Data Modelling

Rohit Budhiraja

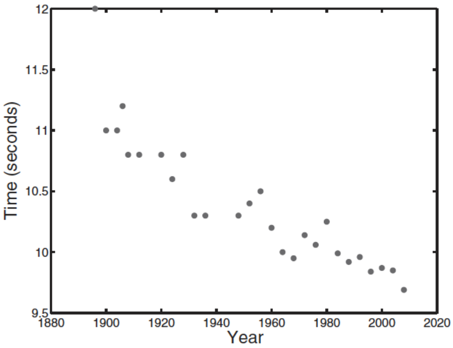
Machine Learning for Wireless Communications (EE798L)

Jan 10, 2024

Today's agenda

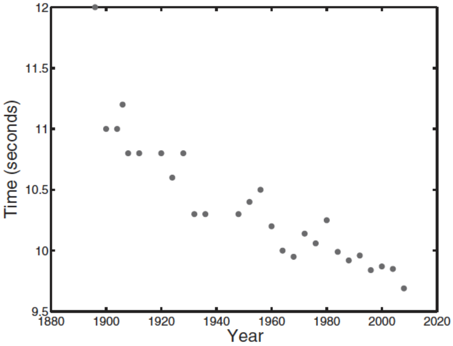
- Data modelling - lots of data and we want to fit a model to it
- Let's start by motivating data modeling
  - Share market - model and predict the market for monetary gains
  - Wireless - model and predict the channel for improving the quality of service
  - Human brain - needs for learning everything
- Today's topics: data modelling - idea and definition of a good model, and linear data modeling
  - Reference: Chapter-1 of FCML

Idea of data modelling



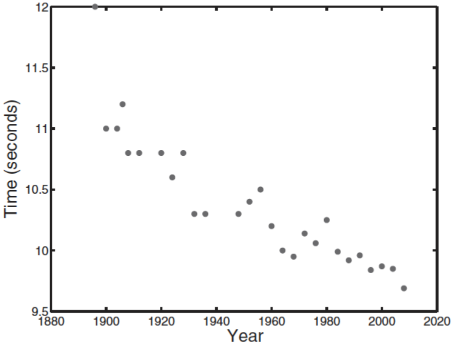
- Figure shows gold medal winning time for men's 100m at each Olympic Games held since 1896
- Aim: use this data to model functional dependence between Olympic year and 100m winning time
  - Use the model to predict winning times in future games
- Clearly the year is not the only factor that affects the winning time
  - if we are interested in using our predictions seriously we may want to take other things into account
- We can see that there is at least a statistical dependence between year and winning time

Model definition (1)



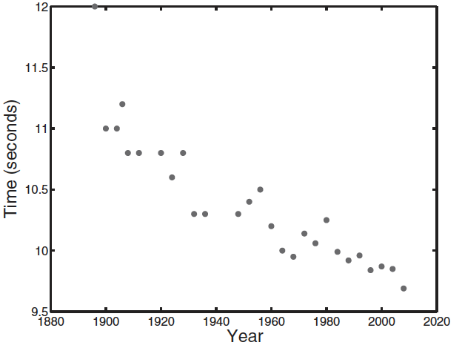
- Enough to help us introduce and develop the main ideas of linear modelling
- Modelling – learn a relationship between input attributes (Olympic year) and output (winning time)
- Many functions (models) that could be used to define this relationship
- Function takes input x (Olympic year) and will return t (winning time in sec)
  - Mathematically $t = f(x)$. For example, consider a simple linear function $f(x) = x$

Modelling assumption – linear (1)



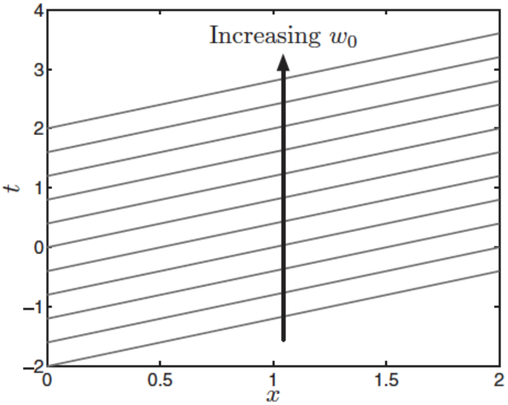
- Relationship between x and t is linear $t = f(x) = x$. What does it mean?
  - Data could be adequately modelled with a straight line
  - Winning time increases by same amount every M years
- Examining figure we can see that this assumption is not perfectly satisfied
  - Winning time reduces as the year increases

Modelling assumption – linear (2)

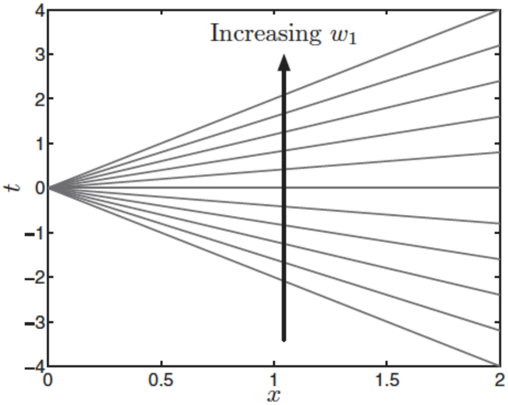


- Adding a single parameter results in $t = f(x; w) = wx$ where w can be either positive or negative
  - Enhances the model – lets us produce a straight line with any gradient through the choice of w
- This is an increase in flexibility but the flexibility is limited
  - At year 0, model predicts a winning time of $w \times 0 = 0$
  - Looking at data, we can see that this is not realistic
- Adding one more parameter to the model overcomes this limitation: $t = f(x; w_0, w_1) = w_0 + w_1 x$

Effect of varying w_0 and w_1



(a) Increasing w_0 changes the point at which the line crosses the t axis.



(b) Increasing w_1 changes the gradient of the line.

Defining a good model

- To choose values of w_0 and w_1 that are somehow best, we need to defined what best means
  - Common sense – values of w_0 and w_1 that produce a line as close as possible to all data points
- Common way of measuring how close a particular model gets to one of the data points
  - Squared difference between true winning time and winning time predicted by the model
- Using x_n, t_n to denote nth Olympic year and winning time respectively, squared difference is

$$\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) = (t_n - f(x_n; w_0, w_1))^2$$

- Squaring is important – without it, we could simply reduce it by increasing $f(x_n; w_0, w_1)$
- Expression is known as squared loss function
  - Describes how much accuracy we are losing through the use of $f(x_n; w_0, w_1)$ to model t_n
- Loss is always positive – lower the loss, better our function describes the data
- As we want a low loss for all N years, we consider average loss across the whole dataset, given as

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n(t_n, f(x_n; w_0, w_1))$$

Loss function

- Tune w_0 and w_1 to produce the model that results in lowest value of the average loss $\mathcal{L}$

$$\underset{w_0, w_1}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n(t_n, f(x_n; w_0, w_1))$$

- Minimization of squared loss is the basis of least-squares errors method of function approximation
- Squared loss is a very common choice because we can derive an analytical solution
    - Modern computational power has reduced the importance of mathematical convenience
    - No longer any excuse for choosing a convenient loss function over one more suited to the data
- Common alternative is the absolute loss:

$$\mathcal{L}_n = |t_n - f(x_n; w_0, w_1)|$$

Minimizing the squared loss – Least squares solution (1)

- We focus on squared loss function and simplify it

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n(t_n, f(x_n; w_0, w_1)) = \frac{1}{N}\sum_{n=1}^{N}(t_n - f(x_n; w_0, w_1))^2 = \frac{1}{N}\sum_{n=1}^{N}(t_n - (w_0 + w_1 x_n))^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n w_0 - 2w_1 x_n t_n + w_0^2 - 2w_0 t_n + t_n^2) \\
&= \frac{1}{N}\sum_{n=1}^{N}(w_1^2 x_n^2 + 2w_1 x_n (w_0 - t_n) + w_0^2 - 2w_0 t_n + t_n^2)
\end{aligned}
$$

- Need to calculate w_1 and w_0 to minimize the loss function
- Calculate partial derivatives of loss function wrt w_1 and w_0 and set them to zero

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_1} &= \frac{2w_1}{N}\left(\sum_{n=1}^{N} x_n^2\right) + \frac{2}{N}\left(\sum_{n=1}^{N} x_n (w_0 - t_n)\right) \\
\frac{\partial \mathcal{L}}{\partial w_0} &= 2w_0 + \frac{2w_1}{N}\left(\sum_{n=1}^{N} x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N} t_n\right)
\end{aligned}
$$

Least squares solution -calculating value of w_0 and w_1

- Setting $\frac{\partial \mathcal{L}}{\partial w_0} = 0$ we have

$$\widehat{w_0} = \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right) - \frac{w_1}{N}\left(\sum_{n=1}^{N} x_n\right) = \bar{t} - w_1\bar{x}$$

- Denote avg. winning time $\bar{t} = \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)$,avg. Olympic year as $\bar{x} = \frac{1}{N}\left(\sum_{n=1}^{N} x_n\right)$. We have

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_1} &= \frac{2w_1}{N}\left(\sum_{n=1}^{N} x_n^2\right) + \frac{2}{N}\left(\sum_{n=1}^{N} x_n(\widehat{w_0} - t_n)\right) \\
&= \frac{2w_1}{N}\left(\sum_{n=1}^{N} x_n^2\right) + \frac{2}{N}\left(\sum_{n=1}^{N} x_n(\bar{t} - w_1\bar{x} - t_n)\right) \\
&= \frac{2w_1}{N}\left(\sum_{n=1}^{N} x_n^2\right) + \frac{2\bar{t}}{N}\left(\sum_{n=1}^{N} x_n\right) - \frac{2w_1\bar{x}}{N}\left(\sum_{n=1}^{N} x_n\right) - \frac{2}{N}\left(\sum_{n=1}^{N} x_n t_n\right) \\
&= 2w_1\left[\frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right) - \bar{x}\bar{x}\right] + 2\bar{t}\bar{x} - \frac{2}{N}\left(\sum_{n=1}^{N} x_n t_n\right)
\end{aligned}$$

Least squares solution -calculating value of w_1

- Setting it to zero

$$
\begin{aligned}
2w_1\left[\frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right) - \bar{x}\bar{x}\right] &+ 2\bar{t}\bar{x} - 2\frac{1}{N}\left(\sum_{n=1}^{N} x_n t_n\right) = 0 \\
2w_1\left[\frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right) - \bar{x}\bar{x}\right] &= 2\frac{1}{N}\left(\sum_{n=1}^{N} x_n t_n\right) - 2\bar{t}\bar{x} \\
\widehat{w_1} &= \frac{\frac{1}{N}\left(\sum_{n=1}^{N} x_n t_n\right) - \bar{t}\bar{x}}{\frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right) - \bar{x}\bar{x}}
\end{aligned}
$$

- By denoting $\overline{x^2} = \frac{1}{N}\left(\sum_{n=1}^{N} x_n^2\right)$ and $\overline{xt} = \frac{1}{N}\left(\sum_{n=1}^{N} x_n t_n\right)$, we have

$$
\widehat{w_1} = \frac{\overline{xt} - \bar{x}\bar{t}}{\overline{x^2} - (\bar{x})^2}
$$

Calculation for Olympic data– least squares fit

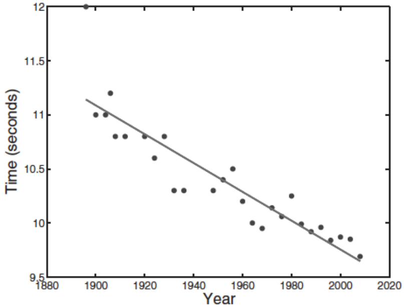
n	x_n	t_n	$x_n t_n$	x_n^2
1	1896	12.00	22752.0	3.5948×10^6
2	1900	11.00	20900.0	3.6100×10^6
3	1904	11.00	20944.0	3.6252×10^6
4	1906	11.20	21347.2	3.6328×10^6
5	1908	10.80	20606.4	3.6405×10^6
6	1912	10.80	20649.6	3.6557×10^6
7	1920	10.80	20736.0	3.6864×10^6
8	1924	10.60	20394.4	3.7018×10^6
9	1928	10.80	20822.4	3.7172×10^6
10	1932	10.30	19899.6	3.7326×10^6
11	1936	10.30	19940.8	3.7481×10^6
12	1948	10.30	20064.4	3.7947×10^6
13	1952	10.40	20300.8	3.8103×10^6
14	1956	10.50	20538.0	3.8259×10^6
15	1960	10.20	19992.0	3.8416×10^6
16	1964	10.00	19640.0	3.8573×10^6
17	1968	9.95	19581.6	3.8730×10^6
18	1972	10.14	19996.1	3.8888×10^6
19	1976	10.06	19878.6	3.9046×10^6
20	1980	10.25	20295.0	3.9204×10^6
21	1984	9.99	19820.2	3.9363×10^6
22	1988	9.92	19721.0	3.9521×10^6
23	1992	9.96	19840.3	3.9681×10^6
24	1996	9.84	19640.6	3.9840×10^6
25	2000	9.87	19740.0	4.0000×10^6
26	2004	9.85	19739.4	4.0160×10^6
27	2008	9.69	19457.5	4.0321×10^6
$(1/N)\sum_{n=1}^{N}$	1952.37	10.39	20268.1	3.8130×10^6

- Calculation of weights

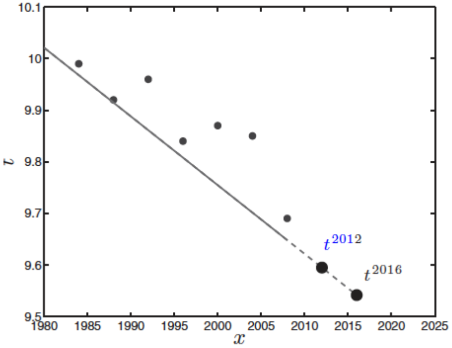
$$w_1 = \frac{20268.1 - 1952.37 \times 10.39}{3.8130 \times 10^6 - 1952.37 \times 1952.37} = -0.0133$$

$$w_0 = 10.39 - (-0.0133) \times 1952.37 = 36.416$$

- Function is $f(x; w_0, w_1) = 36.416 - 0.013x$
- Plot of function:



Making predictions using the developed model (1)

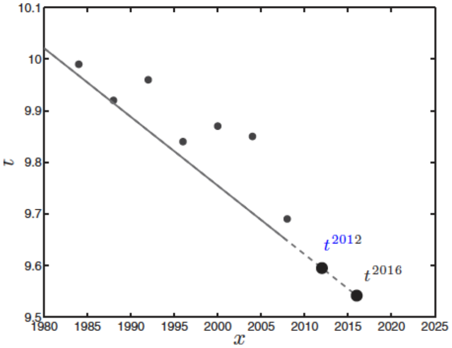


- Predict winning time for a year that we have not yet observed – for 2012 and 2016

$$
\begin{aligned}
f(x; w_0 = 36.416, w_1 = -0.0133) &= 36.416 - 0.0133x \\
t^{2012} = f(2012; w_0, w_1) &= 36.416 - 0.0133 \times 2012 = 9.595s \\
t^{2016} = f(2016; w_0, w_1) &= 36.416 - 0.0133 \times 2016 = 9.541s
\end{aligned}
$$

- Values are very precise

Making predictions using the developed model (2)



- Unlikely that any model can predict outcome of a complex event to such a high degree of accuracy
  - Least of all one based on nothing more than a straight line
- Our model is not even able to predict data that it has seen very precisely
  - Can be seen by the distance of some points to the line
- Precise predictions are of limited use where our model is not perfect (almost all situations)
  - More useful to express a range of values rather than any particular one