This portfolio aims to cover the main concepts and results related to decision making and statistical learning. The notation to be used in the rest of this report will remain consistent and is described in Appendix A.

A statistical learning problem can typically be boiled down to a quantity y we want to predict based on a vector of explanatory variables $\mathbf{x}$, using a function $f(\mathbf{x})$. Often we are given a dataset $D \coloneqq \{(\mathbf{x_i}, y_i)\}_{i=1}^n$ which helps us determine a sensible prediction function.

1 Linear Regression

One way we can define $f(\mathbf{x})$ is as a linear function:

$$f(\mathbf{x}; \mathbf{w}) \coloneqq \langle \mathbf{w_1}, \mathbf{x} \rangle + w_0 = [\mathbf{x},\ 1]\ \mathbf{w},$$
$$\text{where } \mathbf{w} \coloneqq [\mathbf{w_1}, w_0]^T$$

Then our problem is reduced to finding an estimate for $\mathbf{w}$.

1.1 Least Squares Estimation

A common approach to finding an appropriate $\mathbf{w}$ is called least squares, where the general idea is to choose the $\mathbf{w}$ that minimises the squared errors across the dataset. Mathematically, we define this as:

$$\mathbf{w}_{LS} \coloneqq \arg\min_{\mathbf{w}} \sum_{i \in D} [y_i - f(\mathbf{x_i}; \mathbf{w})]^2$$

The main benefit of this is that we can obtain the closed form solution for this as

$$\mathbf{w}_{LS} = (\mathbf{X}\mathbf{X}^\mathbf{T})^{-1}\mathbf{X}\mathbf{y}^\mathbf{T}, \text{ where } \mathbf{X} = \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \\ 1 & \cdots & 1 \end{bmatrix}, \text{ and } \mathbf{y} = (y_1, \ldots, y_n)$$

The derivation of this expression is provided in Appendix B.1

1.2 Maximum Likelihood

Another approach to estimating $\mathbf{w}$ would be to choose the one that was most likely to produce our specific dataset, i.e. has the highest likelihood. In the linear regression context, we assume y_i are independent with distribution $\mathcal{N}_{y_i}(f(\mathbf{x_i}; \mathbf{w}), \sigma^2)$.

Then in regression the maximum likelihood function is given by

$$\mathbf{w}_{ML} \coloneqq \arg\max_{\mathbf{w}}\ \log \prod_{i=1}^{n} \mathcal{N}_{y_i}(f(\mathbf{x_i}; \mathbf{w}), \sigma^2)$$

Rearranging this we actually see that $\mathbf{w}_{ML} = \mathbf{w}_{LS}$, this calculation is shown in Appendix B.2.

1.3 Transformed Linear Regression

Sometimes the response is not linear with respect to the predictors. We can combat this by transforming the data before fitting $f(\mathbf{x}; \mathbf{w})$ while still having a closed form solution using least squares.

$$\mathbf{w}_{LS} \coloneqq \arg\min_{\mathbf{w}} \sum_{i \in D} \left[y_i - f'(\mathbf{x_i}, \mathbf{w})\right]^2$$
$$f'(\mathbf{x_i}, \mathbf{w}) \coloneqq \langle \mathbf{w_1}, \boldsymbol{\phi}(\mathbf{x}) \rangle + w_0,$$
$$\text{where } w \coloneqq [\mathbf{w_1}, w_0]^T \text{ as before}$$

We call the map $\boldsymbol{\phi}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^b$ a feature transform. The closed form solution for $\mathbf{w}_{LS}$ then becomes

$$\mathbf{w}_{LS} = \left(\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^T\right)^{-1} \boldsymbol{\phi}(\mathbf{X})\mathbf{y}^T$$

Remark 1.1 *If $\boldsymbol{\phi}$ is symmetric we can simplify the $\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^T$ term to $\boldsymbol{\phi}(\mathbf{X})^2$. Further, if $\boldsymbol{\phi}$ is invertible, our expression for $\mathbf{w}_{LS}$ becomes:*

$$\mathbf{w}_{LS} = \left(\boldsymbol{\phi}(\mathbf{X})^2\right)^{-1} \boldsymbol{\phi}(\mathbf{X})\mathbf{y}^T = \boldsymbol{\phi}^{-1}(\mathbf{X}) \left(\boldsymbol{\phi}^{-1}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})\right) \mathbf{y}^T = \boldsymbol{\phi}^{-1}(\mathbf{X})\mathbf{y}^T$$

or the rest of this chapter, we will only consider polynomial transforms, i.e., functions of the form $\boldsymbol{\phi}(x) \coloneqq \left[x, x^2, x^3, \ldots, x^b\right]^T$. To select b we refer to the following section.

1.4 Testing

The basis for most testing methods involves the following step:

1. We randomly split our data D into a training set D_0 and testing set D_1
2. We fit $\mathbf{w}_{LS}$ using D_0
3. Evaluate the testing error $E(D_1, \mathbf{w}_{LS}) = \sum_{i \in D_1} \left[y_i - f(\mathbf{x_i}; \mathbf{w})\right]^2$

The aim of the last step is to test the performance of our fit on unseen data. A low testing error would imply that the model is generalisable. An obvious way of selecting b then is to choose the value that produces the lowest testing error, then fit f_{LS} on the full dataset for that value of b. The issue with this is that the D_1 might contain pertinent information to the fit of the data that isn't incorporated with this fitting method.

1.4.1 Cross-Validation

A natural extension of this is to use k-fold cross validation, where k go up to $n-1$. The broad steps for this are:

1. Form a disjoint union of D, $D_0, \ldots, D_k$
2. For $i = 0, \ldots, k$ and for each b, fit $f_{LS}^{(i)}(b)$ on $D \backslash D_i$ and compute $E\left(D_i, f_{LS}^{(i)}(b)\right)$
3. Choose b such that it minimises $\frac{\Sigma_i E^{(i)}(b)}{k+1}$

While cross-validation is easy to implement, the computational cost is high. It also fails when the data is not iid, which occurs often in real world data.

1.5 High Dimensional Data

Often we have a situation where we have multiple variables contributing to an outcome y. For this we can introduce some new notation, we write $\mathbf{h}(t) \coloneqq \left[t^1, \ldots, t^b\right]$ as a polynomial transform and $\boldsymbol{\phi}(\mathbf{x}) \coloneqq \left[h\left(x^{(1)}\right), \ldots, h\left(x^{(d)}\right)\right]^T$ for a vector of transformed $\mathbf{x}$

Remark 1.2 *In this case $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{db}$ is still a vector since we are just increasing the number of covariates we are considering, and still only relates to one observation.*

Other terms we might consider are interaction terms, as the response might depend on the joint value of multiple inputs. In the case where we want pairwise cross-dimensional polynomials, our $\boldsymbol{\phi}(\mathbf{x})$ becomes:

$$\boldsymbol{\phi}(\mathbf{x}) \coloneqq \left[\mathbf{h}\left(x^{(1)}\right), \ldots, \mathbf{h}\left(x^{(d)}\right), x^{(u)}x^{(v)}_{u<v}\right] \in \mathbb{R}^{db+\binom{d}{2}}$$

The dimensionality of $\boldsymbol{\phi}$ actually increases exponentially with d, and since we need $n > d$ to perform least squares methods, we also need n to increase exponentially with d, which is often not reasonable so we look to a probabilistic approach to find other ways to increase the flexibility of $f(\mathbf{x})$.

1.6 Regularisation

If we want the more flexible fit to our predictions provided by high order polynomials, we can consider using a regularisation term. This is a penalty of the magnitude of the vector $\mathbf{w}^T\mathbf{w}$ (other examples of terms to use are in Appendix B.3):

$$\mathbf{w}_{LS-R} \coloneqq \arg\min_{\mathbf{w}} \sum_{i \in D} \left[y_i - f(\mathbf{x_i}, \mathbf{w})\right]^2 + \lambda \mathbf{w}^T\mathbf{w}$$

which has the solution

$$\mathbf{w}_{LS-R} = (\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^{\mathbf{T}} + \lambda\mathbf{I})^{-1}\boldsymbol{\phi}(\mathbf{X})\mathbf{y}^{\mathbf{T}}.$$

for which the proof is given in Appendix B.3. To see how this regularisation term helps reduce overfitting, when $\lambda \to \infty$ we have $f(x; \mathbf{w}_{LS-R}) \to 0$, i.e. our prediction function reduces in complexity and gets flatter as λ increases. Note, it still remains to choose a suitable λ. If we have a large iid data set, we can choose λ using cross validation. If however we have limited data or non iid data, we need to consider probabilistic methods.

1.7 Maximum A Posteriori

Note this section makes use of the Bayesian approach to linear regression, details for which are given in Appendix B.4. One way of choosing $\mathbf{w}$ is to select the value most likely given our dataset, i.e. the point $\mathbf{w}_{MAP}$ at which $p(\mathbf{w}|D)$ is highest

$$\mathbf{w}_{MAP} \coloneqq \arg\max_{\mathbf{w}} \prod_{i \in D} \mathcal{N}_{y_i}(f(\mathbf{x_i}, \mathbf{w}), \sigma^2) \cdot \mathcal{N}_{\mathbf{w}}(0, \sigma^2_{\mathbf{w}}\mathbf{I})$$

As shown in B.4, this is actually equivalent to finding $\mathbf{w}_{LS-R}$ for $\lambda = \frac{\sigma^2}{\sigma^2_{\mathbf{w}}}$.

1.8 Full Bayesian Approach

One flaw of MAP is that the uncertainty $p(\mathbf{w}|D)$ is not captured. We can remedy this by providing a predictive distribution $p(\hat{y}|\mathbf{x}, D)$ rather than a point estimate of y for any given $\mathbf{x}$. Using the definition of conditional distributions, we can write

$$\begin{aligned} p(\hat{y}|\mathbf{x}, D) &= \int p(\hat{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w} \\ &= \mathcal{N}_{\hat{y}}\left[f(\mathbf{x}, \mathbf{w}_{LS-R}), \sigma^2 + \boldsymbol{\phi}^T(\mathbf{x})\sigma^2\left(\boldsymbol{\phi}\boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma^2_{\mathbf{w}}}\boldsymbol{\phi}(\mathbf{x})\right)\right] \end{aligned}$$

With the proof for the latter equality provided in Appendix B.4

2 Binary Classification

With this type of problem, we are given an observed vector $\mathbf{x} \in \mathbb{R}^d$ and want to predict a binary output $y \in \{-1, +1\}$. We do this by choosing an f such that $f(x) = 0$ is a boundary separating the space of $\mathbf{x}$ into two regions R_+ and R_-, relating to a prediction of $+1$ and -1 respectively.

We can choose the decision boundary by minimising the number of misclassifications. For any observed $\mathbf{x}$ we can calculate the probability of such of errors as

$$\mathbb{P}(\mathbf{x} \text{ is misclassified}|f) = \int_{R_+} p(\mathbf{x}, y = -1)d\mathbf{x} + \int_{R_-} p(\mathbf{x}, y = +1)d\mathbf{x}$$

The f that minimises this is $f = p(\mathbf{x}, y = +1) - p(\mathbf{x}, y = -1)$, and is known as the Bayes optimal classifier. The proof for which is given in Appendix C.

2.1 Loss in Classification

The consequences of a false positive (FP) and a false negative (FN) might be different, which we can model using a loss matrix.Then we might instead want to minimise the expected loss of a decision, i.e. the optimal decision is given by

$$\arg\min_{y_0} \mathbb{E}_{p(y|\mathbf{x},D)}[L(y, y_0)|\mathbf{x}] = \arg\min_{y_0} \sum_y p(y|\mathbf{x}, D)L(y, y_0); \text{ where } L \coloneqq \begin{bmatrix} 0 & \text{Cost of FN} \\ \text{Cost of FP} & 0 \end{bmatrix}$$

The issue with this is that we do not have direct access to $p(y|\mathbf{x}, D)$, we can take two approaches with this:

1. **Discriminative**: this takes the direct approach where we model $p(y|\mathbf{x})$ with $p(y|\mathbf{x}, \mathbf{w})$
2. **Generative**: this is an indirect approach where we generate an input $\mathbf{x}$ given output y from $p(\mathbf{x}|y)$ and use this to obtain $p(y|\mathbf{x})$

Remark 2.1 *It is easy to see we will always have $p(y = +1|\mathbf{x}) + p(y = -1|\mathbf{x}) = 1$ but in some cases we may find both these individual probabilities are quite low. We may then want to define a region of $\mathbf{x}$ for which we reject making a decision. For a threshold α, we can define a corresponding region*

$$R = \{\mathbf{x} : \max\{p(y = +1|\mathbf{x}), p(y = -1|\mathbf{x})\} < \alpha\}$$

2.2 Link to Regression

In regression, we don't have a loss matrix, but we can use a loss function to be minimised. One such function, the square loss function, is given below (with the proof of the result given in Appendix C), :

$$\hat{y} \coloneqq \arg\min_{y_0} \mathbb{E}_{p(y|\mathbf{x})} \left[(y - y_0)^2|\mathbf{x}\right] = \mathbb{E}_{p(y|\mathbf{x})}[y]$$

A Notation

The notation used throughout this document is defined below

$$
\begin{aligned}
x, y, z &: \text{ scalars,} \\
\mathbf{x}, \mathbf{y}, \mathbf{z} &: \text{vectors,} \\
\mathbf{x} \in \mathbb{R}^d &: \text{ a vector in } d \text{ dimensional real space,} \\
x^{(i)} &: \text{ the } i^t h \text{ entry in } \mathbf{x}, \\
\mathbf{x_i} \in X &: \text{ the } i^{th} \text{ element of the set X,} \\
\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m &: \text{ takes vector } \mathbf{x} \text{ and maps it to } m \text{ dimensional space,} \\
\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{b \times d} &: \text{ matrices with } b \text{ rows and } d \text{ columns,} \\
\langle \mathbf{x}, \mathbf{y} \rangle &: \text{ the inner product of vectors } \mathbf{x} \text{ and } \mathbf{y}
\end{aligned}
$$

B Linear Regression

B.1 Least Squares

Proof of closed form LS solution:

Supposing $\mathbf{x_1}, \ldots, \mathbf{x_n}$ are column vectors of length d we can construct the matrix

$$\mathbf{X} \coloneqq \begin{bmatrix} \mathbf{x_1} & \ldots & \mathbf{x_n} \\ 1 & \ldots & 1 \end{bmatrix} \in \mathbb{R}^{(d+1) \times n}$$

We can easily see that applying f to $[\mathbf{x_1} \ \ldots \ \mathbf{x_n}]$ column-wise gives us

$$\mathbf{f}([\mathbf{x_1} \ \ldots \ \mathbf{x_n}]; \mathbf{w}) = \mathbf{X^T w}$$

We can also rewrite the expression for $\mathbf{w}_{LS}$ by the definition of a vector norm:

$$
\begin{aligned}
\mathbf{w}_{LS} &= \arg\min_{\mathbf{w}} \|\mathbf{y^T} - \mathbf{X^T w}\| = \arg\min_{\mathbf{w}} \left[(\mathbf{y^T} - \mathbf{X^T w})^\mathbf{T} (\mathbf{y^T} - \mathbf{X^T w})\right] \\
&= \arg\min_{\mathbf{w}} \left[\mathbf{y y^T} - \mathbf{y X^T w} - \mathbf{w^T X y^T} + \mathbf{w^T X X^T w}\right] \\
&= \arg\min_{\mathbf{w}} \left[\mathbf{y y^T} - \mathbf{2 w^T X y^T} + \mathbf{w^T X X^T w}\right]
\end{aligned}
$$

Since we are trying to minimise a function of $\mathbf{w}$, we can simply differentiate the function w.r.t $\mathbf{w}$ and set the result to 0. This yields:

$$-2\mathbf{X y^T} + \mathbf{X X^T w}_{LS} = 0 \implies \mathbf{w}_{LS} = \left(\mathbf{X X^T}\right)^{-1} \mathbf{X y^T}$$

Remark B.1 *If $n < d$ then the matrix $\mathbf{X}$ does not acheive full columnn rank, and thus $\mathbf{X X^T}$ cannot be inverted, rendering our expression for $\mathbf{w}_{LS}$ invalid. One way we might rectify this is by removing covariates that are strongly correlated with each other. The intuition behind this is that the information these two rows contribute to the model is similar enough that only one is enough. We want to remove enough rows such that we have full column rank.*

B.2 Maximum Likelihood

Proof that least squares and maximum likelihood are equivalent:

Under the independence assumption, the ML estimate of $\mathbf{w}$ becomes:

$$\begin{aligned}\mathbf{w}_{ML} &\coloneqq \arg\max_{\mathbf{w}} \log \prod_{i=1}^{n} \mathcal{N}_{y_i}(f(\mathbf{x_i};\mathbf{w}),\sigma^2) \\ &= \arg\max_{\mathbf{w}} \left[\sum_{i=1}^{n} -\frac{(y_i - f(\mathbf{x_i};\mathbf{w})^2}{2\sigma^2}\right] - n\log\sigma\sqrt{2\pi} \\ &= \arg\min_{\mathbf{w}} \sum_{i\in D_0} [y_i - f(\mathbf{x_i};\mathbf{w})]^2 = \mathbf{w}_{LS}\end{aligned}$$

Remark B.2 *The normality assumption on the outcome might not always be valid. For example:*

- *The response might be bounded rather than able to take any values in the real numbers (e.g. exponentially distributed data)*
- *The response might take discrete integer values rather than being continuous (e.g. Poisson distributed data)*

B.3 Regularisation

Proof of expression of regularlised least squares estimate:

We replace $\mathbf{X}$ in the standard least squares expression with $\boldsymbol{\phi}(\mathbf{X})$ since we are considering the transformed $\mathbf{X}$

$$\sum_{i\in D} [y_i - f(\mathbf{x}_i;\mathbf{w})]^2 = \mathbf{y}\mathbf{y}^T - 2\mathbf{w}^T\boldsymbol{\phi}(\mathbf{X})\mathbf{y}^T + \mathbf{w}^T\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^T\mathbf{w},$$

so we can rewrite the objective function as

$$\begin{aligned}\mathbf{w}_{LS-R} &= \arg\min_{\mathbf{w}} \left[\mathbf{y}\mathbf{y}^T - 2\mathbf{w}^T\boldsymbol{\phi}(\mathbf{X})\mathbf{y}^T + \mathbf{w}^T(\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^T + \lambda\mathbf{I})^{-1}\mathbf{w}\right] \\ &= \arg\min_{\mathbf{w}} \left[\mathbf{w}^T(\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^T + \lambda\mathbf{I})^{-1}\mathbf{w} - 2\mathbf{w}^T\boldsymbol{\phi}(\mathbf{X})\mathbf{y}^T\right]\end{aligned}$$

Differentiating with respect to $\mathbf{w}$ and rearranging we reach the closed form solution for regularised least squares as:

$$\mathbf{w}_{LS-R} = (\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^{\mathbf{T}} + \lambda\mathbf{I})^{-1}\boldsymbol{\phi}(\mathbf{X})\mathbf{y}^{\mathbf{T}}.$$

Other Regularisation Terms

In fact, we can use any norm on $\mathbf{w}$. A common class of these is L^p norms, the term we used is in fact one of these; the L^2 norm. The general form for these is:

$$||\mathbf{x}||_p = (|x_1|^p + \ldots + |x_d|^p)^{\frac{1}{p}}$$

These all penalise the size of $\mathbf{w}$ slightly differently, in fact we have for any p $||\mathbf{x}||_p \leq ||\mathbf{x}||_{p+1}$.

B.4 Bayesian Approach

In a generalised scenario, we have a random data set D, and we want to quantify the underlying process. In regression, we observe the output y_i, which we assume is generated by $y = g(x_i) + \epsilon$ where ϵ models the noisiness of the data. The unobserved process g is what we are trying to model. We do this by trying to infer the probability distribution of g given the dataset, $p(g|D)$, which we call the posterior.

To obtain the posterior we use Bayes' Rule to rewrite this conditional probability as

$$p(g|D) = \frac{p(D|g)p(g)}{p(D)}$$

with the necessary functions defined below

$$\begin{aligned} p(D|g) &: \text{the likelihood of observing a specified data set D given} \\ p(g) &: \text{the prior distribution - this contains our beliefs about the data} \\ &\quad \text{generating process before observing the data} \\ p(D) &: \text{the marginal probability of observing the dataset } D \end{aligned}$$

Proof for expression of maximum a posteriori estimate

Since the log is an increasing function we can apply this to get the $\mathbf{w}_{MAP}$. First considering each of the terms we see that

$$\log\left[\prod_{i\in D} \mathcal{N}_{y_i}(f(\mathbf{x_i}, \mathbf{w}), \sigma^2)\right] = -\frac{1}{2\sigma^2}\sum_{i\in D}(y_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i))^2$$

$$\log\left[\mathcal{N}_{\mathbf{w}}(0, \sigma^2_{\mathbf{w}}\mathbf{I})\right] = -\frac{1}{2}\mathbf{w}^T\left(\frac{1}{\sigma^2_{\mathbf{w}}}\mathbf{I}\right)\mathbf{w} = -\frac{1}{2\sigma^2_{\mathbf{w}}}\mathbf{w}^T\mathbf{w}$$

Then returning this into our expression for $\mathbf{w}_{MAP}$ we get

$$\begin{aligned} \mathbf{w}_{MAP} &= \arg\max_{\mathbf{w}}\left[-\frac{1}{2\sigma^2}\sum_{i\in D}(y_i^2 - 2y_i\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}) - \frac{1}{2\sigma^2_{\mathbf{w}}}\mathbf{w}^T\mathbf{w}\right] \\ &= \arg\max_{\mathbf{w}}\left[-\frac{1}{2}\mathbf{w}^T\left(\sum_{i\in D}\frac{1}{\sigma^2}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x})^T + \frac{1}{\sigma^2_{\mathbf{w}}}\mathbf{I}\right)\mathbf{w} - \frac{1}{2}\left(\sum_{i\in D} -\frac{2}{\sigma^2}y_i\boldsymbol{\phi}(\mathbf{x}_i)^T\mathbf{w}\right)\right] \\ &= \arg\max_{\mathbf{w}}\left[\mathbf{w}^T\left(\boldsymbol{\phi}(\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^T + \frac{\sigma^2}{\sigma^2_{\mathbf{w}}}\mathbf{I}\right)\mathbf{w} - 2\mathbf{w}^T\boldsymbol{\phi}(\mathbf{X})\mathbf{y}^T\right] \end{aligned}$$

Which if we take $\lambda = \frac{\sigma^2}{\sigma^2_{\mathbf{w}}}$, we can recognise this as the expression for $\mathbf{w}_{LS-R}$, hence we conclude $\mathbf{w}_{LS-R} = \mathbf{w}_{MAP}$ as desired.

Proof for distribution of a prediction

If we assume $g(\mathbf{x}) = f(\mathbf{x}; \mathbf{w})$, we can replace g with the $\mathbf{w}$ vector we are trying to estimate. Further assuming $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and D is iid, we can write:

$$p(D|\mathbf{w}) = \prod_{i\in D} p(y_i|\mathbf{x_i}, \mathbf{w}, \sigma^2) = \prod_{i\in D}\mathcal{N}_{y_i}(f(\mathbf{x_i}; \mathbf{w}), \sigma^2)$$

We also need the prior $p(\mathbf{w})$, which we take to be a normal distribution: $\mathbf{w} \sim \mathcal{N}(0, \sigma_{\mathbf{w}}^2 \mathbf{I})$. Our expression for the posterior then becomes

$$\begin{aligned} p(\mathbf{w}|D) &= \frac{\prod_{i \in D} \mathcal{N}_{y_i}(f(\mathbf{x_i}; \mathbf{w}), \sigma^2) \cdot \mathcal{N}_{\mathbf{w}}(0, \sigma_{\mathbf{w}}^2 \mathbf{I})}{p(D)} \\ &\propto \prod_{i \in D} \mathcal{N}_{y_i}(f(\mathbf{x_i}; \mathbf{w}), \sigma^2) \cdot \mathcal{N}_{\mathbf{w}}(0, \sigma_{\mathbf{w}}^2 \mathbf{I}) \end{aligned}$$

where we can disregard $p(D)$ since it does not depend on $\mathbf{w}$. We can ignore any constants not dependent on $\mathbf{w}$. We can also take the log for simplicity. Since we have already performed this simplification above for $\mathbf{w}_{MAP}$, we know

$$\log(p(\mathbf{w}|D) = \mathbf{w}^T \left(\boldsymbol{\phi}(\mathbf{X}) \boldsymbol{\phi}(\mathbf{X})^T + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2} \mathbf{I} \right) \mathbf{w} - 2 \mathbf{w}^T \boldsymbol{\phi}(\mathbf{X}) \mathbf{y}^T$$

Then by matching powers we can see (and denoting $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{X})$)

$$p(\mathbf{w}|D) = \mathcal{N} \left(\mathbf{y}^T \boldsymbol{\phi} \sigma^2 \left[\boldsymbol{\phi} \boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2} \mathbf{I} \right]^{-1}, \sigma^2 \left[\boldsymbol{\phi} \boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2} \mathbf{I} \right]^{-1} \right)$$

Using results regarding conditional and marginal distributions given in Section 2.3 of Bishop (2006)[1], which states that given distributions:

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}), \quad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L})$$

the marginal distribution of $\mathbf{y}$ is

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L} + \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T)$$

Since we have access to $p(\mathbf{w}|D)$ as above and by assumption $p(\hat{y}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}, \sigma^2)$, we can see

$$\begin{aligned} \boldsymbol{\mu}_{\hat{y}} &= \boldsymbol{\phi}(\mathbf{x})^T \mathbf{y}^T \boldsymbol{\phi} \sigma^2 \left[\boldsymbol{\phi} \boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2} \mathbf{I} \right]^{-1} = \boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}_{LS-R} = f(\mathbf{x}, \mathbf{w}_{LS-R}) \\ \boldsymbol{\Sigma}_{\hat{y}} &= \sigma^2 + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{y}^T \boldsymbol{\phi} \sigma^2 \left[\boldsymbol{\phi} \boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma_{\mathbf{w}}^2} \mathbf{I} \right]^{-1} \boldsymbol{\phi}(\mathbf{x}) \end{aligned}$$

and so we have obtained the marginal distribution of $\hat{y}$.

C Binary Classification

Proof for Bayes Optimal Classifier

We want to minimise the following expression with respect to f

$$\mathbb{P}(\mathbf{x} \text{ is misclassified}|f) = \int_{R_+} p(\mathbf{x}, y = -1) d\mathbf{x} + \int_{R_-} p(\mathbf{x}, y = +1) d\mathbf{x}$$

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

We can re-write this as

$$\begin{aligned}
\mathbb{P}(\mathbf{x} \text{ is misclassified}|f) &= \int_{R_+} p(\mathbf{x}, y=-1)d\mathbf{x} + \int_{R_-} p(\mathbf{x}, y=+1)d\mathbf{x} \\
&= \int_{\mathbf{x}} \mathbb{I}\left[f(x) \geq 0\right] p(\mathbf{x}, y=-1)d\mathbf{x} + \int_{\mathbf{x}} \mathbb{I}\left[f(x) < 0\right] p(\mathbf{x}, y=+1)d\mathbf{x} \\
&= \int_{\mathbf{x}} \mathbb{I}\left[f(x) \geq 0\right] \left(p(\mathbf{x}, y=-1)d\mathbf{x} - p(\mathbf{x}, y=+1)\right) + \int_{\mathbf{x}} p(\mathbf{x}, y=+1)d\mathbf{x} \\
&= \int_{\mathbf{x}} \mathbb{I}\left[f(x) \geq 0\right] \left(p(\mathbf{x}, y=-1) - p(\mathbf{x}, y=+1)\right) d\mathbf{x} + p(y=+1)
\end{aligned}$$

Since we are interested in finding the function f that minimises the expression, we can disregard the second term and focus only on the integral term. To minimise the integral term, we want to choose an f such that whenever $(p(\mathbf{x}, y=-1) - p(\mathbf{x}, y=+1)) > 0$ the indicator function is 0 and 1 when the expression is negative. This occurs when $f = p(\mathbf{x}, y=+1) - p(\mathbf{x}, y=-1)$.

Proof of regression predictions with respect to loss function:

In the following I will use $\mathbb{E}$ in place of $\mathbb{E}_{p(y|\mathbf{x})}$ for simplicity since this is the only expectation we consider. In the case of the square loss, our objective function is

$$\mathbb{E}\left[(y - y_0)^2|\mathbf{x}\right] = \mathbb{E}[y^2] - 2y_0\mathbb{E}[y] + y_0^2$$

Which we can differentiate with respect to y_0 to obtain the optimal $\hat{y}$

$$-2\mathbb{E}[y] + 2\hat{y} = 0 \implies \hat{y} = \mathbb{E}[y]$$