BIG DATA ANALYTICS
Basic concepts of Statistical Learning

Olga Klopp
klopp@essec.edu

ESSEC BUSINESS SCHOOL

CentraleSupélec

Statistical Learning

- Suppose that we observe a quantitative response Y and p different predictors $X_1, X_2, ..., X_p$
- We assume that there is some relationship between Y and $X = (X_1, X_2, ..., X_p)$:

$$Y = f(X) + \epsilon$$

- f is some fixed but unknown function of $X_1, ..., X_p$ and ϵ is a random error term
  - ϵ is independent of X and has mean zero
  - f represents the systematic information that X provides about Y

Goal: to estimate f based on the observed points

Statistical Learning

Statistical learning refers to a set of approaches for estimating f

Today:

- Outline some of the key theoretical concepts that arise in estimating f
- Tools for evaluating the estimates

Outline

Outline

Prediction

Two main reasons: *prediction and inference*

Prediction:

- In many situations, a set of inputs X is available, but the output Y cannot be easily obtained
- We can predict Y using $\hat{Y} = \hat{f}(X)$
- $\hat{f}$ is our estimate for f
- $\hat{Y}$ is the resulting prediction for Y
- In this setting, $\hat{f}$ is often treated as a black box:
    - one is not concerned with the exact form of $\hat{f}$ provided that it yields accurate predictions for Y

Reducible and irreducible errors

- The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities: the *reducible error* and the *irreducible error*
- $\hat{f}$ will not be a perfect estimate for f: reducible error
  - we can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique
- Variability associated with ϵ also affects the accuracy of our predictions: irreducible error
  - By definition, ϵ cannot be predicted using X
  - No matter how well we estimate f, we cannot reduce the error introduced by ϵ.

Example

- Suppose that $X_1, ..., X_p$ are characteristics of a patient's blood sample
- Y is a variable encoding the patient's risk for a severe adverse reaction to a particular drug
- We seek to predict Y using X to avoid giving the drug to patients who are at high risk
- The risk of an adverse reaction might vary for a given patient on a given day (irreducible error):
  - manufacturing variation in the drug itself
  - patient's general feeling of well-being on that day

Our focus is on techniques for estimating f with the aim of minimizing the reducible error.

Inference

- We are often interested in understanding the way that Y is affected as $X_1, ..., X_p$ change
- We want to understand the relationship between X and Y
- Now $\hat{f}$ cannot be treated as a black box: we need to know its exact form

Inference

Answering the following questions:

- *Which predictors are associated with the response?*
  - often only a small fraction of the available predictors are substantially associated with Y
  - identifying the few important predictors among a large set of possible variables can be extremely useful
- *What is the relationship between the response and each predictor?*
  - Some predictors may have a positive relationship with Y
  - Other predictors may have the opposite relationship
  - The relationship between the response and a given predictor may also depend on the values of the other predictors

Inference

- *Can the relationship between Y and each predictor be summarized using a linear equation?*
  - Historically, most methods for estimating f have taken a linear form
  - In some situations, it is a reasonable assumption
  - Often the true relationship is more complicated: a linear model may not provide an accurate representation of the relationship between the input and output variables

What method?

Depending on whether our goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate:

- Linear models: relatively simple and interpretable inference
- Drawback: may not yield as accurate predictions as some other approaches
- The highly non-linear approaches can potentially provide quite accurate predictions for Y
- Drawback: a less interpretable model for which inference is more challenging

Outline

Training data

- We observe a set of n different data points
- These observations are called the *training data*: we will use these observations to train, or teach, our method how to estimate f
- Our goal is to apply a learning method to the training data in order to estimate the unknown function f:
  - we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y)
- Statistical learning methods: *parametric or non-parametric*

Parametric Methods

Parametric methods involve a two-step approach:

1. We make an assumption about the functional form of f
   - e.g. f is linear in X
   - the problem of estimating f is simplified: one only needs to estimate the p + 1 coefficients $\beta_0, \beta_1, ..., \beta_p$
2. After a model has been selected, we need a procedure that uses the training data to fit or train the model
   - the most common approach to fitting the linear model is the ordinary least squares
   - other approaches: Lasso, elastic-net...

Parametric approach reduces the problem of estimating f to the one of estimating a set of parameters

Parametric Methods

- It is easier to estimate a set of parameters, than to fit an entirely arbitrary function f
- The model we choose will usually not match the true unknown form of f
  - If the chosen model is too far from the true f: poor estimate
- To address this problem: choosing flexible models that can fit many different possible functional forms for f:
  - estimating a greater number of parameters
  - more complex models can lead to **overfitting** the data: the estimates follow the noise closely.

Non-parametric Methods

- Non-parametric methods do not make explicit assumptions about the functional form of f
- They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly
- Major advantage: they have the potential to accurately fit a wider range of possible shapes for f
- Major disadvantage: a large number of observations is required in order to obtain an accurate estimate for f

The Trade-Off Between Prediction Accuracy and Model Interpretability

- Less flexible methods: can produce a relatively small range of shapes to estimate f
  - e.g. linear regression is a quite inflexible approach
- *Why would we ever choose to use a more restrictive method instead of a very flexible approach?*
- If we are mainly interested in inference: restrictive models are much more interpretable
  - e.g. in the linear model it is quite easy to understand the relationship between Y and $X_1, X_2, ..., X_p$
  - very flexible approaches can lead quite complicated estimates of f $\rightarrow$ difficult to understand how any individual predictor is associated with the response.

The Trade-Off Between Prediction Accuracy and Model Interpretability

- In some settings we are only interested in prediction
- *Will it be best to use the most flexible model available?*
- **Surprisingly, this is not always the case!**
  - often we get more accurate predictions using a less flexible method

Overfitting in highly flexible methods

Outline

Assessing Model Accuracy

- No one method dominates all others over all possible data sets
- Decide for any given set of data which method produces the best results
- Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice

Mean Squared Error

- We need some way to measure how well predictions of a statistical learning method actually match the observed data
- In the regression setting, the most commonly-used measure is the **Mean Squared Error (MSE)**:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation

- The MSE is small if the predicted responses are very close to the true responses

Measuring the Quality of Fit

- The MSE is computed using the training data that was used to fit the model
- We are usually interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data
  - e.g. we are interested in developing an algorithm to predict a stock's price based on previous stock returns
  - we can train the method using stock returns from the past 6 months
  - we don't care how well our method predicts last week's stock price
  - we are interested in how well our model will predict tomorrow's price or next month's price.

Test Mean Squared Error

- Let (x_0, y_0) be a test observation not used to train the statistical learning method
- We want to know whether $\hat{f}(x_0)$ is approximately equal to y_0
- **Select the model for which the test MSE is as small as possible**

Test MSE

- How can we select a method that minimizes the test MSE?
- Available large test data set:
  - evaluate test MSE
  - select the learning method for which the test MSE is smallest
- What if no test observations are available?
- Can we simply select a statistical learning method that minimizes the training MSE?

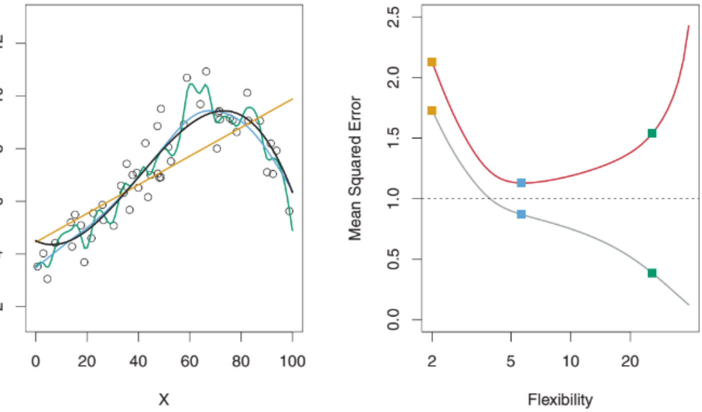
There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE

Test MSE

- Many statistical methods specifically estimate coefficients to minimize the training set MSE

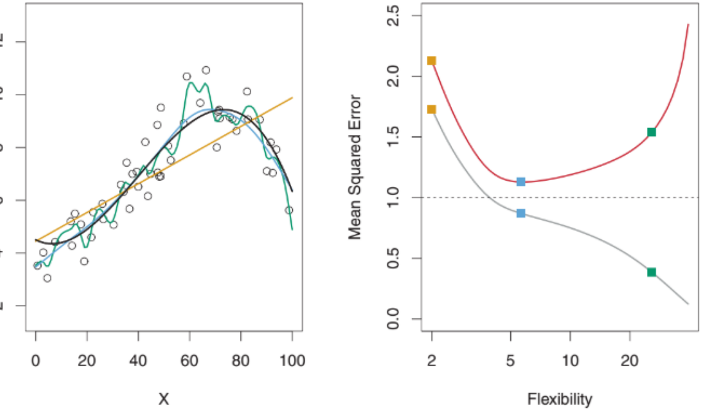
- For these methods, the training set MSE can be quite small, but the test MSE is often much larger.

Measuring the Quality of Fit



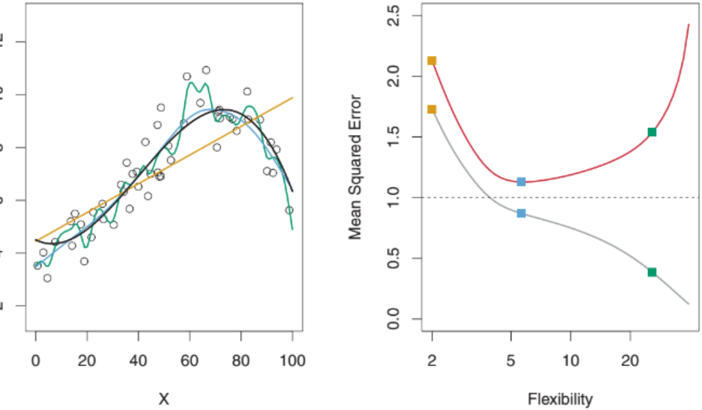
- True f: black curve
- Orange curve: the linear regression fit
- Blue and green curves: smoothing splines with different levels of smoothness

Measuring the Quality of Fit



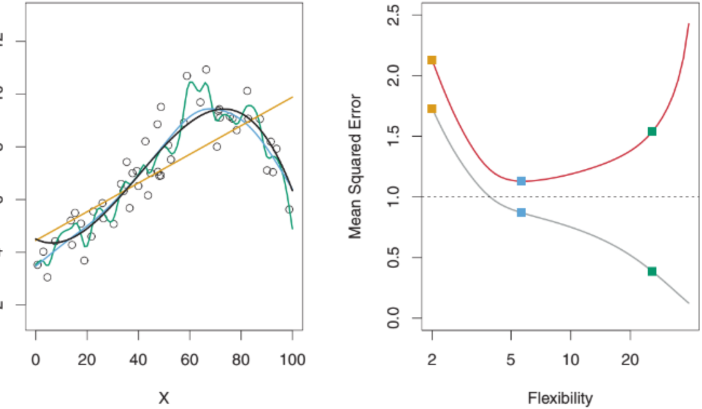
- As the level of flexibility increases, the curves fit the observed data more closely
- The green curve is the most flexible and matches the data very well
- It fits the true f poorly because it is too wiggly

Measuring the Quality of Fit



- The orange, blue and green squares indicate the MSEs associated with the corresponding curves
- The training MSE declines monotonically as flexibility increases: the green curve has the lowest training MSE of all three method

Measuring the Quality of Fit



- Red curve: the test MSE
- The test MSE initially declines as the level of flexibility increases, then it starts to increase again!

Overfitting

- A monotone decrease in the training MSE and a U-shape in the test MSE
- A fundamental property: as model flexibility increases, training MSE will decrease, but the test MSE may not
- When a given method yields a small training MSE but a large test MSE: **overfitting the data**
  - Learning procedure is working too hard to find patterns in the training data, and may pick up some patterns that are just caused by noise
  - When we overfit the training data, the test MSE will be very large because the supposed patterns don't exist in the test data.

Overfitting

- We almost expect the training MSE to be smaller than the test MSE because most statistical learning methods seek to minimize the training MSE
- Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

The Bias-Variance Trade-Off

- The expected test MSE can always be decomposed into the sum of three quantities:
  - the **variance** of $\hat{f}(x_0)$
  - the squared **bias** of $\hat{f}(x_0)$
  - the **variance of the error** terms ϵ

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

The Bias-Variance Trade-Off

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

- The expected test MSE can never lie below $\mathrm{Var}(\epsilon)$, the irreducible error
- In order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias

Variance

- **The variance** is the amount by which $\hat{f}$ would change if we estimated it using a different training data set
- Different training data sets will result in a different $\hat{f}$
- Ideally, the estimate for f should not vary too much between training sets
- If a method has high variance then small changes in the training data can result in large changes in $\hat{f}$
- In general, more flexible statistical methods have higher variance.

Bias

- **Bias** refers to the error that is introduced by approximating a real-life problem by a much simpler model
  - e.g. 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 such a simple linear relationship
  - performing linear regression will undoubtedly result in some bias in the estimate of f
- Generally, more flexible methods result in less bias

The Bias-Variance Trade-Off

- 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 we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases $\implies$ the expected test MSE declines
  - At some point increasing flexibility has little impact on the bias but starts to significantly increase the variance $\implies$ the test MSE increases.

The Bias-Variance Trade-Off

The relationship between bias, variance, and test set MSE is the bias-variance trade-off

- Good test set performance of a statistical learning method requires low variance as well as low bias
- A trade-off: it is easy to obtain a method with extremely low bias but high variance or a method with very low variance but high bias
- The challenge lies in finding a method for which both the variance and the squared bias are low.

The Classification Setting

- Focus on the *regression setting*
- Many of the concepts, such as the bias-variance trade-off, transfer over to the *classification setting*
- In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method
- The bias-variance tradeoff can make this a difficult task
- Next lecture we will discuss some methods for estimating test error rates which allow choosing the optimal level of flexibility