Regression

Hamid Dehghani
School of Computer Science
Birmingham
September 2020

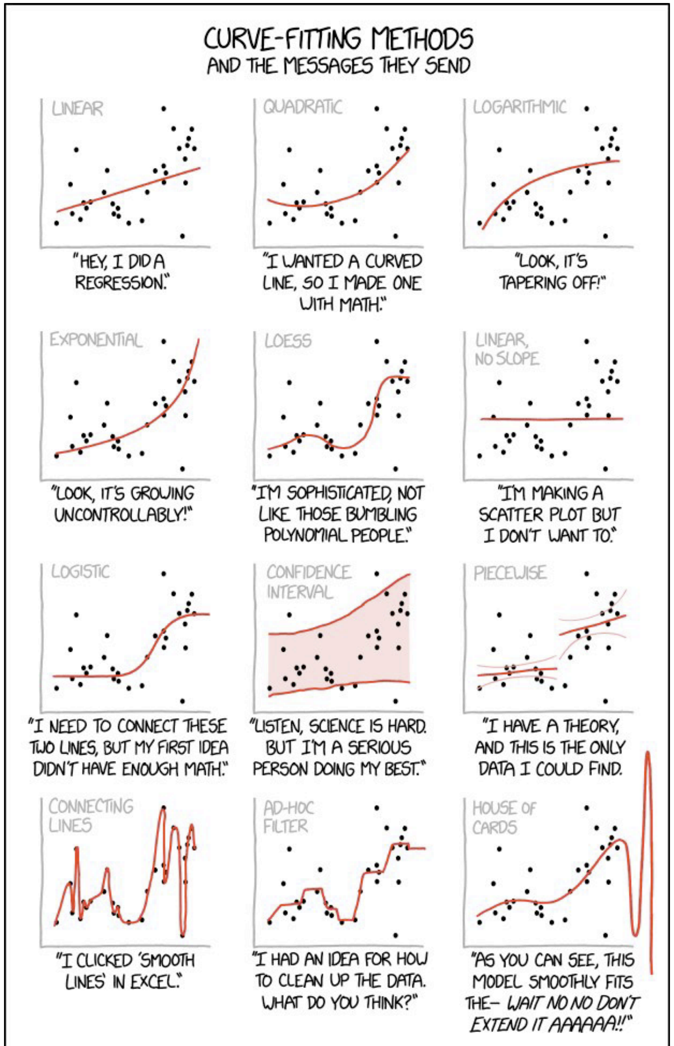
Slides adapted from Iain Styles, School of Computer Science

UNIVERSITY OF BIRMINGHAM

Intended Learning Outcome

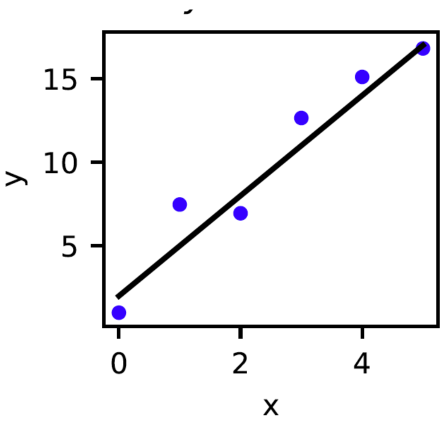
- Understand what type of problems regression is used for
- Understand and explain the concept of a loss of objective function
- Know what linear models are, and why they are linear
- Be able to implement a simple regression algorithm
- Understand and explain some issues that one may face when performing a regression analysis

What is Regression?



- Curve fitting
- Relationship between two continuous variables
- Predict the value of a dependent variable from another independent variable
- Underlying mathematical function describing a relationship given a sample of data points

Simple example



- We have a input variable x
  - Independent variable
- A function takes x and outputs y
  - Dependant variable
- How do we predict y at an x for which we have no data?
  - y(x=2.5)
- Can we use parameters of underlying function, such as gradient, intercept etc?

UNIVERSITY OF BIRMINGHAM

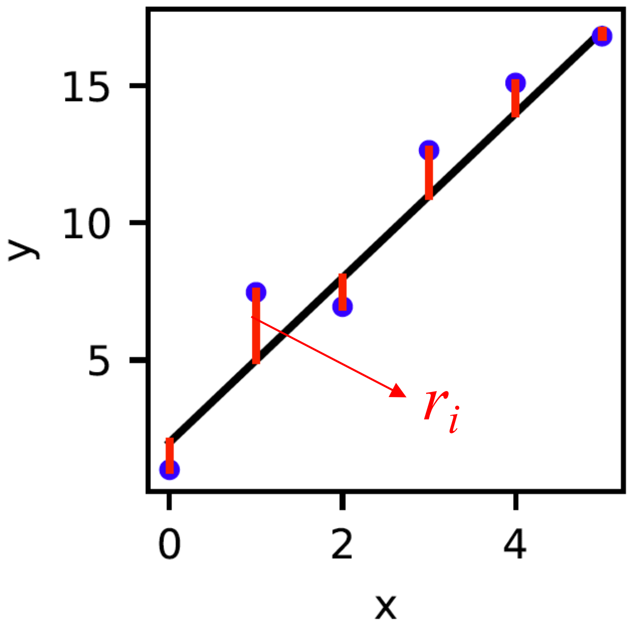
Linear Regression

- More than just a straight line fit
- Take a data set of inputs (x_i) and their corresponding output (y_i)
  - $D=\{(x_0,y_0),\ldots,(x_{N-1}, y_{N-1})\}$
  - $D=\{(x_i,y_i)\}$, *for* $i=0 : N-1$
- We want to model the relationship between x and y as a mathematical function
  - $f(\mathbf{w},x)$

Linear Regression

- So now
  - $y_i \sim= f(\boldsymbol{w}, x_i)$, for some unknown w
- We could also have noise in our data
  - $y_i = f(\boldsymbol{w}, x_i) + n$
- Our aim becomes to find w such that the function f can predict y

Using Least Squares Error (LSE)



- It is an optimization problem
  - A ‘loss/cost’ function such that it minimized the difference between measured and modelled data
- Residual $r_i(\mathbf{w})=y_i-f(\mathbf{w},x_i)$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

$$\mathcal{L}_{\mathrm{LSE}}(\mathbf{w}) = \sum_{i=0}^{N-1} r_i^2 = \mathbf{r}^{\mathrm{T}}\mathbf{r}$$

Linear models

$$f(\mathbf{w}, x) = w_0\phi_0(x) + \cdots + w_{M-1}\phi_{M-1}(x) = \sum_{i=0}^{M-1} w_i\phi_i(x).$$

- For a given input, the output is the "linear combination of basis functions ϕ_i by the free parameter w_i"
- Common choice of basis is polynomials, for example a linear (1^{st} order polynomial)

In matrix form:

$$\boldsymbol{f(w) = \phi w}, \textit{ where } \phi_{ij} = \phi_j(x_i)$$

Linear models

Therefore, $r_i = y_i - \sum_j \Phi_{ij} w_j$ or $\mathbf{r} = \mathbf{y} - \mathbf{\Phi w}$

And the LSE loss becomes $\mathcal{L}_{\mathrm{LSE}}(\mathbf{w}) = (\mathbf{y} - \mathbf{\Phi w})^{\mathrm{T}} (\mathbf{y} - \mathbf{\Phi w})$

Find $\mathbf{w} = \mathbf{w}^*$ that minimises $\mathcal{L}_{\mathrm{LSE}}(\mathbf{w})$ by differentiating w.r.t. $\mathbf{w}$ and setting to zero

- Go back and revisit Lecture on Differentiation
  - We often need to compute how the output of a function changes, when we alter a parameter by a small amount
  - So we differentiate and find the smallest change (i.e. ZERO)

Linear models

Find $\mathbf{w} = \mathbf{w}^*$ that minimises $\mathcal{L}_{\mathrm{LSE}}(\mathbf{w})$ by differentiating w.r.t. $\mathbf{w}$ and setting to zero

Start with the residuals $r_i = y_i - \sum_j \Phi_{ij} w_j$

Differentiate: $\frac{\partial r_i}{\partial w_k} = -\Phi_{ik}$

$\mathcal{L}_{\mathrm{LSE}} = \sum_i r_i^2$ and so $\frac{\mathcal{L}_{\mathrm{LSE}}}{\partial r_l} = 2r_l$

Chain rule:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathrm{LSE}}}{\partial w_k} &= \sum_l \frac{\mathcal{L}_{\mathrm{LSE}}}{\partial r_l} \times \frac{\partial r_l}{\partial w_k} \\ &= -\sum_l 2 r_l \Phi_{lk} \end{aligned}$$

UNIVERSITY OF BIRMINGHAM

Linear models

- Rearrange in matrix form

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{\mathrm{LSE}}}{\partial w_k} &= \sum_l -2 r_l \Phi_{lk} = -2 \sum_l \Phi^{\mathrm{T}}_{kl} r_l \\
\frac{\partial \mathcal{L}_{\mathrm{LSE}}}{\partial \mathbf{w}} &= -2 \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{r} = -2 \boldsymbol{\Phi}^{\mathrm{T}} \left(\mathbf{y} - \boldsymbol{\Phi} \mathbf{w} \right).
\end{aligned}
$$

- Set to zero to find the minimum

$$
\boldsymbol{\Phi}^{\mathrm{T}} \mathbf{y} - \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \mathbf{w}^* = 0
$$

Summary

- Further reading: Sections 1.1 and 3.1 of Bishop, Pattern Recognition and Machine Learning.
- A process for learning a mathematical model from data