Uncertainty in Prediction

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Jan 24, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Showed how parameter estimation in ML model becomes zero forcing receiver in wireless
- Today's class
  - Understand how generative modeling approach will provide uncertainty in prediction
  - Reference is Chap 2 of FCML

Uncertainty in prediction

- Our model responsible for generating the data

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

- Calculated $\hat{\mathbf{w}}$ by maximizing the natural logarithm of the likelihood $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$
  - Approach is called maximum likelihood estimation
  - $\hat{\mathbf{w}}$ is a deterministic function of random variable $\mathbf{t}$, and is thus also a random variable
- We also estimated noise variance σ^2 using maximum likelihood approach
- Suppose we observe a new input $\mathbf{x}_{new}$, we would like to predict the output, t_{new}
- To predict t_{new}, we multiply $\mathbf{x}_{new}$ by the best set of model parameters, $\widehat{\mathbf{w}}$ i.e., $t_{new} = \widehat{\mathbf{w}}^T\mathbf{x}_{new}$
- Since t_{new} is function of random vector $\widehat{\mathbf{w}}$, it is a random variable
  - We calculate the prediction variability by calculating σ^2_{new}, which is also called the predictive variance
- Understand uncertainty in prediction as two-step process
- Step1: Uncertainty in parameter estimate $\hat{\mathbf{w}}$
- Step2: Uncertainty in $\hat{\mathbf{w}}$ will help in capturing uncertainty in prediction
  - Uncertainty in $\hat{\mathbf{w}}$ is mathematically captured using covariance matrix
- Covariance matrix of $\hat{\mathbf{w}}$ is

$$cov\{\hat{\mathbf{w}}\} = \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\} - \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\hat{\mathbf{w}}\}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\hat{\mathbf{w}}\}^T$$

Gaussian random vector (recap)

- Density of Gaussian vector

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- $\boldsymbol{\mu}$ is the mean vector (same size as $\mathbf{x}$) and $\boldsymbol{\Sigma}$ is covariance matrix

$$\begin{aligned}\boldsymbol{\mu} &= [2,1]^T, \boldsymbol{\Sigma} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\\ \boldsymbol{\mu} &= [2,1]^T, \boldsymbol{\Sigma} = \begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix}\end{aligned}$$

- Gaussian vector with $\boldsymbol{\Sigma} = \mathbf{I}$

$$\begin{aligned}p(\mathbf{x}) &= \frac{1}{(2\pi)^{N/2}|\mathbf{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{I}(\mathbf{x}-\boldsymbol{\mu})\right\} = \frac{1}{(2\pi)^{N/2}|\mathbf{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu_n)^2\right\}\\ &= \frac{1}{(2\pi)^{N/2}|\mathbf{I}|^{\frac{1}{2}}} \prod_{n=1}^{N} \exp\left\{-\frac{1}{2}(x_n-\mu_n)^2\right\} = \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x_n-\mu_n)^2\right\}\end{aligned}$$

- Elements of $\mathbf{x}$ are independent with $p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$

Uncertainty in parameter estimate $\hat{\mathbf{w}}$

- For our generative model $t_n = \mathbf{w}^T\mathbf{x}_n + \epsilon_n$

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

- $p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)$ is the generating distribution (or likelihood). We have

$$\begin{aligned}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\hat{\mathbf{w}}\} &= \int \hat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)d\mathbf{t} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \int \mathbf{t}p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)d\mathbf{t} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{w}\end{aligned}$$

- Expectation of $\hat{\mathbf{w}}$ w.r.t. generating distribution will tell us what $\hat{\mathbf{w}}$ on an average will be
- Expected value of $\hat{\mathbf{w}}$ is the true parameter value
  - Estimator, on an average, is nether too big or small – estimator is unbiased
- Covariance matrix of $\hat{\mathbf{w}}$ now is

$$\begin{aligned}cov\{\hat{\mathbf{w}}\} &= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\} - \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\hat{\mathbf{w}}\}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\hat{\mathbf{w}}\}^T \\ &= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\} - \mathbf{w}\mathbf{w}^T\end{aligned}$$

Covariance matrix calculation (2)

- We next simplify the first term

$$\begin{aligned}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\} &= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t})^T\right\} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^T\right\}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \qquad (1)\end{aligned}$$

- We know that $p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$ such that mean of $\mathbf{t}$ is $\mathbf{X}\mathbf{w}$ and covariance $\sigma^2\mathbf{I}$

$$\begin{aligned}cov\{\mathbf{t}\} = \sigma^2\mathbf{I} &= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^T\right\} - \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}^T \\ \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^T\right\} &= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\}^T + \sigma^2\mathbf{I} \\ &= \mathbf{X}\mathbf{w}(\mathbf{X}\mathbf{w})^T + \sigma^2\mathbf{I} = \mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T + \sigma^2\mathbf{I}\end{aligned}$$

- Substituting $\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^T\right\}$ into (1), which is

$$\begin{aligned}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^T\right\}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \mathbf{w}\mathbf{w}^T + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

- Finally, we have

$$cov\{\hat{\mathbf{w}}\} = \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\} - \mathbf{w}\mathbf{w}^T = \mathbf{w}\mathbf{w}^T + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} - \mathbf{w}\mathbf{w}^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Variance calculation

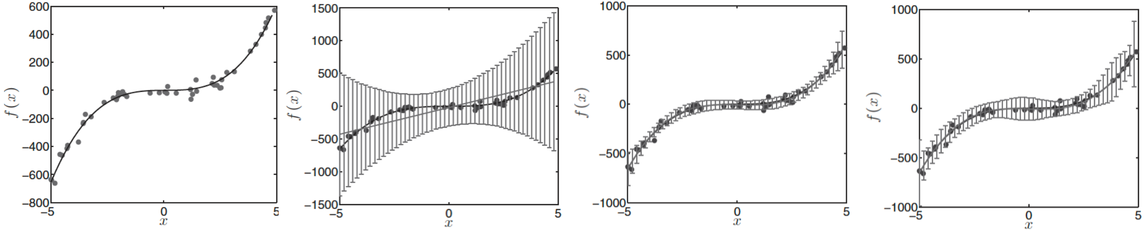
- We now calculate the predictive variance $\sigma^2_{new} = var\{t_{new}\}$, where $t_{new} = \hat{\mathbf{w}}^T\mathbf{x}_{new}$

$$
\begin{aligned}
\sigma^2_{new} &= var\{t_{new}\} = \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{t^2_{new}\} - (\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{t_{new}\})^2 \\
&= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^T\mathbf{x}_{new})^2\right\} - (\mathbf{w}^T\mathbf{x}_{new})^2 \\
&= \mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}^T_{new}\hat{\mathbf{w}}\hat{\mathbf{w}}^T\mathbf{x}_{new}\right\} - \mathbf{x}^T_{new}\mathbf{w}\mathbf{w}^T\mathbf{x}_{new} \\
&= \mathbf{x}^T_{new}\mathbb{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\hat{\mathbf{w}}^T\right\}\mathbf{x}_{new} - \mathbf{x}^T_{new}\mathbf{w}\mathbf{w}^T\mathbf{x}_{new} \\
&= \mathbf{x}^T_{new}(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{w}\mathbf{w}^T)\mathbf{x}_{new} - \mathbf{x}^T_{new}\mathbf{w}\mathbf{w}^T\mathbf{x}_{new} \\
&= \sigma^2\mathbf{x}^T_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new} + \mathbf{x}^T_{new}\mathbf{w}\mathbf{w}^T\mathbf{x}_{new} - \mathbf{x}^T_{new}\mathbf{w}\mathbf{w}^T\mathbf{x}_{new} \\
&= \mathbf{x}^T_{new}cov\{\hat{\mathbf{w}}\}\mathbf{x}_{new}
\end{aligned}
$$

- To summarize, we have

$$
\begin{aligned}
t_{new} &= \widehat{\mathbf{w}}^T\mathbf{x}_{new} = \mathbf{x}^T_{new}\widehat{\mathbf{w}} = \mathbf{x}^T_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \\
\sigma^2_{new} &= \mathbf{x}^T_{new}cov\{\hat{\mathbf{w}}\}\mathbf{x}_{new} = \sigma^2\mathbf{x}^T_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}
\end{aligned}
$$

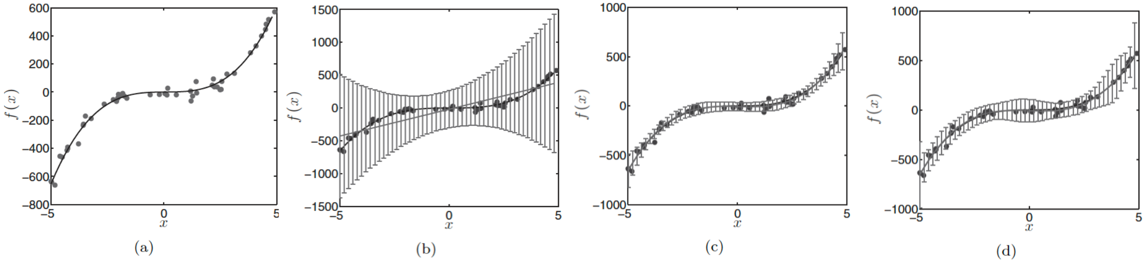
Predictive variability an example (1)



(a) (b) (c) (d)

- Figure (a) shows the function $f(x) = 5x^3 - x^2 + x$ and data points sampled from this function and corrupted by Gaussian noise with mean zero and variance 1000
- Figures (b), (c) and (d) show $t_{new} \pm \sigma^2$ new for linear, cubic and sixth order models
- Linear model has very high predictive variance
  - Unable to model deterministic trend in data very well, and assumes much of data variation as noise

Predictive variability an example (2)



- Cubic model is better able to model the trend
  - It is the correct order and this is reflected in its much more confident predictions
- Sixth-order model is over-complex
  - It has too much freedom and can therefore fits the data well for quite a large range of parameter values
- For all models, predictive variance increases as we move towards the edge of data
- Model is less confident in areas where it has less data – an appealing property

Summary and next agenda

- Summary till now
  - Generative modeling approach tells us how confident the model is about the predictions it is making
  - Maximum likelihood approach favors complex models
- Next agenda
  - Bayesian approach, similar to regularization, can avoid complex models[1]
  - Bayesian approach also allows us to incorporate our prior belief about the model
- Let's re-discuss Facebook example[2]
- We will next see another example of how data can give misleading information

[1] Chap 3 of FCML

[2] The Chaos Machine: The Inside Story of How Social Media Rewired Our Minds and Our World, book by Max Fischer