Aprendizaje Estadístico:

Una Introducción (muy corta)

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Enero de 2023

Contenido

Aprendizaje Estadístico

- Las técnicas de minería de datos se dividen básicamente en dos grandes ramas:
  1. Análisis supervisado: se cuenta con datos de la forma $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ donde las observaciones de y son las variables de interés.
     - El objetivo es estudiar el comportamiento de la variable objetivo y (o respuesta) condicional a las variables independientes x (o predictores).
     - Matemáticamente: estudiar y describir la distribución de y condicional a x.
  2. Análisis no supervisado: se cuentan con datos de la forma $\{x_1, \ldots, x_n\}$, ninguna variables es el centro de atención.
     - El objetivo es estudiar las variables x (o factores), posibles patrones, conglomerados, etc.
     - Matemáticamente: estudiar la distribución de x.

Aprendizaje Estadístico

- Las técnicas de minería de datos se dividen básicamente en dos grandes ramas:
  1. Análisis supervisado: se cuenta con datos de la forma $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ donde las observaciones de y son las variables de interés.
     - El objetivo es estudiar el comportamiento de la variable objetivo y (o respuesta) condicional a las variables independientes x (o predictores).
     - Matemáticamente: estudiar y describir la distribución de y condicional a x.
  2. Análisis no supervisado: se cuentan con datos de la forma $\{x_1, \ldots, x_n\}$, ninguna variables es el centro de atención.
     - El objetivo es estudiar las variables x (o factores), posibles patrones, conglomerados, etc.
     - Matemáticamente: estudiar la distribución de x.

Aprendizaje Estadístico

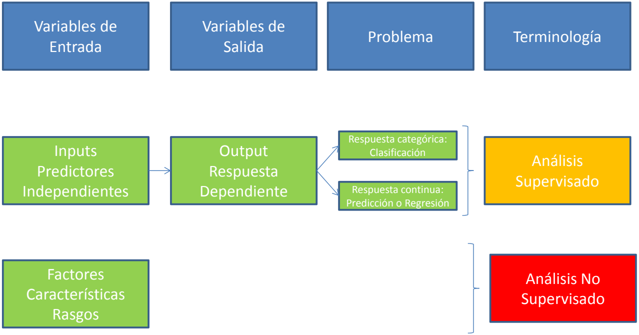
- Las técnicas de minería de datos se dividen básicamente en dos grandes ramas:
  1. Análisis supervisado: se cuenta con datos de la forma $\{(x_1, y_1), ..., (x_n, y_n)\}$ donde las observaciones de y son las variables de interés.
     - El objetivo es estudiar el comportamiento de la variable objetivo y (o respuesta) condicional a las variables independientes x (o predictores).
     - Matemáticamente: estudiar y describir la distribución de y condicional a x.
  2. Análisis no supervisado: se cuentan con datos de la forma $\{x_1, ..., x_n\}$, ninguna variables es el centro de atención.
     - El objetivo es estudiar las variables x (o factores), posibles patrones, conglomerados, etc.
     - Matemáticamente: estudiar la distribución de x.

- Otros problemas son:
  1. Aprendizeje semisupervisado: Se tienen algunos ejemplos etiquetados pero no todos.
  2. Aprendizaje reforzado: Se interactura con un ambiente (i.e., físico, juego, etc) y se tiene un conjunto de acciones. El propósito es, tomar acciones, intereactuar con el ambiente (i.e., explotar y explorar) y maximizar una función de beneficios del agente.
  3. Aprendizaje online: Se reciben nuevos ejemplos de forma secuencial y no es posible reestimar el modelo con todos los datos (*batch learning*).

- Otros problemas son:
  1. Aprendizeje semisupervisado: Se tienen algunos ejemplos etiquetados pero no todos.
  2. Aprendizaje reforzado: Se interactura con un ambiente (i.e., físico, juego, etc) y se tiene un conjunto de acciones. El propósito es, tomar acciones, intereactuar con el ambiente (i.e., explotar y explorar) y maximizar una función de beneficios del agente.
  3. Aprendizaje online: Se reciben nuevos ejemplos de forma secuencial y no es posible reestimar el modelo con todos los datos (*batch learning*).

Aprendizaje Estadístico

- Terminología



Aprendizaje Estadístico

- Para ilustrar algunas de las ideas principales enfoquémonos en el problema de clasificación (que tiene aplicaciones a: otorgamiento de créditos, fraude, caracterización de clientes, etc.)
- Supongamos que tenemos una muestra $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ generada de forma *independiente* de una distribución $P(X, Y)$ donde $y \in \{0, 1\}$.
- La distribución P es desconocida.
- Obsérvese que el supuesto es que la muestra es i.i.d.
- Denotamos por $\mathcal{X}$ el espacio de variables independientes ($x \in \mathcal{X}$) y $\mathcal{Y}$ el espacio de variables dependientes ($y \in \mathcal{Y}$).
- Una función de apredizaje es una función $f : \mathcal{X} \to \mathcal{Y}$. Intuitivamente dada una observación de x la función selecciona un resultado $f(x)$.

- Para ilustrar algunas de las ideas principales enfoquémonos en el problema de clasificación (que tiene aplicaciones a: otorgamiento de créditos, fraude, caracterización de clientes, etc.)
- Supongamos que tenemos una muestra $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ generada de forma *independiente* de una distribución $P(X, Y)$ donde $y \in \{0, 1\}$.
- La distribución P es desconocida.
- Obsérvese que el supuesto es que la muestra es i.i.d.
- Denotamos por $\mathcal{X}$ el espacio de variables independientes ($x \in \mathcal{X}$) y $\mathcal{Y}$ el espacio de variables dependientes ($y \in \mathcal{Y}$).
- Una función de apredizaje es una función $f : \mathcal{X} \to \mathcal{Y}$. Intuitivamente dada una observación de x la función selecciona un resultado $f(x)$.

- Para ilustrar algunas de las ideas principales enfoquémonos en el problema de clasificación (que tiene aplicaciones a: otorgamiento de créditos, fraude, caracterización de clientes, etc.)
- Supongamos que tenemos una muestra $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ generada de forma *independiente* de una distribución $P(X, Y)$ donde $y \in \{0, 1\}$.
- La distribución P es desconocida.
- Obsérvese que el supuesto es que la muestra es i.i.d.
- Denotamos por $\mathcal{X}$ el espacio de variables independientes ($x \in \mathcal{X}$) y $\mathcal{Y}$ el espacio de variables dependientes ($y \in \mathcal{Y}$).
- Una función de apredizaje es una función $f : \mathcal{X} \to \mathcal{Y}$. Intuitivamente dada una observación de x la función selecciona un resultado $f(x)$.

- Para ilustrar algunas de las ideas principales enfoquémonos en el problema de clasificación (que tiene aplicaciones a: otorgamiento de créditos, fraude, caracterización de clientes, etc.)
- Supongamos que tenemos una muestra $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ generada de forma *independiente* de una distribución $P(X, Y)$ donde $y \in \{0, 1\}$.
- La distribución P es desconocida.
- Obsérvese que el supuesto es que la muestra es i.i.d.
- Denotamos por $\mathcal{X}$ el espacio de variables independientes $(x \in \mathcal{X})$ y $\mathcal{Y}$ el espacio de variables dependientes $(y \in \mathcal{Y})$.
- Una función de apredizaje es una función $f : \mathcal{X} \to \mathcal{Y}$. Intuitivamente dada una observación de x la función selecciona un resultado $f(x)$.

Aprendizaje Estadístico

- Para ilustrar algunas de las ideas principales enfoquémonos en el problema de clasificación (que tiene aplicaciones a: otorgamiento de créditos, fraude, caracterización de clientes, etc.)
- Supongamos que tenemos una muestra $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ generada de forma *independiente* de una distribución $P(X, Y)$ donde $y \in \{0, 1\}$.
- La distribución P es desconocida.
- Obsérvese que el supuesto es que la muestra es i.i.d.
- Denotamos por $\mathcal{X}$ el espacio de variables independientes ($x \in \mathcal{X}$) y $\mathcal{Y}$ el espacio de variables dependientes ($y \in \mathcal{Y}$).
- Una función de apredizaje es una función $f : \mathcal{X} \to \mathcal{Y}$. Intuitivamente dada una observación de x la función selecciona un resultado $f(x)$.

- Para ilustrar algunas de las ideas principales enfoquémonos en el problema de clasificación (que tiene aplicaciones a: otorgamiento de créditos, fraude, caracterización de clientes, etc.)
- Supongamos que tenemos una muestra $\tau_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ generada de forma *independiente* de una distribución $P(X, Y)$ donde $y \in \{0, 1\}$.
- La distribución P es desconocida.
- Obsérvese que el supuesto es que la muestra es i.i.d.
- Denotamos por $\mathcal{X}$ el espacio de variables independientes ($x \in \mathcal{X}$) y $\mathcal{Y}$ el espacio de variables dependientes ($y \in \mathcal{Y}$).
- Una función de apredizaje es una función $f : \mathcal{X} \to \mathcal{Y}$. Intuitivamente dada una observación de x la función selecciona un resultado $f(x)$.

Aprendizaje Estadístico: Pérdida

- La forma estándar de evaluar el rendimiento de una función de aprendizaje para el problema de clasificación es usando una función de pérdida, $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$.
- La función de pérdida estándar del problema clasificación binaria: Dado una observación (x, y), si $f(x) \neq y$ entonces $L(x, y, f(x)) = 1$ y $L(x, y, f(x)) = 0$ en caso contrario.
- Cuando el problema es de regresión la forma más común de medir la pérdida es usando el error cuadrático: $L(x, y, f(x)) = (y - f(x))^2$.

Aprendizaje Estadístico: Pérdida

- La forma estándar de evaluar el rendimiento de una función de aprendizaje para el problema de clasificación es usando una función de pérdida, $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$.
- La función de pérdida estándar del problema clasificación binaria: Dado una observación (x, y), si $f(x) \neq y$ entonces $L(x, y, f(x)) = 1$ y $L(x, y, f(x)) = 0$ en caso contrario.
- Cuando el problema es de regresión la forma más común de medir la pérdida es usando el error cuadrático: $L(x, y, f(x)) = (y - f(x))^2$.

Aprendizaje Estadístico: Pérdida

- La forma estándar de evaluar el rendimiento de una función de aprendizaje para el problema de clasificación es usando una función de pérdida, $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$.
- La función de pérdida estándar del problema clasificación binaria: Dado una observación (x, y), si $f(x) \neq y$ entonces $L(x, y, f(x)) = 1$ y $L(x, y, f(x)) = 0$ en caso contrario.
- Cuando el problema es de regresión la forma más común de medir la pérdida es usando el error cuadrático: $L(x, y, f(x)) = (y - f(x))^2$.

Aprendizaje Estadístico: Funciones y máquinas de aprendizaje

- Una máquina o algoritmo de aprendizaje M, es un algoritmo que dada una muestra τ_n nos permite construir una función de aprendizaje $f_n : \mathcal{X} \to \mathcal{Y}$:

$$M : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F} \quad (1)$$

  donde $\mathcal{F}$ es un conjunto de funciones de aprendizaje.
- Obsérvese que la función de aprendizaje f_n depende de una muestra de tamaño n. Para diferentes muestras se obtienen diferentes funciones de aprendizaje.

Aprendizaje Estadístico: Funciones y máquinas de aprendizaje

- Una máquina o algoritmo de aprendizaje M, es un algoritmo que dada una muestra τ_n nos permite construir una función de aprendizaje $f_n : \mathcal{X} \to \mathcal{Y}$:

$$M : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F} \tag{1}$$

  donde $\mathcal{F}$ es un conjunto de funciones de aprendizaje.
- Obsérvese que la función de aprendizaje f_n depende de una muestra de tamaño n. Para diferentes muestras se obtienen diferentes funciones de aprendizaje.

Aprendizaje Estadístico: Ejes de la teoría

- Error de aproximación (sesgo) vrs. error de estimación (varianza).
- Consistencia.
- El problema de minimización de riesgo empírico.
- Capacidad y cotas.

Contenido

Dos Caballitos de Batalla

- Los dos caballitos de batalla que sirven como referencia para gran parte de la teoría son:
  1. Algoritmo del vecino más cercano.
  2. Modelo de regresión lineal.

Dos Caballitos de Batalla: Algoritmo del vecino ms cercano

- Fijemos una noción de distancia entre las variables predictoras.
- Sea k el número de vecinos que la función de aprendizaje utiliza para clasificar.
- Dada una muestra τ_n y un $x \in \mathcal{X}$, calculamos los k puntos $\{x_{i_1}, \dots, x_{i_k}\}$ que tengan menor distancia a x.
- La función de aprendizaje (en el problema de clasificación binario) se define según el número de $\{k : y_{i_k} = 1\}$: voto mayoritario.
- Denotamos esta máquina de aprendizaje por $K - NN_n$.

Dos Caballitos de Batalla: Algoritmo del vecino ms cercano

- Fijemos una noción de distancia entre las variables predictoras.
- Sea k el número de vecinos que la función de aprendizaje utiliza para clasificar.
- Dada una muestra τ_n y un $x \in \mathcal{X}$, calculamos los k puntos $\{x_{i_1}, \ldots, x_{i_k}\}$ que tengan menor distancia a x.
- La función de aprendizaje (en el problema de clasificación binario) se define según el número de $\{k : y_{i_k} = 1\}$: voto mayoritario.
- Denotamos esta máquina de aprendizaje por $K - NN_n$.

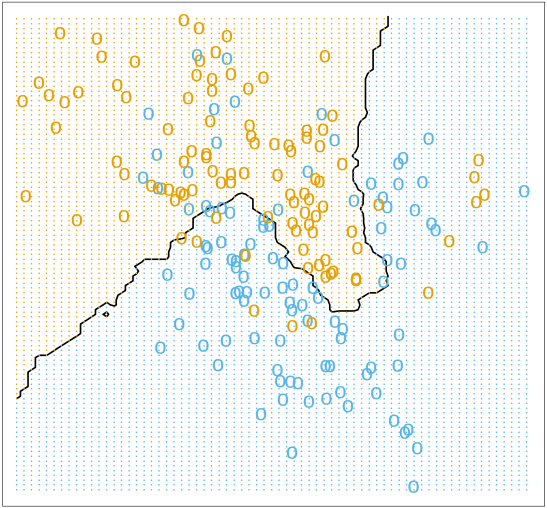
- Fijemos una noción de distancia entre las variables predictoras.
- Sea k el número de vecinos que la función de aprendizaje utiliza para clasificar.
- Dada una muestra τ_n y un $x \in \mathcal{X}$, calculamos los k puntos $\{x_{i_1}, \ldots, x_{i_k}\}$ que tengan menor distancia a x.
- La función de aprendizaje (en el problema de clasificación binario) se define según el número de $\{k : y_{i_k} = 1\}$: voto mayoritario.
- Denotamos esta máquina de aprendizaje por $K - NN_n$.

- Fijemos una noción de distancia entre las variables predictoras.
- Sea k el número de vecinos que la función de aprendizaje utiliza para clasificar.
- Dada una muestra τ_n y un $x \in \mathcal{X}$, calculamos los k puntos $\{x_{i_1}, ..., x_{i_k}\}$ que tengan menor distancia a x.
- La función de aprendizaje (en el problema de clasificación binario) se define según el número de $\{k : y_{i_k} = 1\}$: voto mayoritario.
- Denotamos esta máquina de aprendizaje por $K - NN_n$.

- Fijemos una noción de distancia entre las variables predictoras.
- Sea k el número de vecinos que la función de aprendizaje utiliza para clasificar.
- Dada una muestra τ_n y un $x \in \mathcal{X}$, calculamos los k puntos $\{x_{i_1}, \ldots, x_{i_k}\}$ que tengan menor distancia a x.
- La función de aprendizaje (en el problema de clasificación binario) se define según el número de $\{k : y_{i_k} = 1\}$: voto mayoritario.
- Denotamos esta máquina de aprendizaje por $K - NN_n$.

Dos Caballitos de Batalla: Algoritmo del vecino más cercano

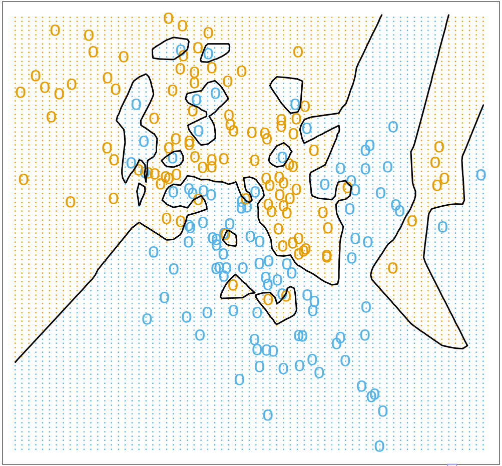
15-Nearest Neighbor Classifier



Dos Caballitos de Batalla: Algoritmo del vecino más cercano

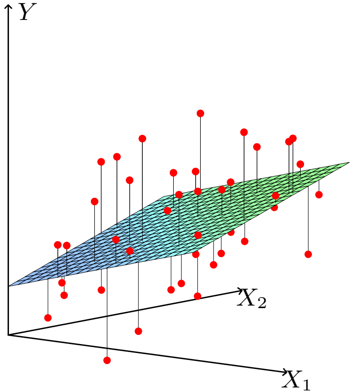
- Obsérvese que este función de aprendizaje ajusta mejor *dentro de muestra* y es más *compleja* que la anterior.

1-Nearest Neighbor Classifier



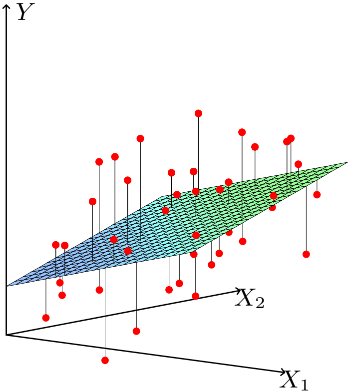
Dos Caballitos de Batalla: Regresión lineal

- Supongamos que $y_i = \beta_n^T x_i$ donde hemos incluido un 1 como primera coordenada en en cada vector x_i con el fin de incluir una constante en el modelo lineal.
- Sea $\hat{\beta}_n$ el estimador de mínimos cuadrados ordinarios.
- Obsérvese que $\hat{\beta}_n$ define una función de aprendizaje $f_n^{OLS}(x) = 1$ si $\beta_n^T x > 0{,}5$ y cero caso contrario.



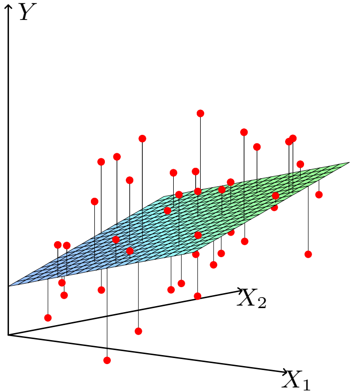
Dos Caballitos de Batalla: Regresión lineal

- Supongamos que $y_i = \beta_n^T x_i$ donde hemos incluido un 1 como primera coordenada en en cada vector x_i con el fin de incluir una constante en el modelo lineal.
- Sea $\hat{\beta}_n$ el estimador de mínimos cuadrados ordinarios.
- Obsérvese que $\hat{\beta}_n$ define una función de aprendizaje $f_n^{OLS}(x) = 1$ si $\beta_n^T x > 0{,}5$ y cero caso contrario.

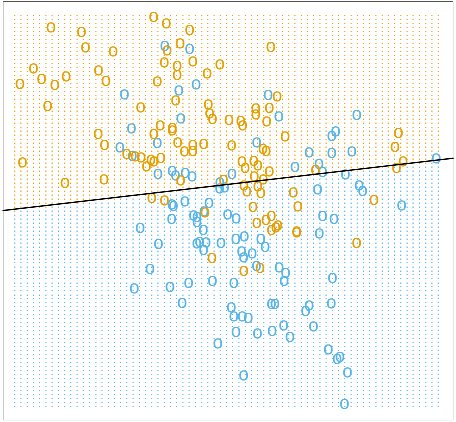


Dos Caballitos de Batalla: Regresión lineal

- Supongamos que $y_i = \beta_n^T x_i$ donde hemos incluido un 1 como primera coordenada en en cada vector x_i con el fin de incluir una constante en el modelo lineal.
- Sea $\hat{\beta}_n$ el estimador de mínimos cuadrados ordinarios.
- Obsérvese que $\hat{\beta}_n$ define una función de aprendizaje $f_n^{OLS}(x) = 1$ si $\beta_n^T x > 0{,}5$ y cero caso contrario.



- La línea corresponde a $\beta_n^T x = 0{,}5$.



Contenido

Aprendizaje Estadístico: Riesgo

- Sea f una función de aprendizaje. Definimos el riesgo de f como:

$$R(f) = E[L(X, Y, f(X))] \tag{2}$$

- El riesgo no se puede calcular porque no conocemos la distribución P.

Aprendizaje Estadístico: riesgo empírico y riesgo

- Dos conceptos claves de la teoría son:
  1. El error de entrenamiento o riesgo empírico es el estimador muestral del riesgo y se denota por $R_{emp}(f)$:
  $$R_{emp}(f) = \sum_{i=1}^{n} \frac{L(X_i, Y_i, f(X_i))}{n} \tag{3}$$
  2. El error de prueba (generalización o predicción) es el riesgo de la función de aprendizaje:
  $$R(f) = E[L(X, Y, f(X))] \tag{4}$$
  donde el valor esperado se toma con respecto a la distribución P.
- Obsérverse que en ambas definiciones podríamos reemplazar $f(x)$ por $f_n(x)$ para indicar que las funciones de aprendizaje dependen de la muestra. En cualquier caso, la muestra τ_n se mantiene fija.

Aprendizaje Estadístico: riesgo empírico y riesgo

- Dos conceptos claves de la teoría son:
  1. El error de entrenamiento o riesgo empírico es el estimador muestral del riesgo y se denota por $R_{emp}(f)$:
  $$R_{emp}(f) = \sum_{i=1}^{n} \frac{L(X_i, Y_i, f(X_i))}{n} \tag{3}$$
  2. El error de prueba (generalización o predicción) es el riesgo de la función de aprendizaje:
  $$R(f) = E[L(X, Y, f(X))] \tag{4}$$
  donde el valor esperado se toma con respecto a la distribución P.
- Obsérverse que en ambas definiciones podríamos reemplazar $f(x)$ por $f_n(x)$ para indicar que las funciones de aprendizaje dependen de la muestra. En cualquier caso, la muestra τ_n se mantiene fija.

Aprendizaje Estadístico: riesgo empírico y riesgo

- Dos conceptos claves de la teoría son:
  1. El error de entrenamiento o riesgo empírico es el estimador muestral del riesgo y se denota por $R_{emp}(f)$:
  $$R_{emp}(f) = \sum_{i=1}^{n} \frac{L(X_i, Y_i, f(X_i))}{n} \tag{3}$$
  2. El error de prueba (generalización o predicción) es el riesgo de la función de aprendizaje:
  $$R(f) = E[L(X, Y, f(X))] \tag{4}$$
  donde el valor esperado se toma con respecto a la distribución P.
- Obsérverse que en ambas definiciones podríamos reemplazar $f(x)$ por $f_n(x)$ para indicar que las funciones de aprendizaje dependen de la muestra. En cualquier caso, la muestra τ_n se mantiene fija.

Aprendizaje Estadístico: riesgo empírico y riesgo

- Dos conceptos claves de la teoría son:
  1. El error de entrenamiento o riesgo empírico es el estimador muestral del riesgo y se denota por $R_{emp}(f)$:

     $$R_{emp}(f) = \sum_{i=1}^{n} \frac{L(X_i, Y_i, f(X_i))}{n} \qquad (3)$$

  2. El error de prueba (generalización o predicción) es el riesgo de la función de aprendizaje:

     $$R(f) = E[L(X, Y, f(X))] \qquad (4)$$

     donde el valor esperado se toma con respecto a la distribución P.
- Obsérverse que en ambas definiciones podríamos reemplazar $f(x)$ por $f_n(x)$ para indicar que las funciones de aprendizaje dependen de la muestra. En cualquier caso, la muestra τ_n se mantiene fija.

- Decimos que una función de aprendizaje generaliza bien si

$$|R_{emp}(f) - R(f)| \tag{5}$$

  es pequeño.
- Obsérvese que la diferencia anterior es una variable aleatoria aún si la función de apredizaje no depende de la muestra ($R_{emp}(f)$ depende de la muestra).
- El error empírico puede ser muy mal indicador de qué también generaliza una función de aprendizaje.
- El problema es que el riesgo empírico se puede controlar con la complejidad (*overfitting*) de la función de aprendizaje pero nada garantiza que generalice bien.

- Decimos que una función de aprendizaje generaliza bien si

$$|R_{emp}(f) - R(f)| \tag{5}$$

  es pequeño.
- Obsérvese que la diferencia anterior es una variable aleatoria aún si la función de apredizaje no depende de la muestra ($R_{emp}(f)$ depende de la muestra).
- El error empírico puede ser muy mal indicador de qué también generaliza una función de aprendizaje.
- El problema es que el riesgo empírico se puede controlar con la complejidad (*overfitting*) de la función de aprendizaje pero nada garantiza que generalice bien.

- Decimos que una función de aprendizaje generaliza bien si

$$|R_{emp}(f) - R(f)| \tag{5}$$

  es pequeño.
- Obsérvese que la diferencia anterior es una variable aleatoria aún si la función de apredizaje no depende de la muestra ($R_{emp}(f)$ depende de la muestra).
- El error empírico puede ser muy mal indicador de qué también generaliza una función de aprendizaje.
- El problema es que el riesgo empírico se puede controlar con la complejidad (*overfitting*) de la función de aprendizaje pero nada garantiza que generalice bien.

Aprendizaje Estadístico: El Problema de Generalización

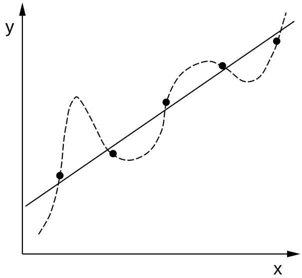
- Decimos que una función de aprendizaje generaliza bien si

$$|R_{emp}(f) - R(f)| \tag{5}$$

  es pequeño.
- Obsérvese que la diferencia anterior es una variable aleatoria aún si la función de apredizaje no depende de la muestra ($R_{emp}(f)$ depende de la muestra).
- El error empírico puede ser muy mal indicador de qué también generaliza una función de aprendizaje.
- El problema es que el riesgo empírico se puede controlar con la complejidad (*overfitting*) de la función de aprendizaje pero nada garantiza que generalice bien.

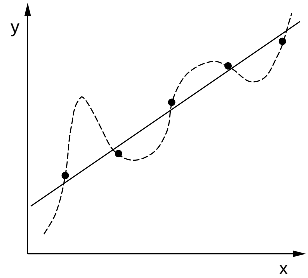
Aprendizaje Estadístico: El Problema de Generalización

- El problema de generalizacion tiene origen en el compromiso entre el error de aproximación (sesgo) y el error de estimación (varianza).
- Obsérvese que si el verdadero modelo es la línea recta, el error empirico de la curva es cero (*overfitting*) pero generaliza muy mal. El error empírico de la línea es mayor que cero pero generaliza mejor.
- La curva es muy compleja, la varianza es alta pero el sesgo es bajo. La línea es poco compleja, varianza baja pero sesgo alto.

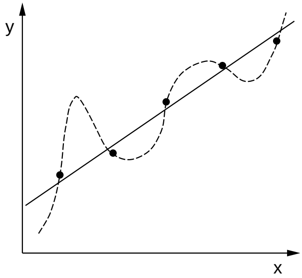


Aprendizaje Estadístico: El Problema de Generalización

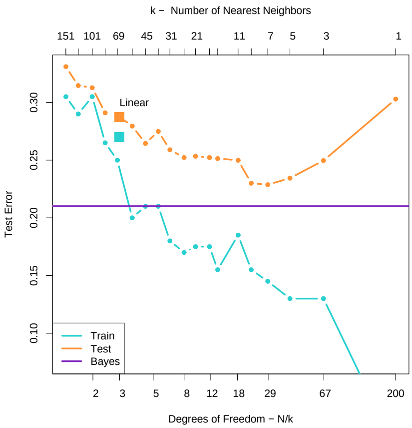
- El problema de generalizacion tiene origen en el compromiso entre el error de aproximación (sesgo) y el error de estimación (varianza).
- Obsérvese que si el verdadero modelo es la línea recta, el error empirico de la curva es cero (*overfitting*) pero generaliza muy mal. El error empírico de la línea es mayor que cero pero generaliza mejor.
- La curva es muy compleja, la varianza es alta pero el sesgo es bajo. La línea es poco compleja, varianza baja pero sesgo alto.



- El problema de generalizacion tiene origen en el compromiso entre el error de aproximación (sesgo) y el error de estimación (varianza).
- Obsérvese que si el verdadero modelo es la línea recta, el error empirico de la curva es cero (*overfitting*) pero generaliza muy mal. El error empírico de la línea es mayor que cero pero generaliza mejor.
- La curva es muy compleja, la varianza es alta pero el sesgo es bajo. La línea es poco compleja, varianza baja pero sesgo alto.



Error de entrenamiento vs error de prueba



- Ejercicio de simulación: Muestra 200 observaciones, prueba 10.000 observaciones.

Contenido

Aprendizaje Estadístico: Formalización

- Sea $\mathcal{F}_0$ un conjunto de funciones y $\mathcal{F}$ el conjunto de todas las funciones.
- El problema de clasificación consiste en:

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{6}$$

  obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones. Cuando queremos hacer explícito el espacio $\mathcal{F}_0$, denotamos f^* por $f_{\mathcal{F}_0}$.
- Cuando $\mathcal{F}_0 = \mathcal{F}$ entonces f^* se llama el clasficador de Bayes y se denota por f_{Bayes}.

Aprendizaje Estadístico: Formalización

- Sea $\mathcal{F}_0$ un conjunto de funciones y $\mathcal{F}$ el conjunto de todas las funciones.
- El problema de clasificación consiste en:

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{6}$$

obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones. Cuando queremos hacer explícito el espacio $\mathcal{F}_0$, denotamos f^* por $f_{\mathcal{F}_0}$.

- Cuando $\mathcal{F}_0 = \mathcal{F}$ entonces f^* se llama el clasficador de Bayes y se denota por f_{Bayes}.

Aprendizaje Estadístico: Formalización

- Sea $\mathcal{F}_0$ un conjunto de funciones y $\mathcal{F}$ el conjunto de todas las funciones.
- El problema de clasificación consiste en:

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{6}$$

  obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones. Cuando queremos hacer explícito el espacio $\mathcal{F}_0$, denotamos f^* por $f_{\mathcal{F}_0}$.
- Cuando $\mathcal{F}_0 = \mathcal{F}$ entonces f^* se llama el clasficador de Bayes y se denota por f_{Bayes}.

Aprendizaje Estadístico: Resumen

- Podemos resumir el objetivo principal del apredizaje de máquinas como: dada una muestra finita τ_n y una función de pérdida, encontrar una espacio de funciones $\mathcal{F}_0$ y un clasificador óptimo $f_{\mathcal{F}_0}$ tal que su riesgo sea lo más cercano posible al clasificador de Bayes.
- El riesgo empírico de un clasificador no es necesarimente un buen estimador del riesgo.

Aprendizaje Estadístico: Resumen

- Podemos resumir el objetivo principal del apredizaje de máquinas como: dada una muestra finita τ_n y una función de pérdida, encontrar una espacio de funciones $\mathcal{F}_0$ y un clasificador óptimo $f_{\mathcal{F}_0}$ tal que su riesgo sea lo más cercano posible al clasificador de Bayes.
- El riesgo empírico de un clasificador no es necesarimente un buen estimador del riesgo.

Contenido

Error de Aproximación y Estimación

- Dado un algoritmo o máquina de aprendizaje M y un espacio de funciones de aprendizaje $\mathcal{F}_0$, definimos el error de M como la variable aleatoria:

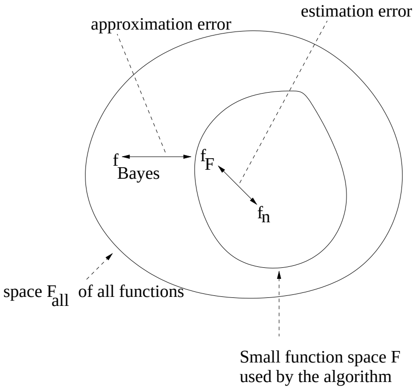
$$R(f_n) - R(f_{Bayes}) \tag{7}$$

que se puede reescribir como:

$$R(f_n) - R(f_{Bayes}) = R(f_{\mathbb{F}_0}) - R(f_{Bayes}) + R(f_n) - R(f_{\mathbb{F}_0}) \tag{8}$$

- El primer término se conoce como error de aproximación (sesgo) y el segundo se conoce como error de estimación (varianza).

Error de Aproximación y Estimación



Sesgo y varianza

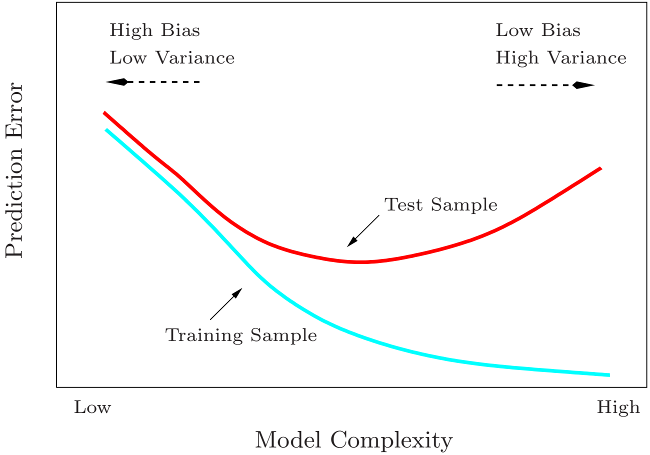
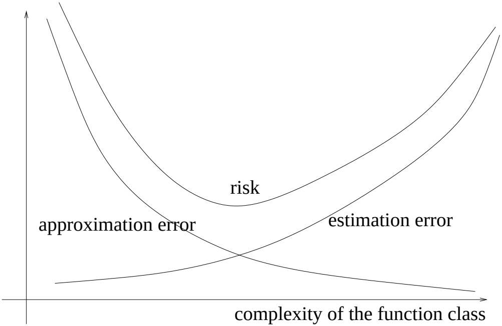
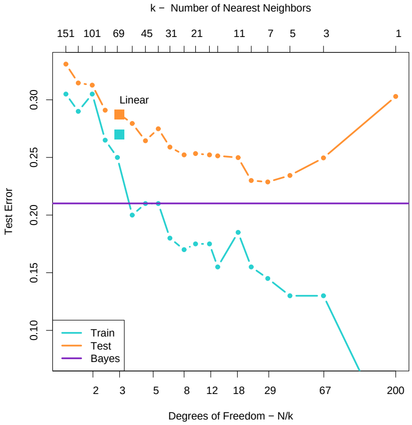


FIGURE 2.11. *Test and training error as a function of model complexity.*

Sesgo y varianza



Sesgo y varianza



- Ejercicio de simulación: Muestra 200 observaciones, prueba 10.000 observaciones.