Linear modelling: Maximum likelihood approach

Rohit Budhiraja

Machine Learning for Wireless Communications (EE 798L)

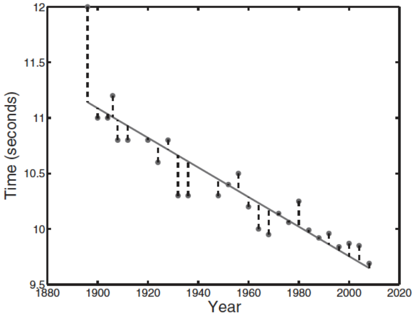
Jan 17, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Discuss generalization and over-fitting
  - Discussed how cross validation and regularized least squares helps avoid overfitting
- Limitation of above model: predicts with absolute certainty
- Today's agenda
  - Learn about generative data modelling and its advantages in expressing the uncertainity of prediction
- Reference Chapter 2 of FCML

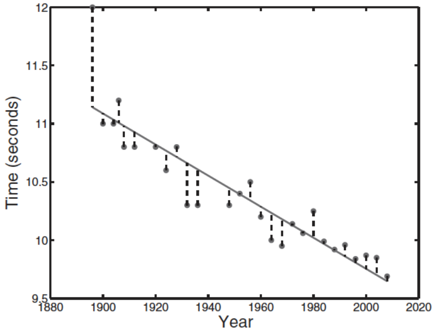
Modelling errors as noise

- Recall we minimized squared loss function to model Olympic 100 m data with a linear model



- Linear model captures downward trend but
  - there are errors between the model and true values, which are highlighted
- Our model assumed a linear relationship between years and winning times
- Model captures general trend in data, but ignores deviation between model and observed data
  - Ignoring these errors is hard to defend from modelling perspective

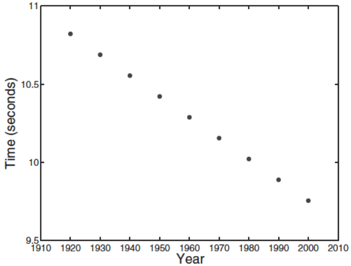
Benefits of modelling errors as noise



- Model error as noise
- Allow us to express level of uncertainty in estimate of model parameters **w**
  - If we change **w** a bit, do we still have a good model?
- Allow us to express a degree of uncertainty in our predictions
  - We believe the winning time will be between ‘a’ and ‘b’ rather than ‘we believe it will be exactly c’.
- Change in view point: think of our modelling problem as a generative one:
  - Can we build a model that could be used to create (or generate) a dataset that looks like ours?

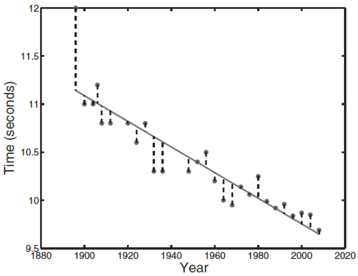
Thinking generatively (1)

- Process that generated this particular dataset is very complex
  - Includes sprinters and the events surrounding their preparation and performance
- We accept that this isn't how data was generated, but we shall see that this is a useful strategy
- How do we generate data from our current model?
  - We have an equation $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$
  - Substitute $\mathbf{w}$ calculated earlier, and it could generate a winning time for any particular year



- Figure shows winning times generated in this way for a number of years between 1920 and 2000
- It doesn't look much like the original data. To make it more realistic, we need to add some errors.

Important features of errors for generative data modeling



- Errors are different in each year– some positive, some negative and all have different magnitudes
- No obvious relationship between the size (or direction) of the error and the year
  - Error does not appear to be a function of x, the Olympic year
- If we generate a random amount of time (in seconds) that could be either positive or negative and was, on average, roughly the same size as above errors
- We could generate one such value for each data point we wished to generate, and add it to $\mathbf{w}^T\mathbf{x}$

Generative data model with error modeling

- Our model now takes the following form

$$t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$$

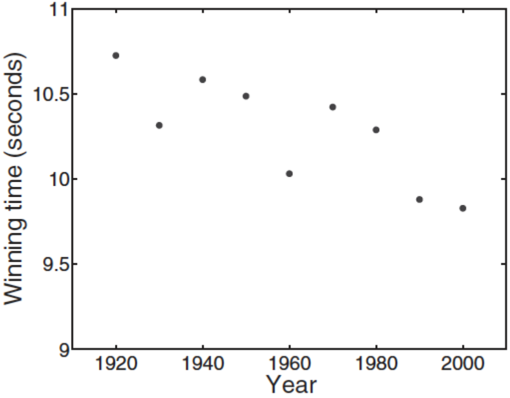
- error ϵ_n – difference between model and actual winning times is a continuous random variable
  - Need to decide its distribution
- Also we do not just have one random variable, but one for each observed Olympic year
  - It seems reasonable to assume that these values are independent:

$$p(\epsilon_1, \cdots, \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n)$$

- We will assume that ϵ_n is Gaussian distributed with pdf $\mathcal{N}(0, \sigma^2)$
  - Distribution allows ϵ_n to be both positive and negative
  - Aside: Distribution has interesting modelling properties that link it to the squared loss discussed earlier
- Model summary: model now consists of two components:
  - Deterministic component ($\mathbf{w}^T \mathbf{x}_n$), sometimes referred to as a trend or drift
  - Random component (ϵ_n), referred to as noise

Olympic data with Gaussian errors

- With $\epsilon_n \sim \mathcal{N}(0, 0.05)$ (don't worry about the particular variance value here for now)



- We obtain a much more realistic looking dataset

Calculation of parameters in generative data modeling

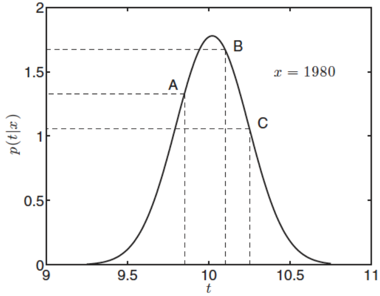
- We need to calculate optimal value of $\mathbf{w}$ and additional parameter σ^2
- $\mathbf{w}$ was earlier calculated to minimize the loss
    - Loss measured difference between observed values of t and those predicted by the model
- With random added noise, t_n is now itself a random variable
    - No single value of t_n for a particular x_n – we cannot use the loss as a means of optimising $\mathbf{w}$ and σ^2
- Basic probability result:
    - If z is a random variable with $p(z) = \mathcal{N}(m, s)$ and $y = a + z$, then $p(y) = \mathcal{N}(m + a, s)$
- Random variable t_n for our model $t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$ has pdf

$$p(t_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$$

- Note conditioning on LHS – pdf of t_n depends on particular values of $\mathbf{x}_n$ and $\mathbf{w}$
    - All t_n are independent, given the conditioning, as noise at each data point is independent
- We will see how to use pdf to calculate optimal values of $\mathbf{w}$ and σ^2

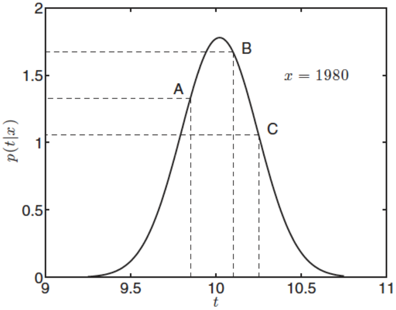
Idea of "likelihood" (1)

- Consider year 1980 from our dataset. We earlier calculated $\mathbf{w} = [36.416, -0.0133]^T$
- pdf of t_n is $p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$
  - With mean $\mathbf{w}^T\mathbf{x}_n = [36.416, -0.0133][1, 1980]^T = 10.02$
- If we assume $\sigma^2 = 0.05$, then pdf of t_n is



- For a continuous random variable, t, $p(t)$ cannot be interpreted as a probability
- Interpretation of height of the curve at a particular value of t
  - How likely it is that we would observe that particular t for $x = 1980$
  - Implies, most likely winning time in 1980 would be 10.02 seconds

Idea of “likelihood” (2)



- But actual winning time in 1980 Olympics is C (10.25 seconds)
- Density $p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2)$ at $t_n = 10.25$ is an important quantity – likelihood of nth data point
- We cannot change $t_n = 10.25$ (this is our data) but we can change $\mathbf{w}$ and σ^2 to move the pdf so as to make it as high as possible at $t_n = 10.25$

Dataset likelihood

- In general, we are not interested in the likelihood of a single data point but that of complete data
- If we have N data points, we are interested in the joint conditional pdf:

$$L = p(t_1, \cdots, t_N | \mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{w}, \sigma^2) = p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \sigma^2)$$

- Evaluating this pdf at the observed data points gives a single likelihood value for the whole dataset, which we can optimise by varying $\mathbf{w}$ and σ^2
- Recall we assume that the noise at each data point is independent for $t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$

$$\begin{aligned} L &= p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T \mathbf{x}_n, \sigma^2) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (t_n - \mathbf{w}^T \mathbf{x}_n)^2 \right\} \end{aligned}$$

- Dataset likelihood can be thought of a generative distribution for dataset
- Will now calculate the values of $\mathbf{w}$ and σ^2 that maximises the likelihood

Maximum likelihood approach to calculate $\mathbf{w}$ and σ^2

- For analytical reasons, we will maximise the natural logarithm of the likelihood

$$
\begin{aligned}
L &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right\} \\
\log L &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right\}\right) \\
&= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right) \\
&= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2 \\
&\overset{(a)}{=} -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{t})^T(\mathbf{X}\mathbf{w} - \mathbf{t})
\end{aligned}
$$

- Equality (a) is derived in earlier lecture by defining

Weight calculation for maximum likelihood approach

- $$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$
- Calculate optimal parameters by taking derivatives and equating them to zero

$$\begin{aligned} \frac{\partial \log L}{\partial \mathbf{w}} &= \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{t} - \mathbf{X}^T\mathbf{X}\mathbf{w}) = \mathbf{0} \\ \Rightarrow \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \end{aligned}$$

- This solution is exactly that the same as that of least squares case
- Minimising the squared loss is equivalent to the maximum likelihood solution
    - If noise is assumed to be Gaussian, otherwise not