Statistical Methods: Portfolio 1

By Henry Bourne

1 Introduction to Decision-Making

Computers are often used to answer complex questions, however, these methods are often a "black-box". We would like to have methods which let us conduct **rational decision-making**:

1. Predictions should be precise (no gibberish)
2. They should be data driven
3. They should take cost (of making the wrong decision) into consideration
4. They should take the random nature of data into consideration

In a **regression problem** we want to predict an outcome given some known inputs. We can use the following as an objective function:

Definition 1.1 ***Least Squares (LS):***

$$\min_f \sum_{i \in D_0} (y_i - f(x_i))^2 \tag{1}$$

where f is the function that gives our prediction for x, where x_i is the i-th input, y_i is the i-th (observed) output and $D_0 \subseteq D$ is the training dataset.

By minimizing this objective function wrt. f, we obtain a function f that minimizes the squared difference between its predictions and our observed values of the target variable.

Let $\boldsymbol{w}$ be a vector parameterising f [1], then the LS solution is $\boldsymbol{w}_{LS} := \text{argmin}_{\boldsymbol{w}} \sum_{i \in D_0} (y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2$. We can prove that:

$$\boldsymbol{w}_{LS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{2}$$

The proof can be found in section A.1.

Alternatively we can find $\boldsymbol{w}$ in another data-driven way but also whilst taking the randomness of the data into account by using a probabilistic approach. We define a new objective function:

Definition 1.2 ***Maximum likelihood estimation:***

$$\max_{\boldsymbol{w}} \log \mathbb{P}(D|\boldsymbol{w}) \tag{3}$$

*we denote the parameters that maximize this as $\boldsymbol{w}_{ML}$, this is called the **Maximum Likelihood Estimator (MLE)**, we can write,*

$$\boldsymbol{w}_{ML} := \underset{\boldsymbol{w}}{\text{argmax}} \log \mathbb{P}(D|\boldsymbol{w}) \tag{4}$$

*where D is the dataset and $\mathbb{P}(D|\boldsymbol{w})$ is called the **likelihood**.*

Note that we can show $\boldsymbol{w}_{ML} = \boldsymbol{w}_{LS}$ [2]. We can also show that $\sigma^2_{ML} = \frac{1}{n} ||\boldsymbol{y} - f(\boldsymbol{x}; \boldsymbol{w}_{ML})||^2$.

1.1 LS with Feature Transform

It is possible to fit non-linear curves to our data using linear LS. All we do is augment our predictor variable, $\boldsymbol{x}$, using a function $\phi(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^{d \times b}$, where:

$$\phi(\boldsymbol{x}) := (\boldsymbol{x}, \boldsymbol{x}^2, ..., \boldsymbol{x}^b)^T \tag{5}$$

With a feature transform we have $\boldsymbol{w}_{LS} = (\phi(\boldsymbol{X})^T \phi(\boldsymbol{X}))^{-1} \phi(\boldsymbol{X})^T \boldsymbol{y}$. A feature transform can let us fit more complex models for our data, however, it can lead to problems...

[1] In the case of linear LS we have: $f(\boldsymbol{x}; \boldsymbol{w}) := \langle \boldsymbol{w_1}, \boldsymbol{x} \rangle + w_0$.

[2] This is true in cases where the underlying data-generating process is normal, ie. error terms Gaussian and iid.

2 Overfitting and the curse of dimensionality

In previous section we introduced the feature transform. By increasing the value of b we can fit increasingly complex models (models with terms with higher powers of x). We note that when we increase b our model can "bend" to better fit our data-points, however there reaches a point where if it can bend too much then our resultant model is going to be too specific to our training dataset and not generalize well when tested on more data.

2.1 Overfitting

Let's split our data into disjoint sets D_0 and D_1 [3]. We'll use LS as our error function and denote it, E_{LS}, so $E_{LS}(D, \boldsymbol{w}) := \sum_{i \in D}(y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2$. The error function quantifies how well our model fits a dataset. We will use our partitioned datasets such that the **training error** is $E(D_0, \boldsymbol{w}_{LS})$ and the **testing error** is $E(D_1, \boldsymbol{w}_{LS})$.

The testing error tells us how well the model fits data it hasn't seen; how well it **generalizes**. Again if we consider fitting models with different values of b what we will see is that as we initially increase b the testing and training error decrease. However, at a certain point the testing error will begin to increase as the training error continues decreasing. What is happening here is called:

Definition 2.1 ***Overfitting:***
Phenomenon where $f(\boldsymbol{x}; \boldsymbol{w}_{LS})$ fits too well on the training set, D_0, while under-performing on unseen (testing) datasets, D_1.

2.2 Cross-Validation

Consider the problem of how best to test the performance of our model given a dataset, D. We would like to both train our dataset on as much data as possible, but also evaluate its performance on unseen (testing) data such that the score quantifies well the model's ability to generalize. With our scenario earlier where we partitioned the dataset into D_0 (training) and D_1 (testing) datasets we can identify some shortcomings:

1. We have wasted D_1 for validation (testing), what if D_1 contains useful information for fitting a good model?
2. The selection of D_0 and D_1 is random, perhaps D_1 is better for training, or perhaps a different selection altogether would be better.

We now introduce a method which aims to solve these shortcomings:

Definition 2.2 ***K-fold Cross-Validation (CV)***
Split D into disjoint $D_0, ..., D_k$,
for $i = 0, ..., k$:
fit $f^{(i)}$ on all subsets but D_i
for all b compute: $E(D_i, f^{(i)})$
select b which minimizes: $\frac{\sum_i E(D_i, f^{(i)})}{k+1}$
Note: the k picked must be leqn -1 and if $k = n - 1$ then we call this ***leave-one-out-validation****.*

Although CV solves the problems stated earlier it does have its problems of its own:

1. Computational cost is high, as $f^{(i)}$ must be fitted and validated for all splits
2. The effectiveness of CV depends on the assumption that the data is iid. and often this assumption doesn't hold (eg. time series data)

2.3 Curse of dimensionality

Let's consider carrying out polynomial transform on higher dimensional datasets. When our input $\boldsymbol{x} \in \mathbb{R}^d$ then $\phi(\boldsymbol{x}) \in \mathbb{R}^{d \times b}$ and $\boldsymbol{w} \in \mathbb{R}^{b+1}$. And this is without considering pairwise cross-dimensional polynomials (eg. $x^{(1)}x^{(2)}$ or $x^{(1)}x^{(2)}x^{(3)}$). We can implement this by redesigning ϕ. Let $\phi(\boldsymbol{x}) := (h(x^{(1)}), ..., h(x^{(d)}), \forall_{u<v} x^{(u)}x^{(v)})$, where $h(t) := (t^1, ..., t^b)$, then $\phi(\boldsymbol{x}) \in \mathbb{R}^{d \times (b + \binom{d}{2})}$.

[3] We assume D contains iid. data-points

We can do the same for cross terms all the way up to d-plets and we know $\binom{d}{1}+\binom{d}{2}+...+\binom{d}{d}=2^d$. The output dimension of $\phi(\boldsymbol{x})$ can grow exponentially with dimensionality[4] and the number of observations much at least match the dimensionality, otherwise we cannot obtain $\boldsymbol{w}_{LS}$.

Definition 2.3 ***The Curse of Dimensionality:***
A phenomenon where the number of observations needed to solve a problem grows exponentially with d, it 'forbids' us from solving high-dimensional problems.

3 Regularization and a Probabilistic View of Regression

3.1 Regularization

A trick for avoiding overfitting is regularization. Let's consider letting our regularization term be $\lambda \boldsymbol{w}^T\boldsymbol{w}$, where $\lambda > 0$. Note that $\boldsymbol{w}^T\boldsymbol{w}$ is the magnitude of $\boldsymbol{w}$, so by including this term in our objective function we are dissuading $\boldsymbol{w}$ taking large values and hence reduces the chance of overfitting. Let $\boldsymbol{w}_{LS-R}$ denote our regularized linear LS solution for the parameters, then:

$$\boldsymbol{w}_{LS-R} := \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i\in D}(y_i - f(\boldsymbol{x}_i;\boldsymbol{w}))^2 + \lambda \boldsymbol{w}^T\boldsymbol{w} \tag{6}$$

We can prove that if we are using regularization term $\lambda \boldsymbol{w}^T\boldsymbol{w}$ then:

$$\boldsymbol{w}_{LS-R} := (\phi(\boldsymbol{X})^T\phi(\boldsymbol{X}) + \lambda \boldsymbol{I})^{-1}\phi(\boldsymbol{X})^T\boldsymbol{y} \tag{7}$$

The proof is in section A.2.

When we increase λ the regularization term gets larger which means the coefficients of $\boldsymbol{w}_{LS-R}$ must get smaller in order to decrease the value of the objective function, this in turn leads to a reduction in complexity of $f(\boldsymbol{x};\boldsymbol{w}_{LS-R})$ and vice-versa. When we increase λ we expect to see at first a decrease in testing error (if the model was overfitted), but after a certain point the testing error to rise again. This is because increasing λ increases the generalization of the model which will reduce overfitting, however, there will reach a point where it begins to over generalize. We can also use different regularization terms, usually we use norms:

Definition 3.1 ***Norm:***
To be a norm a positive function t must satisfy:

1. *If* $t(\boldsymbol{x}) = 0$ *then* $\boldsymbol{x} = 0$
2. $t(\boldsymbol{x}) + t(\boldsymbol{y}) \geq t(\boldsymbol{x}+\boldsymbol{y})$ *(Triangle inequality)*
3. $t(a \cdot \boldsymbol{x}) = |a| \cdot t((x))$

A special class of norm is the $\boldsymbol{L^p}$ **norm** which for a real $p \geq 1$ is $||\boldsymbol{x}||_p := (|x_1|^p + ... + |x_d|^p)^{\frac{1}{p}}$.

3.2 A Probabilistic View of Regression

We are often faced with inverse problems. An **inverse problem** is where we have a dataset, D, of noisy observations, and we want to identify some latent unobserved data generating mechanism. Let's denote our latent function g, then the key to solving an inverse problem is inferring the posterior distribution $\mathbb{P}(g|D)$. If we let $\boldsymbol{w}$ parameterize our latent function then all we must do is infer $\mathbb{P}(\boldsymbol{w}|D)$ or the:

Definition 3.2 ***Maximum A Posteriori (MAP)***
The $\boldsymbol{w}$ *that maximizes* $\mathbb{P}(\boldsymbol{w}|D)$*, we denote the MAP estimator as:*

$$\boldsymbol{w}_{MAP} := \underset{w}{\operatorname{argmax}}\, \mathbb{P}(\boldsymbol{w}|D) \tag{8}$$

$$= \underset{w}{\operatorname{argmax}} \prod_{i\in D} N_{y_i}(f(\boldsymbol{x}_i;\boldsymbol{w}), \sigma^2) \cdot N_{\boldsymbol{w}}(0, \sigma^2_{\boldsymbol{w}}\boldsymbol{I}) \tag{9}$$

[4] We could also include even more complex terms such as $(x^{(u)})^2x^{(v)}$ for example

In section A.3 we prove that $\boldsymbol{w}_{MAP} = \boldsymbol{w}_{LS-R}$ using $\lambda = \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2}$.

This approach of estimating $\boldsymbol{w}$ is probabilistic, but we can go further, we can go fully probabilistic. What we can do is consider finding the distribution of $\mathbb{P}(\hat{y}|\boldsymbol{x}, D)$, of our target variable, the **predictive distribution**. We know that $\mathbb{P}(\hat{y}|\boldsymbol{x}, D) = \int \mathbb{P}(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) \cdot \mathbb{P}(\boldsymbol{w}|D) d\boldsymbol{w}$. We can assume $\mathbb{P}(\hat{y}|\boldsymbol{x}, \boldsymbol{w})$ is distributed $N(f(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$, and we have that $\mathbb{P}(\boldsymbol{w}|D) \propto \prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2) \cdot N_{\boldsymbol{w}}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I})$. In section A.4 we prove:

Suppose $f(\boldsymbol{x}; \boldsymbol{w}) = \langle \boldsymbol{w}, \phi(\boldsymbol{X}) \rangle$, prove the following,

$$\int \mathbb{P}(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) \cdot \mathbb{P}(\boldsymbol{w}|D) d\boldsymbol{w} = N_{\hat{y}}(f(\boldsymbol{x}; \boldsymbol{w}_{LS-R}), \sigma^2 + \boldsymbol{\phi}^T(\boldsymbol{x}) \sigma^2 (\boldsymbol{\phi}\boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I}) \boldsymbol{\phi}(\boldsymbol{x})) \tag{10}$$

where $\boldsymbol{\phi}$ is short for $\phi(\boldsymbol{X})$ and $\boldsymbol{w}_{LS-R}$ is the LS-R solution with $\lambda = \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2}$

4 Risk and Bayes Optimal Prediction

Sometimes we need to make decisions, this is where classification comes into play. In **binary classification** we have an input, $\boldsymbol{x} \in \mathbb{R}^d$, an output, $y \in \{+1, -1\}$[5], and a task where given $\boldsymbol{x}$ we make a prediction y. In classification our aim is to find a **decision boundary**, which carves the input space into areas belonging to a certain prediction (or class). In binary classification we can define the kind of error we make: a **False Positive (FP)** is where a $\boldsymbol{x}$ should be labelled "-1" but is labelled "+1", a **False Negative (FN)** on the other hand is where $\boldsymbol{x}$ is wrongly labelled "-1", similarly we can define **True Positive (TP)** and **True Negative (TN)**. Let f be our **level-set function**, that is the value of $f(\boldsymbol{x})$ determines whether our prediction of $\boldsymbol{x}$ is $+1$ or -1. The optimal level set function is:

Definition 4.1 ***Bayes Optimal Classifier***
$f(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{x}, y = +1) - \mathbb{P}(\boldsymbol{x}, y = -1)$

As $\mathbb{P}(\text{FP or FN}|f)$ is minimized when f defined as above (proof in section A.5). However, this only serves as an ideal optimal classifier, in reality we don't have access to $\mathbb{P}(\boldsymbol{x}, y)$, we only have some data-points.

Sometimes wrong decisions may have different losses associated with them, we can define a **loss matrix**, $\boldsymbol{L}$, where entry i, j of the loss matrix is the loss associated with predicting j instead of the target i. To make an optimal decision whilst taking the risk into account we would like to minimize the expected loss of making a wrong decision: $\text{argmin}_{y_0} \mathbb{E}_{\mathbb{P}(y|\boldsymbol{x})}(L(y, y_0)|\boldsymbol{x}) = \sum_{y \in \{+1, -1\}} \mathbb{P}(y|\boldsymbol{x}) L(y, y_0)$, where y_0 is the decision we make (our prediction). Note, we don't know $\mathbb{P}(y|\boldsymbol{x})$.

4.1 Inference of $\mathbb{P}(y|\boldsymbol{x})$

Instead of $\mathbb{P}(y|\boldsymbol{x})$ we can infer $\mathbb{P}(y|\boldsymbol{x}, D)$. There are two main approaches to doing this:

Definition 4.2 ***Discriminative Approach:***
Model $\mathbb{P}(y|\boldsymbol{x})$ with $\mathbb{P}(y|\boldsymbol{x}; \boldsymbol{w})$. Note that with this method we cannot simulate a new $\boldsymbol{x}$ given a class y, only tells us the difference between $+1$ and -1.

Definition 4.3 ***Generative Approach:***
Note that $\mathbb{P}(y|\boldsymbol{x}, D) \propto \mathbb{P}(\boldsymbol{x}|y, D)\mathbb{P}(y)$. So we can model $\mathbb{P}(\boldsymbol{x}|y)$ with $\mathbb{P}(\boldsymbol{x}|y; \boldsymbol{w})$. With this approach we can generate a new input $\boldsymbol{x}$ given an output $\boldsymbol{y}$.

4.2 Risk in Regression

The output value in regression is a continuous variable, so we cannot have a loss matrix anymore. Instead, we have a loss function, such as squared loss. When we use squared-loss the optimal prediction is $\hat{y} = \mathbb{E}_{\mathbb{P}(y|\boldsymbol{x})}(y)$ (proved in section A.6). If we use the absolute loss then the optimal prediction is the median of $\mathbb{P}(y|\boldsymbol{x})$ (proved in section A.7).

[5] Where $\{+1, -1\}$ are the class labels.

Appendices

A Proofs

A.1 Proof of Equation (2)

Let $\boldsymbol{X}$ be,

$$\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{11}$$

we can write our linear model in matrix form,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \tag{12}$$

note that,

$$\sum_{i \in D_0} (y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2 = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 \tag{13}$$

we can find the minimum by differentiating wrt. $\boldsymbol{w}$ and finding the solution when the gradient equals zero. However, first we expand our expression,

$$||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{14}$$

$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \tag{15}$$

we now can find the derivative wrt. $\boldsymbol{w}$,

$$\frac{\partial}{\partial \boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \tag{16}$$

now setting this to zero and solving,

$$\Rightarrow \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y} \tag{17}$$

$$\Rightarrow \boldsymbol{w} = \boldsymbol{X}^{-1}\boldsymbol{X}^{-T}\boldsymbol{X}^T\boldsymbol{y} \tag{18}$$

$$\Rightarrow \boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{19}$$

hence proven.

A.2 Proof of Equation (7)

We have that,

$$\boldsymbol{w}_{LS-R} := (\phi(\boldsymbol{X})^T\phi(\boldsymbol{X}) + \lambda\boldsymbol{I})^{-1}\phi(\boldsymbol{X})^T\boldsymbol{y} \tag{20}$$

we can rewrite our objective function as:

$$(\boldsymbol{y} - \phi(\boldsymbol{X})\boldsymbol{w})^T(\boldsymbol{y} - \phi(\boldsymbol{X})\boldsymbol{w}) + \lambda\boldsymbol{w}^T\boldsymbol{w} \tag{21}$$

$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\phi(\boldsymbol{X})^T\boldsymbol{y} - \boldsymbol{y}^T\phi(\boldsymbol{X})\boldsymbol{w} + \boldsymbol{w}^T\phi(\boldsymbol{X})^T\phi(\boldsymbol{X})\boldsymbol{w} + \lambda\boldsymbol{w}^T\boldsymbol{w} \tag{22}$$

now taking the partial derivative $\frac{\partial}{\partial \boldsymbol{w}}$ and setting the derivative to zero we have,

$$0 = -\phi(\boldsymbol{X})^T\boldsymbol{y} - \boldsymbol{y}^T\phi(\boldsymbol{X}) + 2\phi(\boldsymbol{X})^T\phi(\boldsymbol{X})\boldsymbol{w} + 2\lambda\boldsymbol{w} \tag{23}$$

$$\rightarrow 2\lambda\boldsymbol{w} + 2\phi(\boldsymbol{X})^T\phi(\boldsymbol{X})\boldsymbol{w} = \phi(\boldsymbol{X})^T\boldsymbol{y} + \boldsymbol{y}^T\phi(\boldsymbol{X}) \tag{24}$$

$$\rightarrow (\phi(\boldsymbol{X})^T\phi(\boldsymbol{X}) + \lambda\boldsymbol{I})\boldsymbol{w} = \phi(\boldsymbol{X})^T\boldsymbol{y} \tag{25}$$

$$\rightarrow \boldsymbol{w} = (\phi(\boldsymbol{X})^T\phi(\boldsymbol{X}) + \lambda\boldsymbol{I})^{-1}\phi(\boldsymbol{X})^T\boldsymbol{y} \tag{26}$$

hence proven.

A.3 Proof of $\boldsymbol{w}_{MAP} = \boldsymbol{w}_{LS-R}$ when $\lambda = \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2}$

Writing down the objective function of regularized LS, we have,

$$\prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2) \cdot N_{\boldsymbol{w}}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}) \tag{27}$$

$$\propto \prod_{i \in D} N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2) \cdot \prod_{i \in \boldsymbol{w}} N_{\boldsymbol{w}_i}(0, \sigma_{\boldsymbol{w}}^2) \tag{28}$$

$$= \frac{1}{2\sigma^2} \sum_{i \in D} (y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2 + \frac{1}{2\sigma_{\boldsymbol{w}}^2} \sum_{i \in \boldsymbol{w}} w_i^2 + c \qquad \propto \sum_{i \in D} (y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2 + \lambda \sum_{i \in \boldsymbol{w}} \boldsymbol{w}_i^2 + c \tag{29}$$

Note: for the second step we took the negative log probability and for the third step we multiplied by $2\sigma^2$. We note that what we ended up with is the objective function for regularized LS.

A.4 Proof of Equation (10)

First we note,

$$\mathbb{P}(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) = N_{\hat{y}}(f(\boldsymbol{x}; \boldsymbol{w}), \sigma^2) \tag{30}$$

also note that,

$$\mathbb{P}(\boldsymbol{w}|D) \propto \mathbb{P}(D|\boldsymbol{w}) \cdot \mathbb{P}(\boldsymbol{w}) \tag{31}$$

$$= N_{\hat{\boldsymbol{y}}}(\boldsymbol{f}(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2 \boldsymbol{I}) \cdot N_{\boldsymbol{w}}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}) \tag{32}$$

$$\propto (\boldsymbol{y} - \boldsymbol{\phi w})^T (\sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{\phi w}) + \boldsymbol{w}^T (\sigma_{\boldsymbol{w}}^2 \boldsymbol{I})^{-1} \boldsymbol{w} \tag{33}$$

$$= \frac{1}{\sigma^2} \boldsymbol{y}^T \boldsymbol{I} \boldsymbol{y} - \frac{2}{\sigma^2} \boldsymbol{y}^T \boldsymbol{\phi w} + \boldsymbol{w}^T \boldsymbol{\phi}^T \frac{1}{\sigma^2} \boldsymbol{\phi w} + \frac{1}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{w}^T \boldsymbol{I w} \tag{34}$$

$$\propto -\frac{2}{\sigma^2} \boldsymbol{y}^T \boldsymbol{\phi w} + \frac{1}{\sigma^2} \boldsymbol{w}^T \boldsymbol{\phi}^T \boldsymbol{\phi w} + \frac{1}{\sigma_{bmw^2}} \boldsymbol{w}^T \boldsymbol{w} \tag{35}$$

$$= \boldsymbol{w}^T (\frac{1}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{\phi}^T \boldsymbol{\phi}) \boldsymbol{w} - \frac{2}{\sigma^2} \boldsymbol{w}^T \boldsymbol{\phi}^T \boldsymbol{y} \tag{36}$$

$$= \frac{1}{\sigma^2} (\boldsymbol{w}^T (\frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I} + \boldsymbol{\phi}^T \boldsymbol{\phi}) \boldsymbol{w} - 2 \boldsymbol{w}^T \boldsymbol{\phi}^T \boldsymbol{y}) \tag{37}$$

$$= \frac{1}{\sigma^2} ((\boldsymbol{w}^T - (\lambda \boldsymbol{I} + \boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \boldsymbol{y})^T (\frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I} + \boldsymbol{\phi}^T \boldsymbol{\phi}) (\boldsymbol{w} - (\lambda \boldsymbol{I} + \boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \boldsymbol{y}) \tag{38}$$

$$- (\boldsymbol{\phi}^T \boldsymbol{y})^T (\frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I} + \boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \boldsymbol{y}) \tag{39}$$

$$= N_{\boldsymbol{w}}(\boldsymbol{\phi w}_{LS-R}, \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I} + \boldsymbol{\phi}^T \boldsymbol{\phi}) \tag{40}$$

therefore, with the help of 2.115 in PRML, we have,

$$\int \mathbb{P}(\hat{y}|\boldsymbol{x}, \boldsymbol{w}) \cdot \mathbb{P}(\boldsymbol{w}|D) d\boldsymbol{w} \tag{41}$$

$$= \mathbb{P}(\hat{y}|\boldsymbol{x}, D) \tag{42}$$

$$= N_{\hat{y}}(f(\boldsymbol{x}; \boldsymbol{w}_{LS-R}), \sigma^2 + \boldsymbol{\phi}^T(\boldsymbol{x}) \sigma^2 (\boldsymbol{\phi} \boldsymbol{\phi}^T + \frac{\sigma^2}{\sigma_{\boldsymbol{w}}^2} \boldsymbol{I}) \boldsymbol{\phi}(\boldsymbol{x})) \tag{43}$$

A.5 $\mathbb{P}(\textbf{FP or FN}|f)$ minimized when $f(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{x}, y = +1) - \mathbb{P}(\boldsymbol{x}, y = -1)$

We can write our objective as,

$$\int_{\{\boldsymbol{x}: f(\boldsymbol{x}) \geq 0\}} \mathbb{P}(\boldsymbol{x}, y = -1) d\boldsymbol{x} + \int_{\{\boldsymbol{x}: f(\boldsymbol{x}) \leq 0\}} \mathbb{P}(\boldsymbol{x}, y = +1) d\boldsymbol{x} \tag{44}$$

We want to pick $f(\boldsymbol{x})$ that minimizes the above, the $f(\boldsymbol{x})$ that minimizes the above objective is the one such that,

$$f(\boldsymbol{x}) \geq 0 \text{ if } \mathbb{P}(\boldsymbol{x}, y = -1) \leq \mathbb{P}(\boldsymbol{x}, y = +1) \tag{45}$$

and

$$f(\boldsymbol{x}) \leq 0 \text{ if } \mathbb{P}(\boldsymbol{x}, y=+1) \leq \mathbb{P}(\boldsymbol{x}, y=-1) \tag{46}$$

an $f(\boldsymbol{x})$ that satisfies these statements is,

$$f(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{x}, y=+1) - \mathbb{P}(\boldsymbol{x}, y=-1) \tag{47}$$

if $f(\boldsymbol{x}) \geq 0$, then $\mathbb{P}(\boldsymbol{x}, y=+1) \geq \mathbb{P}(\boldsymbol{x}, y=-1)$,
if $f(\boldsymbol{x}) \leq 0$, then $\mathbb{P}(\boldsymbol{x}, y=+1) \leq \mathbb{P}(\boldsymbol{x}, y=-1)$. Hence, proven.

A.6 The optimal prediction using squared loss is $\hat{y} = \mathbb{E}_{\mathbb{P}(y|\boldsymbol{x})}(y)$

We can rewrite the objective function as,

$$\int \mathbb{P}(y|\boldsymbol{x}) L(y, y_0) dy = \int \mathbb{P}(y|\boldsymbol{x})(y - y_0)^2 dy \tag{48}$$

$$= \int \mathbb{P}(y|\boldsymbol{x})(y^2 + y_0^2 - 2yy_0) dy \tag{49}$$

taking the derivative of $\frac{\partial}{\partial y_0}$ and setting to 0 we get,

$$0 = \int \mathbb{P}(y|\boldsymbol{x})(2y_0 - 2y) dy \tag{50}$$

$$\Rightarrow 0 = 2y_0 \int \mathbb{P}(y|\boldsymbol{x}) dy - 2 \int \mathbb{P}(y|\boldsymbol{x}) y dy \tag{51}$$

$$\Rightarrow 0 = 2y_0 - 2\mathbb{E}_{\mathbb{P}(y|\boldsymbol{x})}(y) \tag{52}$$

$$\Rightarrow y_0 = \mathbb{E}_{\mathbb{P}(y|\boldsymbol{x})}(y) \tag{53}$$

so $\hat{y} = \mathbb{E}_{\mathbb{P}(y|\boldsymbol{x})}(y)$.

A.7 The optimal prediction using absolute loss is the median of $\mathbb{P}(y|\boldsymbol{x})$

The objective function is,

$$\int \mathbb{P}(y|\boldsymbol{x}) |y - y_0| dy \tag{54}$$

taking the derivative of $\frac{\partial}{\partial y_0}$ and setting to 0 we get,

$$\int -\frac{\mathbb{P}(y|\boldsymbol{x})(y - y_0)}{|y - y_0|} dy = 0 \tag{55}$$

note that,

$$\frac{y - y_0}{|y - y_0|} = \begin{cases} 1, & y > y_0 \\ \text{-1}, & y < y_0 \end{cases} \tag{56}$$

so we have,

$$\Rightarrow \int_{y > y_0} \mathbb{P}(y|\boldsymbol{x}) dy - \int_{y < y_0} \mathbb{P}(y|\boldsymbol{x}) dy = 0 \tag{57}$$

$$\Rightarrow \int_{y_0}^{\infty} \mathbb{P}(y|\boldsymbol{x}) dy = \int_{-\infty}^{y_0} \mathbb{P}(y|\boldsymbol{x}) dy \tag{58}$$

so $\hat{y} = m$ where m is the medium of $\mathbb{P}(y|\boldsymbol{x})$

B Homeworks

B.1 For Section 1

Question B.1 ***Prove*** $\boldsymbol{w}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
Answer is section A.1

Question B.2 ***Why is the solution of*** w_{LS} ***useless if*** $n<d$***?***
If $n < d$ *then the columns of the model matrix don't have full rank, we know that for a matrix* $A \in \mathbb{R}^{n\times d}$, $rank(\boldsymbol{A}) \leq \min(n, d)$. *Therefore, for model matrix X with n, d we have* $rank(X) \leq n < d$ *and so X is not full rank. Since X is not full rank we have that* $\boldsymbol{X}^T\boldsymbol{X}$ *is not invertible.*

Question B.3 ***In what scenarios is the use of the Normal dist. to model*** $\mathbb{P}(y|\boldsymbol{x}, \boldsymbol{w}, \sigma)$ ***a bad idea?***
When the errors aren't normally distributed, which could arise when the predictor or target variables are non-normal or when outliers disrupt the model prediction. When this is the case it can cause lots of problems as we often use this assumption of normality when computing confidence intervals and to carry out hypothesis testing for example. So when our assumption of normality is incorrect it can lead to incorrect analyses.

Question B.4 ***Prove*** $\boldsymbol{w}_{LS} = (\phi(\boldsymbol{X}))^{-1}\boldsymbol{y}$.
We have,

$$\boldsymbol{w}_{LS} = (\phi(\boldsymbol{X})^T\phi(\boldsymbol{X}))^{-1}\phi(\boldsymbol{X})^T\boldsymbol{y} \tag{59}$$
$$= \phi(\boldsymbol{X})^{-1}\phi(\boldsymbol{X})^{-T}\phi(\boldsymbol{X})^T\boldsymbol{y} \tag{60}$$
$$= \phi(\boldsymbol{X})^{-1}\boldsymbol{y} \tag{61}$$

Question B.5 ***If we increase b of*** $\phi(\boldsymbol{x})$ ***by 2-fold, by how many folds will the computation time of*** $\boldsymbol{w}_{LS}$ ***increase?***
Let's consider computing the solution of $\boldsymbol{w}_{LS}$ *using the shortened form we found in question B.4. Increasing b by 2-fold will increase the number of rows in the model matrix by 2-fold, finding the inverse of a matrix is* $O(n^3)$ *(if using Gaussian elimination), so this would lead to a* 2^3*-fold increase in computation of the inverse. We also will have a 2-fold increase in the number of computations during the matrix multiplication.*

B.2 For Section 2

Question B.6 ***Why do machine learning algorithms still work on high-dimensional datasets (such as images), despite the curse of dimensionality telling us that the number of observations needed for solving high dimensional problems should grow exponentially with dimensionality?***
This is down to machine learning algorithms being able to learn something about the underlying structure of the data. In learning about the underlying structure it reduces the dimensionality of the problem, as it can use features within the data. For example in convolutional neural networks it's been shown that the network uses sets of features built up hierarchically in order to help it classify images. Instead of having to directly learn what every combination of pixels contained within an image should be classified as it simply learns that certain low level-patterns of pixels are important for classification, it then learns that these low-level patterns put together in different ways create other larger features which are important and so on...

B.3 For Section 3

Question B.7 ***The proof in section A.2***

Question B.8 ***Is the statement "the solution of*** $\boldsymbol{w}_{LS}$ ***is useless if*** $n < d$***" still true for*** $\boldsymbol{w}_{LS-R}$***?***
If $n < d$ *then* $\boldsymbol{X}$ *not full rank and neither is* $\boldsymbol{X}^T\boldsymbol{X}$, *denote* $\boldsymbol{A} := \boldsymbol{X}^T\boldsymbol{X}$. $\boldsymbol{A}$ *is not full rank and therefore its columns are not linearly independent, that means there exist* $\mu_1, ..., \mu_d$ *such that there exists* $i \in \{1, ..., d\}$ *with* $\mu_i \neq 0$ *and* $\mu_1\boldsymbol{A}^{(1)} + ... + \mu_d\boldsymbol{A}^{(d)} = 0$, *where* $\boldsymbol{A}^{(i)}$ *denotes the i-th column of A. Let* $\mu_1, ..., \mu_d$ *be any reals such that the above is true, then we have for any*

$j \in \{1, ..., n\}$, $\sum_{i \in \{1,...,d\}} \mu_i \boldsymbol{A}_{ji} = 0$. Where $\boldsymbol{A}_{ji}$ denotes the entry found at the j-th row and i-th column of A. Now consider the same linear combination for any row, j, of $\boldsymbol{A} + \lambda \boldsymbol{I}$, we have $(\sum_{i \in \{1,...,d\}/\{j\}} \mu_i \boldsymbol{X}_{ji}) + \mu_j \boldsymbol{X}_{jj} + \lambda = \lambda$. As $\lambda > 0$ we have that no linear combination of columns equals zero, unless $\mu_1 = ... = \mu_d = 0$. Hence, the columns of $(\boldsymbol{A} + \lambda \boldsymbol{I}) = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})$ are linearly independent and is therefore also invertible. So there does exist a solution to $\boldsymbol{w}_{LS-R}$ even if $n < d$.

Question B.9 ***The proof in section A.3***

Question B.10 ***The proof in section A.4***