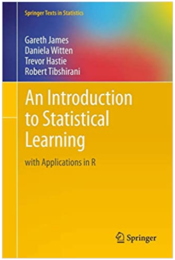
# Introduction to Statistical Learning

Haixia Liu

September 7, 2020

# Textbook and Reference Books

- Textbook: An introduction to Statistical Learning (ISLR)




- Reference: Elements of Statistical Learning (ESL) and machine learning by Zhihua Zhou (Nanjing U).
- Programming languages: R or Python
- Acknowledge the use of the graphics in the textbook/refernce for only the purpose of presentation.

# Example: Stylometry

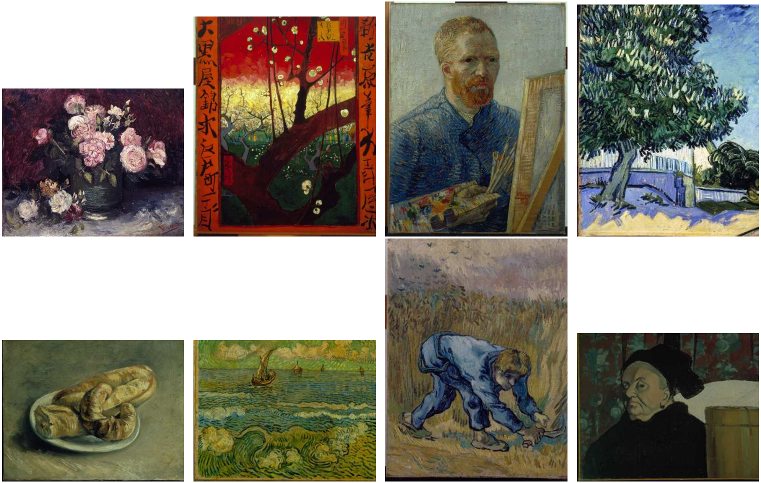
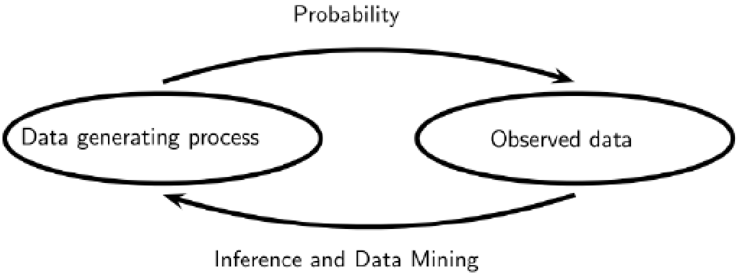


Figure: First line: Genuine paintings. Second line: fakes.

# Example: Face Clustering

Figure: Extended YaleB data.

# Probability vs. Statistcal Learning



1. What is Statiscal/Machine Learning?

2. Assessing Model Accuracy

3. The Bias-Variance Trade-Off

# Statiscal/Machine Learning

- Data:
  - Quantitative response $y$
  - $p$ different prdictors $x_1, x_2, \cdots, x_p$

# Statiscal/Machine Learning

- Data:
  - Quantitative response $y$
  - $p$ different prdictors $x_1, x_2, \cdots, x_p$
- There is some relationship between $y$ and $\mathbf{x} = (x_1, \cdots, x_p)$

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

  - $f$ is some fixed but unknown function of $x_1, \cdots, x_p$,
  - $\epsilon$ is a random error term, which is independent of $\mathbf{x}$ with mean zero.

## Statiscal/Machine Learning

- Data:
  - Quantitative response $y$
  - $p$ different prdictors $x_1, x_2, \cdots, x_p$
- There is some relationship between $y$ and $\mathbf{x} = (x_1, \cdots, x_p)$

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

  - $f$ is some fixed but unknown function of $x_1, \cdots, x_p$,
  - $\epsilon$ is a random error term, which is independent of $\mathbf{x}$ with mean zero.
- In essence, statistical learning refers to a set of approaches for estimating $f$.
- AIM: Estimate $f$ and Evaluate the estimates obtained.

# Why to estimate $f$?

There are TWO main reasons to estimate $f$:

- *Prediction.*
  - a set of inputs $\mathbf{x} = (x_1, \cdots, x_p)$ are readily available,
  - but the output $y$ cannot be easily obtained.
  - $x_1, \cdots, x_p$ are characteristics of a patient's blood sample that can be easily measured in a lab, and $y$ is a variable encoding the patient's risk for a severe adverse reaction to a particular drug.
  - understand the relationship between $\mathbf{x}$ and $y$.

# Why to estimate $f$?

There are TWO main reasons to estimate $f$:

- *Prediction.*
  - a set of inputs $\mathbf{x} = (x_1, \cdots, x_p)$ are readily available,
  - but the output $y$ cannot be easily obtained.
  - $x_1, \cdots, x_p$ are characteristics of a patient's blood sample that can be easily measured in a lab, and $y$ is a variable encoding the patient's risk for a severe adverse reaction to a particular drug.
  - understand the relationship between $\mathbf{x}$ and $y$.
- *Inference.*
  - $\mathbf{x}$ and $y$ are both available
  - understanding the way that $y$ is affected as $x_1, \cdots, x_p$ change.

# Why to estimate $f$?

There are TWO main reasons to estimate $f$:

- *Prediction.*
  - a set of inputs $\mathbf{x} = (x_1, \cdots, x_p)$ are readily available,
  - but the output $y$ cannot be easily obtained.
  - $x_1, \cdots, x_p$ are characteristics of a patient's blood sample that can be easily measured in a lab, and $y$ is a variable encoding the patient's risk for a severe adverse reaction to a particular drug.
  - understand the relationship between $\mathbf{x}$ and $y$.
- *Inference.*
  - $\mathbf{x}$ and $y$ are both available
  - understanding the way that $y$ is affected as $x_1, \cdots, x_p$ change.
  - understand how $y$ changes as a function of $x_1, \cdots, x_p$.
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?
  - Can the relationship is linear more complicated?

# The problems we focus: Regression,Classfication and Clustering.

- Data:
  - Quantitative response $y$
  - $p$ different prdictors $x_1, x_2, \cdots, x_p$

# The problems we focus: Regression,Classfication and Clustering.

- Data:
  - Quantitative response $y$
  - $p$ different prdictors $x_1, x_2, \cdots, x_p$
- There is some relationship between $y$ and $\mathbf{x} = (x_1, \cdots, x_p)$

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

## The problems we focus: Regression,Classfication and Clustering.

- Data:
  - Quantitative response $y$
  - $p$ different prdictors $x_1, x_2, \cdots, x_p$
- There is some relationship between $y$ and $\mathbf{x} = (x_1, \cdots, x_p)$

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

- Classfication vs. Clustering (supervised vs unsupervised learning).
  - Whether responser $y$ is available or not?
- Regression vs. classification.
  - they are all supervised learning.
  - Difference: *quantitative* or *qualitative*.
  - Data: *training data* (train the model $f$) and *testing data* (validate $f$).

# How to estimate $f$?–Parametric Methods.

Parametric methods

- Make an assumption about the functional form, or shape, of $f$. For example, Assume $f$ is linear about $\mathbf{x}$:

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

# How to estimate $f$?–Parametric Methods.

Parametric methods

- Make an assumption about the functional form, or shape, of $f$. For example, Assume $f$ is linear about $\mathbf{x}$:

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

- Once we have assumed that $f$ is linear, the problem of estimating $f$ is greatly simplified.
- One only needs to estimate $\beta_0, \beta_1, \cdots, \beta_p$.

# How to estimate $f$?–Parametric Methods.

Parametric methods

- Make an assumption about the functional form, or shape, of $f$. For example, Assume $f$ is linear about $\mathbf{x}$:

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

- Once we have assumed that $f$ is linear, the problem of estimating $f$ is greatly simplified.
- One only needs to estimate $\beta_0, \beta_1, \cdots, \beta_p$.

**Disadvantage** of a parametric approach

- model we choose may not match the true unknown form of $f$.
- If the chosen model is too far from the true $f$, estimate is poor.

# How to estimate $f$?–Non-parametric Methods.

- NOT make explicit assumptions about the functional form of $f$.
- Seek an estimate of $f$ that gets as close to the data points as possible without being too rough or wiggly.
- **Advantage** over parametric approaches:
  - avoiding the assumption of a particular functional form for $f$.
  - accurately fit a wider range of possible shapes for $f$.
- **disadvantage**:
  - NOT reduce the problem of estimating $f$ to a small number of parameters,
  - a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for $f$.

## The accuracy of $\hat{y}$ as a estimation for $y$.

Let $\hat{f}$ represents our estimate for $f$. We can estimate $y$ using

$$\hat{y} = \hat{f}(\mathbf{x}),$$

# The accuracy of $\hat{y}$ as a estimation for $y$.

Let $\hat{f}$ represents our estimate for $f$. We can estimate $y$ using

$$\hat{y} = \hat{f}(\mathbf{x}),$$

Recall the relationship between $y$ and $\mathbf{x} = (x_1, \cdots, x_p)$

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

$$\begin{aligned}\mathbb{E}(y-\hat{y})^2 =& E[f(\mathbf{x}) + \epsilon - \hat{f}(\mathbf{x})]^2 \\ =& \underbrace{[f(\mathbf{x}) - \hat{f}(\mathbf{x})]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} .\end{aligned}$$

# Reducible error and irreducible error.

- Reducible error:
  - $\hat{f}$ will not be a perfect estimate for $f$,
  - This inaccuracy will introduce some error.
  - This error is reducible by improving the accuracy of $\hat{f}$.

# Reducible error and irreducible error.

- Reducible error:
  - $\hat{f}$ will not be a perfect estimate for $f$,
  - This inaccuracy will introduce some error.
  - This error is reducible by improving the accuracy of $\hat{f}$.
- Irreducible error:
  Even if it were possible to form a perfect estimate for $f$, so that our estimated response took the form $\hat{y} = f(\mathbf{x})$, our prediction would still have some error in it!

# Reducible error and irreducible error.

- Reducible error:
  - $\hat{f}$ will not be a perfect estimate for $f$,
  - This inaccuracy will introduce some error.
  - This error is reducible by improving the accuracy of $\hat{f}$.
- Irreducible error:
  Even if it were possible to form a perfect estimate for $f$, so that our estimated response took the form $\hat{y} = f(\mathbf{x})$, our prediction would still have some error in it!
  - $y$ is also a function of $\epsilon$, which, by definition, cannot be predicted using $\mathbf{x}$.
  - variability associated with $\epsilon$ also affects the accuracy of our predictions.
  - no matter how well we estimate $f$, we cannot reduce the error introduced by $\epsilon$.

## No free lunch!

- *Why is it necessary to introduce so many different statistical learning approaches?*

# No free lunch!

- *Why is it necessary to introduce so many different statistical learning approaches?*
    - NO free lunch in statistics:
    - No one method dominates all others over all possible data sets.
    - On a particular data set, one method may work best, but some other methods may work better on a similar but different data set.

## No free lunch!

- *Why is it necessary to introduce so many different statistical learning approaches?*
    - NO free lunch in statistics:
    - No one method dominates all others over all possible data sets.
    - On a particular data set, one method may work best, but some other methods may work better on a similar but different data set.
- *Decide for any given set of data which method produces the best results.*

## No free lunch!

- *Why is it necessary to introduce so many different statistical learning approaches?*
    - NO free lunch in statistics:
    - No one method dominates all others over all possible data sets.
    - On a particular data set, one method may work best, but some other methods may work better on a similar but different data set.
- *Decide for any given set of data which method produces the best results.*

Figure: George Box: “Essentially, all models are wrong, but some are useful.”

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

- The model is fit by the training data.

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

- The model is fit by the training data.
- The MSE is computed using the training data.

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

- The model is fit by the training data.
- The MSE is computed using the training data.
- So we should more accurately be referred to as the training MSE.

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

- The model is fit by the training data.
- The MSE is computed using the training data.
- So we should more accurately be referred to as the training MSE.
- But, we do not care about training MSE.

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

- The model is fit by the training data.
- The MSE is computed using the training data.
- So we should more accurately be referred to as the training MSE.
- But, we do not care about training MSE.
- We are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

# Measuring the Quality of Fit

Let $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ be the obeservations and $\hat{f}(\mathbf{x})$ be the estimate for $f(\mathbf{x})$, then **mean squared error (MSE)** is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2.$$

- The model is fit by the training data.
- The MSE is computed using the training data.
- So we should more accurately be referred to as the training MSE.
- But, we do not care about training MSE.
- We are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.

Let $(\mathbf{x}_0, y_0)$ be a test observation, not used to train $f$.

- we want to know whether $\hat{f}(\mathbf{x}_0)$ is approximately equal to $y_0$?

# How to select the model?

- If we had a large number of test observations, we check the average squared prediction error for these test observations $(\mathbf{x}_0; y_0)$– the test MSE

$$\mathrm{Ave}(\hat{f}(\mathbf{x}_0) - y_0)^2.$$

choose model with the small test MSE.

# How to select the model?

- If we had a large number of test observations, we check the average squared prediction error for these test observations $(\mathbf{x}_0; y_0)$– the test MSE

$$\text{Ave}(\hat{f}(\mathbf{x}_0) - y_0)^2.$$

  choose model with the small test MSE.
- If no test observations are available?
  In that case, one might imagine simply selecting a statistical learning method that minimizes the training MSE. This seems like it might be a sensible approach, since the training MSE and the test MSE appear to be closely related.

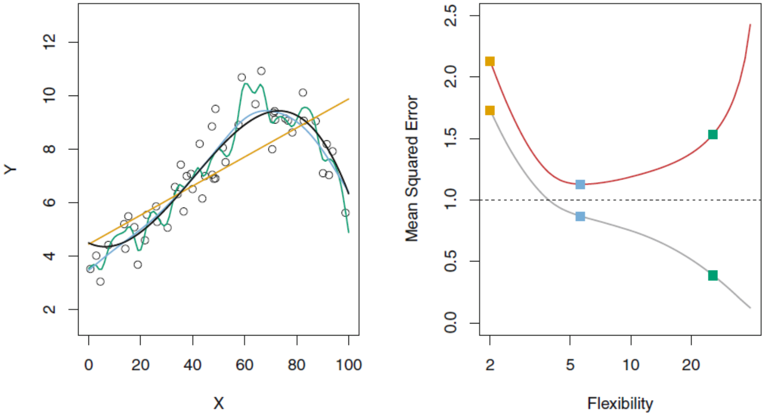



Figure: Left: Data simulated from $f$, shown in black. Three estimates of $f$ are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

# Comment on the etimate of $f$.

- The orange, blue and green curves illustrate three possible estimates for $f$ obtained using methods with increasing levels of flexibility.
- The orange line is the linear regression fitting, which is relatively inflexible.
- The blue and green curves were produced using smoothing splines with different levels of smoothness.

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

where $\lambda$ is a nonnegative tuning parameter. The function $g$ is known as a smoothing spline.

# Training MSE

- The grey curve displays the average training MSE as a function of flexibility, or more formally, the degrees of freedom which is a quantity that summarizes the flexibility of a model.
- The orange, blue and green squares indicate the MSEs associated with the corresponding curves in the left-hand panel.
- A more restricted and hence smoother curve has fewer degrees of freedom than a wiggly curve, linear regression is at the most restrictive end, with two degrees of freedom.
- The training MSE declines monotonically as flexibility increases.
- In this example the true $f$ is non-linear, and so the orange linear fit is not flexible enough to estimate $f$ well.
- The green curve has the lowest training MSE of all three methods, since it corresponds to the most exible of the three curves fit in the left-hand panel.

# Test MSE

- In this example, we know the true function $f$, and so we can also compute the test MSE over a very large test set, as a function of flexibility. (Of course, in general $f$ is unknown, so this will not be possible.)
- As with the training MSE, the test MSE initially declines as the level of flexibility increases. However, at some point the test MSE levels off and then starts to increase again.
- Consequently, the orange and green curves both have higher test MSE. The blue curve minimizes the test MSE, which should not be surprising given that visually it appears to estimate $f$ the best.
- The horizontal dashed line indicates $\mathrm{Var}(\epsilon)$, the irreducible error, which corresponds to the lowest achievable test MSE among all possible methods.

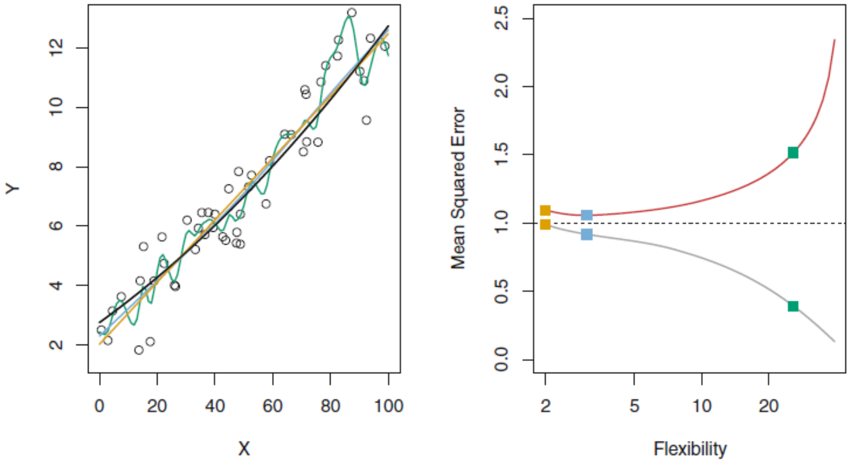



Figure: Using a different true $f$ that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

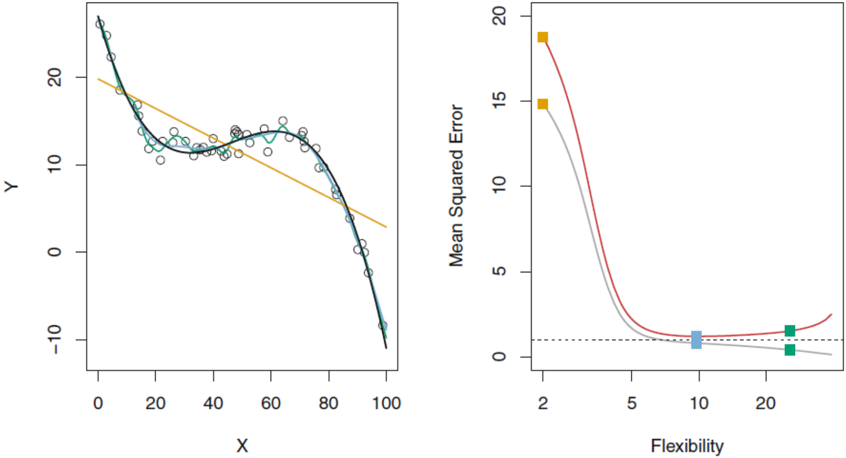
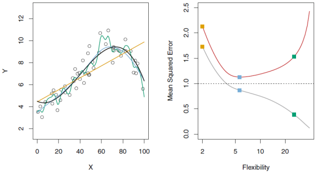



Figure: Using a different $f$ that is far from linear. In this setting, linear regression provides a very poor fit to the data.

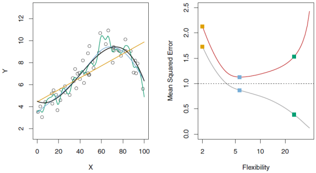
# Overfitting

When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data.



# Overfitting

When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data.




- statistical learning procedure is working too hard to find patterns in the training data,
- may be picking up some patterns that are just caused by **random changes** rather than by true properties of the unknown function $f$ .
- When we overfit the training data, the test MSE will be very large
- because the supposed patterns that the method found in the training data simply don't exist in the test data.

- In practice, one can usually compute the training MSE with relative ease, but estimating test MSE is considerably more difficult because usually no test data are available.
- As the previous three examples illustrate, the flexibility level corresponding to the model with the minimal test MSE can vary considerably among data sets.
- In Chapter 3, we discuss some approaches that can be used in practice to estimate this minimum point, such as **Cross-validation** which is method for estimating test MSE using the training data.

# The Bias-Variance Trade-Off

- Let $f(\mathbf{x})$ be the true function which we aim at estimating from a training data set $\mathcal{D}$.
- Let $\hat{f}(\mathbf{x};\mathcal{D})$ be the estimated function from training data set $\mathcal{D}$.
- Are we really interested in

$$\min_{\hat{f}}(f(\mathbf{x}) - \hat{f}(\mathbf{x};\mathcal{D}))^2 ?$$

# The Bias-Variance Trade-Off

- Let $f(\mathbf{x})$ be the true function which we aim at estimating from a training data set $\mathcal{D}$.
- Let $\hat{f}(\mathbf{x};\mathcal{D})$ be the estimated function from training data set $\mathcal{D}$.
- Are we really interested in

$$\min_{\hat{f}}(f(\mathbf{x}) - \hat{f}(\mathbf{x};\mathcal{D}))^2?$$

- **Fisher's view**: the measurements are a *random* selection from the set of all possible measurements which form the true distribution!

# The Bias-Variance Trade-Off

- Let $f(\mathbf{x})$ be the true function which we aim at estimating from a training data set $\mathcal{D}$.
- Let $\hat{f}(\mathbf{x};\mathcal{D})$ be the estimated function from training data set $\mathcal{D}$.
- Are we really interested in

$$\min_{\hat{f}}(f(\mathbf{x}) - \hat{f}(\mathbf{x};\mathcal{D}))^2?$$

- **Fisher's view**: the measurements are a *random* selection from the set of all possible measurements which form the true distribution!
- What we really care is

$$\min_{\hat{f}} \mathbb{E}_{\mathcal{D}}[f(\mathbf{x}) - \hat{f}(\mathbf{x};\mathcal{D})]^2.$$

where randomness caused by *random selection* has been taken into account.

- If we add and subtract $\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))$ inside the braces and then expand, we obtain

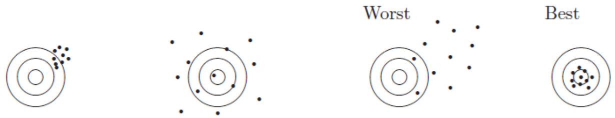
$$
\begin{aligned}
&[f(\mathbf{x}) - \hat{f}(\mathbf{x};\mathcal{D})]^2 \\
=&[f(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D})) + E_{\mathcal{D}}(\hat{f}(X;\mathcal{D})) - \hat{f}(\mathbf{x};\mathcal{D})]^2 \\
=&[f(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))]^2 + [\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D})) - \hat{f}(\mathbf{x};\mathcal{D})]^2 \\
&+ 2[f(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})]][\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})] - \hat{f}(\mathbf{x};\mathcal{D})]
\end{aligned}
$$

- If we add and subtract $\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))$ inside the braces and then expand, we obtain

$$
\begin{aligned}
&[f(\mathbf{x})-\hat{f}(\mathbf{x};\mathcal{D})]^2\\
=&[f(\mathbf{x})-\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))+E_{\mathcal{D}}(\hat{f}(X;\mathcal{D}))-\hat{f}(\mathbf{x};\mathcal{D})]^2\\
=&[f(\mathbf{x})-\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))]^2+[\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))-\hat{f}(\mathbf{x};\mathcal{D})]^2\\
&+2[f(\mathbf{x})-\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})]][\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})]-\hat{f}(\mathbf{x};\mathcal{D})]
\end{aligned}
$$

- Now we take the expectation of this expression with respect to $\mathcal{D}$ and note that the final term will vanish, giving

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[f(\mathbf{x})-\hat{f}(\mathbf{x};\mathcal{D})]^2\\
=&\underbrace{[f(\mathbf{x})-\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))]^2}_{\text{Bias}^2}+\underbrace{\mathbb{E}_{\mathcal{D}}\left[[\mathbb{E}_{\mathcal{D}}(\hat{f}(\mathbf{x};\mathcal{D}))-\hat{f}(\mathbf{x};\mathcal{D})]^2\right]}_{\text{Variance}}
\end{aligned}
$$




(a) High bias, Low variance (high precision) (b) Low bias, High variance (low precision) (c) High bias, High variance (low precision) (d) Low bias Low variance (high precision)

- *Bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- *Variance* refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\hat{f}$.
- Ideally the estimate for $f$ should not vary too much between training sets.
- *Bias and variance trade-off*: The optimal predictive cpability is the one that leads to balance between bias and variance.

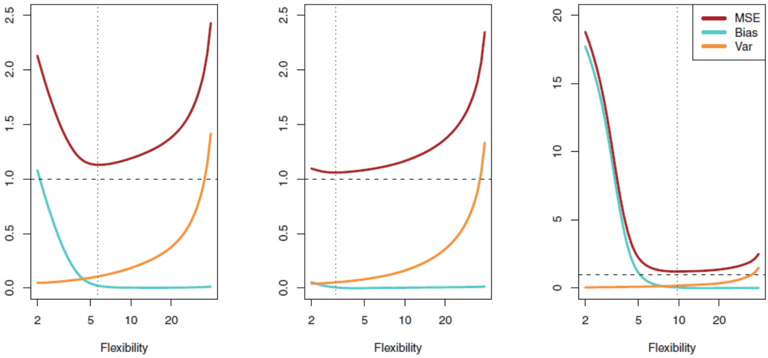



Figure: Squared bias (blue curve), variance (orange curve), $\mathrm{Var}(\epsilon)$ (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

The focus of this course:

- estimate $f$ with the aim of minimizing the reducible error.
- keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for $y$. This bound is almost always unknown in practice.