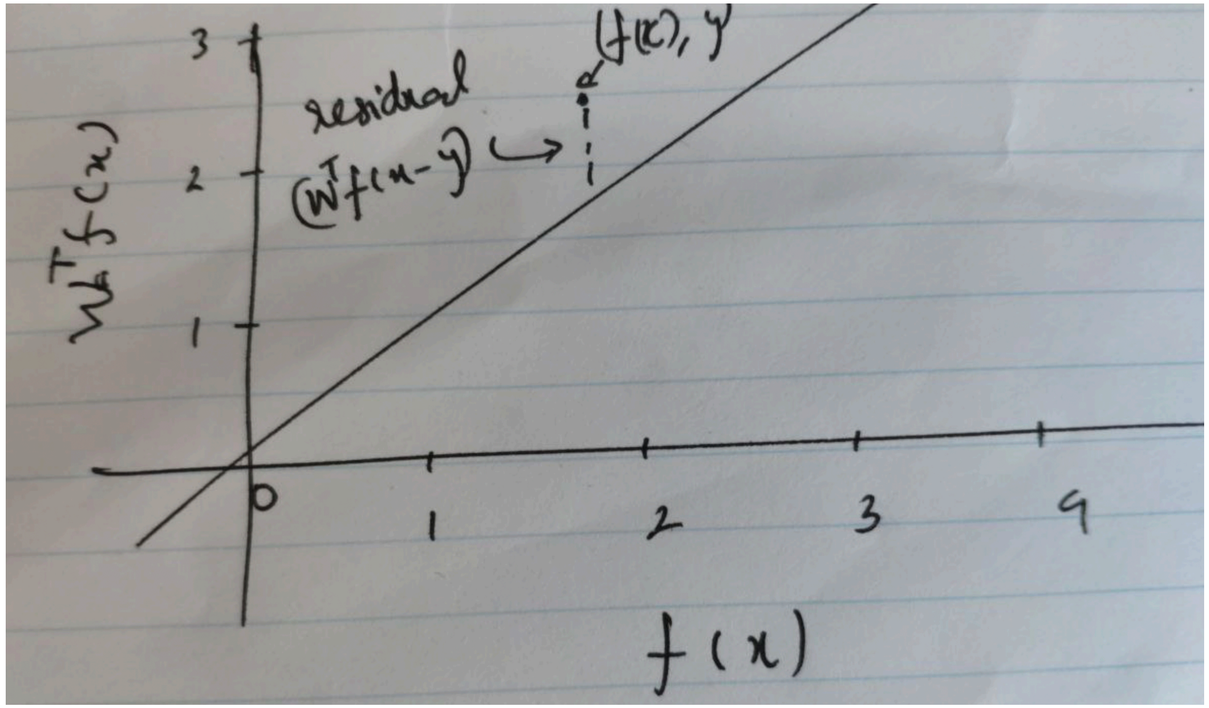
- a variable that responds to the explanatory variables
- the expected effect

Residual

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{x}$$

The residual is $((\mathbf{w} \cdot f(x)) - y)$

Which is the amount by which the prediction overshoots the target.

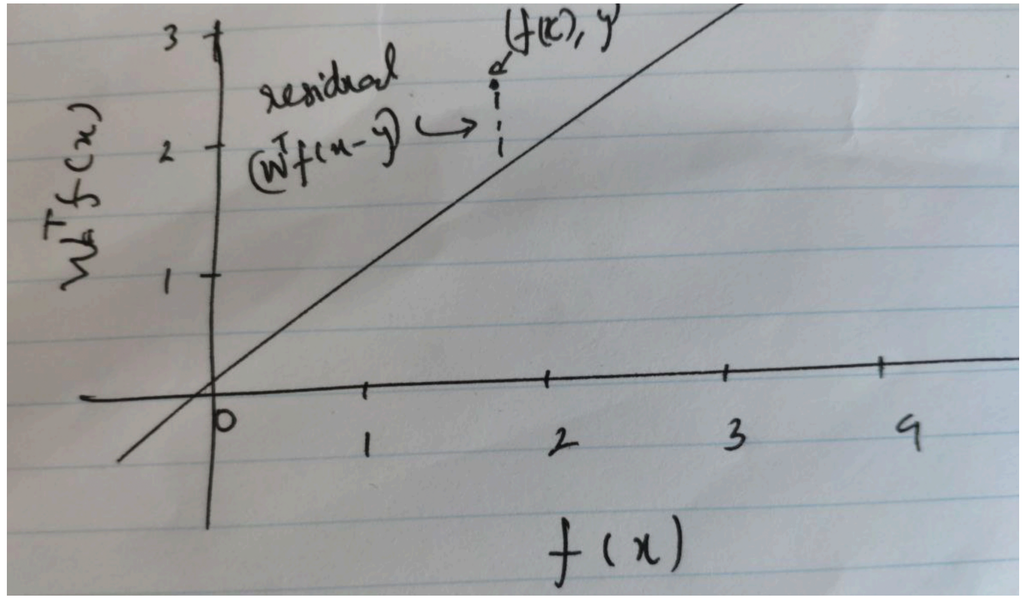


Squared loss

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{x}$$

$$\text{Loss}_{sq}(x, y, \mathbf{w}) = (\mathbf{w} \cdot f(x) - y)^2$$

residual



Squared loss minimisation

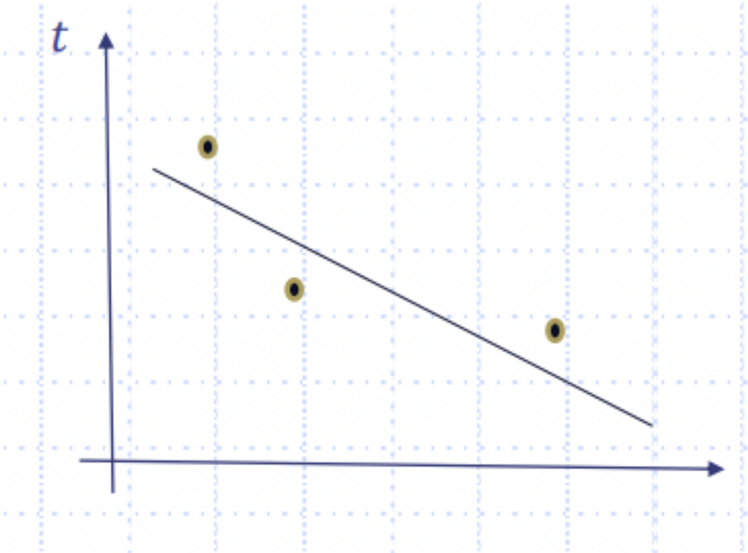
$$\mathbf{y} = \mathbf{w} \cdot \mathbf{x}$$

$\text{Loss}_{sq}(x, y, \mathbf{w}) = (\mathbf{w} \cdot f(x) - y)^2$ - this is per observation

$$\text{TrainingLoss}(\mathbf{w}) = \sum_{(x,y)} Loss_{sq}(x, y, \mathbf{w}))$$

Example

When 2 highway patrol cars are deployed, the average speed on the freeway is 75mph. When 4 patrol cars are deployed the average speed is 45mph. When 10 patrol cars are deployed the average speed is 35mph. Using linear regression and **least squares**, what will be the average speed when 5 cars are deployed?



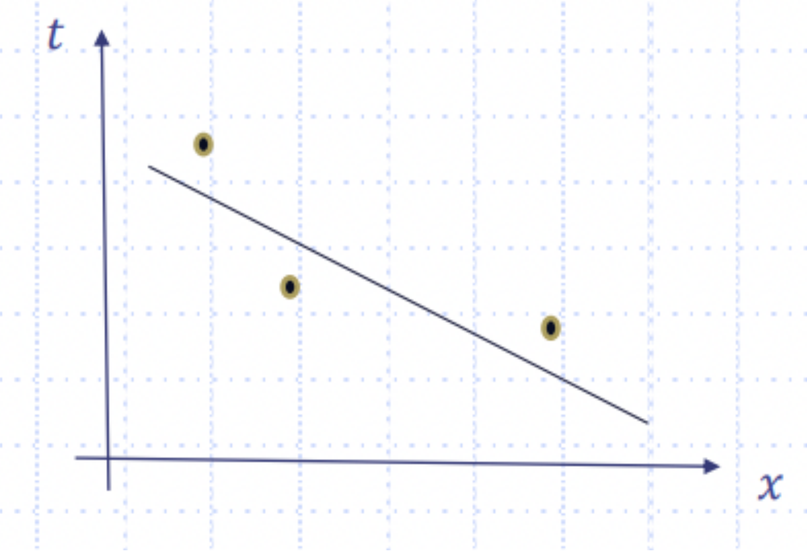
Hints: rename variables, let's assume, $\mathbf{y}(x, \mathbf{w})$ and t are our variables,

$$\text{Loss}_{lsq}(x_i, \mathbf{w}, t_i) = \frac{1}{2}(\mathbf{w} \cdot x_i - t_i)^2 \qquad \mathbf{w} = \{w_0, w_1\}$$

$$\frac{\partial \text{Loss}_{lsq}}{\partial \mathbf{w}} = 0$$

Example

When 2 highway patrol cars are deployed, the average speed on the freeway is 75mph. When 4 patrol cars are deployed the average speed is 45mph. When 10 patrol cars are deployed the average speed is 35mph. Using linear regression and **least squares**, what will be the average speed when 5 cars are deployed?



x_i	t_i	y_i
2	75	65.7
4	45	57.3
10	35	31.8
5	?	53

Example

$$\text{Loss}_{lsq}(x_i, \mathbf{w}, t_i) = \sum_i ((w_1 x_i + w_0) - t_i)^2$$

$$\frac{\partial \text{Loss}_{lsq}(x_i, \mathbf{w}, t_i)}{\partial w_1} = \sum_i ((w_1 x_i + w_0) - t_i) x_i)$$

$$120 w_1 + 16 w_0 - 680 = 0$$

$$\frac{\partial \text{Loss}_{lsq}(x_i, \mathbf{w}, t_i)}{\partial w_0} = \sum_i ((w_1 x_i + w_0) - t_i) 1)$$

$$16 w_1 + 3 w_0 - 155 = 0$$

x_i	t_i	y_i
2	75	65.7
4	45	57.3
10	35	31.8
5	?	53

Example

Learned parameters: $w_1 = -4.237$; $w_0 = 74.264$

Learned model: $y_i = -4.24x_i + 74.26$

x_i	t_i	y_i
2	75	65.7
4	45	57.3
10	35	31.8
5	?	53

Calculate the training set error!

This should be the sum of the squared differences between t_i and y_i

Maximum Likelihood for Regression with Gaussian Noise and Least Squares

Assume observations from a deterministic function with added Gaussian noise:

$$t = \mathbf{y}(x, w) + \epsilon$$

where, the noise distribution $P(\epsilon \,|\, \beta) = \mathcal{N}(\epsilon \,|\, 0, \beta^{-1})$

This is equivalent to:

$$P(t \,|\, \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t \,|\, y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

Maximum Likelihood for Regression with Gaussian Noise and Least Squares

Given observed inputs $X = \{x_1, \cdots, x_m\}$ and targets $t = [t_1, \cdots, t_m]$

We obtain the likelihood function:

$$P(t \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{n} \mathcal{N}(t_n \mid \mathbf{w} \cdot x_n), \beta^{-1})$$

Maximum Likelihood for Regression with Gaussian Noise and Least Squares

We don’t like products, we like sums, so:

$$\ln P(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \sum_{n=1}^{m} \mathcal{N}(t_n \mid \mathbf{w} \cdot x_n), \beta^{-1})$$

$$= \frac{m}{2} \ln \beta - \frac{m}{2} \ln(2\pi) - \frac{\beta}{2} \sum_{n=1}^{m} \{t_n - \mathbf{w} \cdot x_n\}^2$$

Maximum Likelihood for Regression with Gaussian Noise and Least Squares

Taking gradients:

$$\nabla_w \ln P(\mathbf{t} \mid \mathbf{w}, \mathbf{x}, \beta) = \beta \sum_{n=1}^{m} \{t_n - \mathbf{w}^\top x_n\} x_n^\top$$

Solving for $\mathbf{w}$, we get:

$$\mathbf{w}_{ML} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top$$

If we can calculate the inverse, we have closed form solution!

Assumptions

The design matrix $\mathbf{x} \in \mathbb{R}^{m \times p}$ is deterministic and rank($\mathbf{x}$) = p

The model is homoskedastic: i.e., the variance of the residual, or loss term is constant.

The noise is constant for some unknown β

Linear regression exhibits correlations and not causation!

Let’s see linear regression again

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{x}$$

Learning as an optimisation problem

Goal:

a) define an objective function corresponding to the training loss

b) use an optimisation algorithm that obtains the best parameters $\mathbf{w}$ where the objective function achieves the minimum value

Let’s see linear regression again

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{x}$$

Learning as an optimisation problem

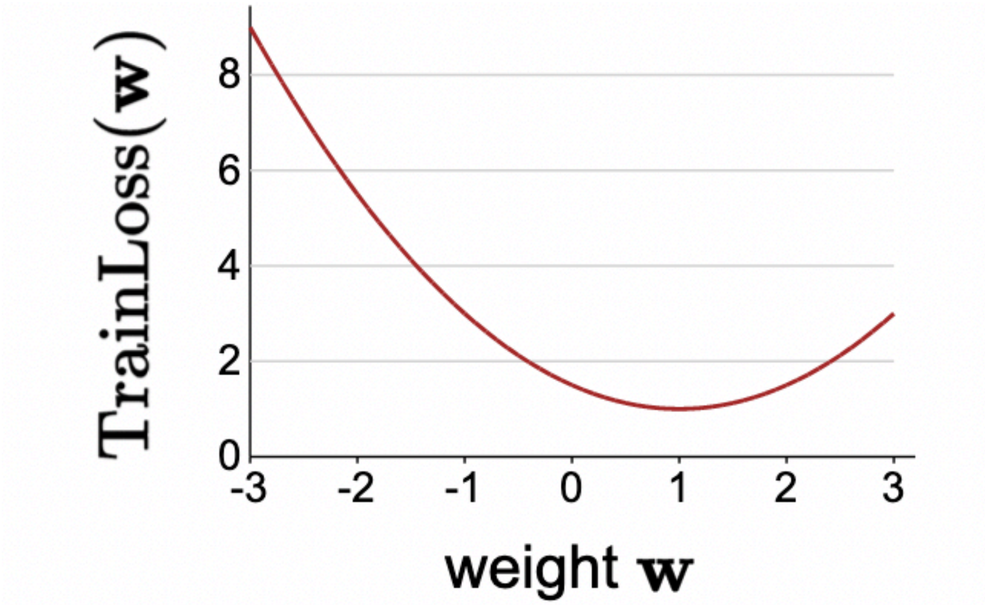
Goal:

- define an objective function corresponding to the training loss
- use an optimisation algorithm that obtains the best parameters $\mathbf{w}$ where the objective function achieves the minimum value

Optimisation problem

Objective is:

$$\min_{\mathbf{w}} \text{TrainingLoss}(\mathbf{w})$$



Gradient descent

A general approach is to use iterative optimisation, which essentially starts at some starting point, and tries to update the parameters such that the objective function value decreases.

The gradient of the objective function informs us the direction to move to decrease the objective the most.

The iterative optimisation procedure is called gradient descent.

Gradient descent has two hyperparameters, the step size η and the K number of iterations (see the data multiple number of times)

Least squares regression with GD

$$\text{Loss}_{lsq}(x_i, \mathbf{w}, t_i) = \frac{1}{2}(\mathbf{w} \cdot x_i - t_i)^2$$

$$\text{Loss}_{lsq}(\mathbf{x}, \mathbf{w}, \mathbf{t}) = \frac{1}{2}\sum_i ((\mathbf{w}^\top \mathbf{x}) - \mathbf{t})^2$$

The gradient of the loss is obtained by taking the first derivative:

$$\nabla_{\mathbf{w}} \text{Loss}_{lsq}(\mathbf{x}, \mathbf{w}, \mathbf{t}) = \sum_i ((\mathbf{w}^\top \mathbf{x}) - \mathbf{t})\mathbf{x}$$

Least squares regression with GD

For epoch 1 to K:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \text{Loss}_{lsq}$$

Each epoch takes a swipe at the entire training data

Least squares regression with Stochastic GD

A faster approach

For epoch 1 to K:

For each (x_i, t_i):

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \text{Loss}_{lsq}(x_i, t_i, \mathbf{w})$$

Summary

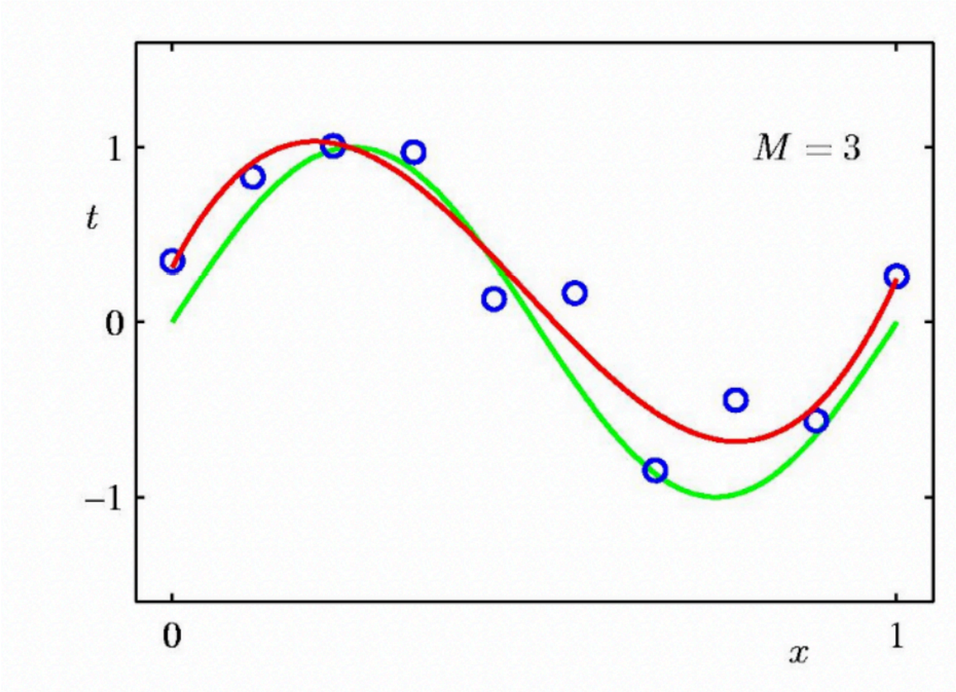
We have so far seen: linear regression

Multivariate linear regression

Closed form solution

Capturing non-linear relationships

$$\mathbf{y} = \sum_{j=0}^{m-1} w_j f_j(x)$$



Basis function

$$\mathbf{y} = \sum_{j=0}^{m-1} w_j f_j(x)$$

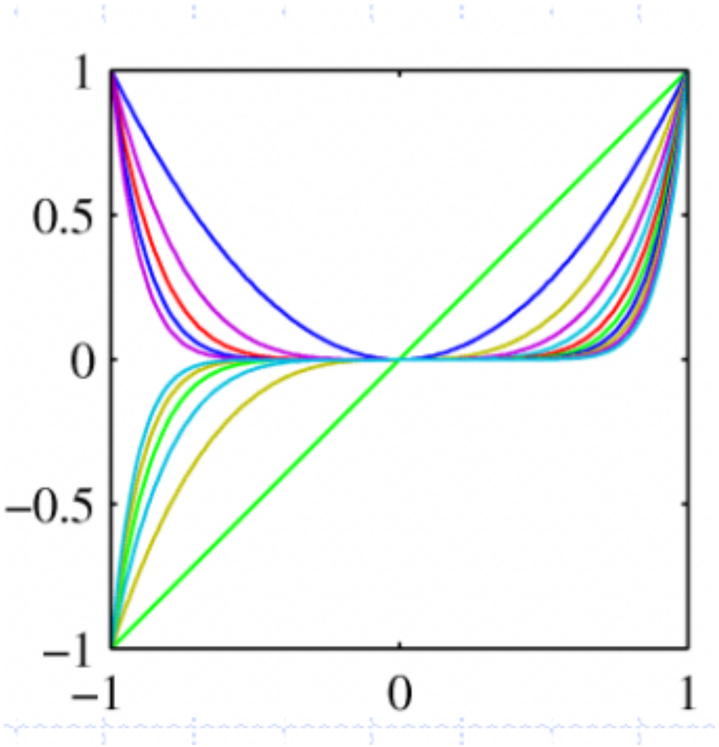
$\mathbf{f}(x)$: basis functions

We can map x over several basis functions: these are over the input.

Basis function: Polynomial

$$\mathbf{y} = \sum_{j=0}^{m-1} w_j f_j(x)$$

Possible choices: $f_j(x) = x^j$

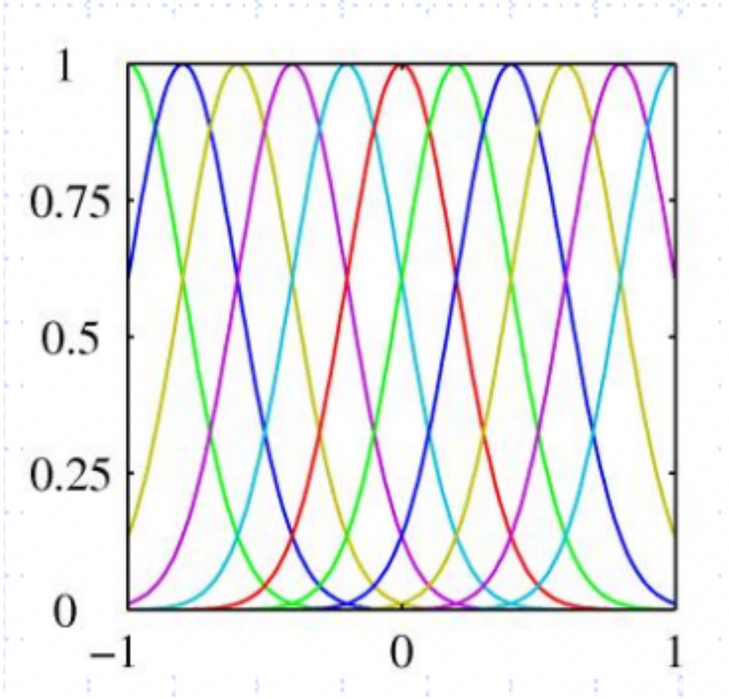


These are global whereby a small change in input can affect all basis functions.

These are nonlinear basis functions: with this it is possible to capture non-linear relationships of x with the output

Basis function: Gaussian basis functions

Possible choices: $f_j(x) = \exp\{-\frac{(x-\mu_j)}{2s^2}\}$



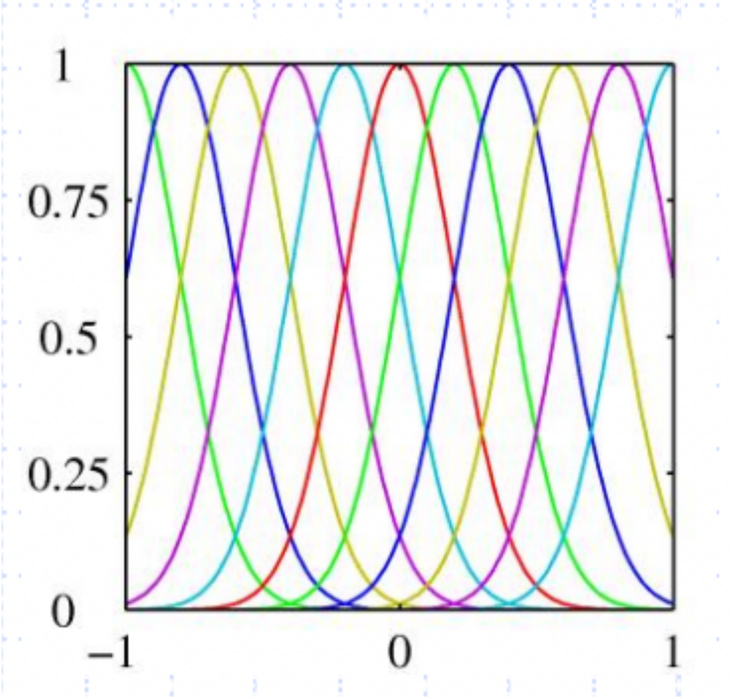
These are local whereby a small change in input can affect nearby basis functions.

(Radial basis function)

Basis function: Sigmoidal basis functions

Possible choices: $f_j(x) = \sigma(\frac{x - \mu_j}{s})$

where, $\sigma(a) = \frac{1}{1 + exp(-a)}$



These are local whereby a small change in input can affect nearby basis functions.

(Radial basis function)

Linear Models (GLM)

A linear model assumes:

$$Y \mid X \sim \mathcal{N}(\mathbf{f}(\mathbf{x}), \beta^{-1} I)$$

And

$$\mathbb{E}(Y \mid X) = \mathbf{w} \cdot \mathbf{f}(\mathbf{x})$$

Generalisation

A generalised linear model (GLM) generalises normal linear regression models in the following directions.

a) No assumption on the residuals

b) The outcome variable can be of any form: binary, integer, strictly positive, etc and need not come from a normal distribution. They can be from any exponential distribution.

c) No assumptions on variance

Basic assumptions

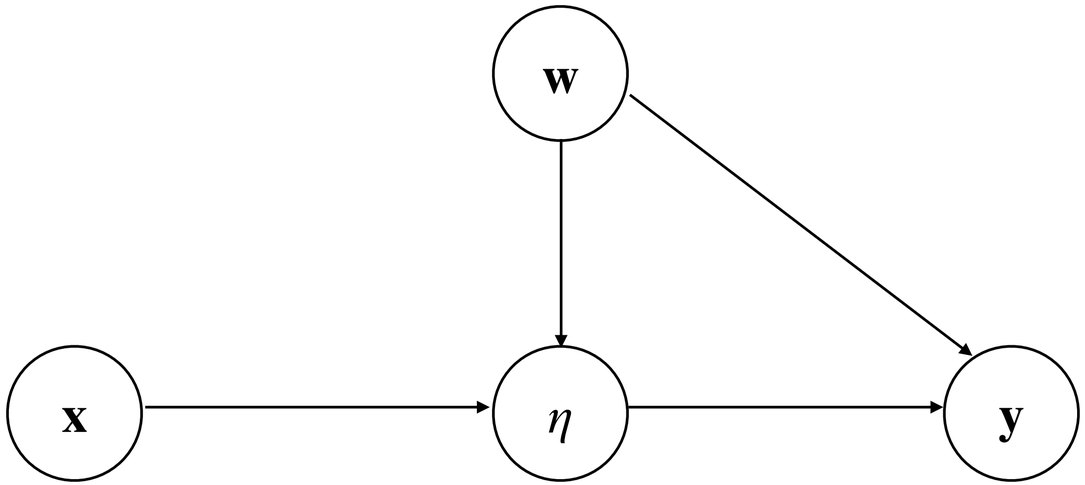
The influences of the $\{x_i\}$ variables on y_i can be summarised into an intermediate form called a Linear Predictor η

η is a linear combination of $\{x_i\}$

We have a function g mapping g to the expected value (μ) of $\mathbf{y}$

We also assume conditional independence of $\mathbf{y}$ from $\{x_i\}$ given η

Generalised linear model



Here:

$\eta = w_0 + w_i x_i + \cdots w_n x_n \leftarrow$ linear predictor

Link function takes linear predictor output and confines it in some way to a different scale.

Consider the case of binomially distributed data

Range of possible $Y \in [0,1]$

$$P(Y = y; \mu) \propto \mu^{ny}(1-\mu)^{r(1-y)}$$

we have $\eta = w_0 + w_i x_i + \cdots w_n x_n$

Now, we have to choose the relationship between choosing a relationship for linear predictor η and the mean μ.

There are many link functions that can be chosen to make this mapping valid, but here we will use the most popular link function, the logit transform

Logit transform

$$\log \frac{\mu}{1-\mu} = \eta$$

The inverse of this is:

$$\mu = \frac{e^{\eta}}{1-e^{\eta}}$$

Replace η with the full expression

Check the lab exercise

Slightly different way of conveying the same thing.

Link functions are from exponential families

Please read Chapter 4 from PRML

Also Refer to Chapters 10, 11, 12 ML:APP