Lecture 03

Random Variables, Probability and Likelihood

May 23, 2025

1 Generalization and Overfitting

Fitting a model perfectly to the training data leads to poor predictions as there will almost be noise present.

There is a common trade-off between generalization and overfitting. We would prefer to generalize our model instead of overfitting it so that it can be practical.

Here is how a generalized and overfitted model looks like (the data points used are not the best to show an overfitted model..):

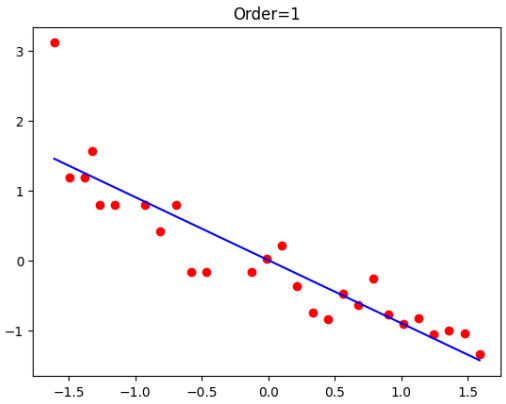


Figure 1: Generalized model

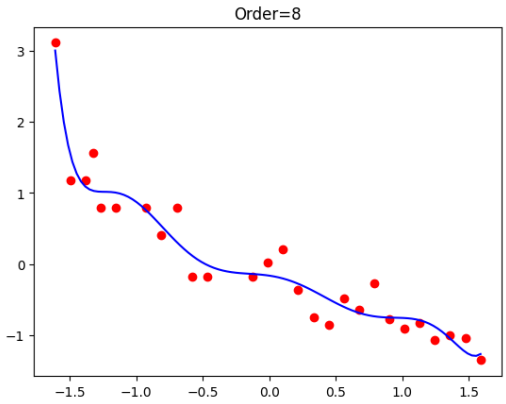
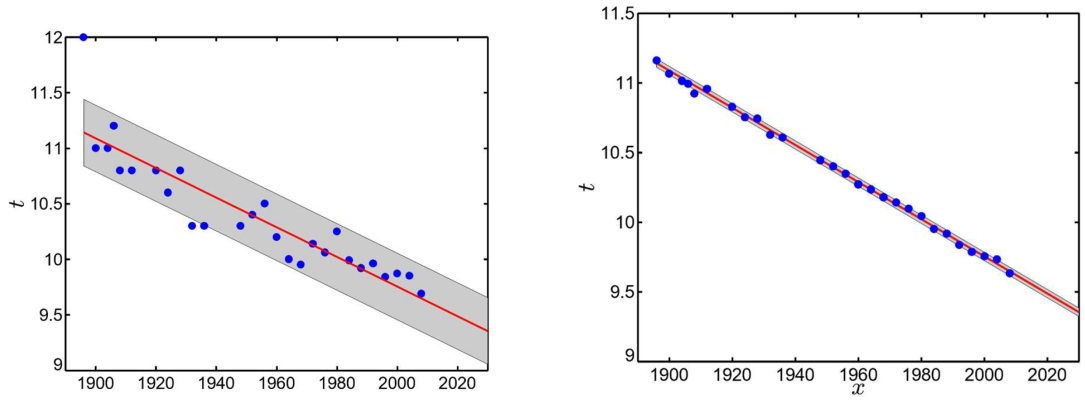


Figure 2: Overfitted model

2 Modelling Errors

Errors exist in all models, and we should not ignore them. Instead, we can use them as an indication of how confident are the model predictions. That is, how certain we feel about the predicted values.



The narrower the shaded region, the more **precise** the model predictions are, **but not necessarily accurate**.

2.1 Additive Errors

We can include noise in our training models by adding an additive term:

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

Since noise is random, we can model it as a random variable. We can model noise in a Gaussian (normal) distribution as below:

$$p(\epsilon \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\epsilon - \mu)^2\right\}$$

By modelling noise using the Gaussian distribution, we are able to quantify uncertainty, generate synthetic data and use probabilistic interpretations like confidence intervals and **likelihood**. (On a side note, modelling noise using Gaussian, the output becomes a random variable too.)

3 Likelihood

The likelihood value is a measure of how likely an observed outcome under a model, or evaluates how well a model explains an observed outcome. Likelihood is also the value of the **probability density function** evaluated from a given dataset and a fixed model (weight parameters, mean and variance values are fixed).

$$\text{Likelihood} = p(t \mid \mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$

We can use likelihood to evaluate model quality across different models. The model with higher total likelihood (log-likelihood) across all data is better.

3.1 Simple Analog

To further solidify our understanding of likelihood, imagine a weather model predicts:

"Tomorrow's temperature should be around 25°C with ±1°C variance"

But then, it turns out to be **35°C**.

The **likelihood** of 35°C under that model is **very low**, which means that this model is not good at explaining this event. Maybe the model needs an update, by changing the mean and variance values, or even the weight parameters.

3.2 Likelihood Optimization

The mean, variance and weight parameters that maximizes the likelihood value (under Gaussian noise assumption) will also minimize the squared error of a linear regression model. It is the **maximum likelihood estimation**.

To compute the optimum weight ($\hat{\boldsymbol{w}}$), we can use this equation:

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

And we use this equation to compute the optimum variance ($\hat{\sigma}^2$):

$$\hat{\sigma}^2 = \frac{1}{N} (\mathbf{t} - \mathbf{X}\hat{\boldsymbol{w}})^\top (\mathbf{t} - \mathbf{X}\hat{\boldsymbol{w}})$$