INF367A Project 1

Naphat Amundsen

March 30, 2020

Introduction

This project is about linear regression using automatic relevance determination (ARD) when doing linear regression. The idea of ARD is to use a flexible prior to push weights of the irrelevant features to zero while not penalizing relevant features, as opposed to L2 regularization and similar methods which suppresses the weights to all features equally. ARD can be viewed as weighted regularization with respect to the features. The weights are then parameters that needs to be optimized as well.

Assume that we have observed n pairs (x_i, y_i) where x_i are the feature values and y_i is the label. Let $w \in \mathbb{R}^d$ be the regression weights. The likelihood is

$$P(y|x, w, \beta) = \prod_{i=1}^{n} N(y_i|w^T x_i, \beta^{-1}) \tag{1}$$

The noise precision β, is modelled with a Gamma prior:

$$P(\beta) = \text{Gamma}(\beta|a_0, b_0) \tag{2}$$

where a_0 and b_0 are user defined hyperparameters.

The prior for the weights w is a Gaussian distribution

$$P(w|\alpha_1, \ldots, \alpha_d) = \prod_{j=1}^{d} N(w_j|0, \alpha_j^{-1}) \tag{3}$$

Comparing this to the standard way, the precision of the weights is not just a prior scalar, but an actual distribution. The prior for α_j is a Gamma distribution.

$$P(\alpha_j) = \text{Gamma}(\alpha_j|c_0, d_0) \quad \forall j = 1, \ldots, d \tag{4}$$

where c_0 and d_0 are user defined hyperparameters.

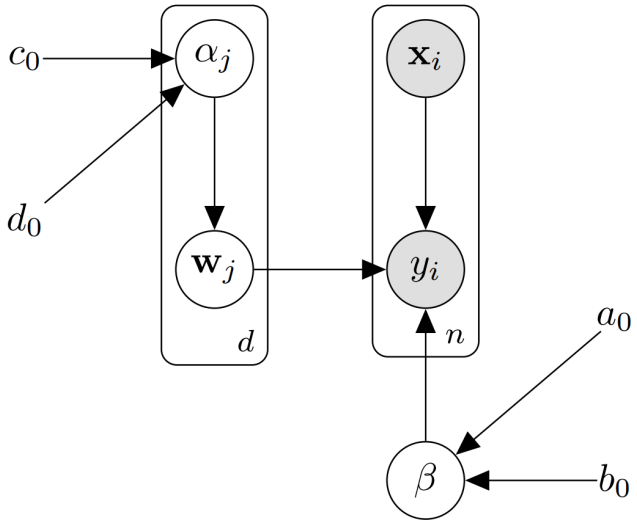

Figure 1: Plate diagram of the regression model with ARD prior. Variables x_i , and y_i are observed; We are interested in the posterior distribution of parameters w, α and β. The constants a_0 , b_0, c_0 , and d_0 are user-defined hyperparameters

The components so far can be represented as a Bayesian network, illustrated in Figure 1. The posterior can be conveniently deducted from the diagram to be

$$P(w, \alpha, \beta | y, x) \propto \prod_{i=1}^{n} P(y_i | w, x_i, \beta) \prod_{j=1}^{d} \left[P(w_j | \alpha_j) P(\alpha_j) \right] P(\beta). \tag{5}$$

1 Gibbs sampling and conjugate priors

Gibbs sampling will be used to approximate the optimal ARD regression model parameters. Gibbs sampler requires the full conditional distributions for the model parameters; $P(w|y, x, \alpha, \beta)$, $P(\alpha|y, x, w, \beta)$, $P(\beta|y, x, w, \alpha)$, where $\alpha = (\alpha_1, \ldots, \alpha_d)$.

1.1 Deriving conjugate prior for w

$$\begin{aligned}
P(w|y, x, \alpha, \beta) &\propto \prod_{i=1}^{n} P(y_i | w, x_i, \beta) \prod_{j=1}^{d} P(w_j | \alpha_j) \\
&= N(y_i | X^T w) \prod_{j=1}^{d} N(w_j | 0, (\sqrt{\alpha_j})^{-1}) \\
&\propto \exp\left(-\frac{1}{2} \beta (y - X^T w)^T (y - X^T w) \right) \prod_{j=1}^{d} \exp\left(-\frac{1}{2} w^2 (\sqrt{\alpha_j})^2 \right) \\
&= \exp\left(-\frac{1}{2} \beta (y - X^T w)^T (y - X^T w) - \frac{1}{2} \sum_{j=1}^{d} w^2 \alpha \right)
\end{aligned}$$

$$= \exp\left(-\frac{1}{2}\beta(y - X^T w)^T(y - X^T w) - \frac{1}{2}w^T D w\right)$$

where D is a $d \times d$ diagonal matrix with the values $\alpha_1, \alpha_2, \ldots, \alpha_d$

Completing the square seems reasonable at this point, now the goal is to complete the square, that is get the exponent in the form $\frac{1}{2}w^T A w + k^T w$. For convenience we will take the logarithm to remove the base exponential:

$$\begin{aligned}
\Rightarrow \log P(w|y, x, \alpha, \beta) &\propto -\frac{1}{2}\beta\left(y^T y - y^T X^T w - w^T X y + w^T X X^T w\right) - \frac{1}{2}w^T D w \\
&\propto -\frac{1}{2}\left[-2\beta y^T X^T w + \beta w^T X X^T w\right] - \frac{1}{2}w^T D w \\
&= \beta y^T X^T w - \frac{1}{2}\beta w^T X X^T w - \frac{1}{2}w^T D w \\
&= -\frac{1}{2}\left[w^T \beta X X^T w + w^T D w\right] + \beta y^T X^T w \\
&= -\frac{1}{2}w^T \underbrace{\left(\beta X X^T + D\right)}_{=A} w + \underbrace{\beta y^T X^T}_{=k^T} w
\end{aligned}$$

Then by completing the square we obtain the parameters S (covariance) and m (mean) for a Multivariate Gaussian that is proportional to $P(w|y, x, \alpha, \beta)$:

$$\begin{aligned}
S &= A^{-1} = \left(\beta X X^T + D\right)^{-1} \\
m &= A^{-1}k = S\beta X y \\
&\underline{\underline{\Rightarrow P(w|y, x, \beta) \propto \mathcal{N}(w|m, S)}}
\end{aligned} \tag{6}$$

1.2 Deriving conjugate prior for α

$$\begin{aligned}
P(\alpha|y, x, w, \beta) &\propto \prod_{j=1}^{d} P(w_j|\alpha_j)P(\alpha_j) \\
&= \prod_{j=1}^{d} N(w_j|0, \alpha^{-1}) \,\mathrm{Gamma}(a_j|c_0, d_0) \\
&\propto \prod_{j=1}^{d} \frac{1}{\sqrt{\alpha_j^- 1}} \exp\left(-\frac{1}{2}\left[\frac{w_j}{\sqrt{\alpha}^{-1}}\right]^2\right) \alpha_j^{c_0 - 1} \exp\left(-d_0 \alpha_j\right) \\
&= \prod_{j=1}^{d} \exp\left(-\frac{w_j^2}{2}\alpha_j - d_0\alpha_j\right) \alpha_j^{c_0 - 1 + 1/2} \\
&= \prod_{j=1}^{d} \exp\left(-\alpha_j\left(\frac{w_j^2}{2} + d_0\right)\right) \alpha_j^{c_0 - 1 + 1/2}
\end{aligned}$$

$$\propto \underline{\underline{\prod_{j=1}^{d} \text{Gamma}(\alpha_j | c_0 + \frac{1}{2}, \frac{w_j^2}{2} + d_0)}} \tag{7}$$

To sample an α_j from the distribution we simply sample from $\text{Gamma}(\alpha_j | c_0 - \frac{1}{2}, \frac{w_j^2}{2} + d_0)$.

1.3 Deriving conjugate prior for β

$$\begin{aligned}
P(\beta | y, x, w, \alpha) &\propto P(\beta) \prod_{i=1}^{n} P(y_i | w, x_i, \beta) \\
&= \text{Gamma}(\beta | a_0, b_0) \prod_{i=1}^{n} N(y_i | w^T x_i, \sqrt{\beta^{-1}}) \\
&= \beta^{a_0 - 1} \exp(-b_0 \beta) \prod_{i=1}^{n} \frac{1}{\sqrt{\beta^{-1}}} \exp\left(-\frac{1}{2} \left(\frac{y_i - w^T x_i}{\sqrt{\beta^{-1}}} \right)^2 \right) \\
&= \beta^{a_0 - 1} \exp(-b_0 \beta) \left(\beta^{1/2}\right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \beta (y_i - w^T x_i)^2 \right) \\
&= \beta^{a_0 - 1} \beta^{n/2} \exp\left(-b_0 \beta - \frac{\beta}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 \right) \\
&= \beta^{a_0 - 1 + n/2} \exp\left(-\beta \left(b_0 + \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 \right) \right) \\
&\propto \underline{\underline{\text{Gamma}(\beta | a_0 + n/2, b_0 + \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2)}}
\end{aligned} \tag{8}$$

1.4 Implementation

Only polynomial kernel lololol xD xD xD

2 Simulation study

To test the regression using ARD against the counterpart which does not use ARD we simulate some datasets to test on. The generated data will be generated using multivariate polynomials, as the implementation of the regression models are only capable of using polynomial features.

2.1 The datasets

The simulated data will be generated from 6 polynomial functions:

$$f_1(x, y) = -4x + 20y^2 x + 69 \tag{9}$$

$$f_2(x, y) = 69x + 69y - 420 \tag{10}$$
$$f_3(x, y) = x^2 - 2y^2 \tag{11}$$
$$f_4(x, y) = xy - x^2 + y^2 - 420 \tag{12}$$
$$f_5(x, y) = -x^3 + 42x^2 - 20x - y^3 + 42y^2 - 20y + 69 \tag{13}$$
$$f_6(x, y) = 2x^3 - y^3 - 3xy^2 + 3x^2y + x^3 - 3yx + 69 \tag{14}$$

The method of generating the data is simple. For each function f_i do:

1. Randomly generate some data X from the domain of f_i
2. Obtain response values y by inputting X into f_i
3. Offset each value of y with gaussian noise (same noise distribution for all elements in y)

The domain sampling distribution and noise distribution with respect for each function are as follows:

Table 1: Note that the domain distributions are univariate even though the functions are multivariate. Values from the domain distributions are simply sampled repeatedly for each variable, in this case x and y.

function	domain distribution	noise distribution
$f_1(x, y)$	$\text{Exp}(\text{rate} = 5)$	$N(0, 22)$
$f_2(x, y)$	$N(\text{mean} = 69, \text{std} = 4)$	$N(0, 420)$
$f_3(x, y)$	$\text{Unif}(\text{min} = -2, \text{max} = 20)$	$N(0, 69)$
$f_4(x, y)$	$N(\text{mean} = 0, \text{std} = 8)$	$N(0, 22)$
$f_5(x, y)$	$\text{Exp}(\text{rate} = 5)$	$N(0, 420)$
$f_6(x, y)$	$\text{Unif}(\text{min} = -6, \text{max} = 6)$	$N(0, 69)$

Then generating 128 data samples for each function produces the following plots:

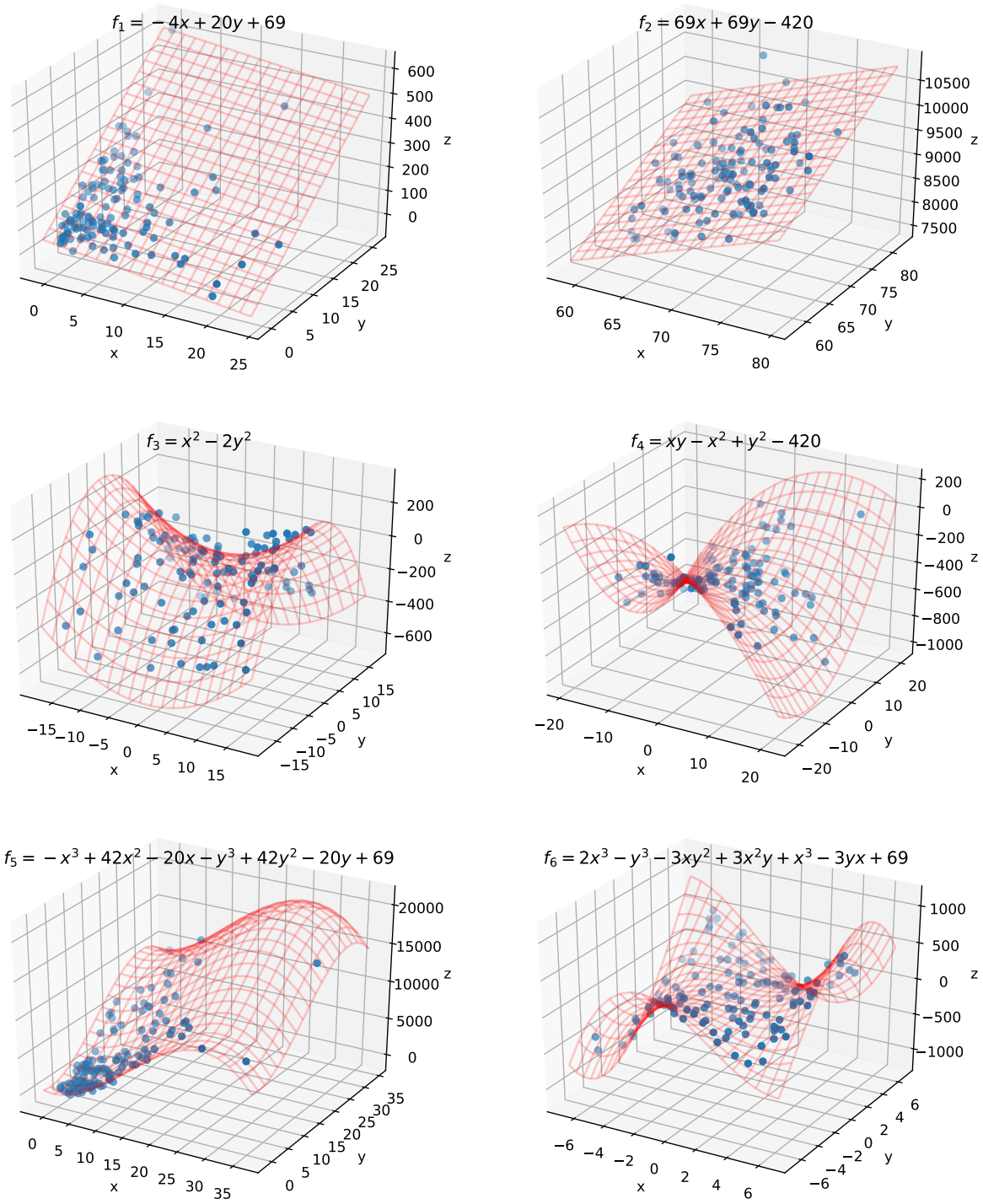
$f_1 = -4x + 20y + 69$
$f_2 = 69x + 69y - 420$
$f_3 = x^2 - 2y^2$
$f_4 = xy - x^2 + y^2 - 420$
$f_5 = -x^3 + 42x^2 - 20x - y^3 + 42y^2 - 20y + 69$
$f_6 = 2x^3 - y^3 - 3xy^2 + 3x^2y + x^3 - 3yx + 69$
x
y
z

Figure 2

2.2 Regression on the generated data

We fit regression models using ARD and without ARD to assess the differences in performance. We use the mean squared error between the true and estimated parameters. To assess the models' performances under presence of irrelevant features, we simply a add few extra features to the data that does not affect the response variables.

Table 2: "noise_features" column indicates number of added irrelevant features to the data. The degree column indicates which degree the polynomial regression was done in.

noise features	function	degree	ARD MAE	Regular MAE	Regular over ARD
0	f_1	1	18.27	109.55	6.00
	f_2	1	336.33	8914.98	26.51
	f_3	2	48.91	48.86	1.00
	f_4	2	17.65	249.43	14.13
	f_5	3	314.27	2326.37	7.40
	f_6	3	49.44	59.51	1.20
1	f_1	1	18.57	82.77	4.46
	f_2	1	347.24	8848.73	25.48
	f_3	2	54.41	55.03	1.01
	f_4	2	15.18	232.57	15.32
	f_5	3	322.23	1954.00	6.06
	f_6	3	53.24	56.67	1.06
2	f_1	1	16.91	72.16	4.27
	f_2	1	312.72	8816.07	28.19
	f_3	2	51.41	51.67	1.01
	f_4	2	16.46	218.45	13.27
	f_5	3	297.41	1975.98	6.64
	f_6	3	54.04	59.68	1.10
3	f_1	1	16.69	69.08	4.14
	f_2	1	325.36	8756.69	26.91
	f_3	2	55.46	55.37	1.00
	f_4	2	17.47	177.05	10.13
	f_5	3	335.03	2139.92	6.39
	f_6	3	54.70	59.95	1.10