A Probabilistic Interpretation of Regularization

Brian Keng — 2016-08-29 08:52

This post is going to look at a probabilistic (Bayesian) interpretation of regularization. We'll take a look at both L1 and L2 regularization in the context of ordinary linear regression. The discussion will start off with a quick introduction to regularization, followed by a back-to-basics explanation starting with the maximum likelihood estimate (MLE), then on to the maximum a posteriori estimate (MAP), and finally playing around with priors to end up with L1 and L2 regularization.

Regularization

Regularization (https://en.wikipedia.org/wiki/Regularization_(mathematics)) is the process of introducing additional information in order to solve ill-posed problems or prevent overfitting. A trivial example is when trying to fit a simple linear regression but you only have one point. In this case, you can't estimate both the slope and intercept (you need at least two points) so any MLE estimate (which *only* uses the data) will be ill-formed. Instead, if you provide some "additional information" (i.e. prior information [1]), you can get a much more reasonable estimate.

To make things a bit more concrete, let's talk about things in the context of a ordinary linear regression (https://en.wikipedia.org/wiki/Linear_regression). Recall from my previous post on linear regression (../a-probabilistic-view-of-regression) (Equation 11 in that post) that the maximum likelihood estimate for ordinary linear regression is given by:

$$\hat{\beta_{\mathbf{MLE}}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}))^2$$
$$= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \quad (1)$$

The estimate is quite intuitive: pick the coefficients (β_j) that minimizes the squared error between the observed values (y_i) and those generated by our linear model ($\hat{y_i}$).

In a similar vein as above, consider what happens when we only have one data point ($y_0, \mathbf{x_0}$) but more than one coefficient. There are any number of possible "lines" or equivalently coefficients that we could draw to minimize Equation 1. Thinking back to high school math, this is analogous to estimating the slope and intercept for a line but with only one point. Definitely a problem they didn't teach you in high school. I'm using examples where we don't have enough data but there could other types of issues such as colinearity (https://en.wikipedia.org/wiki/Multicollinearity) that may not outright prevent fitting of the model but will probably produce an unreasonable estimate.

Two common schemes for regularization add a simple modification to Equation 1:

$$\hat{\beta_{\mathbf{L1}}} = \arg\min_{\beta} \Big(\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} |\beta_j| \Big) \quad (2)$$

$$\hat{\beta_{\mathbf{L2}}} = \arg\min_{\beta} \Big(\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} |\beta_j|^2 \Big) \quad (3)$$

L1 regularization (https://en.wikipedia.org/wiki/Regularization_(mathematics)#Regularizers_for_sparsity) (also known as LASSO in the context of linear regression) promotes sparsity of coefficients. Sparsity translates to some coefficients having values, while others are zero (or closer to zero). This can be seen as a form of feature selection. L2 regularization (https://en.wikipedia.org/wiki/Regularization_(mathematics)#Tikhonov_regularization) (also known as ridge regression in the context of linear regression and generally as

Tikhonov regularization) promotes smaller coefficients (i.e. no one coefficient should be too large). This type of regularization is pretty common and typically will help in producing reasonable estimates. It also has a simple probabilistic interpretation (at least in the context of linear regression) which we will see below. (You can skip the next two sections if you're already familiar with the basics of MLE and MAP estimates.)

The Likelihood Function

Recall Equation 1 can be derived from the likelihood function (without $\log(\cdot)$) for ordinary linear regression:

$$\begin{aligned}\mathcal{L}(\beta|\mathbf{y}) &:= P(\mathbf{y}|\beta) \\ &= \prod_{i=1}^{n} P_Y(y_i|\beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2}{2\sigma^2}}\end{aligned} \tag{4}$$

where $\mathbf{y}$ are our observed data points $(y_1, \ldots, y_N)$ and $P_Y(y_i|\mu, \sigma^2)$ is the probability of observing the data point y_i. The implicit assumption in linear regression is that the data points are normally distributed about the regression line (see my past post on linear regression (../a-probabilistic-view-of-regression) for more on this).

Classical statistics focuses on maximizing this likelihood function, which *usually* provides a pretty good estimate -- except when it doesn't like in the case of "small data". However, looking at the problem from a more probabilistic point of view (i.e. Bayesian), we don't just want to maximize the likelihood function but rather the posterior probability. Let's see how that works.

The Posterior

We'll start by reviewing Bayes Theorem (https://en.wikipedia.org/wiki/Bayesian_inference#Formal) where we usually denote the parameters we're trying to estimate by θ and the data as y:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \tag{5}$$

In Bayesian inference, we're primarily concerned with the posterior: "the probability of the parameters given the data". Put in another way, we're looking to estimate the probability distribution of the parameters (θ) given the data we have observed (y). Contrast this with classical methods which instead try to find the best parameters to maximize likelihood: the probability of observing data (y) given a different values of the parameters. Definitely a subtle difference but I think most would agree the Bayesian interpretation is much more natural [2].

Looking at Equation 5 in more detail, we already know how to compute the likelihood but the two new parts are the prior and the evidence. This is where proponents of frequentist statistics usually have a philosophical dilemma. The prior is actually something we (the modeler) explicitly choose that is **not** based on the data [3] (y) *gasp*!

Without getting into a whole philosophical spiel, adding some additional prior information is exactly what we want in certain situations! For example when we don't have enough data, we probably have some idea about what is reasonable given our knowledge of the problem. This prior allows us to encode this knowledge. Even in cases where we don't explicitly have this problem, we can choose a "weak" prior (https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors) which will only bias the result slightly from the MLE estimate. In cases where we have lots of the data, the likelihood dominates Equation 5 anyways so the result will be similar in these cases.

From Equation 5, a full Bayesian analysis would look at the distribution of the parameters (θ). However, most people will settle for something a bit less involved: finding the maximum of the posterior (which turns out to be an easier problem in

most cases). This is known as the maximum a posteriori probability estimate (https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation), usually abbreviated by MAP. This simplifies the analysis and in particular allows us to ignore the evidence ($P(y)$), which is constant relative to the parameters we're trying to estimate (θ).

Formalizing the MAP estimate, we can write it as:

$$\begin{aligned}\hat{\theta}_{\mathbf{MAP}} &= \arg\max_{\theta} P(\theta|y) \\ &= \arg\max_{\theta} \frac{P(y|\theta)P(\theta)}{P(y)} \\ &= \arg\max_{\theta} P(y|\theta)P(\theta) \\ &= \arg\max_{\theta} \log(P(y|\theta)P(\theta)) \\ &= \arg\max_{\theta} \log P(y|\theta) + \log P(\theta) \end{aligned} \tag{6}$$

Notice that we can get rid of the evidence term ($P(y)$) because it's constant with respect to the maximization and also take the $\log$ of the inner function because it's monotonically increasing. Contrast this with the MLE estimate:

$$\hat{\theta}_{\mathbf{MLE}} = \arg\max_{\theta} \log P(y|\theta) \tag{7}$$

Selecting Priors for Linear Regression

The main idea is to select Bayesian priors on the coefficients of linear regression that get us to L1 and L2 regularization (Equation 2 and 3). Let's see how this works.

Normally Distributed Priors

We'll start with our good old friend the normal distribution and place a zero-mean normally distributed prior on *each* β_i value, all with identical variance τ^2. From Equation 6 and filling in the likelihood function from Equation 4 and our prior:

$$
\begin{aligned}
&\arg\max_{\beta} \Big[\log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{\beta_j^2}{2\tau^2}} \Big] \\
&= \arg\max_{\beta} \Big[-\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2}{2\sigma^2} - \sum_{j=0}^{p} \frac{\beta_j^2}{2\tau^2} \Big] \\
&= \arg\min_{\beta} \frac{1}{2\sigma^2} \Big[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2 + \frac{\sigma^2}{\tau^2} \sum_{j=0}^{p} \beta_j^2 \Big] \\
&= \arg\min_{\beta} \Big[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} \beta_j^2 \Big] \qquad (8)
\end{aligned}
$$

Notice that we dropped many of the constants (with respect to β) and factored a bit to simplify the expression. You can see this is the same expression as Equation 3 (L2 Regularization) with $\lambda = \sigma^2/\tau^2$ (remember σ is assumed to be constant in ordinary linear regression, and we get to pick τ for our prior). We can adjust the amount of regularization we want by changing λ. Equivalently, we can adjust how much we want to weight the priors carry on the coefficients (β). If we have a very small variance (large λ) then the coefficients will be very close to 0; if we have a large variance (small λ) then the coefficients will not be affected much (similar to as if we didn't have any regularization).

Laplacean Priors

Let's first review the density of the Laplace distribution (https://en.wikipedia.org/wiki/Laplace_distribution) (something that's usually not introduced in beginner probability classes):

$$
Laplace(\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}
$$

This is sometimes called the "double exponential" distribution because it looks like two exponential distributions placed back to back (appropriately scaled with a location parameter). It's also quite similar to our Gaussian in form, perhaps you can see how we can get to L1 regularization already?

Starting with a zero-mean Laplacean prior on all the coefficients like we did in the previous subsection:

$$
\begin{aligned}
&\arg\max_{\beta} \Big[\log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} \frac{1}{2b} e^{-\frac{|\beta_j|}{2b}} \Big] \\
&= \arg\max_{\beta} \Big[-\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2}{2\sigma^2} - \sum_{j=0}^{p} \frac{|\beta_j|}{2b} \Big] \\
&= \arg\min_{\beta} \frac{1}{2\sigma^2} \Big[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2 + \frac{\sigma^2}{b} \sum_{j=0}^{p} |\beta_j| \Big] \\
&= \arg\min_{\beta} \Big[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1}+\ldots+\beta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} |\beta_j| \Big] \qquad (9)
\end{aligned}
$$

Again we can see that Equation 9 contains the same expression as L1 Regularization in Equation 2.

The Laplacean prior has a slightly different effect compared to L2 regularization. Instead of preventing any of the coefficients from being too large (due to the squaring), L1 promotes sparsity. That is, zeroing out some of the coefficients. This makes some sense if you look at the density of a Laplacean prior where there is a sharp increase in the density at its mean.

Another way to intuitively see this is to compare two solutions [4]. Let's imagine we are estimating two coefficients in a regression. In L2 regularization, the solution $\beta = (1, 0)$ has the same weight as $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ so they are both treated equally. In L1 regularization, the same two solutions favor the sparse one:

$$
||(1,0)||_1 = 1 < ||(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})||_1 = \sqrt{2} \qquad (10)
$$

So L2 regularization doesn't have any specific built in mechanisms to favor zeroed out coefficients, while L1 regularization actually favors these sparser solutions.

Conclusion

L1 and L2 regularization are such intuitive techniques when viewed shallowly as just extra terms in the objective function (i.e. "shrink the coefficients"). However, I was delighted to find out that it also has a Bayesian interpretation, it just seems so much more elegant that way. As I have mentioned before, I'm a big fan probabilistic interpretations of machine learning and you can expect many more posts on this subject!

Further Reading

- Wikipedia: Regularization (https://en.wikipedia.org/wiki/Regularization_(mathematics)), ordinary linear regression (https://en.wikipedia.org/wiki/Linear_regression), Bayes Theorem (https://en.wikipedia.org/wiki/Bayesian_inference#Formal), "weak" prior (https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors), maximum a posteriori probability estimate (https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation)
- A previous post on linear regression (../a-probabilistic-view-of-regression)
- Machine Learning: A Probabilistic Perspective, Kevin P. Murphy

[1] One philosophical counterpoint is that we should "let the data speak for itself". Although superficially satisfying, it is almost always the case where you inject "prior" knowledge into interpreting the data. For example, selecting a linear regression model already adds some prior knowledge or intuition to the data. In the same way, if we have some vague idea that the mean of the data should be close to zero, why not add that information into the problem? If there's enough data and we've coded things right, then the prior isn't that impactful anyways.

[2] As you might have guessed, I'm fall into the Bayesian camp. Although I would have to say that I'm much more of a pragmatist above all else. I'll use whatever works, frequentist, Bayesian, no theoretical basis, doesn't really matter as long as I can solve the desired problem in a reasonable manner. It just so happens Bayesian methods produce reasonable estimates very often.

[3] Well that's not exactly true. As a modeler, we can pick a prior that does

depend on the data, although that's a bit of "double dipping". These methods are generally known as empirical Bayes methods (https://en.wikipedia.org/wiki/Empirical_Bayes_method).

[4] I got this example from Machine Learning: A Probabilistic Perspective. It has great explanations on both L1 and L2 regularization as long as you have some moderate fluency in probability. I highly recommend this textbook if you want to dig into the meat of ML techniques. It's very math (probability) heavy but really provides good high level explanations that help with intuition.

Bayesian **probability** **regularization**

I'm Brian Keng (http://www.briankeng.com/about), a former academic, current data scientist and engineer. This is the place (../../) where I write about all things technical.

Twitter: @bjlkeng (http://www.twitter.com/bjlkeng)

Archive (../../archive.html)

Tags (../../categories/index.html)

RSS feed (../../rss.xml)