Foundations of Machine Learning

Alvaro J. Riascos Villegas
University of los Andes and Quantil

July 6 2018

Contenido

Statistical Learning

- Most Statistical Learning techniques fall into one of the following two categories:
  1. Supervised Learning: Data take the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where y are the output variables.
     - The aim is to study the behavior of the output variable y (response variable) conditional on the independent variables x (predictor variables).
     - Mathematically: study and describe the distribution of y conditional on x.
  2. Unsupervised Learning: Data takes the form $\{x_1, \ldots, x_n\}$, there are inputs but no output to supervise.
     - The aim is study the x variables (inputs), its patterns, clusters, etc.
     - Mathematically: study the distribution of x.

Statistical Learning

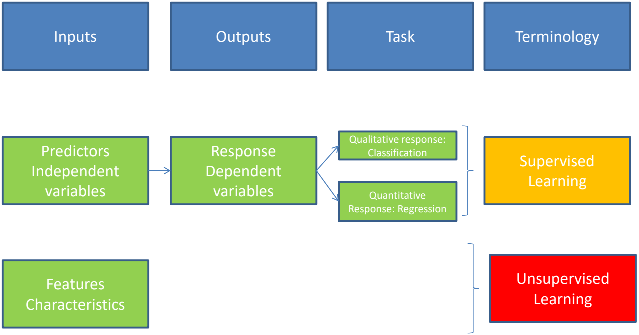
- Most Statistical Learning techniques fall into one of the following two categories:
  1. Supervised Learning: Data take the form $\{(x_1, y_1), ..., (x_n, y_n)\}$ where y are the output variables.
     - The aim is to study the behavior of the output variable y (response variable) conditional on the independent variables x (predictor variables).
     - Mathematically: study and describe the distribution of y conditional on x.
  2. Unsupervised Learning: Data takes the form $\{x_1, ..., x_n\}$, there are inputs but no output to supervise.
     - The aim is study the x variables (inputs), its patterns, clusters, etc.
     - Mathematically: study the distribution of x.

Statistical Learning

- Most Statistical Learning techniques fall into one of the following two categories:
  1. Supervised Learning: Data take the form $\{(x_1, y_1), ..., (x_n, y_n)\}$ where y are the output variables.
     - The aim is to study the behavior of the output variable y (response variable) conditional on the independent variables x (predictor variables).
     - Mathematically: study and describe the distribution of y conditional on x.
  2. Unsupervised Learning: Data takes the form $\{x_1, ..., x_n\}$, there are inputs but no output to supervise.
     - The aim is study the x variables (inputs), its patterns, clusters, etc.
     - Mathematically: study the distribution of x.

Statistical Learning

- Terminology



Statistical Learning

- To ilustrate the main ideas, let's focus on the classification task (applications: credit granting, fraud, customer profiling, etc).
- Suppose we have a sample $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ *independently* generated by a distribution $P(X, Y)$ where $y \in \{0, 1\}$.
- The distribution P is unknown.
- Assumption: sample is i.i.d.
- Denote Ξ the space of independent variables ($x \in \Xi$) and Υ the space of dependent variables ($y \in \Upsilon$).
- A **learning function** is a function $f : \Xi \to \Upsilon$. Intuitively, given an observation of x, the function selects a response $f(x)$.

- To ilustrate the main ideas, let's focus on the classification task (applications: credit granting, fraud, customer profiling, etc).
- Suppose we have a sample $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ *independently* generated by a distribution $P(X, Y)$ where $y \in \{0, 1\}$.
- The distribution P is unknown.
- Assumption: sample is i.i.d.
- Denote Ξ the space of independent variables ($x \in \Xi$) and Υ the space of dependent variables ($y \in \Upsilon$).
- A **learning function** is a function $f : \Xi \rightarrow \Upsilon$. Intuitively, given an observation of x, the function selects a response $f(x)$.

Statistical Learning

- To ilustrate the main ideas, let's focus on the classification task (applications: credit granting, fraud, customer profiling, etc).
- Suppose we have a sample $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ *independently* generated by a distribution $P(X, Y)$ where $y \in \{0, 1\}$.
- The distribution P is unknown.
- Assumption: sample is i.i.d.
- Denote Ξ the space of independent variables ($x \in \Xi$) and Υ the space of dependent variables ($y \in \Upsilon$).
- A **learning function** is a function $f : \Xi \to \Upsilon$. Intuitively, given an observation of x, the function selects a response $f(x)$.

Statistical Learning

- To ilustrate the main ideas, let's focus on the classification task (applications: credit granting, fraud, customer profiling, etc).
- Suppose we have a sample $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ *independently* generated by a distribution $P(X, Y)$ where $y \in \{0, 1\}$.
- The distribution P is unknown.
- Assumption: sample is i.i.d.
- Denote Ξ the space of independent variables ($x \in \Xi$) and Υ the space of dependent variables ($y \in \Upsilon$).
- A **learning function** is a function $f : \Xi \rightarrow \Upsilon$. Intuitively, given an observation of x, the function selects a response $f(x)$.

- To ilustrate the main ideas, let's focus on the classification task (applications: credit granting, fraud, customer profiling, etc).
- Suppose we have a sample $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ *independently* generated by a distribution $P(X, Y)$ where $y \in \{0, 1\}$.
- The distribution P is unknown.
- Assumption: sample is i.i.d.
- Denote Ξ the space of independent variables ($x \in \Xi$) and Υ the space of dependent variables ($y \in \Upsilon$).
- A **learning function** is a function $f : \Xi \rightarrow \Upsilon$. Intuitively, given an observation of x, the function selects a response $f(x)$.

Statistical Learning

- To ilustrate the main ideas, let's focus on the classification task (applications: credit granting, fraud, customer profiling, etc).
- Suppose we have a sample $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ *independently* generated by a distribution $P(X, Y)$ where $y \in \{0, 1\}$.
- The distribution P is unknown.
- Assumption: sample is i.i.d.
- Denote Ξ the space of independent variables ($x \in \Xi$) and Υ the space of dependent variables ($y \in \Upsilon$).
- A **learning function** is a function $f : \Xi \to \Upsilon$. Intuitively, given an observation of x, the function selects a response $f(x)$.

Statistical Learning: Loss function

- The standard way to evaluate the performance of a learning function for the **classification** task is using a loss function.
- Let $L : \Xi \times \Upsilon \times \Upsilon \rightarrow \{0, 1\}$. Given an observation (x, y), if $f(x) \neq y$ then $L(x, y, f(x)) = 1$ and $L(x, y, f(x)) = 0$ otherwise (standard Loss function for binary classification tasks).
- The most common way to measure the loss in a **regression** task is using the squared error: $L(x, y, f(x)) = (f(x) - y)^2$

Statistical Learning: Loss function

- The standard way to evaluate the performance of a learning function for the **classification** task is using a loss function.
- Let $L : \Xi \times \Upsilon \times \Upsilon \to \{0, 1\}$. Given an observation (x, y), if $f(x) \neq y$ then $L(x, y, f(x)) = 1$ and $L(x, y, f(x)) = 0$ otherwise (standard Loss function for binary classification tasks).
- The most common way to measure the loss in a **regression** task is using the squared error: $L(x, y, f(x)) = (f(x) - y)^2$

Statistical Learning: Loss function

- The standard way to evaluate the performance of a learning function for the **classification** task is using a loss function.
- Let $L : \Xi \times \Upsilon \times \Upsilon \to \{0, 1\}$. Given an observation (x, y), if $f(x) \neq y$ then $L(x, y, f(x)) = 1$ and $L(x, y, f(x)) = 0$ otherwise (standard Loss function for binary classification tasks).
- The most common way to measure the loss in a **regression** task is using the squared error: $L(x, y, f(x)) = (f(x) - y)^2$

- A Machine Learning algorithm M, is an algorithm that enable us to build a learning function form each sample τ_n.
- Let $\mathbb{F}$ be the set of all learning functions (i.e., $f_n : \Xi \to \Upsilon$), then:

$$M : (\Xi \times \Upsilon)^n \to \mathbb{F} \quad (1)$$

is a machine learning algorithm.

- $M(\tau_n)$ is denoted by f_n. The learning function f_n depends on the size of the sample n. For different samples we get different Learning functions.

- A Machine Learning algorithm M, is an algorithm that enable us to build a learning function form each sample τ_n.
- Let $\mathbb{F}$ be the set of all learning functions (i.e., $f_n : \Xi \to \Upsilon$), then:

$$M : (\Xi \times \Upsilon)^n \to \mathbb{F} \tag{1}$$

is a machine learning algorithm.

- $M(\tau_n)$ is denoted by f_n. The learning function f_n depends on the size of the sample n. For different samples we get different Learning functions.

- A Machine Learning algorithm M, is an algorithm that enable us to build a learning function form each sample τ_n.
- Let $\mathbb{F}$ be the set of all learning functions (i.e., $f_n : \Xi \to \Upsilon$), then:

$$M : (\Xi \times \Upsilon)^n \to \mathbb{F} \tag{1}$$

  is a machine learning algorithm.
- $M(\tau_n)$ is denoted by f_n. The learning function f_n depends on the size of the sample n. For different samples we get different Learning functions.

Statistical Learning: Pillars

- Approximation Error (bias) vs Estimation Error (Variance).
- Consistency.
- The problem of empirical risk minimization.
- Capacity.

Contenido

Two Workhorses of Statistical Learning

- Two workhorses (algorithms) of statitical learning are:
  1. K-Nearest Neighbors.
  2. Linear Regression Model.

Two Workhorses: K-Nearest Neighbors

- Suppose you have a concept of distance between predictors.
- Define k as the number of neighbors that the Learning function uses to classify.
- Given a sample τ_n and $x \in \Xi$, we identify the k points $\{x_{i_1}, ..., x_{i_k}\}$ that are closest to x.
- The learning function in binary classification tasks is defined based on the number of $\{k : y_{i_k} = 1\}$: majority vote.
- For regression tasks we estimate the average.
- We denote these learning machines as $K - NN_n$.

Two Workhorses: K-Nearest Neighbors

- Suppose you have a concept of distance between predictors.
- Define k as the number of neighbors that the Learning function uses to classify.
- Given a sample τ_n and $x \in \Xi$, we identify the k points $\{x_{i_1}, ..., x_{i_k}\}$ that are closest to x.
- The learning function in binary classification tasks is defined based on the number of $\{k : y_{i_k} = 1\}$: majority vote.
- For regression tasks we estimate the average.
- We denote these learning machines as $K - NN_n$.

Two Workhorses: K-Nearest Neighbors

- Suppose you have a concept of distance between predictors.
- Define k as the number of neighbors that the Learning function uses to classify.
- Given a sample τ_n and $x \in \Xi$, we identify the k points $\{x_{i_1}, ..., x_{i_k}\}$ that are closest to x.
- The learning function in binary classification tasks is defined based on the number of $\{k : y_{i_k} = 1\}$: majority vote.
- For regression tasks we estimate the average.
- We denote these learning machines as $K - NN_n$.

Two Workhorses: K-Nearest Neighbors

- Suppose you have a concept of distance between predictors.
- Define k as the number of neighbors that the Learning function uses to classify.
- Given a sample τ_n and $x \in \Xi$, we identify the k points $\{x_{i_1}, ..., x_{i_k}\}$ that are closest to x.
- The learning function in binary classification tasks is defined based on the number of $\{k : y_{i_k} = 1\}$: majority vote.
- For regression tasks we estimate the average.
- We denote these learning machines as $K - NN_n$.

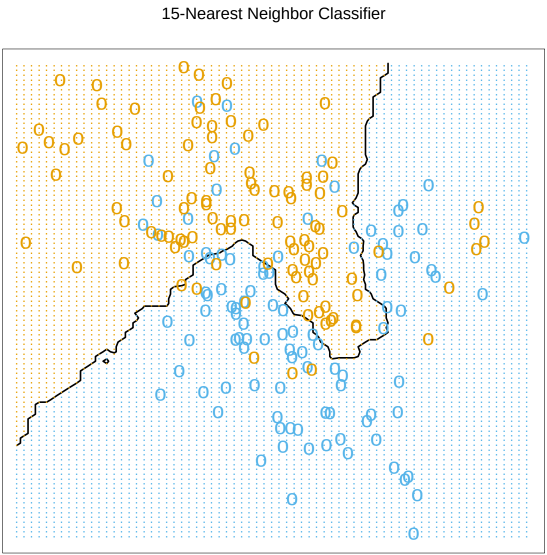
Two Workhorses: K-Nearest Neighbors

- Suppose you have a concept of distance between predictors.
- Define k as the number of neighbors that the Learning function uses to classify.
- Given a sample τ_n and $x \in \Xi$, we identify the k points $\{x_{i_1}, ..., x_{i_k}\}$ that are closest to x.
- The learning function in binary classification tasks is defined based on the number of $\{k : y_{i_k} = 1\}$: majority vote.
- For regression tasks we estimate the average.
- We denote these learning machines as $K - NN_n$.

Two Workhorses: K-Nearest Neighbors

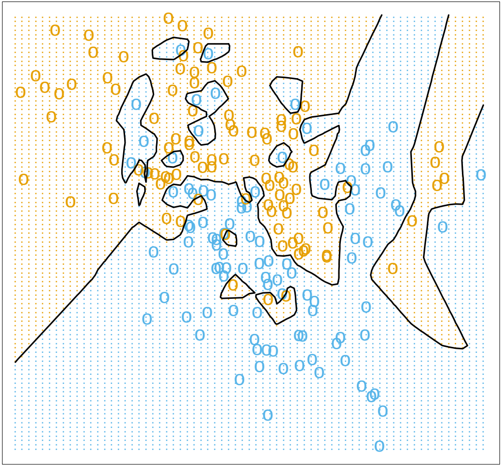
- Suppose you have a concept of distance between predictors.
- Define k as the number of neighbors that the Learning function uses to classify.
- Given a sample τ_n and $x \in \Xi$, we identify the k points $\{x_{i_1}, \ldots, x_{i_k}\}$ that are closest to x.
- The learning function in binary classification tasks is defined based on the number of $\{k : y_{i_k} = 1\}$: majority vote.
- For regression tasks we estimate the average.
- We denote these learning machines as $K - NN_n$.

Two Workhoses: K-Nearest Neighbors



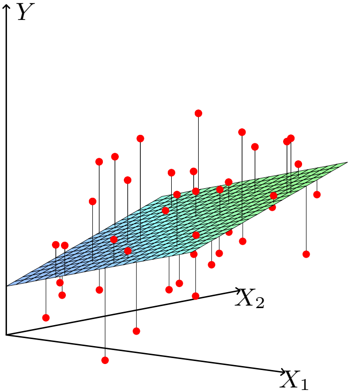
- Note that this learning function fits better *in-sample* and is more *complex* than the previous one.

1-Nearest Neighbor Classifier



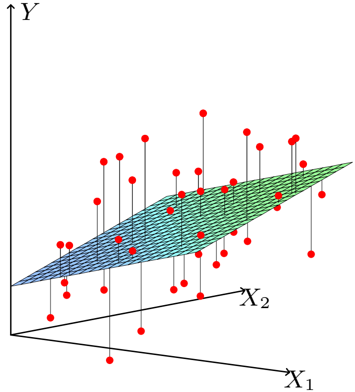
Two Workhorses: Linear Regression

- Suppose that $y_i = \beta_n^T x_i$ where we include a 1 as the first coordinate in every vector x_i (the constant in the linear regression model).
- Defining $\hat{\beta}_n$ as the estimator of ordinary least squares.
- Note that $\hat{\beta}_n$ defines a learning function $f_n^{OLS}(x) = 1$ if $\beta_n^T x > 0{,}5$ and zero otherwise.



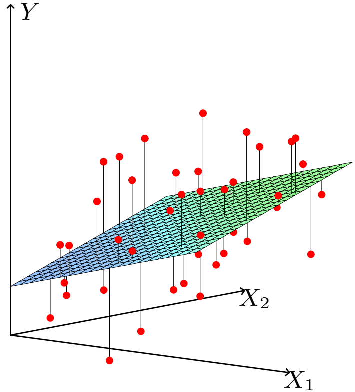
Two Workhorses: Linear Regression

- Suppose that $y_i = \beta_n^T x_i$ where we include a 1 as the first coordinate in every vector x_i (the constant in the linear regression model).
- Defining $\hat{\beta}_n$ as the estimator of ordinary least squares.
- Note that $\hat{\beta}_n$ defines a learning function $f_n^{OLS}(x) = 1$ if $\beta_n^T x > 0{,}5$ and zero otherwise.



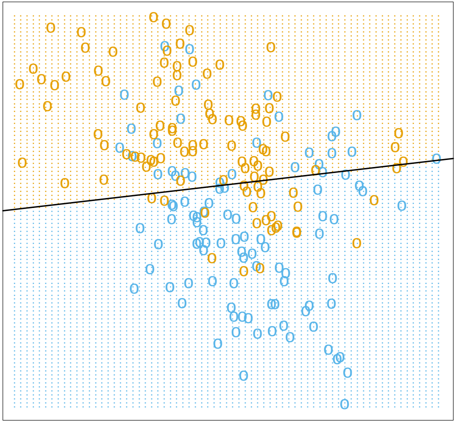
Two Workhorses: Linear Regression

- Suppose that $y_i = \beta_n^T x_i$ where we include a 1 as the first coordinate in every vector x_i (the constant in the linear regression model).
- Defining $\hat{\beta}_n$ as the estimator of ordinary least squares.
- Note that $\hat{\beta}_n$ defines a learning function $f_n^{OLS}(x) = 1$ if $\beta_n^T x > 0{,}5$ and zero otherwise.



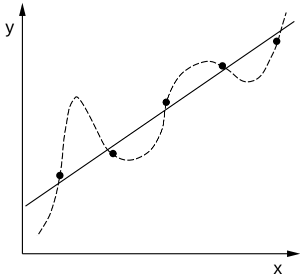
Two Workhorses: Linear Regression Model

- Black line corresponds to $\beta_n^T x_i = 0{,}5$.



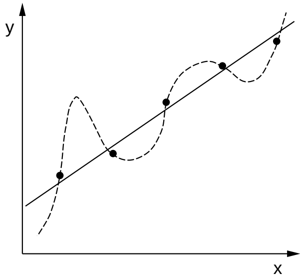
Statistical Learning: The Classification Task

- The Overfitting problem.
- Note that if the true model is the straight line, the empirical error of the curve is zero but the curve generalizes badly. The line's empirical error is larger than zero but the line generalizes better. And viceversa.
- In the first case, the curve is more complex, the variance is high and the bias is low. In the second case, the line is less complex, the variance is low but the bias is high.

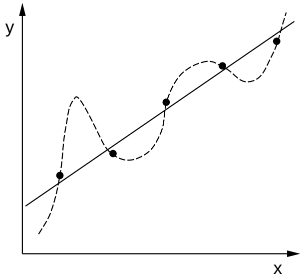


Statistical Learning: The Classification Task

- The Overfitting problem.
- Note that if the true model is the straight line, the empirical error of the curve is zero but the curve generalizes badly. The line's empirical error is larger than zero but the line generalizes better. And viceversa.
- In the first case, the curve is more complex, the variance is high and the bias is low. In the second case, the line is less complex, the variance is low but the bias is high.



- The Overfitting problem.
- Note that if the true model is the straight line, the empirical error of the curve is zero but the curve generalizes badly. The line's empirical error is larger than zero but the line generalizes better. And viceversa.
- In the first case, the curve is more complex, the variance is high and the bias is low. In the second case, the line is less complex, the variance is low but the bias is high.



Contenido

Risk

- Given a learning function f, the risk of f is defined as:

$$R(f) = E_P[L(X, Y, f(X))]$$

  where P is the joint distribution of X, Y, the true data generatig process.
- Notice that P is unknown so you cannot actually estimate the risk of a learning function. We will develop techniques to estimate this risk.
- Notice also that $R(f)$ is just the out of sample error of the learning function f (also called test error).
- The analogous in-sample error (also called training error) is the called the empirical risk.

Risk

- Given a learning function f, the risk of f is defined as:

$$R(f) = E_P[L(X, Y, f(X))]$$

  where P is the joint distribution of X, Y, the true data generatig process.
- Notice that P is unknown so you cannot actually estimate the risk of a learning function. We will develop techniques to estimate this risk.
- Notice also that $R(f)$ is just the out of sample error of the learning function f (also called test error).
- The analogous in-sample error (also called training error) is the called the empirical risk.

Risk

- Given a learning function f, the risk of f is defined as:

$$R(f) = E_P[L(X, Y, f(X))]$$

  where P is the joint distribution of X, Y, the true data generatig process.
- Notice that P is unknown so you cannot actually estimate the risk of a learning function. We will develop techniques to estimate this risk.
- Notice also that $R(f)$ is just the out of sample error of the learning function f (also called test error).
- The analogous in-sample error (also called training error) is the called the empirical risk.

Risk

- Given a learning function f, the risk of f is defined as:

$$R(f) = E_P[L(X, Y, f(X))]$$

  where P is the joint distribution of X, Y, the true data generatig process.
- Notice that P is unknown so you cannot actually estimate the risk of a learning function. We will develop techniques to estimate this risk.
- Notice also that $R(f)$ is just the out of sample error of the learning function f (also called test error).
- The analogous in-sample error (also called training error) is the called the empirical risk.

Learning Problem

- Define $\mathbb{F}_0$ as a set of functions (i.e., a subset of $\mathbb{F}$ the set of all functions).
- The learning problem is to solve:

$$f^* = argmin_{f \in \mathbb{F}_0} R[f] \tag{2}$$

  Note that $\mathbb{F}_0$ can be the set of all functions. When we want to make the space $\mathbb{F}_0$ explicit, we define f^* as $f_{\mathbb{F}_0}$.
- If $\mathbb{F}_0 = \mathbb{F}$ then f^* is called the Bayes Classifier (f_{Bayes}).
- Notice that since P is unknown you cannot actually solve the learning function. You can do experimental simulations.

Learning Problem

- Define $\mathbb{F}_0$ as a set of functions (i.e., a subset of $\mathbb{F}$ the set of all functions).
- The learning problem is to solve:

$$f^* = argmin_{f \in \mathbb{F}_0} R[f] \tag{2}$$

  Note that $\mathbb{F}_0$ can be the set of all functions. When we want to make the space $\mathbb{F}_0$ explicit, we define f^* as $f_{\mathbb{F}_0}$.
- If $\mathbb{F}_0 = \mathbb{F}$ then f^* is called the Bayes Classifier (f_{Bayes}).
- Notice that since P is unknown you cannot actually solve the learning function. You can do experimental simulations.

Learning Problem

- Define $\mathbb{F}_0$ as a set of functions (i.e., a subset of $\mathbb{F}$ the set of all functions).
- The learning problem is to solve:

$$f^* = argmin_{f \in \mathbb{F}_0} R[f] \tag{2}$$

  Note that $\mathbb{F}_0$ can be the set of all functions. When we want to make the space $\mathbb{F}_0$ explicit, we define f^* as $f_{\mathbb{F}_0}$.
- If $\mathbb{F}_0 = \mathbb{F}$ then f^* is called the Bayes Classifier (f_{Bayes}).
- Notice that since P is unknown you cannot actually solve the learning function. You can do experimental simulations.

Learning Problem

- Define $\mathbb{F}_0$ as a set of functions (i.e., a subset of $\mathbb{F}$ the set of all functions).
- The learning problem is to solve:

$$f^* = argmin_{f \in \mathbb{F}_0} R[f] \tag{2}$$

  Note that $\mathbb{F}_0$ can be the set of all functions. When we want to make the space $\mathbb{F}_0$ explicit, we define f^* as $f_{\mathbb{F}_0}$.
- If $\mathbb{F}_0 = \mathbb{F}$ then f^* is called the Bayes Classifier (f_{Bayes}).
- Notice that since P is unknown you cannot actually solve the learning function. You can do experimental simulations.

Learning Problem

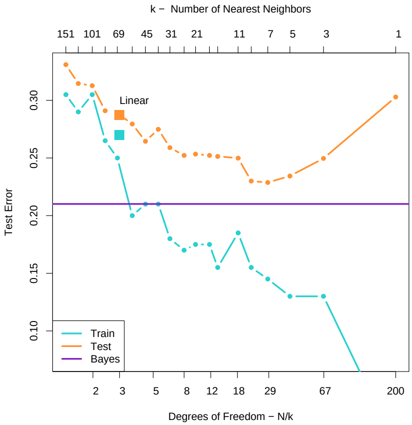
- We can summarize the primary objetive of statistical learning as follows: Given a finite sample τ_n and a loss function L, we want to find a space of functions $\mathbb{F}_0$ and a optimal classifier $f_{\mathbb{F}_0}$ such that its risk is as close as possible to the Bayes Classifier.
- The empirical risk of a classifier is not necessarily a good estimator of risk.

Learning Problem

- We can summarize the primary objetive of statistical learning as follows: Given a finite sample τ_n and a loss function L, we want to find a space of functions $\mathbb{F}_0$ and a optimal classifier $f_{\mathbb{F}_0}$ such that its risk is as close as possible to the Bayes Classifier.
- The empirical risk of a classifier is not necessarily a good estimator of risk.

Contenido

Bias vs. Variance Tradeoff



- Simulation exercise: 200 examples (training set), 10.000

Bias vs. Variance Tradeoff

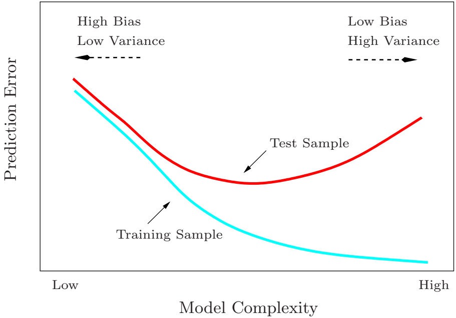
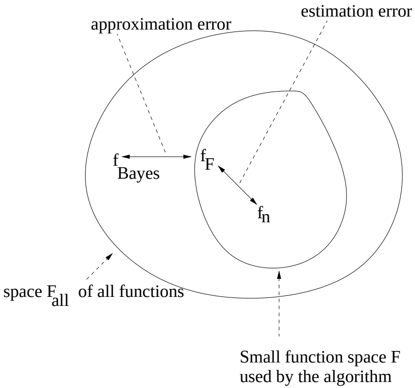
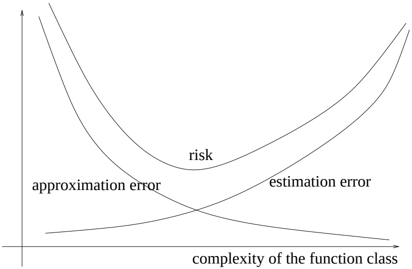


FIGURE 2.11. *Test and training error as a function of model complexity.*

Aproximation Error vs. Estimation Error



Aproximation Error vs. Estimation Error



Contenido

Prediction vs Causality

- Kleinberg, Ludwig, Mullainnathan, Obermeyer. 2015. Prediction Policy Problems. AER.
- Sometimes, the identification of a causal effect is irrelevant.
- This paper introduces a conceptual framework to think about the relationship between the prediction problem and causality.

Prediction vs Causality

- Kleinberg, Ludwig, Mullainnathan, Obermeyer. 2015. Prediction Policy Problems. AER.
- Sometimes, the identification of a causal effect is irrelevant.
- This paper introduces a conceptual framework to think about the relationship between the prediction problem and causality.

Prediction vs Causality

- Kleinberg, Ludwig, Mullainnathan, Obermeyer. 2015. Prediction Policy Problems. AER.
- Sometimes, the identification of a causal effect is irrelevant.
- This paper introduces a conceptual framework to think about the relationship between the prediction problem and causality.

Introduction

- Defining y as the interest variable, we don't know how y relates to x_0 (exogenous variable: economic policy) and x (covariates).
- The aim is to maximize a known function $\Pi(x_0, y)$.
- The decision depends on:

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y) + \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

  1. Even though Π is known, the effect of x_0 depends on y (prediction problem).
  2. The second term depends on how x_0 affects y (causality problem).
- Note that both effects depend on the prediction y.
- Therefore, a policymaker must solve both problems.

Introduction

- Defining y as the interest variable, we don't know how y relates to x_0 (exogenous variable: economic policy) and x (covariates).
- The aim is to maximize a known function $\Pi(x_0, y)$.
- The decision depends on:

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y) + \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

  1. Even though Π is known, the effect of x_0 depends on y (prediction problem).
  2. The second term depends on how x_0 affects y (causality problem).
- Note that both effects depend on the prediction y.
- Therefore, a policymaker must solve both problems.

Introduction

- Defining y as the interest variable, we don't know how y relates to x_0 (exogenous variable: economic policy) and x (covariates).
- The aim is to maximize a known function $\Pi(x_0, y)$.
- The decision depends on:

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y) + \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

1. Even though Π is known, the effect of x_0 depends on y (prediction problem).
2. The second term depends on how x_0 affects y (causality problem).

- Note that both effects depend on the prediction y.
- Therefore, a policymaker must solve both problems.

- Defining y as the interest variable, we don't know how y relates to x_0 (exogenous variable: economic policy) and x (covariates).
- The aim is to maximize a known function $\Pi(x_0, y)$.
- The decision depends on:

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y) + \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

  1. Even though Π is known, the effect of x_0 depends on y (prediction problem).
  2. The second term depends on how x_0 affects y (causality problem).
- Note that both effects depend on the prediction y.
- Therefore, a policymaker must solve both problems.

Introduction

- Defining y as the interest variable, we don't know how y relates to x_0 (exogenous variable: economic policy) and x (covariates).
- The aim is to maximize a known function $\Pi(x_0, y)$.
- The decision depends on:

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y) + \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

  1. Even though Π is known, the effect of x_0 depends on y (prediction problem).
  2. The second term depends on how x_0 affects y (causality problem).

- Note that both effects depend on the prediction y.
- Therefore, a policymaker must solve both problems.

Introduction

- Defining y as the interest variable, we don't know how y relates to x_0 (exogenous variable: economic policy) and x (covariates).
- The aim is to maximize a known function $\Pi(x_0, y)$.
- The decision depends on:

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y) + \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

  1. Even though Π is known, the effect of x_0 depends on y (prediction problem).
  2. The second term depends on how x_0 affects y (causality problem).

- Note that both effects depend on the prediction y.
- Therefore, a policymaker must solve both problems.

Example

- Suppose that the interest variable y is defined based on whether it rains or not. The decision in a problem could be a ritual so that it does not rain. In another problem the decision could be to carry or not umbrella.
- The objective function can be the utility generated by going to the park a Sunday.
- The task of doing a ritual is a causality problem:

$$\frac{\partial \Pi}{\partial x_0}(x_0, y) = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

- The ritual has no direct effect on the utility function.
- It may be the case that $\frac{\partial y}{\partial x_0}(x_0) \neq 0$.

Example

- Suppose that the interest variable y is defined based on whether it rains or not. The decision in a problem could be a ritual so that it does not rain. In another problem the decision could be to carry or not umbrella.
- The objective function can be the utility generated by going to the park a Sunday.
- The task of doing a ritual is a causality problem:

$$\frac{\partial \Pi}{\partial x_0}(x_0, y) = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

- The ritual has no direct effect on the utility function.
- It may be the case that $\frac{\partial y}{\partial x_0}(x_0) \neq 0$.

Example

- Suppose that the interest variable y is defined based on whether it rains or not. The decision in a problem could be a ritual so that it does not rain. In another problem the decision could be to carry or not umbrella.
- The objective function can be the utility generated by going to the park a Sunday.
- The task of doing a ritual is a causality problem:

$$\frac{\partial \Pi}{\partial x_0}(x_0, y) = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

- The ritual has no direct effect on the utility function.
- It may be the case that $\frac{\partial y}{\partial x_0}(x_0) \neq 0$.

Example

- Suppose that the interest variable y is defined based on whether it rains or not. The decision in a problem could be a ritual so that it does not rain. In another problem the decision could be to carry or not umbrella.
- The objective function can be the utility generated by going to the park a Sunday.
- The task of doing a ritual is a causality problem:

$$\frac{\partial \Pi}{\partial x_0}(x_0, y) = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

- The ritual has no direct effect on the utility function.
- It may be the case that $\frac{\partial y}{\partial x_0}(x_0) \neq 0$.

Example

- Suppose that the interest variable y is defined based on whether it rains or not. The decision in a problem could be a ritual so that it does not rain. In another problem the decision could be to carry or not umbrella.
- The objective function can be the utility generated by going to the park a Sunday.
- The task of doing a ritual is a causality problem:

$$\frac{\partial \Pi}{\partial x_0}(x_0, y) = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial y}(x_0, y)\frac{\partial y}{\partial x_0}(x_0)$$

- The ritual has no direct effect on the utility function.
- It may be the case that $\frac{\partial y}{\partial x_0}(x_0) \neq 0$.

Example

- The problem of carrying an umbrella is a pure problem of prediction:

$$\frac{\partial y}{\partial x_0} = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y)$$

- The utility of carrying an umbrella depends on whether it rains or not.

Example

- The problem of carrying an umbrella is a pure problem of prediction:

$$\frac{\partial y}{\partial x_0} = 0$$

$$\frac{\partial \Pi}{\partial x_0} = \frac{\partial \Pi}{\partial x_0}(x_0, y)$$

- The utility of carrying an umbrella depends on whether it rains or not.