Week4 - Bias-Variance Analysis

Regularisation can effectively control over-fitting, but it also has side effect of limiting the flexibility of the model. To determine a suitable value for the regularisation coefficient λ , considering frequentist viewpoint of the model complexity issue, known as **Bias-Variance trade-off**, can help.

Note that the phenomenon of over-fitting is an property of maximum likelihood (systematically involve bias in estimates), but **does not arise when we marginalise over parameter in a Bayesian setting**.

1. Components

As discussed in decision theory, the decision stage consists of choosing a specific estimate $y(\boldsymbol{x})$ of the value of t for each input $\boldsymbol{x}$, and also the loss matrix $\ell\{t,(y(\boldsymbol{x}))\}$. The common choice of loss function in regression problem is the squared loss:

$\ell\{t,(y(\boldsymbol{x}))\} = \{y(\boldsymbol{x}) - t\}^2$

$E[\ell] = \int\int \ell\{t,(y(\boldsymbol{x}))\}p(\boldsymbol{x},t)d\boldsymbol{x}dt$

Hence, we need to find the $y(\boldsymbol{x})$ that minimise $E[\ell]$, then we make use of derivative and set it equal to zero:

$$\frac{\partial E[\ell]}{\partial y(\boldsymbol{x})} = 2\int \{y(\boldsymbol{x}) - t\}p(\boldsymbol{x},t)dt = 0$$

$$\Rightarrow y(\boldsymbol{x}) = \frac{\int tp(\boldsymbol{x},t)dt}{p(\boldsymbol{x})} = \int tp(t|\boldsymbol{x})dt = \mathbb{E}_t(t|\boldsymbol{x})$$

$\mathbb{E}_t(t|\boldsymbol{x})$ represents the average prediction over all data sets (aka the optimal prediction)

Then from $\{y(\boldsymbol{x}) - t\}^2$ we can develop:

$\{y(\boldsymbol{x}) - t\}^2 = \{y(\boldsymbol{x}) - \mathbb{E}_t(t|\boldsymbol{x}) + \mathbb{E}_t(t|\boldsymbol{x}) - t\}^2$

$\Rightarrow \{y(\boldsymbol{x}) - \mathbb{E}_t(t|\boldsymbol{x})\}^2 + \{\mathbb{E}_t(t|\boldsymbol{x}) - t\}^2 + 2\{y(\boldsymbol{x}) - \mathbb{E}_t(t|\boldsymbol{x})\}\{\mathbb{E}_t(t|\boldsymbol{x}) - t\}$

Hence:

$$E[\ell] = \int \{y(\boldsymbol{x}) - \mathbb{E}_t(t|\boldsymbol{x})\}^2 p(\boldsymbol{x})d\boldsymbol{x} + \int \{\mathbb{E}_t(t|\boldsymbol{x}) - t\}^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt$$

- The second term is known as **noise**, which depends on data (intrinsic noise), representing the minimum achievable value of the expected loss.
- The first term is known as **generalisation error**, It is known as a measure of how accurately an algorithm can predict outcome of unseen data. Since $\mathbb{E}_t(t|\boldsymbol{x})$ focus on over all data sets, we can further develop it by considering the data set D in hand: (Now we use $h(\boldsymbol{x})$ replace $\mathbb{E}_t(t|\boldsymbol{x})$ to avoid cluster)

$$\{y(\boldsymbol{x},D) - h(\boldsymbol{x})\}^2 = \{y(\boldsymbol{x},D) - \mathbb{E}_D[y(\boldsymbol{x},D)] + \mathbb{E}_D[y(\boldsymbol{x},D)] - h(\boldsymbol{x})\}^2$$

$\Rightarrow \{y(\boldsymbol{x}, D) - \mathbb{E}_D[y(\boldsymbol{x}, D)]\}^2 + \{\mathbb{E}_D[y(\boldsymbol{x}, D)] - h(\boldsymbol{x})\}^2$

$+2\{y(\boldsymbol{x}, D) - \mathbb{E}_D[y(\boldsymbol{x}, D)]\}\{\mathbb{E}_D[y(\boldsymbol{x}, D)] - h(\boldsymbol{x})\}$

Hence:

$E_D[\{y(\boldsymbol{x}, D) - h(\boldsymbol{x})\}^2] = E_D[\{y(\boldsymbol{x}, D) - \mathbb{E}_D[y(\boldsymbol{x}, D)]\}^2] + E_D[\{\mathbb{E}_D[y(\boldsymbol{x}, D)] - h(\boldsymbol{x})\}^2]$

$= E_D[\{y(\boldsymbol{x}, D) - \mathbb{E}_D[y(\boldsymbol{x}, D)]\}^2] + \{\mathbb{E}_D[y(\boldsymbol{x}, D)] - h(\boldsymbol{x})\}^2$

- The first term is known as **variance error**, it measures how the solution for individual data sets D varies around the average $\mathbb{E}_t(t|\boldsymbol{x})$ (denotes as $h(\boldsymbol{x})$).
    - **Indicate how consistent are the predictions from one another over different training sets**
- The second term is known as **squared bias error** $(bias)^2$, It represent the difference between the average prediction of our model based on D and the correct value of the true model.
    - **Indicatie the accuracy of the model against the true function**

Therefore, we can clearly see that: $generalisation\ error = (bias)^2 + variance$

And: $expected\ loss = generalisation\ error + noise = (bias)^2 + variance + noise$

- $(bias)^2$ error: $\int \{\mathbb{E}_D[y(\boldsymbol{x}, D)] - h(\boldsymbol{x})\}^2 p(\boldsymbol{x}) d\boldsymbol{x}$
- Variance error : $\int \mathbb{E}_D[\{y(\boldsymbol{x}, D) - \mathbb{E}_D[y(\boldsymbol{x}, D)]\}^2] p(\boldsymbol{x}) d\boldsymbol{x}$
- Noise: $\int\int \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t) d\boldsymbol{x} dt$

Understanding bias and variance can help us

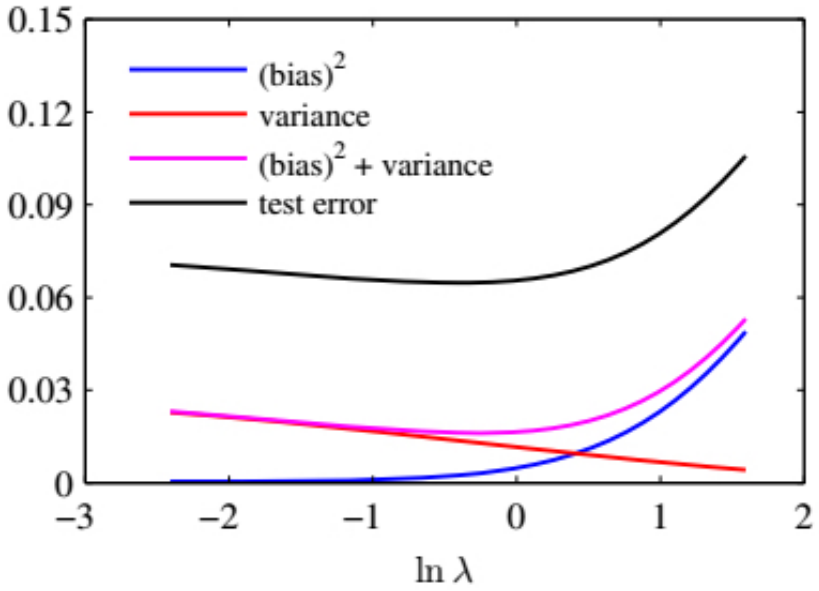
- diagnose the model performance,
- Avoid the mistake of over-fitting / under-fitting

> Bias Error 考察的是不同模型的预测与真实数据的偏差程度
>
> Variance Error 考察的是对于某一模型，若分别使用不同的训练集进行拟合后，每次训练后所得到的预测值之间的偏离程度，从而对模型进行评估。换句话说，Variance高的模型，其模型参数则需要再优化，因为其对不同的训练集过于敏感，即模型不稳定，过于反覆无常。

Since our goal is to minimise the generalization error, we need to do some trade-off between squared bias and variance.

- **Flexible** model:
    - Low bias, high variance
    - High complexity - too complex
- **Rigid** model:
    - High bias, low variance
    - Low complexity - too simple

Quantitative analysis:



- As λ becomes larger, the effect of the penalty term affect the model complexity:
  - If λ is very large, the penalty term largely affect the model parameters, as well as the model complexity. Since the model is too simple, it fails to capture the underlying trend in the data. Both the squared bias and variance are large, as well as the test error (**Under-fitting**)
  - If λ is very small, the penalty term has limited effect on the model parameters. The model is of high complexity. Although squared bias decreases to almost zero, but both test error and variance are large. That indicate that not only fitting the model to the training data set, but also fitting the model to the noise which is intrinsically exist in the data. (**Over-fitting**)
  - At the middle range, a sweet point (suitable value for λ) can be found where both the test error and the generalisation error are small.
- The generalisation error derived from frequentist view is similar to the test error

> The noise in the data is irreducible bias. It is the minimum error that can be achieved in the error function definition. Hence, when we consider bias-variance trade-off, we should consider the overall expect error at the expense of some bias or variance exist in the model.

Model	Bias	Variance	MSE
0-order Poly	0.9117	0.0052	1.0361
1-order Poly	0.3353	0.0039	0.4539
3-order Poly	0.0039	0.0028	0.1032
15-order Poly	0.0008	0.0121	0.1069

Worst=Blue, Best=Red

Learning Curves

Assume that we have L data sets for each having N data points, using $D^{(l)}$ to denote each data set. Then we can derive the expression of these error:

- The average perdiction: $\bar{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x_n)$
- The squared bias: $(bias)^2 = \frac{1}{N}\sum_{n=1}^{N}\{\bar{y}(x_n) - h(x_n)\}^2$
- The variance: $variance = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{L}\sum_{l=1}^{L}\{y^{(l)}(x_n) - \bar{y}(x_n)\}^2\right)$
- The test error: $test\ error = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{1}{N}\sum_{n=1}^{N}\{y^{(l)}(x_n) - h(x_n)\}^2\right)$