Bayesian Linear Regression

Contents

Probabilistic regression I (maximum likelihood)

I will now show you how to derive least squares from the maximum likelihood principle. Recall that the maximum likelihood principle states that you should pick the model parameters that maximize the probability of the data conditioned on the parameters.

Just like before assume that we have N observations of inputs $\mathbf{x}_{1:N}$ and outputs $\mathbf{y}_{1:N}$. We model the map between inputs and outputs using a generalized linear model with M basis functions:

$$y(\mathbf{x};\mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

Now here is the difference with what we did before. Instead of directly picking a loss function to minimize we come up with a probabilistic description of the measurement process. In particular, we *model the measurement process* using a **likelihood** function:

$$\mathbf{y}_{1:N}|\mathbf{x}_{1:N},\mathbf{w} \sim p(\mathbf{y}_{1:N}|\mathbf{x}_{1:N},\mathbf{w}).$$

What is the interpretation of the likelihood function? Well, $p(\mathbf{y}_{1:N}|\mathbf{x}_{1:N},\mathbf{w})$ tells us how plausible is it to observe $\mathbf{y}_{1:N}$ at inputs $\mathbf{x}_{1:N}$, if we know that the model parameters are $\mathbf{w}$.

The most common choice for the likelihood of a single measurement is to pick it to be Normal. This corresponds to the belief that our measurement is around the model prediction $\mathbf{w^T}\boldsymbol{\phi}(\mathbf{x})$ but it is contaminated with Gaussian noice of variance σ^2. Mathematically, we have:

$$\begin{aligned} p(y_i|\mathbf{x}_i,\mathbf{w},\sigma) &= N\left(y_i|y(\mathbf{x}_i;\mathbf{w}),\sigma^2\right) \\ &= N\left(y_i|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i),\sigma^2\right), \end{aligned}$$

where σ^2 models the **variance of the measurement noise**. Note that here I used the notation $N(y|\mu,\sigma^2)$ to denote the PDF of a Normal with mean μ and variance σ^2, i.e.,

$$N(y|\mu,\sigma^2) := (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

Since, in almost all the cases we encounter, the measurements are independent conditioned on the model, then likelihood of the data factorizes as follows:

$$\begin{aligned} p(\mathbf{y}_{1:N}|\mathbf{x}_{1:N},\mathbf{w}) &= \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\mathbf{w}) \\ &= \prod_{i=1}^{N} N\left(y_i|\mathbf{w^T}\boldsymbol{\phi}(\mathbf{x_i}),\sigma^2\right) \\ &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{\left[y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right]^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N}\frac{\left[y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right]^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\right]^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2} \parallel \mathbf{y}_{1:N} - \boldsymbol{\Phi}\mathbf{w} \parallel^2\right\}, \end{aligned}$$

where $\mathbf{\Phi}$ is the $N \times M$ design matrix.

Now we are ready to apply the maximum likelihood function to find all the parameters. This includes both the weight vector $\mathbf{w}$ and the measurement variance σ^2. We need to solve this:

$$\max_{\mathbf{w},\sigma^2} \log p(\mathbf{y}_{1:N}|\mathbf{x}_{1:N}, \mathbf{w}) = \max_{\mathbf{w},\sigma^2} \left\{ -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \parallel \mathbf{y}_{1:N} - \mathbf{\Phi}\mathbf{w} \parallel^2 \right\}$$

Notice that the rightmost part is actually the negative of the sum of square errors. So, by maximizing the likelihood with respect to $\mathbf{w}$ we are actually minimizing the sum of square errors. This means that the maximum likelihood weights and the least square weights are exactly the same! We do not even have to do anything further. The weights should satisfy this linear system:

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} = \mathbf{\Phi}^T\mathbf{y}_{1:N}.$$

This is nice. The probabilistic interpretation above gives the same solution as least squares! But there is more. Notice that it can also give us an estimate for the measurement noise variance σ^2. All you have to do is maximize likelihood with respect to σ^2. If we take the derivative of the log-likelihood with respect to σ^2, set it equal to zero and solve for σ^2 you get:

$$\sigma^2 = \frac{\parallel \mathbf{\Phi}\mathbf{w} - \mathbf{y}_{1:N} \parallel^2}{N}.$$

Finally, you can incorporate this measurement uncertainty when you are making predictions. This is done through the **point predictive distribution**, which is Normal in our case:

$$p(y|\mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}\left(y \middle| \mathbf{w}^T\phi(\mathbf{x}), \sigma^2\right).$$

In other words, your prediction about the measured output y is that it will be Normally distributed around your model prediction with a variance σ^2. You can use this to find a 95% credible interval.

Examples

See this example.

Probabilistic regression II (maximum a posteriori estimates)

This version of probabilistic is similar to maximum likelihood in the sense that you maximum the log probability of something (the posterior instead of the likelihood) and it has the adendum that it can help you avoid overfitting.

Just like before, we wish to model the data using some **fixed** basis/features:

$$y(\mathbf{x}; \mathbf{w}) = \sum_{j=1}^{m} w_j\phi_j(\mathbf{x}) = \mathbf{w^T}\boldsymbol{\phi}(\mathbf{x})$$

Again, we *model the measurement process* using a **likelihood** function:

$$\mathbf{y}_{1:n}|\mathbf{x}_{1:n}, \mathbf{w}, \sigma \sim N(\mathbf{w^T}\boldsymbol{\phi}(\mathbf{x}), \sigma^2).$$

The new ingredient is that we *model the uncertainty in the model parameters* using a **prior**:

$$\mathbf{w} \sim p(\mathbf{w}).$$

Gaussian Prior on the Weights

The Gaussian prior is the simplest possible choice for the weights. It is:

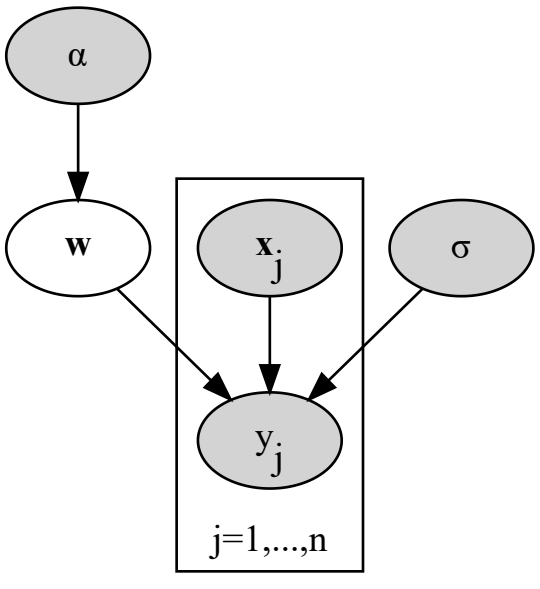
$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right) = \left(\frac{\alpha}{2\pi}\right)^{\frac{m}{2}} \exp\left\{-\frac{\alpha}{2}\|\mathbf{w}\|^2\right\}.$$

The interpretation is that, before we see the data, we beleive that $\mathbf{w}$ must be around zero with a precision of α. This push to the weights to be towards zero is exactly what helps us avoid overfitting. The bigger the precision parameter α the more the weights are pushed towards zero.

Graphical representation of the model

Let's visualize the regression model as a graph. Remember that the shaded nodes are assumed to be observed (so below we are assuming that we know α and σ). Another thing to observe is that the nodes that are inside the box are repeated as many times as indicated. This is the so-called plate notation for graphical models and it saves from the troubkle of drawing n input-output nodes.

```
from graphviz import Digraph
g = Digraph('bayes_regression')
g.node('alpha', label='<&alpha;>', style='filled')
g.node('w', label='<<b>w</b>>')
g.node('sigma', label='<&sigma;>', style='filled')
with g.subgraph(name='cluster_0') as sg:
    sg.node('xj', label='<<b>x</b><sub>j</sub>>', style='filled')
    sg.node('yj', label='<y<sub>j</sub>>', style='filled')
    sg.attr(label='j=1,...,n')
    sg.attr(labelloc='b')
g.edge('alpha', 'w')
g.edge('sigma', 'yj')
g.edge('w', 'yj')
g.edge('xj', 'yj')
g.render('bayes_regression', format='png')
g
```



The Posterior of the Weights

Combining the likelihood and the prior, we get using Bayes' rule:

$$p(\mathbf{w}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \sigma, \alpha) = \frac{p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}, \mathbf{w}, \sigma)p(\mathbf{w}|\alpha)}{\int p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}, \mathbf{w}', \sigma)p(\mathbf{w}'|\alpha)d\mathbf{w}'}.$$

The posterior summarizes our state of knowledge about $\mathbf{w}$ after we see the data, if we know α and σ.

Maximum Posterior Estimate

We can find a point estimate of $\mathbf{w}$ by solving:

$$\mathbf{w}_{\text{MPE}} = \arg\max_{\mathbf{w}} p(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}, \mathbf{w}, \sigma)p(\mathbf{w}|\alpha).$$

For Gaussian likelihood and weight prior, the logarithm of the posterior is:

$$\log p(\mathbf{w}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \sigma, \alpha) = -\frac{1}{2\sigma^2}\|\mathbf{\Phi}\mathbf{w} - \mathbf{y}_{1:n}\|^2 - \frac{\alpha}{2}\|\mathbf{w}\|^2.$$

Taking derivatives with respect to $\mathbf{w}$ and setting them equal to zero (necessary condition), we find:

$$\mathbf{w}_{\text{MPE}} = \sigma^{-2}\left(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \alpha\mathbf{I}\right)^{-1}\mathbf{\Phi}^T\mathbf{y}_{1:n}.$$

Unfortunately, we no longer have an analytic formula for σ... (we will fix that later).

Examples

See this example

See this example.

Probabilistic regression III (Bayesian linear regression)

This has the same setup version III of probabilistic regression but we do not just get a point estimate for the weights. We retain the posterior of the weughts in its full complexity. The adendum is that we can now quantify the epistemic uncertainty induced by the limited number of observations used to estimate the weights.

For Gaussian likelihood and weight prior, the posterior of the weights is Gaussian:

$$p(\mathbf{w}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \sigma, \alpha) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right),$$

where

$$\mathbf{S} = \left(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \alpha\mathbf{I}\right)^{-1},$$

and

$$\mathbf{m} = \sigma^{-2}\mathbf{S}\Phi^T\mathbf{y}_{1:n}.$$

In the general case of non-Gaussian likelihood (and non-linear models), the posterior will not be analytically available. We will learn how to deal with these cases in Lectures 27 and 28 when we talk about generic ways to characterize posteriors.

Posterior Predictive Distribution

Using probability theory, we ask: What do we know about y at a new $\mathbf{x}$ after seeing the data. To answer this question, we just use the sum rule:

$$p(y|\mathbf{x}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \sigma, \alpha) = \int p(y|\mathbf{x}, \mathbf{w}, \sigma)p(\mathbf{w}|\mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \sigma, \alpha)d\mathbf{w}.$$

See this page for more information on the Law of Total Probability: https://en.wikipedia.org/wiki/Law_of_total_probability

For the all-Gaussian case, this is analytically available:

$$p(y|\mathbf{x}, \mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \sigma, \alpha) = \mathcal{N}\left(y|m(\mathbf{x}), s^2(\mathbf{x})\right),$$

where

$$m(\mathbf{x}) = \mathbf{m}^T\boldsymbol{\phi}(\mathbf{x}) \text{ and } s(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}\boldsymbol{\phi}(\mathbf{x}) + \sigma^2.$$

Notice that the **predictive uncertainty** is:

$$s^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}\boldsymbol{\phi}(\mathbf{x}) + \sigma^2,$$

where:

- σ^2 corresponds to the measurement noise.
- $\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}\boldsymbol{\phi}(\mathbf{x})$ is the epistemic uncertainty induced by limited data.

Examples

See this example.

By Ilias Bilionis (ibilion[at]purdue.edu)