Inteligencia Artificial

UTN – FRVM

5º Año Ing. en Sistemas de Información

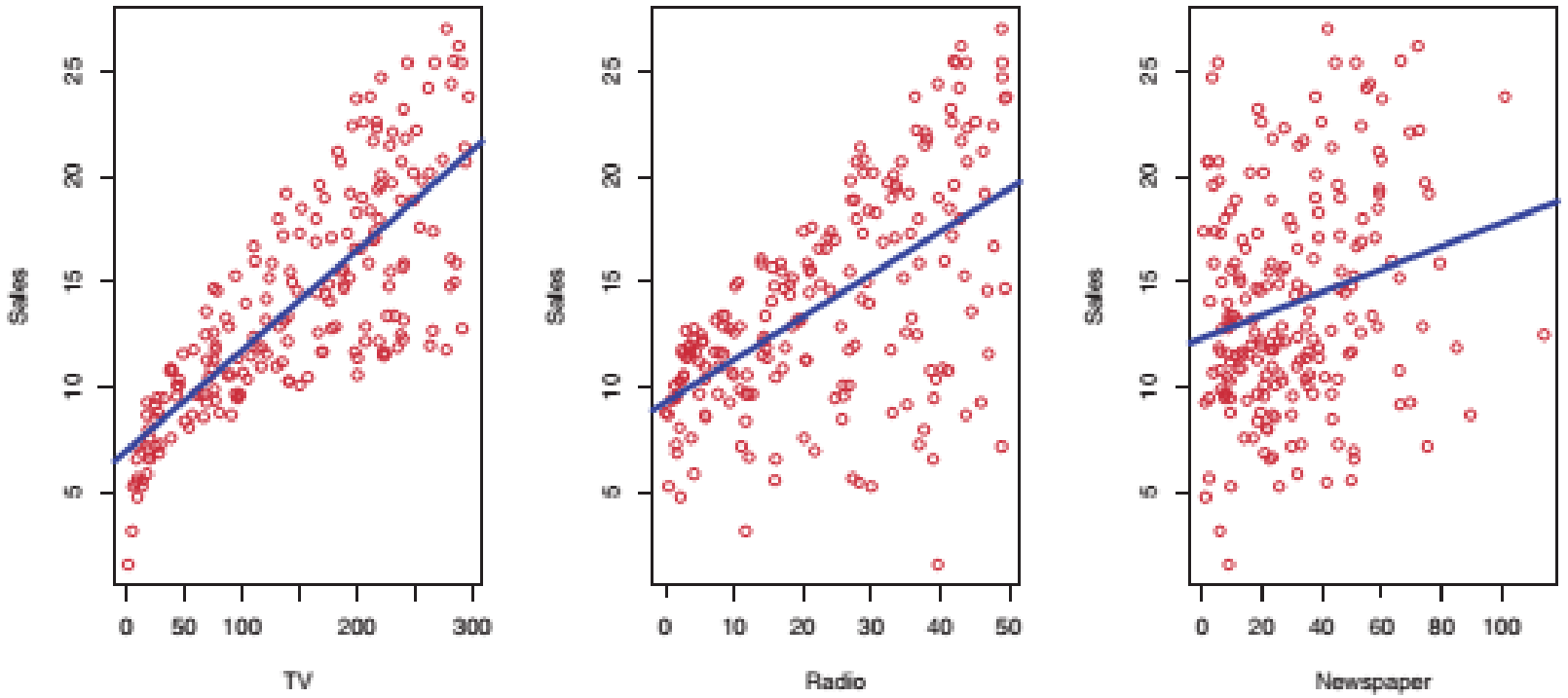
Agenda

- Introducción al Aprendizaje Supervisado y No Supervisado.
- Regresión y Clasificación.
- Training/Test MSE
- Equilibrio entre Varianza y Sesgo.

Supervised/Unsupervised Learning

- ***Simple Example***: provide advice on how to improve sales of a particular product.
- The ***Advertising*** data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: ***TV, radio, and newspaper***.
- It is not possible for our client to **directly increase sales** of the product. On the other hand, they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is ***an association between advertising and sales***, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
- In other words, our goal is ***to develop an accurate model that can be used to predict sales on the basis of the three media budgets***.

Supervised/Unsupervised Learning



- *The* Advertising *data set. The plot displays* sales*, in thousands of units, as a function of* TV*,* radio*, and* newspaper *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* sales *to that variable.*
- *In other words, each blue line represents a simple model that can be used to predict* sales *using* TV*,* radio*, and* newspaper*, respectively*

Input and Output Variables

- The advertising budgets are ***input*** *variables* while sales is an ***output*** *variable*.
- The input variables are typically denoted using the variable symbol X, with a subscript to distinguish them. So X_1 might be the TV budget, X_2 the radio budget, and X_3 the newspaper budget.
- The inputs go by different names, such as ***predictors***, ***independent variables***, ***features***, or sometimes just ***variables***.
- The output variable—in this case, sales—is often called the ***response*** **or** ***dependent*** *variable*, and is typically denoted using the symbol Y.

Function *f*

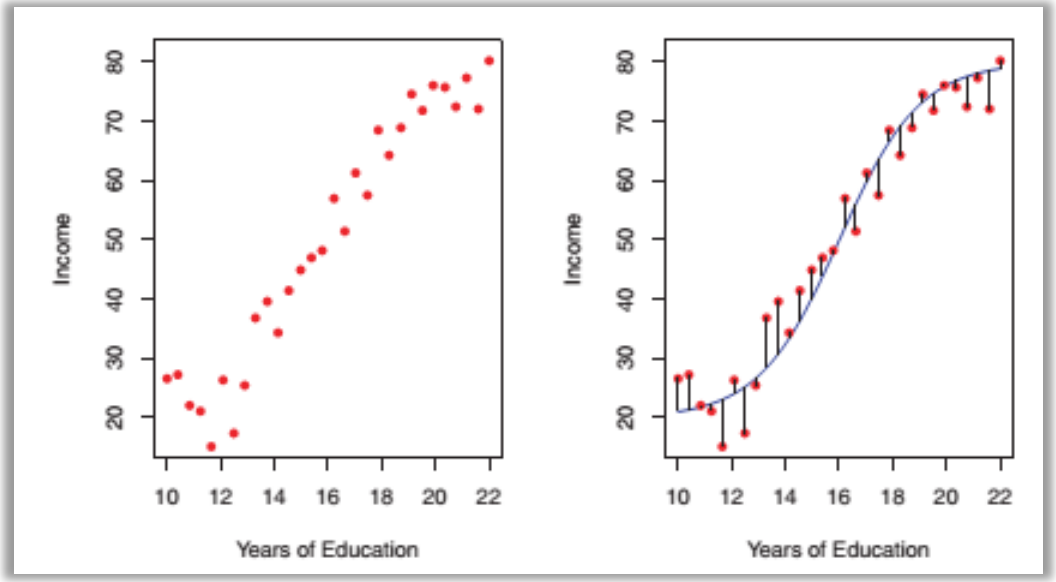
- More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$.
- We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

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

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

Income Dataset

- As another example, consider the left-hand panel, a plot of ***income versus years of education*** for 30 individuals in the Income data set.
- The plot suggests that one might be able to predict income using years of education. However, the function ***f*** that connects the input variable to the output variable is in general unknown.
- In this situation one must estimate *f* based on the observed points. Since Income is a simulated data set, *f* is known and is shown by the blue curve in the right-hand panel.
- The vertical lines represent the error terms ε. We note that some of the 30 observations lie above the blue curve and some lie below it; overall, the errors have approximately mean zero.



Approaches for estimating *f*

?

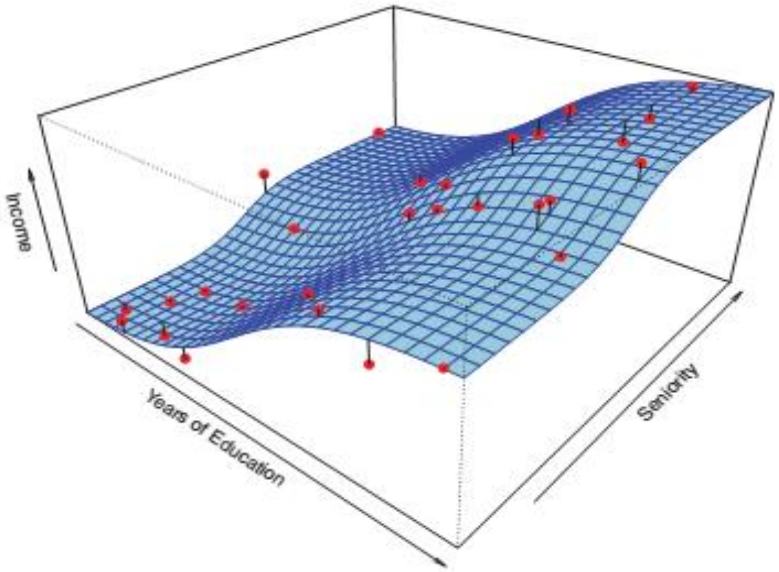
Why Estimate f?

- ***Prediction*** and ***inference***
- ***Prediction****:* In many situations, a set of inputs *X* are readily available, but the output *Y* cannot be easily obtained. In this setting, since the error term averages to zero, we can predict *Y* using

$$\hat{Y} = \hat{f}(X)$$

- where $\hat{f}$ represents our estimate for ***f***, and $\hat{Y}$ represents the resulting prediction for *Y*. In this setting, $\hat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for *Y.*

f for ***prediction***



The plot displays **income** as a function of **years of education** and **seniority** in the Income data set. The blue surface represents the true underlying relationship between income and years of education and seniority, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Reducible and Irreducible error

- Suppose that $X_1,...,X_p$ 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.
- It is natural to seek to predict Y using X, since we can then avoid giving the drug in question to patients who are at high risk of an adverse reaction—that is, patients for whom the estimate of Y is high.
- The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities, which we will call the *reducible error* and the *irreducible error*. In general, $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error.
- This error is *reducible* because we can potentially improve the accuracy of $\hat{f}$ by using the most appropriate learning technique to estimate f.
- However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\hat{Y} = f(X)$, our prediction would still have some error in it!
- This is because Y is also a function of ε, which, by definition, cannot be predicted using X. Therefore, variability associated with ε also affects the accuracy of our predictions.
- This is known as the *irreducible* error, because no matter how well we estimate f, we cannot reduce the error introduced by ε.

Reducible and Irreducible error

- Why is the irreducible error larger than zero?
- The quantity ε may contain unmeasured variables that are useful in predicting Y: since we don't measure them, f cannot use them for its prediction.
- The quantity ε may also contain unmeasurable variation.
- For example, the risk of an adverse reaction might vary for a given patient on a given day, depending on manufacturing variation in the drug itself or the patient's general feeling of well-being on that day.
- Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed.
- Then, it is easy to show that

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\
&= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

Reducible and Irreducible error

- Where $E(Y - \hat{Y})^2$ represents the average, or *expected value*, of the squared expected difference between the predicted and actual value of *Y*, and Var(ε) represents the *variance* associated with the error term ε.

- The focus of *ML* is on techniques for estimating ***f*** with the aim of minimizing the reducible error. It is important to 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.

f for ***inference***

- We are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change.
- In this situation we wish to estimate ***f***, to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, \ldots, X_p$.
- Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in answering the following questions

f for ***inference***

- ***Which predictors are associated with the response?*** *O*nly a small fraction of the available predictors are substantially associated with *Y*?

- ***What is the relationship between the response and each predictor?*** Some predictors may have a positive relationship with *Y*, in the sense that increasing the predictor is associated with increasing values of *Y*.

- ***Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?***

How Do We Estimate f?

- We will always assume that we have observed a set of n different data points.
- These observations are called the ***training data*** because we will use these training data observations to train, or teach, our method how to estimate ***f***.
- Let x_{ij} represent the value of the jth predictor, or input, for observation i, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.
- Correspondingly, let y_i represent the response variable for the ith observation.
- Then our training data consist of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.
- Our goal is to apply a learning method to the training data in order to estimate the unknown ***function f***.
- In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y). Broadly speaking, most learning methods for this task can be characterized as either *parametric* or *non-parametric*.

Parametric Methods

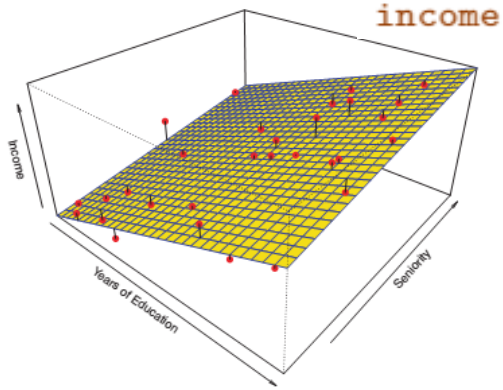
1. First, we make an assumption about the functional form, or shape, of *f*. For example, one very simple assumption is that *f* is linear in *X*: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

This is a *linear model*. Once we have assumed that *f* is linear, the problem of estimating *f* is greatly simplified. Instead of having to estimate an entirely arbitrary *p*-dimensional function *f*(*X*), one only needs to estimate the $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

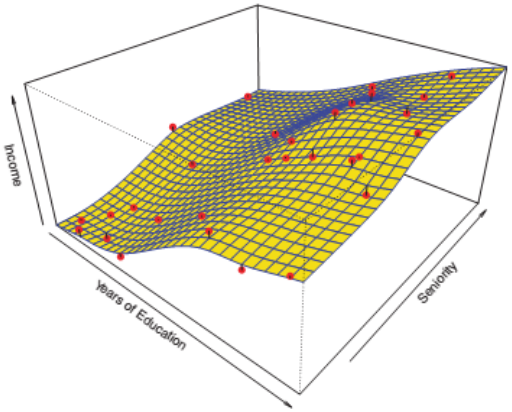
2. After a model has been selected, we need a procedure that uses the training data to *fit* or *train* the model. In the case of the linear model we need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$. That is, we want to find values of these parameters such that $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$.

The most common approach to fitting this model is referred to as ***(ordinary) least squares***. However, least squares is one of many possible ways to fit the linear model.

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}.$$

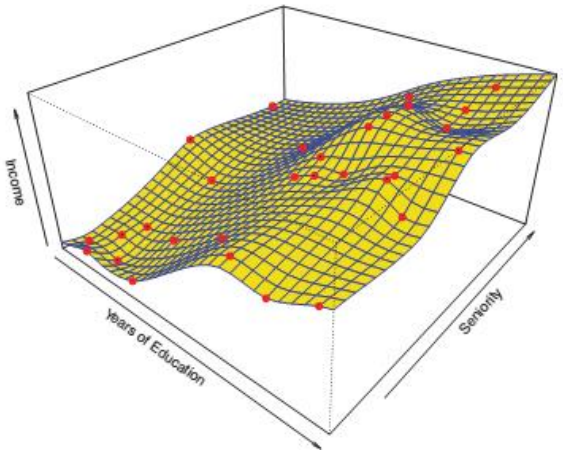


Overfitting?



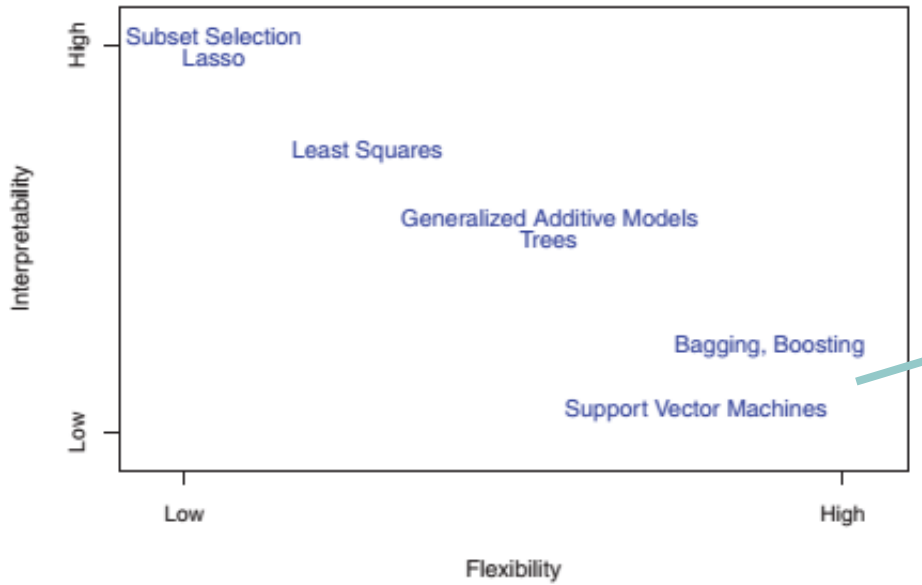
Non-parametric Methods

- Non-parametric methods **do not make explicit assumptions** about the functional form of *f*. Instead they seek an estimate of *f* that **gets as close to the data points as possible** without being too rough or wiggly.
- A major advantage over parametric approaches: by avoiding the assumption of a particular functional form for *f*, they have the potential **to accurately fit a wider range of possible shapes for *f***.
- Any parametric approach brings with it the possibility that the functional form used to estimate *f* is very different from the true *f*, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of *f* is made.
- **Major disadvantage**: since they do 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 Trade - Off Between Prediction Accuracy and Model Interpretability

- Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating *f* may be appropriate.
- For example, *linear models* allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches.
- In contrast, some of the ***highly non-linear approaches*** can potentially provide quite accurate predictions for *Y*, but this comes at the expense of a less interpretable model for which inference is more challenging.

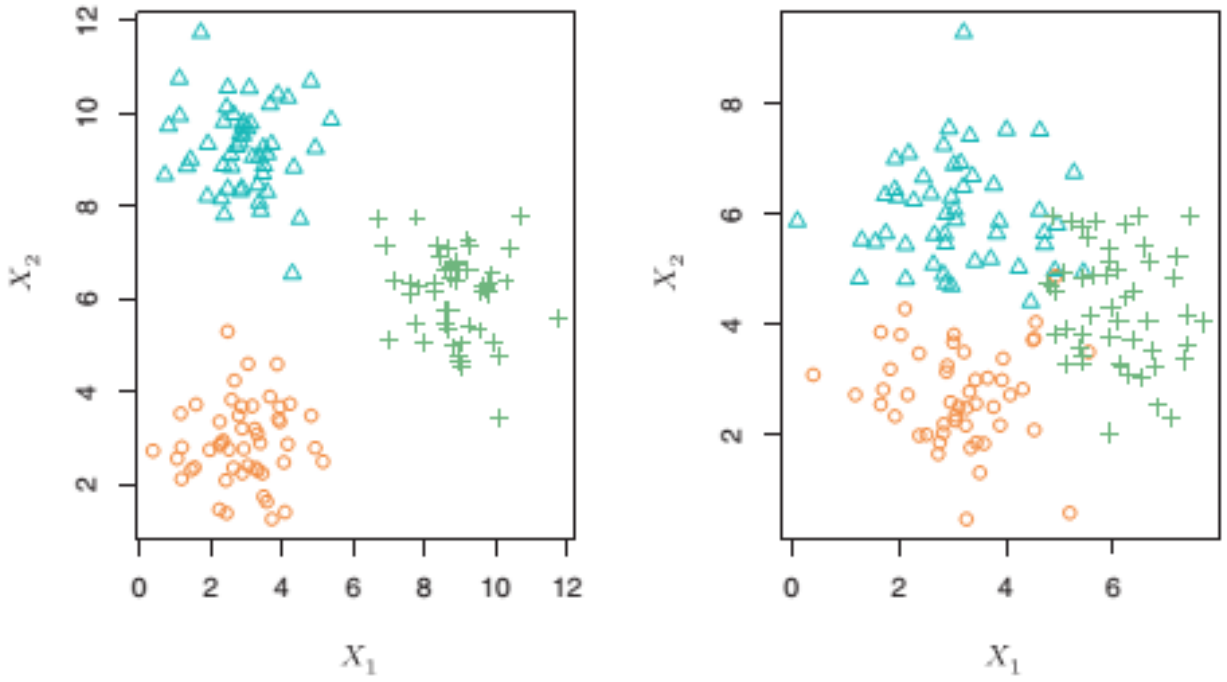


Non-linear kernels

Supervised/Unsupervised Learning

- Most learning problems fall into one of two categories: ***supervised or unsupervised***.
- The examples that we have discussed fall into the ***supervised learning*** domain. For each observation of the predictor measurement(s) $x_{i,\,i} = 1, \ldots, n$ there is an associated response measurement y_i.
- We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).
- Many classical statistical learning methods such as linear regression and *logistic regression* as well as more modern approaches such as GAM, boosting, and support vector machines, operate in the ***supervised learning*** domain.
- In contrast, ***unsupervised learning*** describes the somewhat more challenging situation in which for every observation $i = 1, \ldots, n$, we observe a vector of measurements $\boldsymbol{x}_i$ but no associated response $\boldsymbol{y}_i$.
- It is not possible to fit a linear regression model, since ***there is no response variable to predict***. In this setting, we are in some sense working blind; the situation is referred to as ***unsupervised*** because we lack a response variable that can supervise our analysis.

Clustering



- *A clustering data set involving three groups. Each group is shown using a different colored symbol.* **Left**: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* **Right**: *There is some overlap among the groups. Now the clustering task is more challenging.*

Regression Versus Classification Problems

- Variables can be characterized as either *quantitative* or *qualitative* (also quantitative known as *categorical*). Quantitative variables take on numerical values.
- Examples include a **person's age, height, or income, the value of a house, and the price of a stock**. In contrast, qualitative variables take on values in one of *K* different *classes*, or categories. Examples of qualitative class variables include a **person's gender (male or female), the brand of product purchased (brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia)**.
- Quantitative response: *regression* problems
- Qualitative response: *classification* problems.
- **Least squares linear regression** is used with a quantitative response, whereas **logistic regression** is typically used with a qualitative (two-class, or *binary*) response. But since it estimates class probabilities, it can be thought of as a regression method as well.
- So methods, such a s *K*-nearest neighbors and boosting, can be used in the case of either quantitative or qualitative responses.
- Wether the *predictors* are qualitative or quantitative is generally considered less important.

Assessing Model Accuracy

- Why is it necessary to introduce so many different statistical learning approaches, rather than just a single ***best*** method?
- ***There is no free lunch in statistics****:* no one method dominates all others over all possible data sets.
- On a particular data set, one specific method may work best, but some other method may work better on a **similar but different data set**.
- Hence it is an important task to 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.

Measuring the Quality of Fit

- In order to evaluate the performance of a statistical learning method on a given data set, we need **some way to measure** how well its predictions actually match the observed data.
- That is, we need to quantify the extent to which the **predicted response value** for a given observation is **close to the true response value** for that observation.
- In the regression setting, the most commonly-used measure is the ***mean squared error***.

MSE

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

- The MSE will be small if the predicted responses are very close to the true responses, and will be large if for some of the observations, the predicted and true responses differ substantially.

MSE

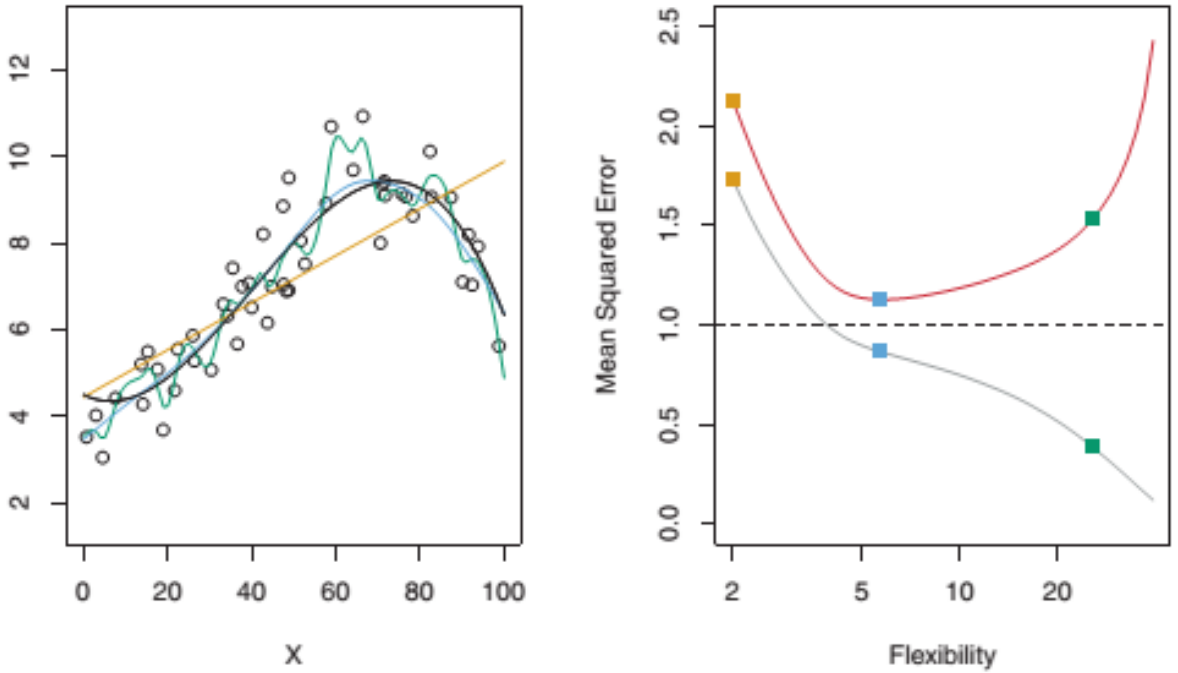
- The MSE is computed using the training data that was used to fit the model, and so should more accurately be referred to as the *training MSE*.
- But in general, we do not really care how well the method works on the training data. Rather, *we are interested in the accuracy of the pre dictions that we obtain when we apply our method to previously unseen test data*. **Why is this what we care about?**

Test MSE

- Suppose that we fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, and we obtain the estimate $\hat{f}$.
- We can then compute $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$. If these are approximately equal to $y_1, y_2, \ldots, y_n$, then the training MSE is small.
- However, we are really not interested in whether $\hat{f}(x_i) \approx y_i$; instead, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a ***previously unseen test observation not used to train the statistical learning method***. We want to choose the method that gives the lowest ***test MSE***, as opposed to the lowest training MSE. In other words, if we had a large number of test observations, we could compute

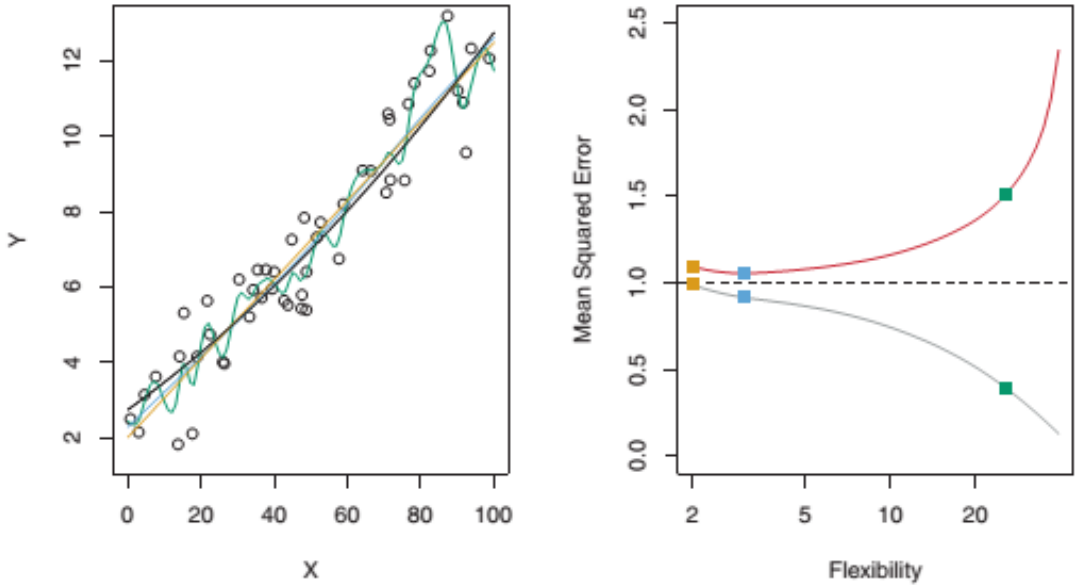
$$\text{Ave}(y_0 - \hat{f}(x_0))^2$$

Training vs. Test MSE



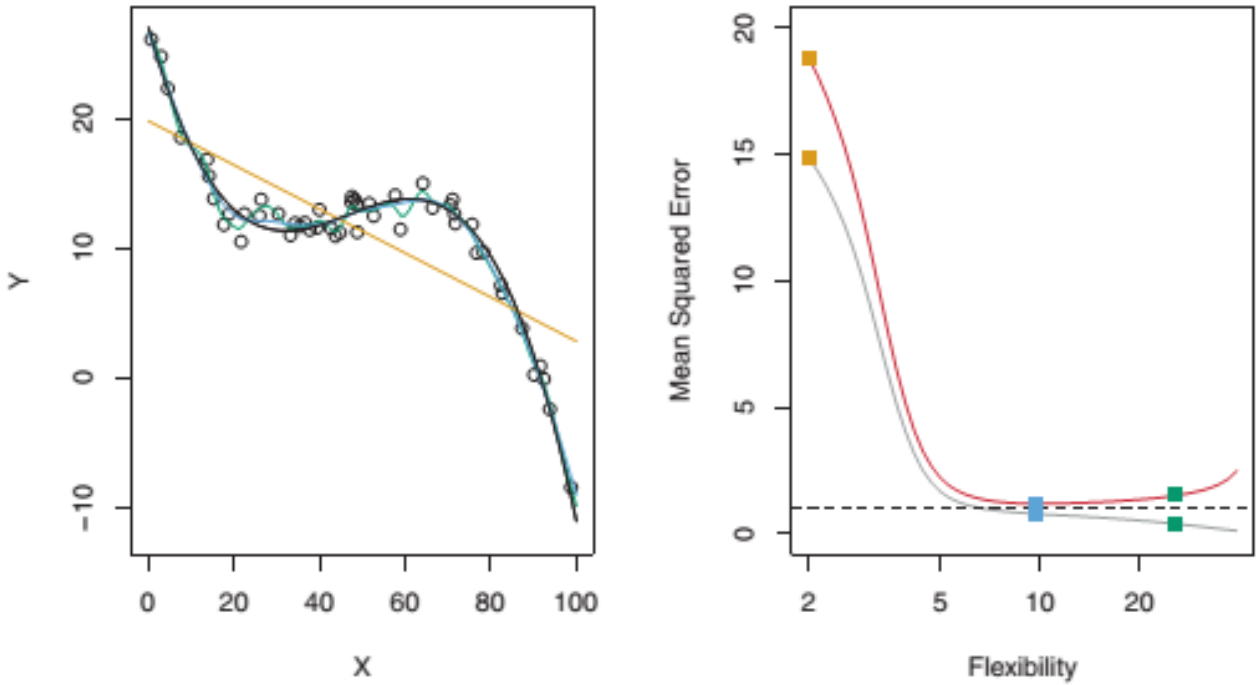
- ***Left***: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).*
- ***Right***: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed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

Training vs. Test MSE



- *In this setting, linear regression provides a very good fit to the data.*

Training vs. Test MSE



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

The Bias - Variance Trade - Off

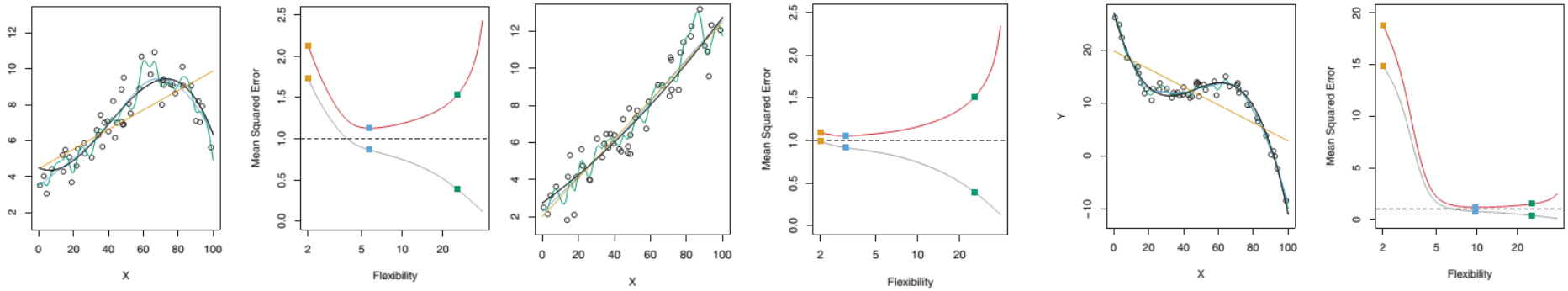
- The U-shape observed in the test MSE curves turns out to be the result of two competing properties of statistical learning methods.
- The expected test MSE, for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the *variance* of $\hat{f}(x_0)$, the squared *bias*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

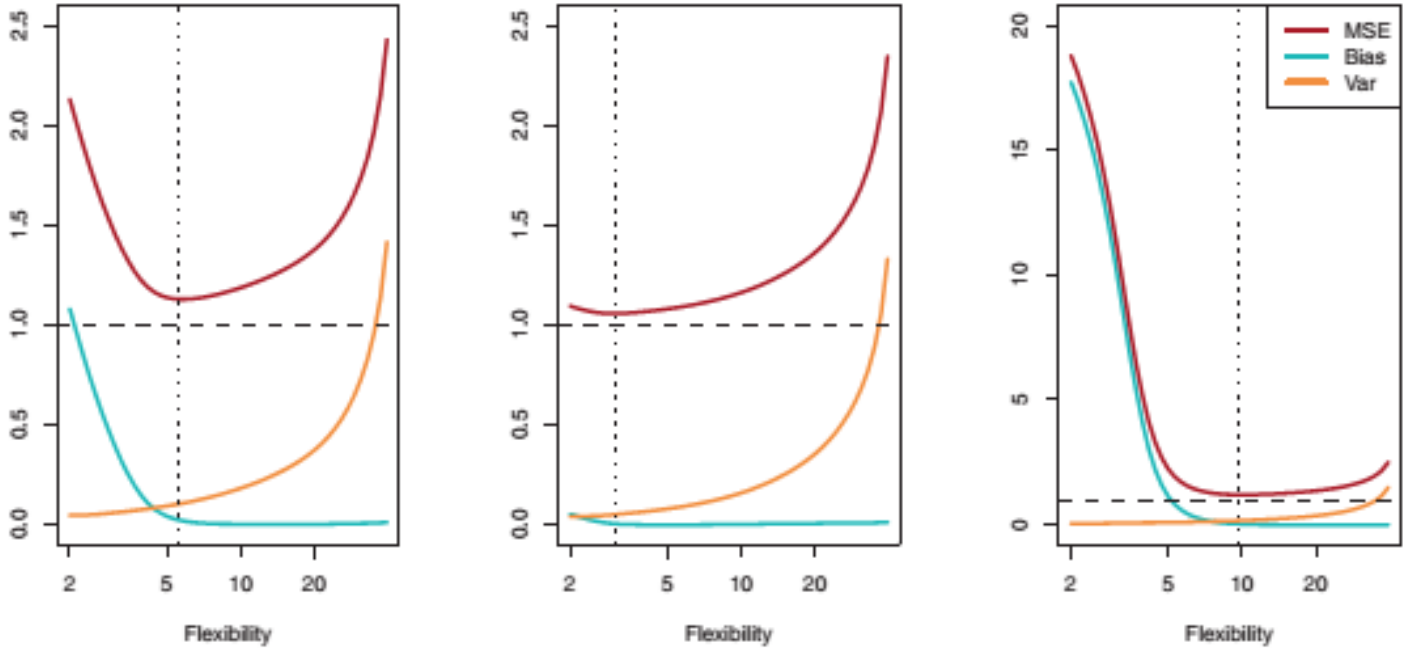
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

Expected test MSE

- Here the notation $E(y_0 - \hat{f}(x_0))^2$ defines the ***expected test MSE***, and refers to the average test MSE that we would obtain if we repeatedly estimated test MSE *f* using a large number of training sets, and tested each at x_0
- 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.

Expected test MSE





The Classification Setting

- Suppose that we seek to estimate f on the basis of training observations $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ where now $y_1,\ldots,y_n$ are qualitative.
- The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the ***training error rate***, the proportion of mistakes that are made if we apply $\hat{f}$ to the training observations

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

As in the error regression setting, we are most interested in the error rates that result from applying our classifier to test observations that were not used in training.
The ***test error rate*** associated with a set of test observations of the form
test error (x_0,y_0) is given by

$$\text{Ave}\left(I(y_0 \neq \hat{y}_0)\right)$$