Lecture 4: A Bayesian View of Regression

Attendance code: R9Q86YEV

Iain Styles

22 October 2019

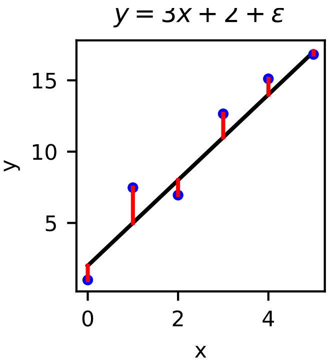
Learning Outcomes

By the end of this lecture you should be able to:

- Reason about regression using methods of probability
- Understand how likelihood maximisation and least-squares fitting are related
- Understand the role of prior information in machine learning

Least squares fitting

- Least squares error function is intuitive, but has no formal justification
- Why choose this approach? Why not some other form of the loss?
- Probabilistic approach will help us understand



Modelling the data-generating process

- Starting point: model the underlying data-generating process
- Assume data points generated by process that has a deterministic component, and some associated sampling uncertainty.

$$y = \mathbf{f}(x, \mathbf{w}) + \epsilon$$

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Modelling the data-generating process

- Starting point: model the underlying data-generating process
- Assume data points generated by process that has a deterministic component, and some associated sampling uncertainty.

$$y = \mathbf{f}(x, \mathbf{w}) + \epsilon$$

- $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $y(x)$ drawn from a normal distribution with mean $f(x, \mathbf{w})$ and variance σ^2

Modelling the data-generating process

- We can write the distribution of y as

$$p(y|x, \mathbf{w}, \sigma^2) = \mathcal{N}(y|f(x, \mathbf{w}), \sigma^2)$$

- Normal distribution with mean $f(x, \mathbf{w})$, variance σ^2
- Note that it is conditional on x, $\mathbf{w}$, and σ

Forming the joint distribution

- Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ which we will write as $(\mathbf{x}, \mathbf{y})$.
- Assume the y_i are sampled independently normal distributions with the same variance σ^2
- Joint PDF is then

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i | f(x_i, \mathbf{w}), \sigma^2)$$

Forming the joint distribution

- Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ which we will write as $(\mathbf{x}, \mathbf{y})$.
- Assume the y_i are sampled independently normal distributions with the same variance σ^2
- Joint PDF is then

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i | f(x_i, \mathbf{w}), \sigma^2)$$

- The *likelihood* of y
- PDF of measurements given parameters

Maximum Likelihood

- Can now ask "what are the most likely measurements"
- Maximise the likelihood
- Substitute in the full form of the normal distribution $\mathcal{N}(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-(x-\mu)^2/(2\sigma^2))$

$$p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{N} \exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2))$$

Maximum Likelihood

- Can now ask "what are the most likely measurements"
- Maximise the likelihood
- Substitute in the full form of the normal distribution $\mathcal{N}(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp(-(x-\mu)^2/(2\sigma^2))$

$$p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}}\prod_{i=1}^{N}\exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2))$$

- Take the logarithm (log is monotonic so has same maximum)

$$\begin{aligned}\ln p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) &= \ln(2\pi\sigma^2)^{-\frac{N}{2}}) \\ &\quad + \ln\left(\prod_{i=1}^{N}\exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2))\right)\end{aligned}$$

Maximum Likelihood

- Can now ask "what are the most likely measurements"
- Maximise the likelihood
- Substitute in the full form of the normal distribution $\mathcal{N}(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-(x-\mu)^2/(2\sigma^2))$

$$p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{N} \exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2))$$

- Take the logarithm (log is monotonic so has same maximum)

$$\begin{aligned}\ln p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) &= \ln(2\pi\sigma^2)^{-\frac{N}{2}}) \\ &\quad + \ln\left(\prod_{i=1}^{N} \exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2))\right)\end{aligned}$$

- Rearrange using $\ln \prod_i a_i = \sum_i \ln a_i$, and $\ln a^b = b \ln a$

$$\ln p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(x_i,\mathbf{w}))^2$$

Maximum Likelihood

$$\ln p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(x_i, \mathbf{w}))^2$$

Maximum Likelihood

$$\ln p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(x_i, \mathbf{w}))^2$$

- First term (negative) maximised by minimising the number of data points or the variance
- More data and/or more noise means less certainty (accumulation of errors)
- Second term: negative least-squares error
- Maximising the likelihood minimises the least-squares error

Including Priors

- Likelihood allows us to apply Bayes rule to include prior knowledge
$$p(a|b) = p(b|a)p(a)/p(b)$$
- $p(a|b)$ is the posterior distribution of a given b, $p(b|a)$ is the likelihood of b given a and $p(a)$ is the prior distribution of a.
- Can now ask: given a set of measurements, how are the weights distributed?

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma^2) = \frac{p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times p(\mathbf{w})}{P(\mathbf{y})}$$

Including Priors

- Likelihood allows us to apply Bayes rule to include prior knowledge
$$p(a|b) = p(b|a)p(a)/p(b)$$
- $p(a|b)$ is the posterior distribution of a given b, $p(b|a)$ is the likelihood of b given a and $p(a)$ is the prior distribution of a.
- Can now ask: given a set of measurements, how are the weights distributed?

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma^2) = \frac{p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times p(\mathbf{w})}{P(\mathbf{y})}$$

- Ignore $P(\mathbf{y})$ for simplicity

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma^2) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times p(\mathbf{w})$$

Simple Prior

- Consider $p(\mathbf{w}) = c$, a constant.
- All parameter values equally likely

Simple Prior

- Consider $p(\mathbf{w}) = c$, a constant.
- All parameter values equally likely

$$\begin{aligned} p(\mathbf{w}|\mathbf{x}, \mathbf{w}, \sigma^2) &\propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times c \\ &\propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \end{aligned}$$

Simple Prior

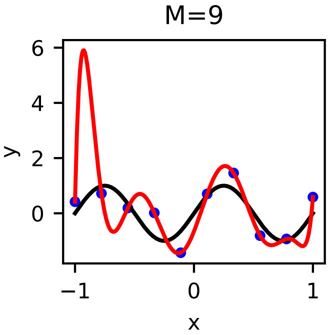
- Consider $p(\mathbf{w}) = c$, a constant.
- All parameter values equally likely

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{x}, \mathbf{w}, \sigma^2) &\propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times c \\
&\propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2)
\end{aligned}
$$

- The same max likelihood problem as before
- The least squares error assigns model weights that are uniformly distributed
- Is this desireable?

Distribution of weights

- Uniform distribution of weights seems reasonable
- But allows very large high-frequency terms to match model noise



M	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
9	-0.66	10.98	25.62	-117.80	-143.29	405.10	246.74	-561.32	-127.91	263.129

Gaussian Prior

- How to make large weights unlikely?
- Gaussian prior: most weights near zero

$$
\begin{aligned}
p(\mathbf{w}|\lambda) &\propto \prod_{i=1}^{M} \exp(-\lambda w_i^2) \\
&\propto \exp(-\lambda \sum_i w_i^2) \\
&\propto \exp(-\lambda \mathbf{w}^{\mathrm{T}} \mathbf{w})
\end{aligned}
$$

- Conditioned on parameters $\lambda = 1/2\sigma^2$ (ie large lambda *mapsto* narrow distribution)

Gaussian Prior

- From Bayes theorem:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma^2, \lambda) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times p(\mathbf{w}|\lambda)$$

- Take logs and maximise likelihood:

$$\mathcal{L} = \sum_{i=1}^{N} (y_i - f(x_i, \mathbf{w}))^2 + \lambda \mathbf{w}^{\mathrm{T}} \mathbf{w}. \qquad (1)$$

- Gaussian prior adds a "penalty" to the least squares loss.
- Proportional to the square of the length of the weight vector
- Minimise $\mathcal{L}$, have to simultaneously minimise model-date mismatch and the length of **w**
- Larger lambda (narrower distribution) $\mapsto$ bigger penalty
- L_2 (or sometimes Tikhonov) *regularisation*

Summary

- Probabilistic formulation of regression
- Maximising likelihood minimises least squares error
- Prior distributions of parameters
- Next lecture: solving regularised problems
- Reading: Bishop, section 1.2.5