

Likelihood and Noise

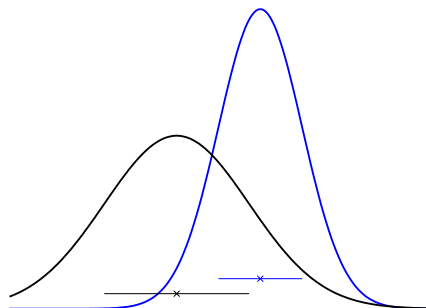
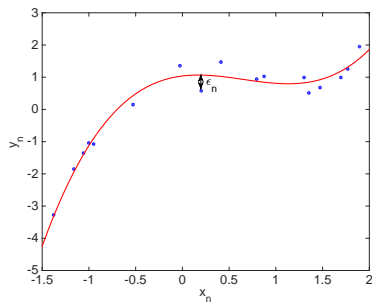
Carl Edward Rasmussen

October 10th, 2022

Key concepts

- Linear in the parameters models
 - the concept of a model
 - making predictions
 - least squares fitting
 - limitation: overfitting
- Likelihood and the concept of noise
 - Gaussian iid noise
 - maximum likelihood fitting
 - equivalence to least squares
 - motivation for inference with multiple hypotheses

Observation noise



- Imagine the data was in reality **generated** by the **red function**.
- But each $f(x_*)$ was independently contaminated by a noise term ϵ_n .
- The observations are noisy: $y_n = f_w(x_n) + \epsilon_n$.
- We can characterise the noise with a probability density function. For example a Gaussian density function, $\epsilon_n \sim \mathcal{N}(\epsilon_n; 0, \sigma_{\text{noise}}^2)$:

$$p(\epsilon_n) = \frac{1}{\sqrt{2\pi\sigma_{\text{noise}}^2}} \exp\left(-\frac{\epsilon_n^2}{2\sigma_{\text{noise}}^2}\right)$$

Probability of the observed data given the model

A vector and matrix notation view of the noise.

- $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_N]^\top$ stacks the **independent** noise terms:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; 0, \sigma_{\text{noise}}^2 \mathbf{I}) \quad p(\boldsymbol{\epsilon}) = \prod_{n=1}^N p(\epsilon_n) = \left(\frac{1}{\sqrt{2\pi \sigma_{\text{noise}}^2}} \right)^N \exp\left(-\frac{\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{2 \sigma_{\text{noise}}^2}\right)$$

- Given that $y = f + \boldsymbol{\epsilon}$ we can write the probability of y given f :

$$\begin{aligned} p(y|f, \sigma_{\text{noise}}^2) &= \mathcal{N}(y; f, \sigma_{\text{noise}}^2) = \left(\frac{1}{\sqrt{2\pi \sigma_{\text{noise}}^2}} \right)^N \exp\left(-\frac{\|y - f\|^2}{2 \sigma_{\text{noise}}^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi \sigma_{\text{noise}}^2}} \right)^N \exp\left(-\frac{\mathbf{E}(\mathbf{w})}{2 \sigma_{\text{noise}}^2}\right) \end{aligned}$$

- $\mathbf{E}(\mathbf{w}) = \sum_{n=1}^N (y_n - f_w(x_n))^2 = \|y - \boldsymbol{\Phi} \mathbf{w}\|^2 = \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}$ is the sum of squared errors
- Since $f = \boldsymbol{\Phi} \mathbf{w}$ we can write $p(y|\mathbf{w}, \sigma_{\text{noise}}^2) = p(y|f, \sigma_{\text{noise}}^2)$ for a given $\boldsymbol{\Phi}$.

Likelihood function

The *likelihood* of the parameters is the probability of the data given parameters.

- $p(y|w, \sigma_{\text{noise}}^2)$ is the probability of the observed data given the weights.
- $\mathcal{L}(w) \propto p(y|w, \sigma_{\text{noise}}^2)$ is the likelihood of the weights.

Maximum likelihood:

- We can fit the model weights to the data by maximising the likelihood:

$$\hat{w} = \operatorname{argmax} \mathcal{L}(w) = \operatorname{argmax} \exp\left(-\frac{E(w)}{2\sigma_{\text{noise}}^2}\right) = \operatorname{argmin} E(w)$$

- With an additive Gaussian independent noise model, the **maximum likelihood** and the **least squares** solutions are the same.
- But... we still have not solved the prediction problem! We still overfit.

Multiple explanations of the data

- We do not **believe** all models are equally probable to explain the data.
- We may **believe** a simpler model is more probable than a complex one.

Model complexity and uncertainty:

- We do not **know** what particular function generated the data.
- More than one of our models can perfectly fit the data.
- We **believe** more than one of our models could have generated the data.
- We want to reason in terms of **a set of possible explanations**, not just one.

