Probabilistic Artificial Intelligence

A. Krause

1 Probability Review

Is not really worth writing out.

2 Bayesian Linear Regression

The idea is simple - ordinary linear regression (and variants thereof) yield a point estimate $\hat{y}$. We'd like to know how uncertain We are about this estimate.

2.1 Ridge Regression as Bayesian Inferece

Recall that We can find the coefficients for the best linear fit by doing

$$X^\top(X\mathbf{w} - \mathbf{y}) = 0 \tag{1}$$
$$X^\top X\mathbf{w} - X^\top\mathbf{y} = 0 \tag{2}$$
$$X^\top X\mathbf{w} = X^\top\mathbf{y} \tag{3}$$
$$\mathbf{w} = (X^\top X)^{-1}X^\top\mathbf{y} \tag{4}$$

Is the way that usually goes. Now, in order to get the solution for ridge regression We do

$$\mathbf{w} = (X^\top X + \lambda I)^{-1}X^\top\mathbf{y} \tag{5}$$

Is the modification for ridge regression to penalize large coefficients in $\mathbf{w}$.

So, now, let's make things probabilistic:

For $\mathbf{w}$, let's assume that $\mathbf{w} \sim \mathcal{N}(0, \sigma^2_{\mathbf{W}})$ and also let's assume that $\mathbf{w} \perp \mathbf{x}_i \; \forall i \in [n]$, so We have a normal prior over the weights and a-priori the weights are independent of the data - without knowing anything about $\mathbf{x}_i$, this is all We've got.

Then let's make the standard

$$P(\mathbf{y}_i|\mathbf{w}, \mathbf{x}_i) \sim \mathcal{N}(\mathbf{w}^\top\mathbf{x}_i, \sigma^2_{\mathbf{y}}) \tag{6}$$

So We're assuming that given the one true weight vector, our labels are normally distributed around the predicted mean. They are also of course independent, and all have the same variance.

So now the idea is - We have a prior on the weights, and We have a likelihood function which involves the weights. By the glory of Bayes' rule, that's enough to try to calculate a posterior on the weights:

$$P(\mathbf{w}|\mathbf{x}_{1\ldots n}, \mathbf{y}_{1\ldots n}) = \frac{1}{z} \cdot P(\mathbf{w}, \mathbf{x}_{1\ldots n}, \mathbf{y}_{1\ldots n}) \tag{7}$$

$$= \frac{1}{z} \cdot P(\mathbf{x}_{1\ldots n}) \cdot p(\mathbf{w}|\mathbf{x}_{1\ldots n}) \cdot P(\mathbf{y}_{1\ldots n}|\mathbf{w}, \mathbf{x}_{1\ldots n}) \tag{8}$$

$$= \frac{1}{z'} \cdot p(\mathbf{w}|\mathbf{x}_{1\ldots n}) \cdot P(\mathbf{y}_{1\ldots n}|\mathbf{w}, \mathbf{x}_{1\ldots n}) \tag{9}$$

$$= \frac{1}{z'} \mathcal{N}(0, \sigma^2_{\mathbf{W}}) \cdot \prod_{i=1}^{n} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2_{\mathbf{y}}) \tag{10}$$

$$= \frac{1}{z'} \frac{1}{z_{\mathbf{w}}} \exp\left(-\frac{1}{\sigma^2_{\mathbf{w}}} \|\mathbf{w}\|^2\right) \cdot \frac{1}{z_{\mathbf{y}}} \prod_{i=1}^{n} \exp\left(-\frac{1}{\sigma^2_{\mathbf{y}}} \cdot \|y_i - \mathbf{w}^\top \mathbf{x}_i\|^2\right) \tag{11}$$

So, at 7 We are simply using the whole $P(A|B) = P(A, B)/P(B)$ thing, to rearrange the terms any way We like - the point is to have one conjunction above and one below, and factorize the above conjunction with the chain rule in a way that can leverage our assumptions.

Then at 9, We absorb that $P(\mathbf{x}_{1\ldots n})$ term since it's kind of irrelevant - just some Gaussian that We'll ultimately not care about.

Finally We use an assumption or two.

Now, note that when fitting $\mathbf{w}$, We're going to maximize our result. Since We just take about an extrema, We can start stripping parts away:

$$\arg\max_{\mathbf{w}} = \frac{1}{z'}\frac{1}{z_{\mathbf{w}}}\exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2\right)\cdot\frac{1}{z_{\mathbf{y}}}\prod_{i=1}^{n}\exp\left(-\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{12}$$

$$= \exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2\right)\cdot\prod_{i=1}^{n}\exp\left(-\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{13}$$

$$= \exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2\right)\cdot\prod_{i=1}^{n}\exp\left(-\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{14}$$

$$= \exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2-\sum_{i=1}^{n}\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{15}$$

$$= -\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2-\sum_{i=1}^{n}\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2 \tag{16}$$

$$\arg\min_{\mathbf{w}} = \frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2+\sum_{i=1}^{n}\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2 \tag{17}$$

Where in 16, We just multiplied by the positive constant $\sigma_{\mathbf{y}}^2$ to get a clean coefficient for the λ term.

Anyway this shows that if We choose $\lambda = \frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{w}}^2}$, then the solution to the ridge regression problem is equivalent to finding the maximum a posteriori solution. ($P(\mathbf{w}|\mathbf{x}_{1\ldots n}, \mathbf{y}_{1\ldots n})$ is the posterior in question).

2.2 Distribution of the weights

Okay, so, We have that under Bayesian regression

$$\mathbf{w} = \mathbf{w} = (X^\top X+\lambda I)^{-1}X^\top\mathbf{y} \tag{18}$$

$$= \mathbf{w} = (X^\top X+\frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{w}}^2}I)^{-1}X^\top\mathbf{y} \tag{19}$$

$$\tag{20}$$