Lecture 1: Regression

Iain Styles

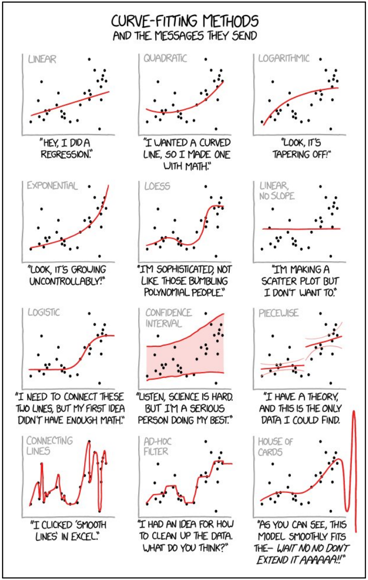
7 October 2019

Learning Outcomes

By the end of this lecture you should be able to:

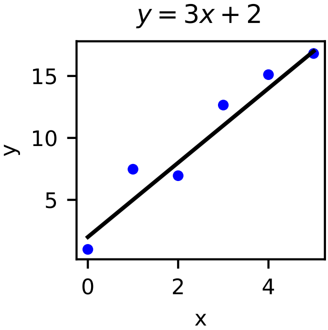
1. Understand what type of problems regression is used for
2. Understand and explain the concept of a loss of objective function
3. Know what linear models are, and why they are linear
4. Be able to implement a simple regression algorithm
5. Understand and explain some issues that one may face when performing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

What is Regression?



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

Visually...

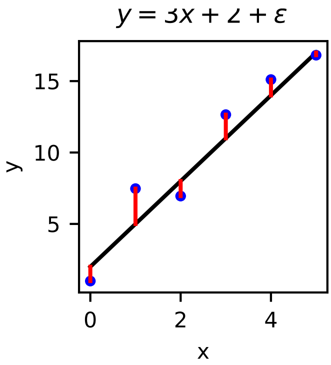


- Independent variable x
- Dependent variable y
- Predictions – $y(x = 2.5) =?$
- Parameters (intercept, gradient) of the underlying function.
- Example: $s = ut + \frac{1}{2}at^2$

Linear Regression

- Not just straight-line fitting...
- Consider a problem with one independent variable x and one dependent variable y.
- Dataset $\mathcal{D} = \{(x_0, y_0), \ldots, (x_{N-1}, y_{N-1})\} = \{(x_i, y_i)\}_{i=0}^{N-1}$
- Model the relationship between x and y as a mathematical function $f(\mathbf{w}, x)$
- $y_i \approx f(\mathbf{w}, x_i)$ with unknown parameters $\mathbf{w}$.
- Measurements of y subject to noise: $y_i = f(\mathbf{w}, x_i) + \epsilon$
- Goal: find $\mathbf{w}$ that allows f to predict y.

The Least-squares Loss



- Optimisation problem
- Define a "loss" function that measures the difference between model and data
- Find the parameters $\mathbf{w} = \mathbf{w}^*$ that minimise the loss:

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

- Residuals $r_i(\mathbf{w}) = y_i - f(\mathbf{w}, x_i)$
- Then *least square loss* is

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

Linear Models

- For simplicity we will restrict attention to *linear models*

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

- *Linear combination* of *basis functions* $\{\phi_i(x)\}_{i=0}^{M-1}$ weighted by the free parameters $\{w_i\}_{i=0}^{M-1}$
- Common choice of basis is the polynomials $\left\{x^i\right\}_{i=0}^{M-1}$
  - $\left\{x^0, x\right\}$ for a straight line
- In matrix form: $\mathbf{f}(\mathbf{w}) = \mathbf{\Phi}\mathbf{w}$ 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 $\mathcal{L}_{\mathrm{LSE}}(\mathbf{w}) = (\mathbf{y} - \mathbf{\Phi w})^{\mathrm{T}} (\mathbf{y} - \mathbf{\Phi w})$
- Unbound above, but bound by zero below so we can minimise
- 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:

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

The Normal Equations

- $$\frac{\partial \mathcal{L}_{\mathrm{LSE}}}{\partial w_k} = -\sum_l 2 r_l \Phi_{lk}$$
- Rearrange in matrix form

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

- Set to zero to find the minimum

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

- **Normal Equations**
- `https://colab.research.google.com/drive/1sHZqzkiDpLgJJmCOodGFo6D4NF9fCgIu`

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
- Simple implementation and an example of how things can fail
- Next lecture: Model selection