Lecture 4: A Bayesian View of Regression

Iain Styles

22 October 2019

Bayesian View of Regression

So far, we have adopted quite an informal approach to regression: we wrote down an error function (least-squares) that made some sense from an intuitive viewpoint that seemed to make logical sense, but we have no formal basis for claiming that the "least squares fitting" method was a correct and valid way to approach the regression problem. Studying the problem from a Bayesian perspective will give us the formal rigour that we need in order to justify the choices we have made.

Our starting point will be to construct a model of the underlying data-generating process. We assume that each data point is the result of some process that has a deterministic component, and some associated sampling uncertainty.

$$y = \mathbf{f}(x, \mathbf{w}) + \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a normal distribution of variance σ^2 such that σ is a measure of the uncertainty in the sampling. That is, when the value of the dependent variable y is sampled for some value of the independent variable x, it will be drawn from a normal distribution with mean $f(x, \mathbf{w})$ and variance σ^2. Under this model, we can write the distribution of y as

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

that is, it is normally distributed with mean $f(x, \mathbf{w})$ and variance σ^2.

Now consider that we have a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ which we will write as $(\mathbf{x}, \mathbf{y})$. We assume that the dependent variables y_i are sampled *independently* from normal distributions with the same variance σ^2. The independence of the sampling means that the joint probability distributions over all data points can be written as the product of the distributions for each point:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i|f(x_i, \mathbf{w}), \sigma^2) \tag{3}$$

This is known as the *likelihood* of y: it is the probability density function of the dependent variables $\mathbf{y}$ conditioned on the set of parameters that describe the data generating function (ie. given some set of parameters, what is the probability of the measurements?).

With the likelihood, we can now approach regression in a different way. Since the likelihood is a proper probability density function, we can ask "what parameters $\mathbf{w}$ maximise it"? In other words, what is the most likely set of measurements, and what are

the parameters that gives rise to the most likely measurements? This is known as *maximum likelihood* inference.

First, we substitute in the full form of the normal distribution $\mathcal{N}(x|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp(-(x-\mu)^2/(2\sigma^2))$

$$p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}}\prod_{i=1}^{N}\exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2)) \quad (4)$$

We now take the logarithm of this to get rid of the exponential terms. Since the logarithm is a monotonic function (it has no maxima or minima of its own), the maximum of the log-likelihood will be at the same value of $\mathbf{w}$ as the maximum of the likelihood.

$$\ln p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = \ln(2\pi\sigma^2)^{-\frac{N}{2}} + \ln\left(\prod_{i=1}^{N}\exp(-(y_i - f(x_i,\mathbf{w}))^2/(2\sigma^2))\right) \quad (5)$$

where we have used $\ln ab = \ln a + \ln b$. We now use the generalisation of this, $\ln\prod_i a_i = \sum_i \ln a_i$, and the identity $\ln a^b = b\ln a$ to obtain the following expression for the log-likelihood:

$$\ln p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2) = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(x_i,\mathbf{w}))^2 \quad (6)$$

This has two terms. The first term (which is negative) is maximised by minimising the number of data points or the variance in the measurement. This is intuitively obvious: more data and/or more noise means less certainty. The second term is exactly the familiar least-squares error term (negated). Maximising the log-likelihood is therefore equivalent to minimising the least-squares error.

We have written down an expression for the likelihood assuming Gaussian noise on the data. We can now use this to perform some rather more sophisticated types of regression. In particular, it allows us to incorporate *prior information* about the problem using Bayes Rule:

$$p(a|b) = p(b|a)p(a)/p(b) \quad (7)$$

where $p(a|b)$ 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.

Given the likelihood $p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2)$, we can use Bayes' rule to compute the probability density function of the model weights:

$$p(\mathbf{w}|\mathbf{x},\mathbf{y},\sigma^2) = \frac{p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2)\times p(\mathbf{w})}{P(\mathbf{y})} \quad (8)$$

That is, the probability density function of the model weights depends on the likelihood of the measurements conditioned on the weight, multiplied by the **prior distribution of the weights**, and the normalised by the distribution of the measurements. We will ignore the normalising factor $p(\mathbf{y})$ for simplicity and consider

$$p(\mathbf{w}|\mathbf{x},\mathbf{y},\sigma^2) \propto p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2)\times p(\mathbf{w}) \quad (9)$$

The simplest case to consider is $p(\mathbf{w}) = c$, a constant. In this case we have that

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma^2) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times c \quad (10)$$
$$\propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \quad (11)$$

and the maximum likelihood solution of this is the same as before – it is the least-squares solution. This solution assumes that all model weights - large or small - are equally likely.

Is this desireable? Sometime, but not necessarily so. One characteristic of overfitting is that the model weights of the high-order terms can be very large. We have seen this previously in our earlier examples, reproduced in Figure 1 and Table 1. Our previous studies have focussed on removing these high order terms from the basis set, but could we control their contribution to the model fitting in a different way?

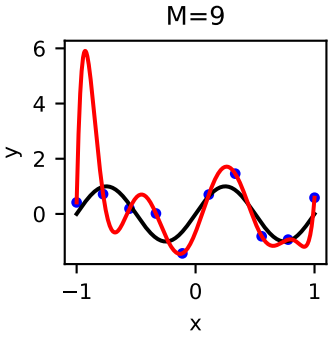

Figure 1: Fitting $y = \sin(2\pi x)$ with a polynomial fit of degree $M = 9$ to data with added noise..

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

Table 1: Coefficients of a high-order polynomial fit to noisy data show characteristic large values of high-order coefficients.

Let us consider another form of prior distribution for the model weights. We assume that they are drawn from a normal distribution with zero mean, and, for convenience, variance $\sigma^2 = 1/2\lambda$. We ignore normalisation constants for simplicity as they will all be absorbed into a single constant of proportionality. The distribution is condition on λ and assuming each of the components is independent, the joint distribution can be written

$$p(\mathbf{w}|\lambda) \propto \prod_{i=1}^{M} \exp(-\lambda w_i^2) \quad (12)$$
$$\propto \exp(-\lambda \sum_i w_i^2) \quad (13)$$
$$\propto \exp(-\lambda \mathbf{w}^{\mathrm{T}} \mathbf{w}) \quad (14)$$

Using Bayes Theorem we have

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, \sigma^2, \lambda) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) \times p(\mathbf{w}|\lambda) \quad (15)$$

and noting that $\ln ab = \ln a + \ln b$, we follow the same process as before and find that this is maximised by the minimum of

$$\mathcal{L} = \sum_{i=1}^{N} (y_i - f(x_i, \mathbf{w}))^2 + \lambda \mathbf{w}^{\mathrm{T}} \mathbf{w}. \quad (16)$$

That is, a Gaussian prior with zero mean and variance $\sigma^2 = 1/2\lambda$ is equivalent to adding a "penalty" term to the least-squares error function. This penalty is proportional to the square of the length of the weight vector and so when we try to minimise $\mathcal{L}$ it will preferentially prefer solutions with small values for its components. This is consistent with the Bayesian prior, which is normally distributed around zero. The most likely values of the weights are those near

to zero, and the least likely are those that are large. The parameter λ controls the width of the Gaussian prior: large λ means low variance and therefore a narrow distribution, and so the larger λ is, the less likely the weights are to take large values. Because this prior distribution results in the model coefficient being kept small, it is known as a *shrinkage* method, and since the penalty term is the L_2 norm (ie the square length of the weight vector, this is often referred to as L_2 regularisation, or sometimes as Tikhonov regularisation (although this is a more general class of methods).

L_2 regularisation is very widely used in regression tasks. In the next section of the module, we will study how to use it effectively

Reading

Sections 1.2.5 of Bishop, Pattern Recognition and Machine Learning.