Linear modelling: Maximum likelihood approach (2)

Rohit Budhiraja

Machine Learning for Wireless Communications (EE 798L)

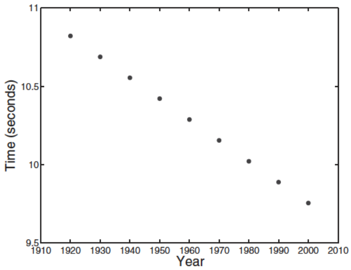
Jan 19, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Discussed generative data modelling approach
    - Approach will help in capturing uncertainty in prediction
- Today's agenda
  - Show that maximum likelihood (ML) model favours complex models
  - Show how parameter estimation in ML model becomes zero forcing receiver in wireless
- Reference Chapter 2 of FCML

Generative data modelling (recap)

- 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



- It doesn't look much like the original data. To make it more realistic, we need to add some errors
- 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

Generative data modelling (recap)

- Error ϵ_n is modelled as continuous random variable, and also independent across Olympic years

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

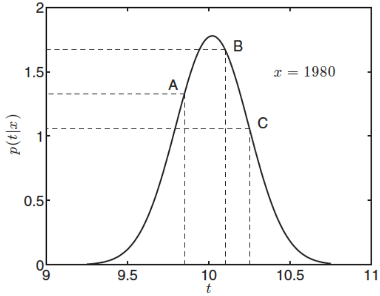
- Assumed ϵ_n to be Gaussian distributed with pdf $\mathcal{N}(0, \sigma^2)$
  - Distribution allows ϵ_n to be both positive and negative
- Model has two components: deterministic ($\mathbf{w}^T\mathbf{x}_n$) and random (ϵ_n), which we need to calculate
- Random variable t_n for our model $t_n = \mathbf{w}^T\mathbf{x}_n + \epsilon_n$ 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 $\mathbf{w}$
  - Also all t_n (conditioned) are independent as noise at each data point is independent

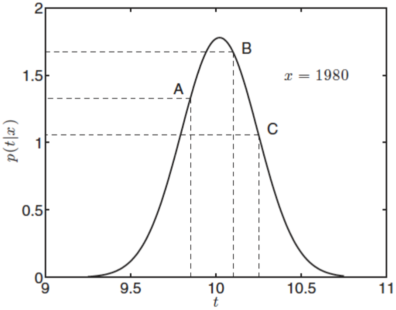
Idea of "likelihood" (recap)

- 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” (recap)



- 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 try and move the pdf so as to make it as high as possible at $t_n = 10.25$

Dataset likelihood and calculation of parameters (recap)

- If we have N data points, we maximize their joint likelihood while calculating $\mathbf{w}$ and σ^2

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

- Recall we assume that the noise at each data point is independent for $t_n = \mathbf{w}^T \mathbf{x}_n + \epsilon_n$

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

- Values of $\mathbf{w}$ that maximises the likelihood

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

Variance calculation for maximum likelihood approach

- We now calculate an expression for σ^2 assuming $\mathbf{w} = \hat{\mathbf{w}}$

$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(t_n - \mathbf{x}^T\hat{\mathbf{w}})^2 = 0$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^T\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t} - \mathbf{X}\hat{\mathbf{w}})^T(\mathbf{t} - \mathbf{X}\hat{\mathbf{w}}) = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\mathbf{X}\hat{\mathbf{w}} + \hat{\mathbf{w}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}})$$

- Above approach is also known as parameter estimation problem for a given distribution
- By substituting $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$, we have

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} + \mathbf{t}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}) \\ &= \frac{1}{N}(\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} + \mathbf{t}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}) \\ &= \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}) = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})\end{aligned}$$

- Using the Olympic 100m data, $\hat{\mathbf{w}} = [36.4165 \;\; -0.0133]^T$ and $\widehat{\sigma}^2 = 0.0503$

Maximum likelihood favours complex models (1)

- Log likelihood expression is

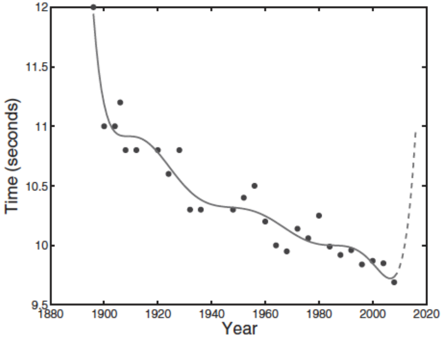
$$\log L = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2 \qquad (1)$$

- Substituting $\mathbf{w} = \hat{\mathbf{w}}$ and $\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2$ into (1) maximizes log likelihood

$$\log L = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} = -\frac{N}{2}(1 + \log 2\pi) - \frac{N}{2}\log\widehat{\sigma^2}$$

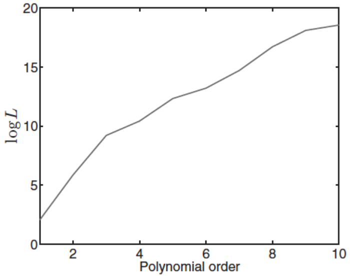
- Note: for fixed data N, maximum value of L will keep increasing as we decrease noise variance σ^2
  - Noise is incorporated into model to capture effects which its deterministic part (i.e. $\mathrm{f}(\mathbf{x};\mathbf{w})$) cannot
- One way to decrease σ^2 is to modify $\mathrm{f}(\mathbf{x};\mathbf{w})$ so that it can capture more variability in data
  - i.e., make it more flexible – by fitting increasingly higher-order polynomial function

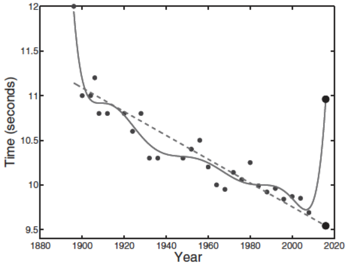
Maximum likelihood favours complex models (2)



- Eighth-order polynomial gets closer to observed data than first-order polynomial
- More complex model is overfitting
  - We have given it too much freedom and it is attempting to make sense out of what is essentially noise

Maximum likelihood favours complex models (3)





- If we use $\log L$ to choose model order, it would always point us to models of increasing complexity
- Simpler model is better able to generalize than the complex one
- Showed how regularisation could be used to penalize over-complex parameter values
  - Same can be done with Bayesian approach through use of prior distributions on parameter values

Maximum likelihood approach in a form suitable for wireless

- Recall our data model form $t_n = \mathbf{w}^T\mathbf{x}_n + \epsilon_n$. We therefore have

$$
\begin{aligned}
t_1 &= w_0 + x_1 w_1 + \epsilon_1 \\
t_2 &= w_0 + x_2 w_1 + \epsilon_2 \\
\vdots &= \vdots \\
t_N &= w_0 + x_N w_1 + \epsilon_N \\
\mathbf{t} &= \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}
\end{aligned}
$$

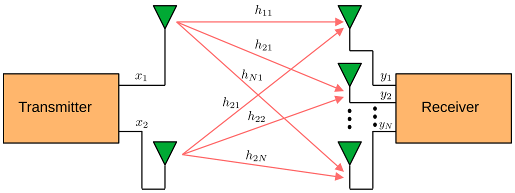
  where $\boldsymbol{\epsilon} = [\epsilon_1 \epsilon_2, \dots, \epsilon_N]^T$
- Joint likelihood of N data points

$$
\begin{aligned}
L &= p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = p(t_1, \cdots, t_N|\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2) \\
&= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right\}
\end{aligned}
$$

- Calculated $\mathbf{w}$ by maximizing the natural logarithm of the likelihood $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$

Multiple-input multiple-output (MIMO) wireless systems

- Consider a transmitter with 2 antennas and receiver with N antennas.



- MIMO system simultaneously transmits 2 different symbols. Received signal is

$$
\begin{aligned}
y_1 &= h_{11}x_1 + h_{12}x_2 + n_1 \\
y_2 &= h_{21}x_1 + h_{22}x_2 + n_2 \\
\vdots &= \vdots \\
y_N &= h_{N1}x_1 + h_{N2}x_2 + n_N \\
\mathbf{y} &= \mathbf{H}\mathbf{x} + \mathbf{n}
\end{aligned}
$$

- Receive signal vector $\mathbf{y} = [y_1, \cdots, y_N]^T$, transmit signal $\mathbf{x} = [x_1, x_2]^T$, receiver noise $\mathbf{n} = [n_1, \cdots, n_N]^T$

Multiple-input multiple-output (MIMO) wireless systems

- Channel

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ \vdots & \vdots \\ h_{N1} & h_{N2} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \vdots \\ \mathbf{h}_N^T \end{bmatrix}$$

- Receive signal vector is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$$

- Two symbols in transmit vector $\mathbf{x}$ interfere with each other at receiver – need to recover $\mathbf{x}$ from $\mathbf{y}$

MIMO systems and machine learning

- Receive signal vector is
$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$$
- Joint likelihood of N receive signals (data points)

$$\begin{aligned} L &= p(\mathbf{y}|\mathbf{H}, \mathbf{x}, \sigma^2) = p(y_1, \cdots, y_N|\mathbf{h}_1, \cdots, \mathbf{h}_N, \mathbf{x}, \sigma^2) = \prod_{n=1}^{N} p(y_n|\mathbf{h}_n, \mathbf{x}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}^T\mathbf{h}_n, \sigma^2) \\ &= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mathbf{x}^T\mathbf{h}_n)^2\right\} \end{aligned}$$

- Calculate $\mathbf{x}$ by maximizing the natural logarithm of the likelihood $\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$
- We denote $\mathbf{W} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ (zero forcing receiver)

$$\begin{aligned} \mathbf{W}\mathbf{y} = \hat{\mathbf{x}} &= \mathbf{W}\mathbf{H}\mathbf{x} + \mathbf{W}\mathbf{n} \\ &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\mathbf{x} + \underbrace{\mathbf{W}\mathbf{n}}_{\tilde{\mathbf{n}}} = \mathbf{x} + \tilde{\mathbf{n}} \end{aligned}$$