1 Regression ([1] Chapter 2)

1.1 Weight-space view

1.1.1 Bayesian linear model

- We're trying to learn the distribution $p(y|X, W)$
  - X is the input data, W is the model parameters, y is the output
  - $p(y|X, W)$ is the conditional distribution of y after everything we know about X and W, distribution of errors
- Standard linear model: $f(X) = X^T W, y = f(X) + \epsilon$ with our Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2 I)$, which produces $p(y|X, W) = \mathcal{N}(y|f(X), \sigma_n^2)$
- Bayesian linear model: firstly, specify the linear prior distribution $p(W)$ over the weights $p(W) \sim \mathcal{N}(0, \Sigma_p)$
  - A prior $p(W)$ expresses our beliefs about the parameters before we see the data
  - Linear model specifies that our weights follow a zero mean Gaussian prior with a covariance matric Σ_p
- Then, we update our beliefs about the weights after seeing the data, using Bayes' theorem

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \tag{1}$$

$$p(W|X, y) = \frac{p(y|X, W)p(W)}{p(y|X)} \tag{2}$$

  - $p(y|X, W)$ is the density of the residuals after applying our priors $p(W)$ to the data X, W under our assumed noise model ϵ
  - $p(W)$ is the prior distribution of the weights
  - $p(y|X)$ is the marginal likelihood, which is the probability of the data given the model

$$p(y|X) = \int p(y|X, W)p(W)dW \tag{3}$$

  - $p(y|X)$ is the normalising constant, ensures the posterior distribution integrates to 1
  - $p(W|x, y)$ is the distribution of the weights given the data - combines the likelihood and the prior, representing everything we know about the parameters
- To understand how our posterior varies with our weights, we can write terms that only depend on weights (i.e. likelihood and prior, not marginal likelihood)

$$p(W|X, y) \propto p(y|X, W)p(W) \tag{4}$$

  - We will adopt the same idea throughout: if a term doesn't depend on weights, we simply remove it

1.1.2 Deriving our posterior

- Given our linear model $f(X) = X^T W$ and our Gaussian noise ϵ, we can write $p(Y|X, W)$ as the distribution of errors for each data point i

$$p(y|X, W) = \prod_{i=1}^{N} \mathcal{N}(y_i|X^T W, \sigma_n^2) \tag{5}$$

- We can find $p(y|X, W)$ by multiplying the Gaussian density $-\frac{1}{2\sigma_n^2}$ and the squared errors from our model $||y - X^T W||^2$, so our final likelihood becomes

$$p(y|X, W) = exp\left(-\frac{1}{2\sigma_n^2}||y - X^T W||^2\right) \tag{6}$$

- Given that our $p(W)$ is a zero mean Gaussian prior with covariance Σ_p, we can substitute this into the Gaussian density function

$$p(W) = \frac{1}{[\sqrt{\sigma_p}]\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{([W] - [0])}{[\Sigma_p]}\right) \tag{7}$$

- The first term of this $p(W)$ is another normalising constant, so rewriting the fraction in the exponent as a negative exponential gives us

$$p(W) \propto exp\left(-\frac{1}{2}W^T \Sigma_p^{-1} W\right) \tag{8}$$

- Putting both expressions for $p(y|X, W)$ and $p(W)$ together, we can write the posterior as

$$p(W|X, y) \propto exp\left(-\frac{1}{2\sigma_n^2}||y - X^T W||^2\right) exp\left(-\frac{1}{2}W^T \Sigma_p^{-1} W\right) \tag{9}$$

- To simplify, first we can expand $||y - X^T W||^2$ to $y^T y - 2y^T XW + W^T X^T XW$, and substitute this expanded expression to get

$$p(W|X, y) \propto exp\left(-\frac{1}{2\sigma_n^2}(y^T y - 2y^T XW + W^T X^T XW)\right) exp\left(-\frac{1}{2}W^T \Sigma_p^{-1} W\right) \tag{10}$$

- Then, we put both exponentials together (by adding their powers)

$$p(W|X, y) \propto exp\left(\frac{1}{\sigma_n^2}(y^T y - 2y^T XW + W^T X^T XW) + \left(-\frac{1}{2}W^T \Sigma_p^{-1} W\right)\right) \tag{11}$$

- We can rearrange the inside term to be a quadratic, linear and constant term in W:

$$p(W|X, y) \propto exp\left(\frac{1}{2}W^T\left(\frac{1}{\sigma_n^2}X^T X + \Sigma_p^{-1}\right)W - \left(\frac{1}{\sigma_n^2}y^T X\right)W + \frac{1}{2}y^T y\right) \tag{12}$$

- We can ignore the constant final term, and introduce $A = \Sigma_p^{-1} + \frac{1}{\sigma_n^2}X^T X$ and $b = \frac{1}{\sigma_n^2}y^T X$ to get

$$p(W|X, y) \propto exp\left(-\frac{1}{2}W^T AW + b^T W\right) \tag{13}$$

1.1.3 Deriving the properties of the posterior by completing the square

- Now we have a simplified form of the posterior density, we need to get it into a Gaussian form to recover the properties of the posterior distribution
- Firstly, we can bring all terms inside the exponential to a single term

$$-\frac{1}{2}W^TAW + b^TW = \frac{1}{2}\left(-W^TAW + 2b^TW\right) \tag{14}$$

- We can "complete the square" on this term $W^TAW - 2b^TW$ to rewrite it in a form that is easier to interpret

$$W^TAW - 2b^TW = (W - A^{-1}b)^TA(W - A^{-1}b) - b^TA^{-1}b \tag{15}$$

- Substituting this back into our posterior density gives us

$$p(W|X, y) \propto exp\left(-\frac{1}{2}\left((W - A^{-1}b)^TA(W - A^{-1}b) - b^TA^{-1}b\right)\right) \tag{16}$$

- If we look at our Gaussian density:

$$N(W|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}exp\left(-\frac{1}{2}(W - \mu)^T\Sigma^{-1}(W - \mu)\right) \tag{17}$$

- We can see that our expression lines up with the RHS Gaussian "kernel" term $exp\left(-\frac{1}{2}(W - \mu)^T\Sigma^{-1}(W - \mu)\right)$, where $\mu = A^{-1}b$ and $\Sigma^{-1} = A$ thus $\Sigma = A^{-1}$
- So we can write our posterior density in Gaussian form

$$p(W|X, y) \sim N(A^{-1}b, A^{-1}) \tag{18}$$

1.1.4 Gaussian posteriors and ridge regression

- For Gaussian posteriors, our mean $A^{-1}b$ is also its mode, called the maximum a posteriori (MAP) estimate of W
  - Due to symmetries in linear model and posterior, not the case in general
- In non-Bayesian settings the MAP point is the MLE estimation
- B is dependent on σ_n^2, y and X - all known
- A is dependent on σ_n^2, X and Σ_p - all known except Σ_p
- Our weight variance Σ_p under the Bayesian linear model is "isotropic", meaning it is the same in all directions

$$\Sigma_p = \tau^2 I \tag{19}$$

  - I is our $D \times D$ correlation matrix, here we assume independence so our correlation matrix is an "identity matrix" (each diagonal element is 1 and all off-diagonal elements are 0)
  - τ^2 is a scalar variance term, chosen as a prior

- We can substitute our new isotropic prior Σ_p into A to get

$$A = \Sigma_p^{-1} + \frac{1}{\sigma_n^2} X^T X = \left[\tau^2 I\right]^{-1} + \frac{1}{\sigma_n^2} X^T X = \frac{1}{\tau^2} I + \frac{1}{\sigma_n^2} X^T X = \frac{1}{\sigma_n^2}\left(X^T X + \frac{\sigma_n^2}{\tau^2} I\right) \tag{20}$$

- Now we have full expressions for A and B, we can substitute them into our MAP estimation for W to get

$$W_{\text{MAP}} = A^{-1} b = \left[\frac{1}{\sigma_n^2}(X^T X + \frac{\sigma_n^2}{\tau^2} I\right]^{-1} \cdot \left[\frac{1}{\sigma_n^2} y^T X\right] \tag{21}$$

- We can compute the LHS inversion of A

$$A^{-1} = \frac{\sigma_n^2}{X^T X + \frac{\sigma_n^2}{\tau^2} I} = \sigma_n^2 \left(X^T X + \frac{\sigma_n^2}{\tau^2} I\right)^{-1} \tag{22}$$

- Substituting this back into W_{MAP} cancels out the σ_n^2 term in A with the $\frac{1}{\sigma_n^2}$ term in B, giving us

$$W_{\text{MAP}} = \sigma_n^2 \left(X^T X + \frac{\sigma_n^2}{\tau^2} I\right)^{-1} \cdot \frac{1}{\sigma_n^2} y^T X = \left(X^T X + \frac{\sigma_n^2}{\tau^2} I\right)^{-1} \cdot y^T X \tag{23}$$

- The solution to ridge regression is very similar

$$W_{\text{ridge}} = \left(X^T X + \lambda I\right)^{-1} X^T y \tag{24}$$

- In ridge regression, λ is a regularisation parameter that controls the amount of shrinkage which is usually selected to maximise likelihood/minimise error
- MAP estimation in Bayesian linear regression with isotropic priors is equivalent to ridge regression with a regularisation parameter $\lambda = \frac{\sigma_n^2}{\tau^2}$
  - The higher our λ, the more biased our model is towards the prior, and the more we shrink our weights towards zero and the prior has more influence
  - A lower τ causes a higher λ - smaller weight variances around zero means lower weights because of higher confidence in priors
  - A higher σ_n also causes a higher λ - larger noise variances means lower weights because of lower confidence in weights forces deference to the prior

1.1.5 Deriving the predictive distribution

- Ultimately, our goal is to approximate a data-generating function f_* (or a new observation y_*) that produced a new X_* given training data X and y and weights W from the same f_*
- In non-Bayesian frameworks, we make predictions by choosing a single parameter value W to maximise the likelihood of the data, which is our MLE estimate

- In a Bayesian framework, we average over all possible parameter values weighted by their posterior probability $p(W|X,y)$, e.g. for a linear model $\hat{W} = \mathbb{E}_{p(W|X,y)}[W] = W_{\text{MAP}} = A^{-1}b$

- In this framework, we can make comments about our uncertainty of W by forming a "predictive distribution" $p(f_*|X_*,X,y)$

$$p(f_*|X_*,X,y) = \int p(f_*|X_*,W) \cdot p(W|X,y)dW \tag{25}$$

  - $p(f_*|X_*,W)$ is what we think the function looks like after producing a prediction using X_* and perfect knowledge of W
  - $p(W|X,y)$ is the posterior distribution of the weights given the training data, e.g. minimised for W_{MAP}
  - $p(f_*|X_*,W) \cdot p(W|X,y)$ is the joint distribution of our predictions and our posterior weights, which gets us the conditional distribution $p(f_*,W|X_*,X,y)$ by definition of conditional probability
  - $p(f_*,W|X_*,X,y)$ relies on our perfect knowledge of W, which we don't have, so we integrate over all possible W to get the predictive distribution $p(f_*|X_*,X,y)$

- We already know $p(W|X,y)$

$$p(W|X,y) \propto exp\left(\frac{1}{2}(-W^TAW + 2b^TW)\right) \tag{26}$$

- $p(f_*|X_*,W)$ is our errors, which we assume to be distributed normally and independently with our I identity matrix:

$$p(f_*|X_*,W) = \mathcal{N}(f_*|W^TX_*,\sigma_n^2 I) \tag{27}$$

- Plugging these into our Gaussian density and ignoring the LHS normalisation term yields

$$p(f_*|X_*,w) \propto exp\left(-\frac{1}{2}\frac{1}{\sigma_n^2}(f_* - W^TX_*)^2\right) \tag{28}$$

- We can multiply $P(f_*|X_*,W)$ and $p(W|X,y)$ to get our conditional $p(f_*,W|X_*,X,y)$, and add the exponents to simplify

$$p(f_*,W|X_*,X,y) \propto exp\left(\frac{1}{2}(-W^TAW + 2b^TW) + \left(-\frac{1}{2}\frac{1}{\sigma_n^2}(f_* - W^TX_*)^2\right)\right) \tag{29}$$

- We can further combine these with a single factor of $\frac{1}{2}$ to get

$$p(f_*,W|X_*,X,y) \propto exp\left(-\frac{1}{2}\left(W^TAW - 2b^TW + \frac{1}{\sigma_n^2}(f_* - W^TX_*)^2\right)\right) \tag{30}$$

- Expanding the squared term gives us

$$p(f_*,W|X_*,X,y) \propto exp\left(-\frac{1}{2}\left(W^TAW - 2b^TW + \frac{1}{\sigma_n^2}(f_*^2 - 2f_*W^TX_* + W^TX_*X_*^TX_*)\right)\right) \tag{31}$$

- Similar to our posterior, we can rearrange this to be a quadratic, linear and constant term in W

$$p(f_*, W|X_*, X, y) \propto exp\left(-\frac{1}{2}\left(W^T\left(A + \frac{1}{\sigma_n^2}X_*X_*^T\right)W - 2\left(b + \frac{1}{\sigma_n^2}f_*X_*\right)^T W + \frac{1}{\sigma_n^2}f_*^2\right)\right) \tag{32}$$

- By defining $A_* = A + \frac{1}{\sigma_n^2}X_*X_*^T$ and $b_* = b + \frac{1}{\sigma_n^2}f_*X_*$, we can rewrite this as

$$p(f_*, W|X_*, X, y) \propto exp\left(-\frac{1}{2}\left(W^T A_* W - 2b_*^T W + \frac{1}{\sigma_n^2}f_*^2\right)\right) \tag{33}$$

- We have to integrate this wrt W to get our predictive distribution $p(f_*|X_*, X, y)$

$$p(f_*|X_*, X, y) = \int p(f_*, W|X_*, X, y)dW \propto \int exp\left(-\frac{1}{2}\left(W^T A_* W - 2b_*^T W + \frac{1}{\sigma_n^2}f_*^2\right)\right)dW \tag{34}$$

- We can factor out the $\frac{1}{\sigma_n^2}f_*^2$ term from the integral, as it does not depend on W so remains the same since $\int exp(X)dX = exp(X)$

$$= exp\left(-\frac{1}{2}\frac{1}{\sigma_n^2}f_*^2\right) \times \int exp\left(-\frac{1}{2}\left(W^T A_* W - 2b_*^T W\right)\right)dW \tag{35}$$

- The RHS term is a multivariate Gaussian integral (beyond your paygrade) which evaluates to:

$$\int exp\left(-\frac{1}{2}\left(W^T A_* W - 2b_*^T W\right)\right)dW = \frac{(2\pi)^{D/2}}{\sqrt{|A_*|}}exp\left(\frac{1}{2}b_*^T A_*^{-1} b_*\right) \tag{36}$$

- Substituting this back into our predictive distribution gets us

$$p(f_*|X_*, X, y) \propto exp\left(-\frac{1}{2}\frac{1}{\sigma_n^2}f_*^2\right) + \frac{(2\pi)^{D/2}}{\sqrt{|A_*|}} \cdot exp\left(\frac{1}{2}b_*^T A_*^{-1} b_*\right) \tag{37}$$

  - Note that no part of our expression is now dependent on W

- Now we need an expression of everything that changes f_*

- Absorb the second term, since it does not depend on f_* into the proportionality constant, and combining the remaining exponential terms by adding their powers gives us

$$p(f_*|X_*, X, y) \propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_n^2}f_*^2 + \frac{1}{2}b_*^T A_*^{-1} b_*\right) \tag{38}$$

- Similar to deriving properties from our posterior, we can rearrange this expression, complete the square and derive the properties of our predictive distribution

$$p(f_*|X_*, W) \sim N(X_*^T A^{-1} b, \sigma_n^2 + X_*^T A^{-1} X_*) \tag{39}$$

- Predictive variance is quadratic form of test input with A^{-1}, showing that predictive uncertainties grow with size of X_*

1.1.6 Projections of inputs into feature space

- Bayesian linear models suffer from limited expressiveness due to the linearity of the model
- To address this, we can project our inputs into a higher dimensional feature space and apply linear model in this space
- e.g. a scalar x could be projected into the space of powers of x: $\phi(x) = [1, x, x^2, \ldots, x^d]^T$ for a polynomial basis expansion of degree d
- How to choose $\phi(x)$? Gaussian process formalism allows us to answer this question, but for now assume $\phi(x)$ is a given
- $\phi(X)$ maps a D-dimensional input vector X into an N dimensional feature space
- So our full model looks like:
$$f(X) = \phi(X)^T W \tag{40}$$
- And our predictive distribution becomes
$$p(f_*|X_*, X, y) = N(\phi(X_*)^T A_\phi^{-1} b_\phi, \sigma_n^2 + \phi(X_*)^T A_\phi^{-1} \phi(X_*)) \tag{41}$$
  - $A_\phi = \Sigma_p^{-1} + \frac{1}{\sigma_n^2} \phi(X)^T \phi(X)$
  - $b_\phi = \frac{1}{\sigma_n^2} \phi(X)^T y$

1.1.7 Avoiding inversion of A_ϕ

- This formulation of our predictive distribution inverts the $N \times N$ matrix A_ϕ to get the expected value and variance
- Inverting matrices is $O(N^3)$ - not feasible for large N - so we need to restate our predictive distribution in a form that avoids this inversion
- We can use the Sherman-Morrison identity (beyond your paygrade) to get an expression for A_ϕ^{-1} directly, where $K = \phi(X)^T \Sigma_p \phi(X)$
$$A_\phi^{-1} = \Sigma_p - \Sigma_p \phi(X)(K + \sigma_n^2 I)^{-1} \phi(X)^T \Sigma_p \tag{42}$$
- For the mean, we can use the Sherman-Morrison identity again to get an expression for $A_\phi^{-1}\phi(X)$
$$A_\phi^{-1}\phi(X) = \sigma_n^2 \Sigma_p \phi(X)(K + \sigma_n^2 I)^{-1} \tag{43}$$
- Substitute this into our predictive distribution mean
$$\mathbb{E}_{p(f_*|X_*,X,y)}[f_*] = \phi(X_*)^T \cdot A_\phi^{-1} \cdot \left[\frac{1}{\sigma_n^2}\phi(X)^T y\right] \tag{44}$$
- Factoring out σ_n^2
$$= \phi(X_*)^T \cdot \sigma_n^{-2}(A_\phi^{-1}\phi(X))y \tag{45}$$
- σ_n^-2 and σ_n^2 cancel out, leaving us with
$$\mathbb{E}_{p(f_*|X_*,X,y)}[f_*] = \phi(X_*)^T \cdot \Sigma_p \phi(X)(K + \sigma_n^2 I)^{-1} y \tag{46}$$

References

[1] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Nov. 2005. ISBN: 9780262256834. DOI: 10.7551/mitpress/3206.001.0001. eprint: https://direct.mit.edu/book-pdf/2514321/book_9780262256834.pdf. URL: https://doi.org/10.7551/mitpress/3206.001.0001.