TODAY’S LECTURE

- Data collection
- Data processing
- Exploratory analysis & Data viz
- Analysis, hypothesis testing, & ML
- Insight & Policy Decision

Flow diagram showing the process from data collection to insight and policy decision.
MODEL ASSESSMENT
Introduction

- Suppose we have a quantitative response $Y$ and $d$ different predictors, $x_1, x_2, \ldots, x_d$.
- There is a relationship between $Y$ and $X$ which can be written in the very general form:
  \[ Y = f(X) + \varepsilon \]
- $f$ is our model (for example, $\beta_0 + \beta_1 x$) and $\varepsilon$ is a random error term which is independent of $X$ and has a mean zero.
Errors

- The accuracy of, \( \hat{y} \), as a prediction of \( y \) depends on two quantities, reducible error and irreducible error.

- In general \( \hat{f} \) will not be a perfect estimate for \( f \)

- This inaccuracy will introduce reducible error.

- We can potentially build a model that would improve the estimate of \( f \)
Error

- However, our prediction would still have some error in it.

- Output $Y$ is a function of $\varepsilon$ as well,
  \[ Y = f(X) + \varepsilon \]

- Variability associated with $\varepsilon$ affects the accuracy of our predictions - this is known as *irreducible* error.
Irreducible Error, $\varepsilon$

- Quantity $\varepsilon$, may contain unmeasured variables, useful in predicting $Y$.

- Since they are not measured, $f$ cannot use them for its prediction.

- Quantity $\varepsilon$ may also contain unmeasurable variation.
Irreducible Error, $\mathcal{E}$

- For example, the risk of an adverse reaction might vary for a given patient on a given day.

- The variability may depend on manufacturing variation in the drug itself, or

- the patient’s general feeling of well-being on that day.
**Errors**

- Consider a given estimate \( \hat{f} \) and a set of predictors \( X \), which yields the prediction

\[
\hat{y} = \hat{f}(x)
\]

- The average or expected value of the squared difference between the predicted and the actual value of \( Y \) is

\[
E (y - \hat{y})^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2
\]

\[
= E [f(X) - \hat{f}(X)]^2 + \text{Var}(\varepsilon)
\]

Reducible Irreducible

Var(\( \varepsilon \)) represents the variance associated with the error term \( \varepsilon \)
Loss Function to Optimize

- We want to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

- The most commonly used measure is the Mean Squared Error (MSE)
Mean Squared Error (MSE)

- MSE is given by:
  \[ \text{MSE Loss function} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y} - y_i)^2 \]

- MSE will be small if the predicted responses are very close to the true responses and large otherwise.

- When MSE is computed using training data, it is called training MSE.
Mean Squared Error (MSE)

- We are, however, interested in knowing whether a previously unseen observation \((x_0, y_0)\) is estimated correctly.
- We want to pick a method that gives the lowest test MSE, as opposed to the lowest training MSE.
- If we had a large number of test observations, we would compute

\[
\text{avg}(f(x_0) - y_0)^2
\]

the average squared prediction error for the test observations \((x_0, y_0)\).
- We want to have this average as small as possible.
Model Fit

source: Introduction to Statistical learning
Linear Model

\[ y = -2.526 \times + 1.191 \]
Quadratic Model

\[ y = 1.583 \, x^2 - 3.983 \, x + 1.386 \]
Degree 4 Model \[ y = 4.165 x^4 - 0.5359 x^3 - 4.369 x^2 - 0.8634 x + 1.139 \]
Degree 8 Model

\[ y = 4722x^8 - 1.81e+04x^7 + 2.872e+04x^6 - 2.444e+04x^5 \\
+ 1.206e+04x^4 - 3464x^3 + 537.3x^2 - 39.02x + 1.84 \]
Degree 16 Model

\[ y = 1.33 \times 10^6 x^{16} - 6.428 \times 10^6 x^{15} + 1.268 \times 10^7 x^{14} - 1.378 \times 10^7 x^{13} + 1.019 \times 10^7 x^{12} - 4.277 \times 10^6 x^{11} - 6.472 \times 10^6 x^{10} + 1.821 \times 10^7 x^9 - 2.086 \times 10^7 x^8 + 1.389 \times 10^7 x^7 - 5.775 \times 10^6 x^6 + 1.504 \times 10^6 x^5 - 2.35 \times 10^5 x^4 + 1.939 \times 10^4 x^3 - 516.8 x^2 - 20.56 x + 1.757 \]
Model Fit

source: Introduction to Statistical learning
As the flexibility of the model increases we observe a monotone decrease in the training MSE and a U-shape in the test MSE.

This is a fundamental property of the learning methods.

As the model complexity increases, the training MSE will decrease but the test MSE may not.
Overfitting

- When a given method yields a small training MSE but a large test MSE, we are overfitting data.

- This is because our model is working too hard to find patterns in the training data.

- It may be picking patterns that are caused by a random chance, rather than by true data.

- However, regardless of overfitting or not test MSE will be larger than the training MSE.
Bias-Variance Trade-Off

- The U-shape observed in the test MSE curves is a result of two competing properties of the learning methods.
- The expected test MSE, for a given value \( x_0 \), can be decomposed into the sum of three fundamental quantities: the variance of \( \hat{f}(x_0) \), the squared bias of \( \hat{f}(x_0) \) and the variance of the error terms \( \varepsilon \)

\[
E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon)
\]

\( E(y_0 - \hat{f}(x_0))^2 \) defines the expected test MSE and refers to the average test MSE that we would obtain if \( f \) were estimated repeatedly.
Bias-Variance Trade-Off

\[ E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon) \]

- The overall expected test MSE can be computed by averaging
  \[ E(y_0 - \hat{f}(x_0))^2 \]
  over all possible values of \( x_0 \) in the test set.

- In order to minimize the expected test error, we need to select a model that simultaneously achieves low variance and low bias.
Bias and Variance

- 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 model, different training sets will result in different \( \hat{f} \).

- Ideally, the estimate for \( f \) should not vary too much between training sets.

- However, a method with high variance will result in large changes in \( f \) for small changes in training data.
Consider the green and the orange curves in our figure. The flexible green curve is following the observations very closely. It has high variance. Changing any one of these data points may cause the estimate to change considerably. In contrast, the orange least squares line is relatively inflexible and has low variance.
Bias

- Bias refers to the error that is introduced by approximating a real-life problem, by a much simpler model.
- For example, linear regression assumes that there is a linear relationship between $Y$ and $X_1, X_2, \ldots, X_p$.
- It is unlikely that any real-life problem truly has a simple linear relationship.
- So performing linear regression will undoubtedly result in some bias in estimating $f$.
- Generally, more flexible methods result in less bias.
Bias and Variance

- As we use more flexible methods the variance will increase and the bias will decrease.
- The relative rate of change of these two quantities determines whether the test MSE increases or decreases.
- As flexibility increases the bias tends to initially decrease faster than the variance increases, decreasing test MSE.
- At some point increasing flexibility has little impact on bias, but starts to increase the variance, increasing test MSE.
Bias-Variance trade-off

- A good model requires low variance as well as low squared bias.
- It is easy to obtain a method with extremely low bias but high variance (for example, by drawing a curve that passes through every observation), or
- a method with low variance but high bias (by fitting a horizontal line).
- The challenge lies in finding a method in which both the variance and the squared bias are low.
Model Fit

source: Introduction to Statistical learning