Likelihood and noise

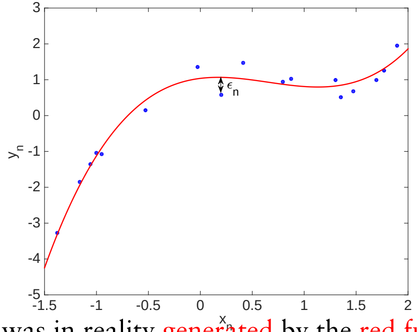
Carl Edward Rasmussen

June 23rd, 2016

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: $y_n = f_{\mathbf{w}}(x_n) + \epsilon_n$.
- We can characterise the noise with a probability density function. For example a Gaussian density function, $\epsilon_n \sim \mathcal{N}(\epsilon_n;\ 0, \sigma^2_{\text{noise}})$:

$$p(\epsilon_n) \;=\; \frac{1}{\sqrt{2\pi\,\sigma^2_{\text{noise}}}}\exp\Big(-\frac{\epsilon_n^2}{2\,\sigma^2_{\text{noise}}}\Big)$$

Probability of the observed data given the model

A vector and matrix notation view of the noise.

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

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

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

$$p(\mathbf{y}|\mathbf{f},\ \sigma^2_{\text{noise}}) = \mathcal{N}(\mathbf{y};\ \mathbf{f},\ \sigma^2_{\text{noise}}) = \Big(\frac{1}{\sqrt{2\pi\,\sigma^2_{\text{noise}}}}\Big)^N \exp\Big(-\frac{\|\mathbf{y}-\mathbf{f}\|^2}{2\,\sigma^2_{\text{noise}}}\Big)$$
$$= \Big(\frac{1}{\sqrt{2\pi\,\sigma^2_{\text{noise}}}}\Big)^N \exp\Big(-\frac{E(\mathbf{w})}{2\,\sigma^2_{\text{noise}}}\Big)$$

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

Likelihood function

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

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

Maximum likelihood:

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

$$\hat{\mathbf{w}} \;=\; \text{argmax}\;\mathcal{L}(\mathbf{w}) \;=\; \text{argmax}\;\exp\big(-\frac{\mathsf{E}(\mathbf{w})}{2\,\sigma^2_{\text{noise}}}\big) \;=\; \text{argmin}\;\mathsf{E}(\mathbf{w})$$

- 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.