CS 337, Fall 2023
MAP, Bias and Variance

Scribes: Aadhitya Narayanan A B (210070001)*, Kadivar Lisan Kumar (210050076)*, Ancha Pranavi (210050012)*, Ameya Vikrama Singh, Ramavath Sai Srithan, Pragnan Maharshi
Edited by: Bhavani Shankar

August 17, 2023

Disclaimer. Please note this document has not received the usual scrutiny that formal publications enjoy. This may be distributed outside this class only with the permission of the instructor.

Recap

We are currently analyzing the probabilistic model of Linear Regression, where the data points are sampled according to

$$y = \mathbf{w}^{\mathsf{T}}\mathbf{x} + \epsilon$$

where ϵ is a zero mean Gaussian noise in the observation with standard deviation σ. Using this, we get

$$\mathcal{P}(y \mid \mathbf{x}, \mathbf{w}) \sim \mathcal{N}\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}, \sigma^2\right)$$

Also, for a given training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, the expression for log-likelihood of this data will be

$$\sum_{i=1}^{n} \log\big(\mathcal{P}(y_i \,|\, \mathbf{x}_i, \mathbf{w})\big)$$

Also we derived the Maximum Likelihood Estimate (MLE) as follows

$$\mathbf{w}' = \arg\max_{\mathbf{w}} \left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right)^2\right) \tag{1}$$

Here, we focus on the Maximum A Posteriori (MAP) estimate of the parameter $\mathbf{w}$

1 Maximum A Posteriori Estimate (MAP)

Unlike the Maximum Likelihood Estimate (MLE), which was only dependent on the data, Maximum A Posteriori Estimate (MAP) incorporates information beyond the data. Here, we consider the parameter that we want to estimate as a random variable, say Θ. Our goal will be to find the parameter that maximizes the probability of that parameter given the data, i.e. the **Posterior Distribution**.

$$\Theta_{\text{MAP}} = \arg\max_{\Theta}\big(\mathcal{P}\left(\Theta \,|\, \mathcal{D}\right)\big) \tag{2}$$

*Larger credit to these scribes for the final notes.

Using Bayes Theorem, we can write the Posterior Distribution as follows

$$\mathcal{P}\left(\Theta \,|\, \mathcal{D}\right) \propto \mathcal{P}\left(\mathcal{D} \,|\, \Theta\right)\, \mathcal{P}\left(\Theta\right)$$

the first term is the probability of observing given data for a particular parameter, i.e., the **Likelihood**. And the second term is called the **Prior Distribution** over the parameter. This is the extra information about the parameter which is being added to improve the estimate when the provided data is limited. Hence, we can rewrite equation 2 as

$$\begin{aligned}\Theta_{\mathrm{MAP}} &= \arg\max_{\Theta}\big(\mathcal{P}\left(\mathcal{D} \,|\, \Theta\right)\, \mathcal{P}\left(\Theta\right)\big) \\ &= \arg\max_{\Theta}\Big(\log\big(\mathcal{P}\left(\mathcal{D} \,|\, \Theta\right)\big) + \log\big(\mathcal{P}\left(\Theta\right)\big)\Big)\end{aligned}$$

1.1 Coin Toss MAP Estimate

We have to estimate the probability of getting a head for a biased coin (say Θ). We are given data of N coin tosses, of which N_H are heads and N_T are tails. As we have derived earlier, the likelihood of this observation will be

$$\mathcal{P}\left(\mathcal{D} \,|\, \Theta\right) = \mathcal{P}\left(N_H, N_T \,|\, \Theta\right) = \Theta^{N_H}\left(1-\Theta\right)^{N_T}$$

For the prior distribution, we can choose a Beta distribution (which is the conjugate prior to Bernoulli Likelihood)

$$\mathcal{P}\left(\Theta\right) = \mathcal{B}\left(\Theta, \alpha, \beta\right) = \frac{1}{c}\,\Theta^{\alpha-1}\left(1-\Theta\right)^{\beta-1}$$

And hence, the Posterior will be

$$\begin{aligned}\mathcal{P}\left(\Theta \,|\, \mathcal{D}\right) &\propto \mathcal{P}\left(\mathcal{D} \,|\, \Theta\right)\, \mathcal{P}\left(\Theta\right) \\ &\propto \frac{1}{c}\,\Theta^{\alpha-1}\left(1-\Theta\right)^{\beta-1}\Theta^{N_H}\left(1-\Theta\right)^{N_T} \\ &\propto \frac{1}{c}\,\Theta^{N_H+\alpha-1}\left(1-\Theta\right)^{N_T+\beta-1} \\ &\propto \mathcal{B}\big(\Theta, N_H+\alpha, N_T+\beta\big)\end{aligned}$$

And,

$$\begin{aligned}\Theta_{\mathrm{MAP}} &= \arg\max_{\Theta}\Big(\log\big(\mathcal{P}\left(\mathcal{D} \,|\, \Theta\right)\big) + \log\big(\mathcal{P}\left(\Theta\right)\big)\Big) \\ &= \arg\max_{\Theta}\Big(\big(N_H+\alpha-1\big)\log(\Theta) + \big(N_T+\beta-1\big)\log(1-\Theta)\Big)\end{aligned}$$

On maximizing it, we get

$$\begin{aligned}\frac{d}{d\Theta}\Big(\big(N_H+\alpha-1\big)\log(\Theta) + \big(N_T+\beta-1\big)\log(1-\Theta)\Big) &= 0 \\ \frac{N_H+\alpha-1}{\Theta} - \frac{N_T+\beta-1}{1-\Theta} &= 0 \\ \Theta_{\mathrm{MAP}} = \frac{N_H+\alpha-1}{N+\alpha+\beta-2}\end{aligned}$$

Compared to the MLE estimate $\Theta_{\mathrm{MLE}} = \frac{N_H}{N}$, MAP estimate has extra counts denoted by α and β.

We can see that as $N \to \infty$, Θ_{MAP} converges to Θ_{MLE}. This represents the fact that prior beliefs become less representative over more data. This is encapsulated in the Bernstein - von Mises Theorem. Further, we also see that in essence, the prior data is enforcing a belief of a “pre-existing” $\alpha - 1$ heads out of $\alpha+\beta-2$ tosses. For this reason, these numbers can also be thought of as pseudo coin counts.

1.2 Linear Regression MAP Estimate

Given the probabilistic setup of Linear Regression, we try to estimate the weights $\mathbf{w}$.

Let's consider zero mean Gaussian Prior on the weights $\mathbf{w}$ with covariance matrix as scaled identity matrix, i.e.

$$\begin{aligned}
\mathcal{P}\left(\mathbf{w}\right) &= \mathcal{N}\left(0, \frac{\mathbf{I}}{\lambda}\right) \\
&= \frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{I}/\lambda)}} \exp\left(-\frac{1}{2}\mathbf{w}^T\left(\frac{\mathbf{I}}{\lambda}\right)^{-1}\mathbf{w}\right) \\
&= \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right) \\
&= \left(\frac{\lambda}{2\pi}\right)^{d/2} \exp\left(-\frac{\lambda}{2}\left\|\mathbf{w}\right\|_2^2\right)
\end{aligned}$$

Now, if we try to find the MAP estimate for $\mathbf{w}$,

$$\begin{aligned}
\mathbf{w}_{\text{MAP}} &= \arg\max_{\mathbf{w}}\big(\mathcal{P}\left(\mathcal{D}\,|\,\mathbf{w}\right)\ \mathcal{P}\left(\mathbf{w}\right)\big) \\
&= \arg\max_{\mathbf{w}}\Big(\log\big(\mathcal{P}\left(\mathcal{D}\,|\,\mathbf{w}\right)\big) + \log\big(\mathcal{P}\left(\mathbf{w}\right)\big)\Big)
\end{aligned}$$

Using equation 1 to rewrite the log-likelihood term and plugging in the prior term gives us,

$$\begin{aligned}
\mathbf{w}_{\text{MAP}} &= \arg\max_{\mathbf{w}}\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 + \frac{d}{2}\log\left(\frac{\lambda}{2\pi}\right) - \frac{\lambda}{2}\left\|\mathbf{w}\right\|_2^2\right) \\
&= \arg\min_{\mathbf{w}}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|_2^2\right)
\end{aligned}$$

Recall that this is same as the L_2-regularised linear regression (or Ridge Regression).

NOTE :- By changing the prior, we can get various types of regularisations. For example, using a Laplace Prior will give L_1 Regularisation (or Lasso Regression)

2 Bias and Variance of Estimates

Natural question that arises after creating different types of Linear Regression models is that "How do we evaluate whether the predictor is good or not?".

One of the metric that we use to generate the model is Training Loss, but this does not tell whether the model will be able to generalize or not. Hence, to define whether a model is good or not, we look at the Test Loss. So, we try to decompose the Expected Test Loss and it turns out that it has 3 components

1. Variance
2. Bias
3. Noise (or Irrecoverable error/ Unavoidable error)

Out of these, Variance and Bias are inherent to the model and Noise is inherent to the data (due to inaccuracies in measurements).

2.1 Bias

Let us assume that we are sampling data from a true distribution

$$y = f(\mathbf{x}) + \epsilon$$

where ϵ is a zero mean Gaussian Noise with standard deviation σ. Let $h_{\mathbf{w}}(.)$ be the predictor for a given data $\mathcal{D}$. Then the bias of the given model for a particular $\mathbf{x}$ can be defined as follows

$$\text{Bias} = \mathbb{E}\left[h_{\mathbf{w}}(\mathbf{x})\right] - f(\mathbf{x}) \tag{3}$$

Note that the expectation is over various choice of training data set $\mathcal{D}$ which will give rise to various $h_{\mathbf{w}}(.)$. Also, $\mathbf{x}$ is fixed value and the expectation does not depend on it. To explain the meaning of bias, let us sample multiple different data sets from the given true distribution and find the predictor function for each of them (figure 1). Then the bias represents how close the average of these predictor functions is to the true distribution.

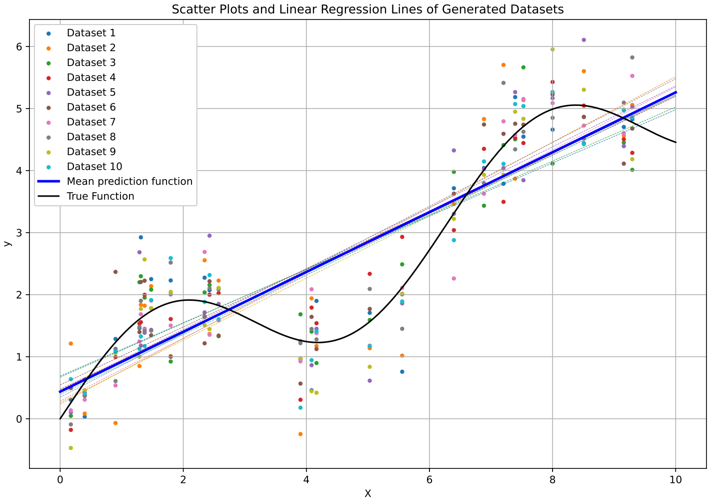


Figure 1: Showing various regression arising from various data-sets and the mean of them

2.2 Variance

Considering the same setup as before, the variance of the given model for a particular $\mathbf{x}$ can be defined as follows

$$\text{Variance} = \mathbb{E}\left[\left(h_{\mathbf{w}}(\mathbf{x}) - \mathbb{E}\left[h_{\mathbf{w}}(\mathbf{x})\right]\right)^2\right] \tag{4}$$

Note that the expectation is over various choices of training data set $\mathcal{D}$ which will give rise to various $h_{\mathbf{w}}(.)$. Intuitively, variance of given model is basically the variance of different predictor functions that we get by varying the data set $\mathcal{D}$. Qualitatively, it measures the spread of predictor functions in figure 1.

2.3 Noise

Noise is the unavoidable error in the data caused by the errors in measurement. The expression for this error is as follows

$$\text{Noise} = \mathbb{E}\left[\left(y - f(\mathbf{x})\right)^2\right] \tag{5}$$

Note that this expectation is over the random noise (i.e. ϵ in this case) and not dependent on the data set. This reduces to

$$\begin{aligned}
\text{Noise} &= \mathbb{E}\left[(\epsilon)^2\right] \\
&= \text{Var}(\epsilon) \\
&= \sigma^2
\end{aligned}$$

2.4 Analysing Unregularized Linear Regression

Let us assume that we have the same model as before, i.e.

$$y = f(\mathbf{x}) + \epsilon \;\text{ and }\; \epsilon \sim \mathcal{N}(0, \sigma^2)$$

As mentioned before, we are trying to find the expected test error. Let us denote the test data point as $(\widetilde{\mathbf{x}}, \widetilde{y})$ i.e. $\widetilde{y} = f(\widetilde{\mathbf{x}}) + \widetilde{\epsilon}$. Then we have to find

$$E_{\text{test error}} = \mathbb{E}_{\widetilde{\epsilon},\mathcal{D}}\left[\left(\widetilde{y} - h_{\mathbf{w}}\left(\widetilde{\mathbf{x}}\right)\right)^2\right] \tag{6}$$

NOTE :- Here, we are assuming that $\widetilde{\mathbf{x}}$ is a constant while taking the expectation, i.e. LHS should ideally be parameterized as $E_{\text{test error}}(\widetilde{\mathbf{x}})$. This would also mean that $\widetilde{y}$ (which is a random variable) is only dependent on $\widetilde{\epsilon}$.

$$\begin{aligned}
E_{\text{test error}} &= \mathbb{E}\left[\left(\widetilde{y} - h\left(\widetilde{\mathbf{x}}\right)\right)^2\right] \\
&= \mathbb{E}\left[\widetilde{y}^2\right] + \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)^2\right] - 2\,\mathbb{E}\left[\widetilde{y}\right]\,\mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right] \\
&= \mathbb{E}\left[\left(\widetilde{y} - \mathbb{E}\left[\widetilde{y}\right]\right)^2\right] + \mathbb{E}\left[\widetilde{y}\right]^2 + \mathbb{E}\left[\left(h\left(\widetilde{\mathbf{x}}\right) - \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right]\right)^2\right] + \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right]^2 - 2\,\mathbb{E}\left[\widetilde{y}\right]\,\mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right] \\
&= \mathbb{E}\left[\left(\widetilde{y} - f(\widetilde{\mathbf{x}})\right)^2\right] + \mathbb{E}\left[\left(h\left(\widetilde{\mathbf{x}}\right) - \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right]\right)^2\right] + \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right]^2 + \mathbb{E}\left[\widetilde{y}\right]^2 - 2\,\mathbb{E}\left[\widetilde{y}\right]\,\mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right] \\
&= \mathbb{E}\left[\left(\widetilde{y} - f(\widetilde{\mathbf{x}})\right)^2\right] + \mathbb{E}\left[\left(h\left(\widetilde{\mathbf{x}}\right) - \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right]\right)^2\right] + \left(\mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right] - \mathbb{E}\left[\widetilde{y}\right]\right)^2 \\
&= \mathbb{E}\left[\left(\widetilde{y} - f(\widetilde{\mathbf{x}})\right)^2\right] + \mathbb{E}\left[\left(h\left(\widetilde{\mathbf{x}}\right) - \mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right]\right)^2\right] + \left(\mathbb{E}\left[h\left(\widetilde{\mathbf{x}}\right)\right] - f(\widetilde{\mathbf{x}})\right)^2 \\
&= \text{Noise} + \text{Variance} + \text{Bias}^2
\end{aligned}$$

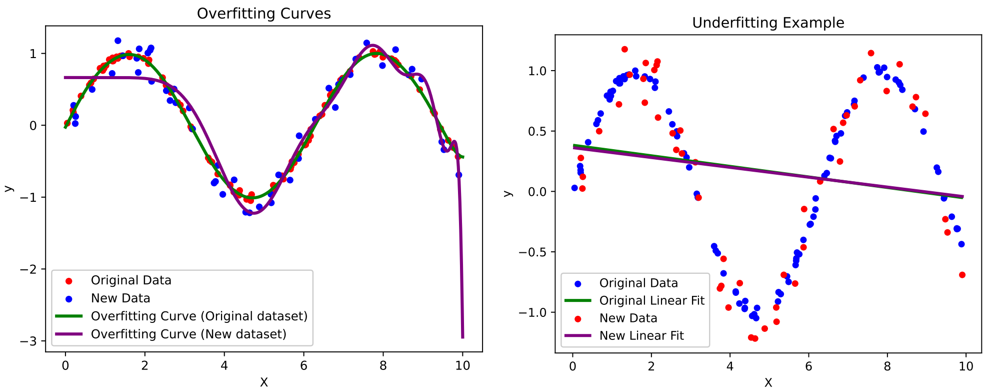
Explanations of steps :-

- The predicted value $h(\widetilde{\mathbf{x}})$ depends only on the data set and hence it is independent of the actual value $\widetilde{y}$ (which is only dependent on $\widetilde{\epsilon}$)
- Using the definition of variance we can get $\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2\right] + \mathbb{E}\left[Y\right]^2$
- Also, note that $\mathbb{E}\left[\widetilde{y}\right] = \mathbb{E}\left[f(\widetilde{\mathbf{x}}) + \widetilde{\epsilon}\right] = \mathbb{E}\left[f(\widetilde{\mathbf{x}})\right] + \mathbb{E}\left[\widetilde{\epsilon}\right] = f(\widetilde{\mathbf{x}})$ as the noise is zero mean and $\widetilde{\mathbf{x}}$ is a constant (as mentioned before)

2.5 Variance-Bias Trade-Off

Let us look at the use of the above-mentioned formula.

Consider the following case where we have used a very high complexity model for linear regression (as in Figure 2a).



(a) Over-fitting - High Variance, Low Bias (b) Under-fitting - Low Variance, High Bias

Figure 2: Analysing models using bias and variance

It is evident that the bias of this high-complexity model is very low. But, as we change the dataset slightly, we can see that there is too much variation in the newly fitted curve. This means that this high-complexity model also has high variance. Similarly, we can consider a low-complexity model such as linear (as in Figure 2b). Here, the predicted values deviate significantly from the actual function, and hence the bias is high. But, changing the dataset has very less effect on the predictor function, suggesting that there is low variance.

The ideal model is one that strikes a balance between bias and variance. Highly complex models, like high-order polynomials, may overfit, resulting in **low bias but high variance**. In contrast, lower complexity models might not capture all data complexities, leading to **high bias and low variance**. Both can yield high test errors.

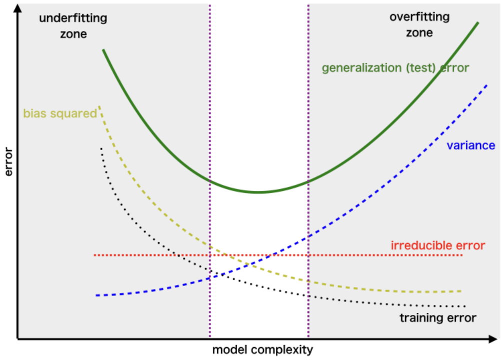


Figure 3: Variance Bias Trade-off with Expected Test Error/Generalisation Error

Hence, there is an inherent trade-off between variance and bias. We can't make both of them arbitrarily small at the same time. This is qualitatively shown in figure 3.

Recall L_2-Regularised model where, $\mathbf{w}_{\text{ridge}} = \arg\min_{\mathbf{w}} \left(\|y - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2 \right)$. As λ increases, the loss penalises high variation of $\mathbf{w}$ and hence the variance decreases. However, this is also reducing the flexibility of model and hence leading to increased bias.