Aprendizaje Estadístico

Riesgo y el Problema de Error de Aproximación y Estimación

Alvaro J. Riascos Villegas

Junio de 2020

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 $\{(x_1, y_1), ..., (x_n, y_n)\}$ 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.

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.

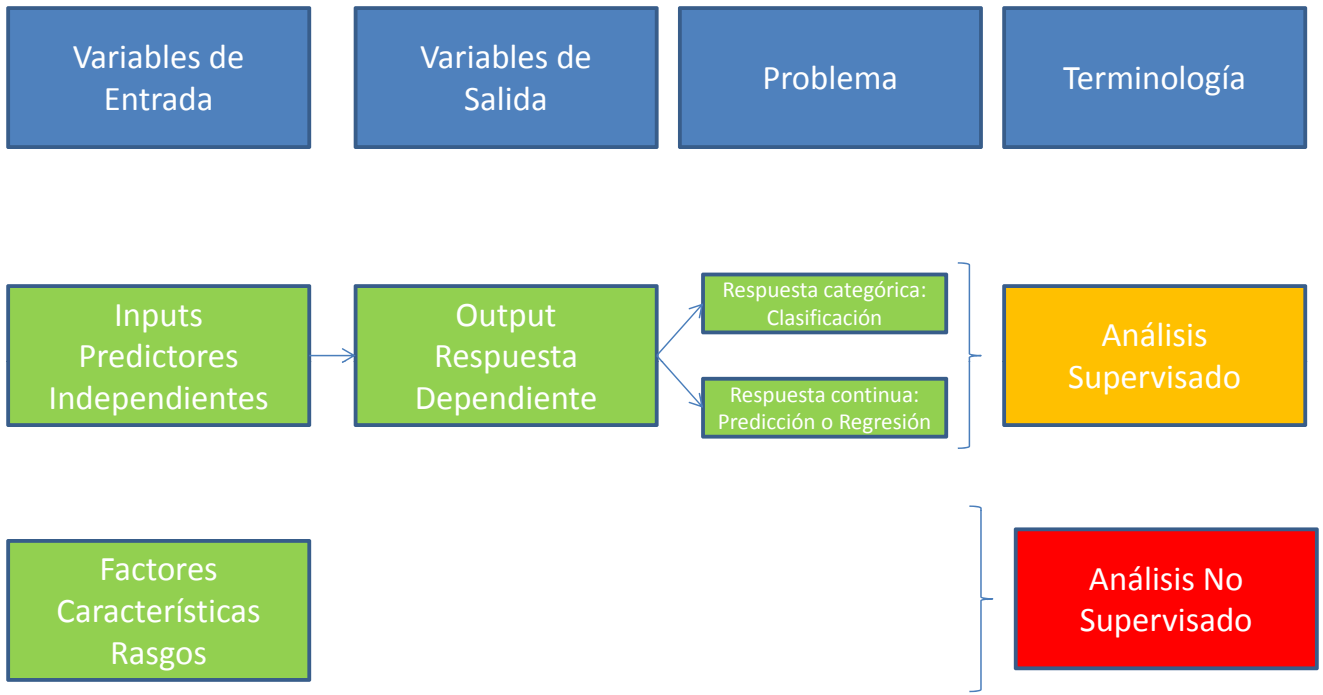
Aprendizaje Estadístico

- 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. Aprendizaje reforzado: Se interactura con un ambiente (i.e., físico, juego, etc) y se tiene un conjunto de acciones. El propósito es maximizar una función de beneficios del agente tomando las mejores acciones.

- Otros problemas son:
  1. Aprendizaje reforzado: Se interactura con un ambiente (i.e., físico, juego, etc) y se tiene un conjunto de acciones. El propósito es maximizar una función de beneficios del agente tomando las mejores acciones.

- 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 $f(x)$.

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 $f(x)$.

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 $f(x)$.

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 $f(x)$.

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 $f(x)$.

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 $f(x)$.

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 $f_n : \mathcal{X} \to \mathcal{Y}$:

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

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: 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 $\{x_{i_1}, ..., x_{i_k}\}$ 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 $\{x_{i_1}, ..., x_{i_k}\}$ 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 $\{x_{i_1}, ..., x_{i_k}\}$ 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 $K - NN_n$.

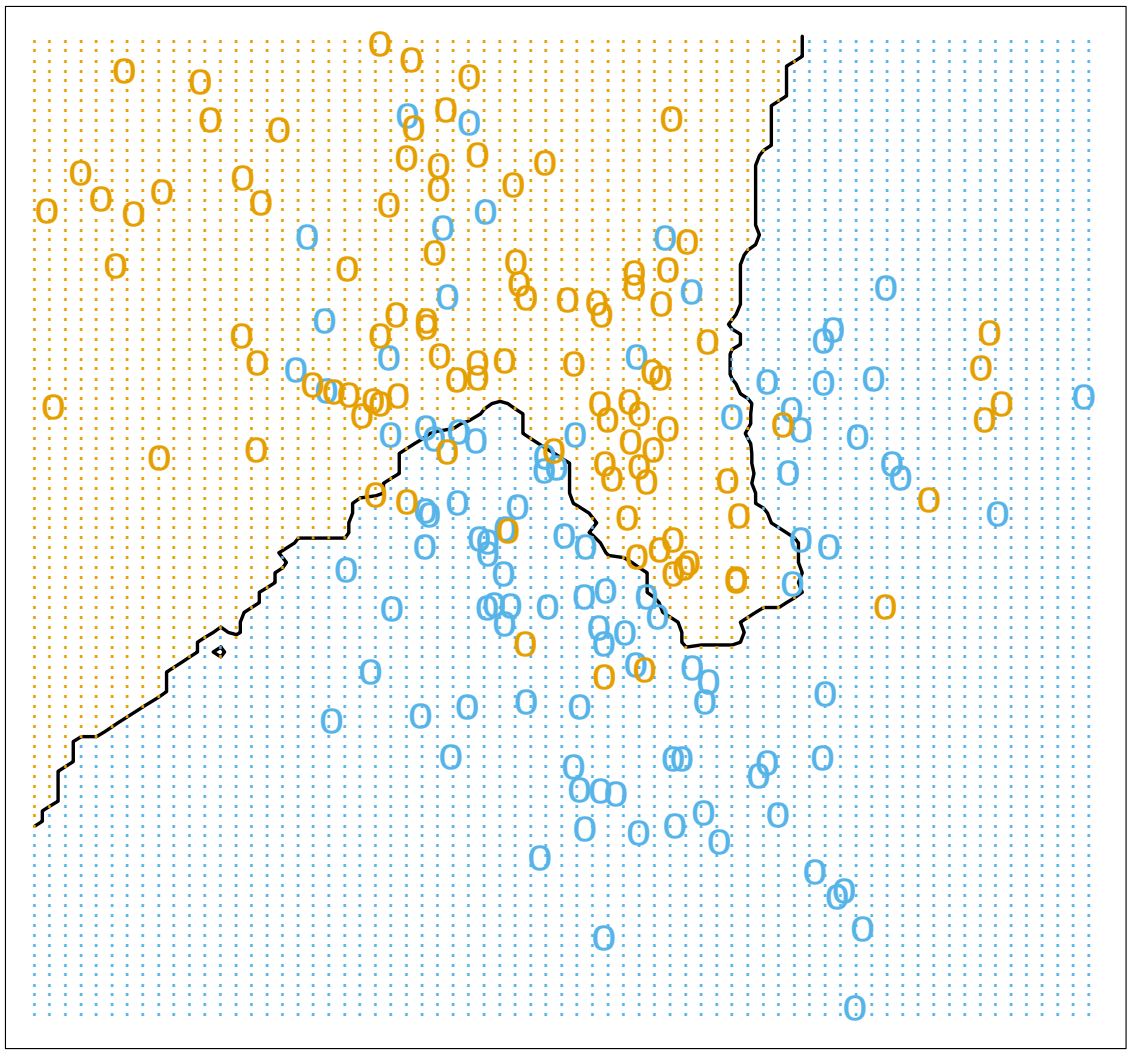
Dos Caballitos de Batalla: Algoritmo del vecino ms cercano

- 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 $\{x_{i_1}, ..., x_{i_k}\}$ 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

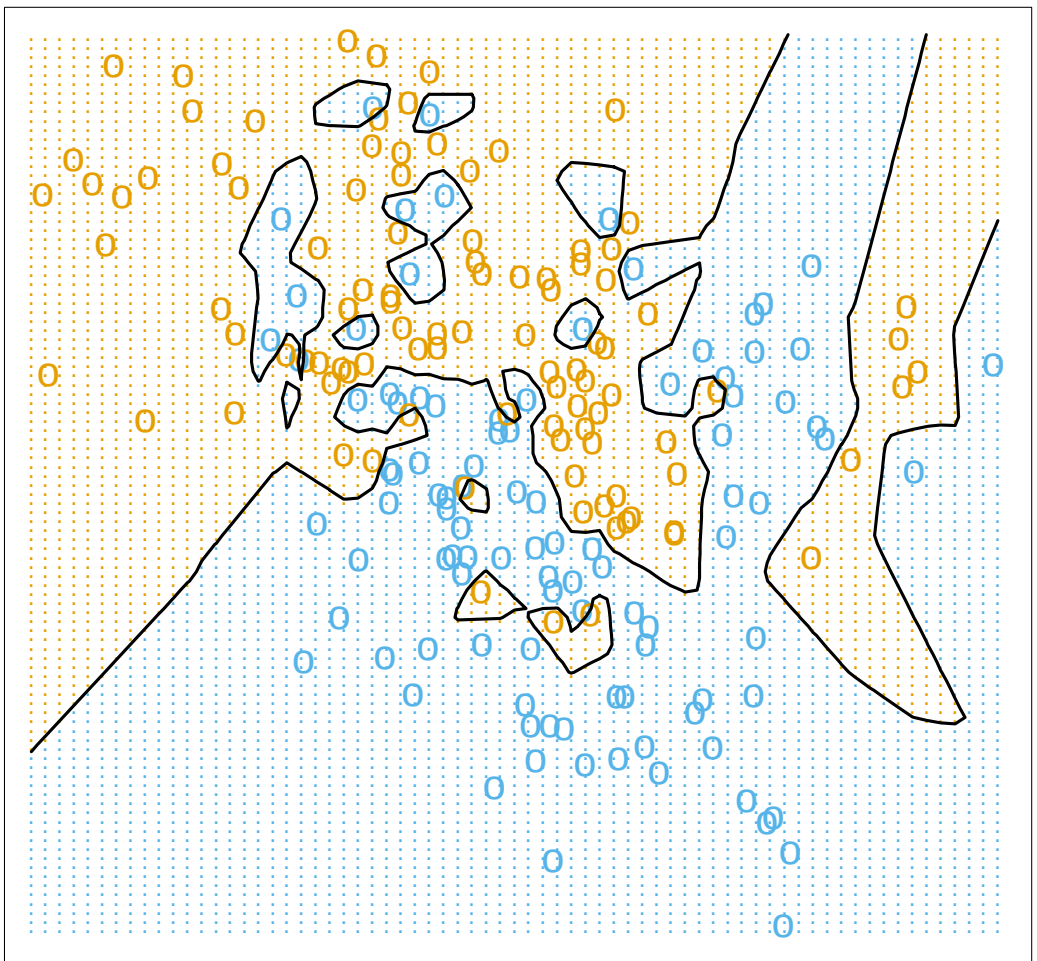
- Fijemos una noción de distancia entre las variables predictoras.
- Sea k el número de vecinos que la función de aprendizaje utiliza para clasificar.
- Dada una muestra τ_n y un $x \in \mathcal{X}$, calculamos los k puntos $\{x_{i_1}, ..., x_{i_k}\}$ que tengan menor distancia a x.
- La función de aprendizaje (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



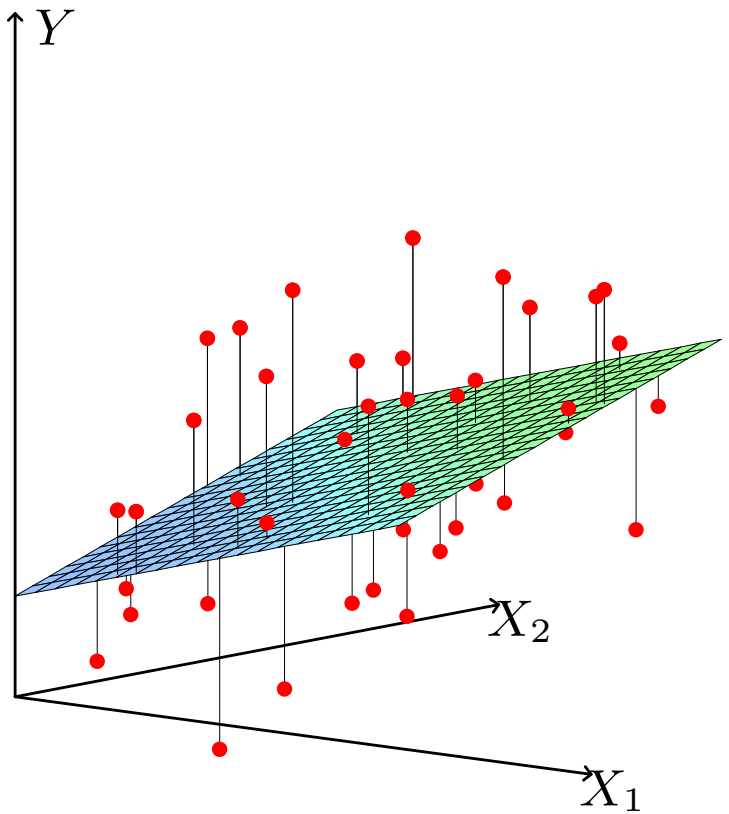
Dos Caballitos de Batalla: Algoritmo del vecino más cercano

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



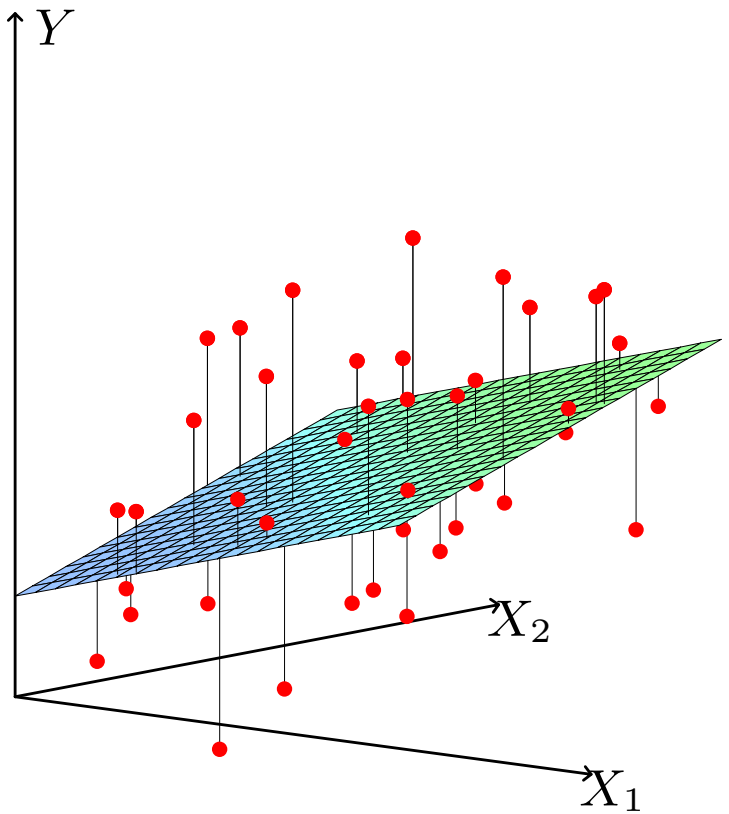
Dos Caballitos de Batalla: Regresión lineal

- 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.



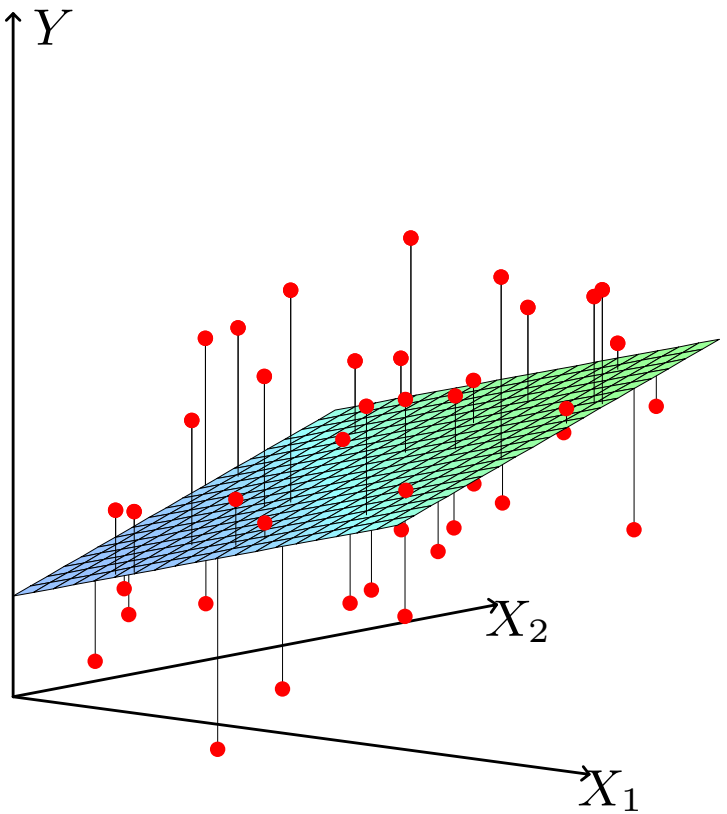
Dos Caballitos de Batalla: Regresión lineal

- 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.



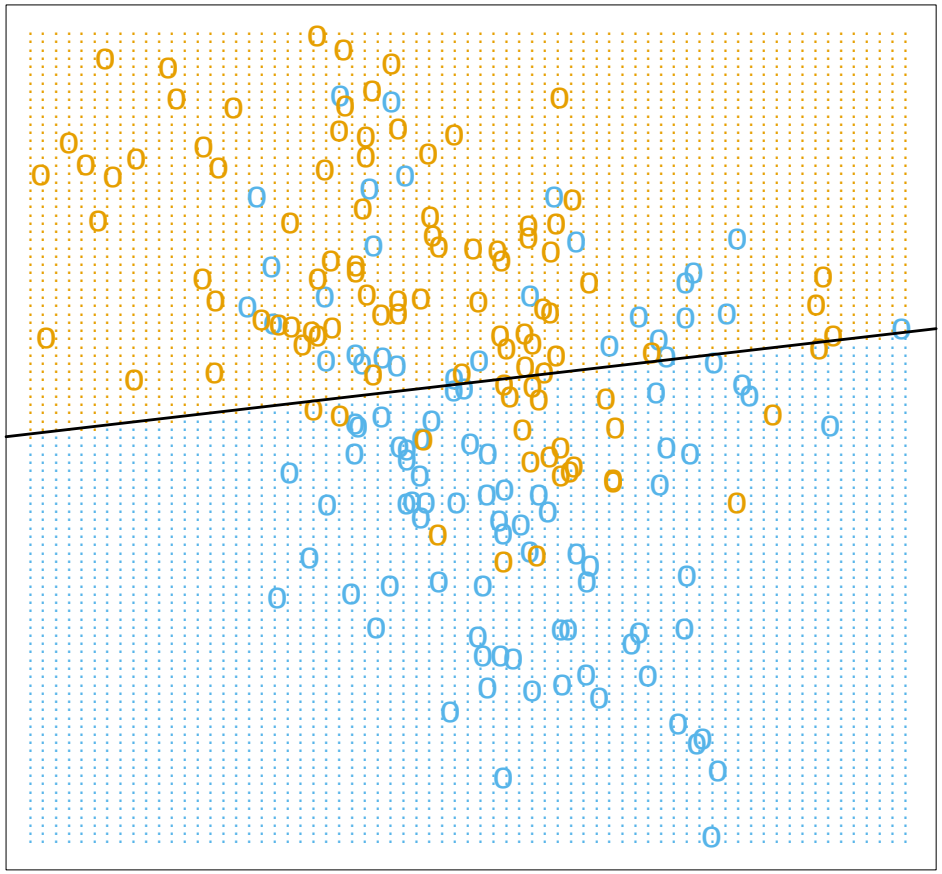
Dos Caballitos de Batalla: Regresión lineal

- 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.



Dos Caballitos de Batalla: Regresión lineal

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



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 $R_{emp}(f)$:

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

  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 esperado

- El error de prueba (generalización o predicción) esperado o el riesgo esperado es:

$$E[R(f_n)] \tag{5}$$

  donde el valor esperado se toma con respecto a todas las muestras τ_n (obsérvese que P define una distribución de probabilidad sobre $(\mathcal{X} \times \mathcal{Y})^n$).

- Esta definición aplica solo cuando la función de aprendizaje depende de la muestra.

Aprendizaje Estadístico: riesgo esperado

- El error de prueba (generalización o predicción) esperado o el riesgo esperado es:

$$E[R(f_n)] \tag{5}$$

  donde el valor esperado se toma con respecto a todas las muestras τ_n (obsérvese que P define una distribución de probabilidad sobre $(\mathcal{X} \times \mathcal{Y})^n$).
- Esta definición aplica solo cuando la función de aprendizaje depende de la muestra.

Aprendizaje Estadístico: El Problema de Generalización

- Decimos que una función de aprendizaje generaliza bien si

$$|R_{emp}(f) - R(f)| \tag{6}$$

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

$$|R_{emp}(f) - R(f)| \tag{6}$$

  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

$$|R_{emp}(f) - R(f)| \tag{6}$$

  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

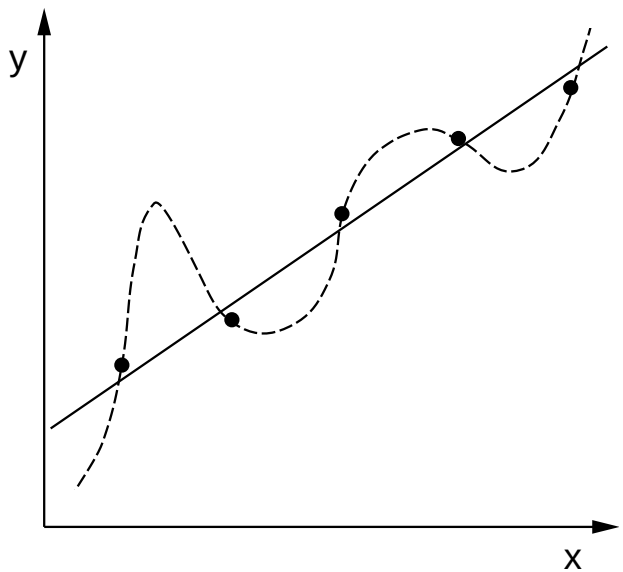
- Decimos que una función de aprendizaje generaliza bien si

$$|R_{emp}(f) - R(f)| \tag{6}$$

  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.

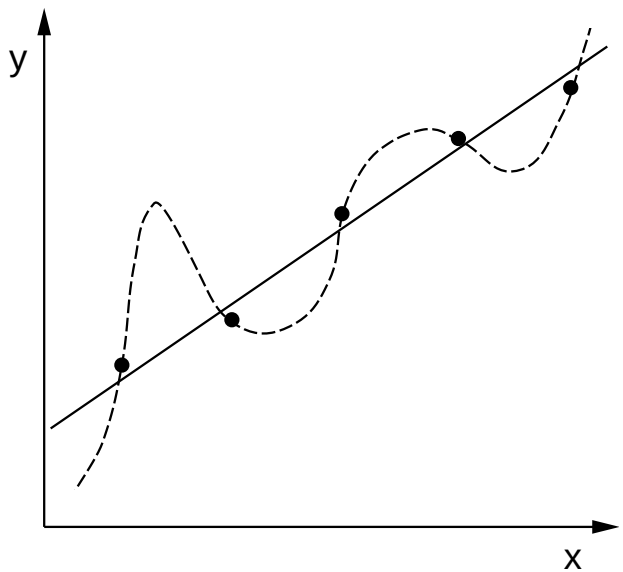
Aprendizaje Estadístico: El Problema de Generalización

- El problema de (*overfitting*).
- Obsérvese que si el verdadero modelo es la línea recta, el error empirico de la curva es cero 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.



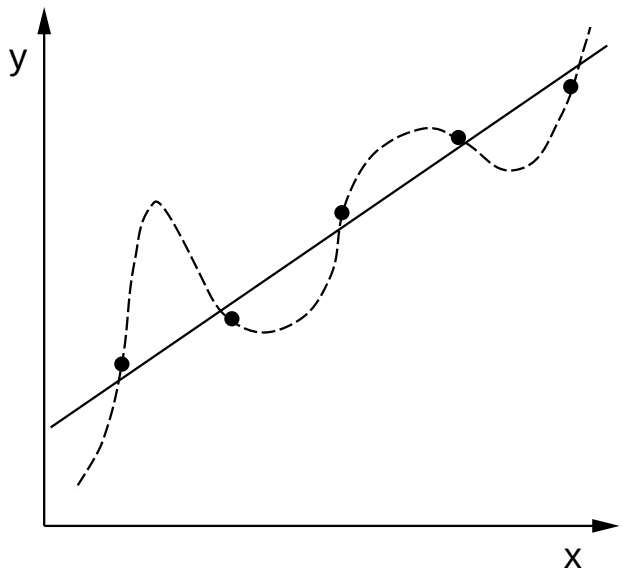
Aprendizaje Estadístico: El Problema de Generalización

- El problema de (*overfitting*).
- Obsérvese que si el verdadero modelo es la línea recta, el error empirico de la curva es cero 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 (*overfitting*).
- Obsérvese que si el verdadero modelo es la línea recta, el error empirico de la curva es cero 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.



Contenido

Aprendizaje Estadístico: El problema de clasificación formalmente

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

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{7}$$

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: El problema de clasificación formalmente

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

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{7}$$

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: El problema de clasificación formalmente

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

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{7}$$

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.

Funciones de aprendizaje óptimas: Variables categóricas

- Supongamos que G es una variable categórica que toma valores un número finito K de valores. Denotamos estos valores por g de tal forma que $g = 1, ..., K$.
- Definamos una función de pérdida de ceros y unos L donde $L(i, j)$ es cero si la observación de la clase i es clasificada erroneamente en la clase $j \neq i$.
- Sea $\widehat{G}$ una función de aprendizaje para clasificar nuevos datos.
- Entonces el riesgo de $\widehat{G}$ es:

$$E[L(G, \hat{G}(X))] = E_X \sum_{k=1}^{K} L(k, \widehat{G}(X)) P(G = k \mid X)$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables categóricas

- Supongamos que G es una variable categórica que toma valores un número finito K de valores. Denotamos estos valores por g de tal forma que $g = 1, ..., K$.
- Definamos una función de pérdida de ceros y unos L donde $L(i, j)$ es cero si la observación de la clase i es clasificada erroneamente en la clase $j \neq i$.
- Sea $\widehat{G}$ una función de aprendizaje para clasificar nuevos datos.
- Entonces el riesgo de $\widehat{G}$ es:

$$E[L(G, \hat{G}(X))] = E_X \sum_{k=1}^{K} L(k, \widehat{G}(X)) P(G = k \mid X)$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables categóricas

- Supongamos que G es una variable categórica que toma valores un número finito K de valores. Denotamos estos valores por g de tal forma que $g = 1, ..., K$.
- Definamos una función de pérdida de ceros y unos L donde $L(i, j)$ es cero si la observación de la clase i es clasificada erroneamente en la clase $j \neq i$.
- Sea $\widehat{G}$ una función de aprendizaje para clasificar nuevos datos.
- Entonces el riesgo de $\widehat{G}$ es:

$$E[L(G, \hat{G}(X))] = E_X \sum_{k=1}^{K} L(k, \widehat{G}(X)) P(G = k \mid X)$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables categóricas

- Supongamos que G es una variable categórica que toma valores un número finito K de valores. Denotamos estos valores por g de tal forma que $g = 1, ..., K$.
- Definamos una función de pérdida de ceros y unos L donde $L(i, j)$ es cero si la observación de la clase i es clasificada erroneamente en la clase $j \neq i$.
- Sea $\widehat{G}$ una función de aprendizaje para clasificar nuevos datos.
- Entonces el riesgo de $\widehat{G}$ es:

$$E[L(G, \hat{G}(X))] = E_X \sum_{k=1}^{K} L(k, \widehat{G}(X)) P(G = k \mid X)$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables categóricas

- Como ilustración veamos de donde sale la anterior fórmula cuando la variable independiente X también es categórica: $X \in \{x_1, ..., x_N\}$.
- En este caso el riesgo de $\widehat{G}$ es:

$$E[L(G, \hat{G}(X))] = \tag{8}$$

$$\sum_i \sum_k L(k, \widehat{G}(x_i)) P(X = x_i, \widehat{G} = k) \tag{9}$$

$$\sum_i \sum_k L(k, \widehat{G}(x_i)) P(\widehat{G} = k \mid X = x_i) P(X = x_i) \tag{10}$$

$$\sum_i P(X = x_i) \sum_k L(k, \widehat{G}(x_i) P(\widehat{G} = k \mid X = x_i) \tag{11}$$

$$E_X \sum_k L(k, \widehat{G}(X) P(\widehat{G} = k \mid X) \tag{12}$$

Funciones de aprendizaje óptimas: Variables categóricas

- Como ilustración veamos de donde sale la anterior fórmula cuando la variable independiente X también es categórica: $X \in \{x_1, ..., x_N\}$.
- En este caso el riesgo de $\widehat{G}$ es:

$$E[L(G, \hat{G}(X))] = \tag{8}$$

$$\sum_i \sum_k L(k, \widehat{G}(x_i)) P(X = x_i, \widehat{G} = k) \tag{9}$$

$$\sum_i \sum_k L(k, \widehat{G}(x_i)) P(\widehat{G} = k \mid X = x_i) P(X = x_i) \tag{10}$$

$$\sum_i P(X = x_i) \sum_k L(k, \widehat{G}(x_i) P(\widehat{G} = k \mid X = x_i) \tag{11}$$

$$E_X \sum_k L(k, \widehat{G}(X) P(\widehat{G} = k \mid X) \tag{12}$$

Aprendizaje Estadístico: El problema de clasificación formalmente

- Ahora minimizamos esta función sobre el espacio de las funciones de aprendizaje $\widehat{G}$ y obsérvese que basta con minimizar puntualmente. Dado X resolver:

$$\min_g \sum_{k=1}^{K} L(k,g) P(G = k \mid X) \tag{13}$$

$$= \min_g P(G \neq g \mid X) \tag{14}$$

$$= \min_g (1 - P(G = g \mid X)) \tag{15}$$

- Luego es suficiente resolver este problema:

$$= \max_g P(G = g \mid X) \tag{16}$$

Aprendizaje Estadístico: El problema de clasificación formalmente

- Ahora minimizamos esta función sobre el espacio de las funciones de aprendizaje $\widehat{G}$ y obsérvese que basta con minimizar puntualmente. Dado X resolver:

$$\min_g \sum_{k=1}^{K} L(k,g) P(G = k \mid X) \tag{13}$$

$$= \min_g P(G \neq g \mid X) \tag{14}$$

$$= \min_g (1 - P(G = g \mid X)) \tag{15}$$

- Luego es suficiente resolver este problema:

$$= \max_g P(G = g \mid X) \tag{16}$$

Aprendizaje Estadístico: El problema de clasificación formalmente

- Ahora minimizamos esta función sobre el espacio de las funciones de aprendizaje $\widehat{G}$ y obsérvese que basta con minimizar puntualmente. Dado X resolver:

$$\min_g \sum_{k=1}^{K} L(k,g) P(G = k \mid X) \tag{13}$$

$$= \min_g P(G \neq g \mid X) \tag{14}$$

$$= \min_g (1 - P(G = g \mid X)) \tag{15}$$

- Luego es suficiente resolver este problema:

$$= \max_g P(G = g \mid X) \tag{16}$$

Aprendizaje Estadístico: El problema de clasificación formalmente

- Ahora minimizamos esta función sobre el espacio de las funciones de aprendizaje $\widehat{G}$ y obsérvese que basta con minimizar puntualmente. Dado X resolver:

$$\min_g \sum_{k=1}^{K} L(k,g) P(G = k \mid X) \tag{13}$$

$$= \min_g P(G \neq g \mid X) \tag{14}$$

$$= \min_g (1 - P(G = g \mid X)) \tag{15}$$

- Luego es suficiente resolver este problema:

$$= \max_g P(G = g \mid X) \tag{16}$$

Aprendizaje Estadístico: El problema de clasificación formalmente

- En conclusión el clasificador de Bayes es:

$$\widehat{G}_{Bayes}(X) = arg\ max_g P(G = g \mid X) \tag{17}$$

Aprendizaje Estadístico: El problema de clasificación formalmente

Example (Clasificador de Bayes binario con pérdida estándar)

En este caso el mejor clasificador es:

$$f_{\text{Bayes}}(x) = 1 \tag{18}$$

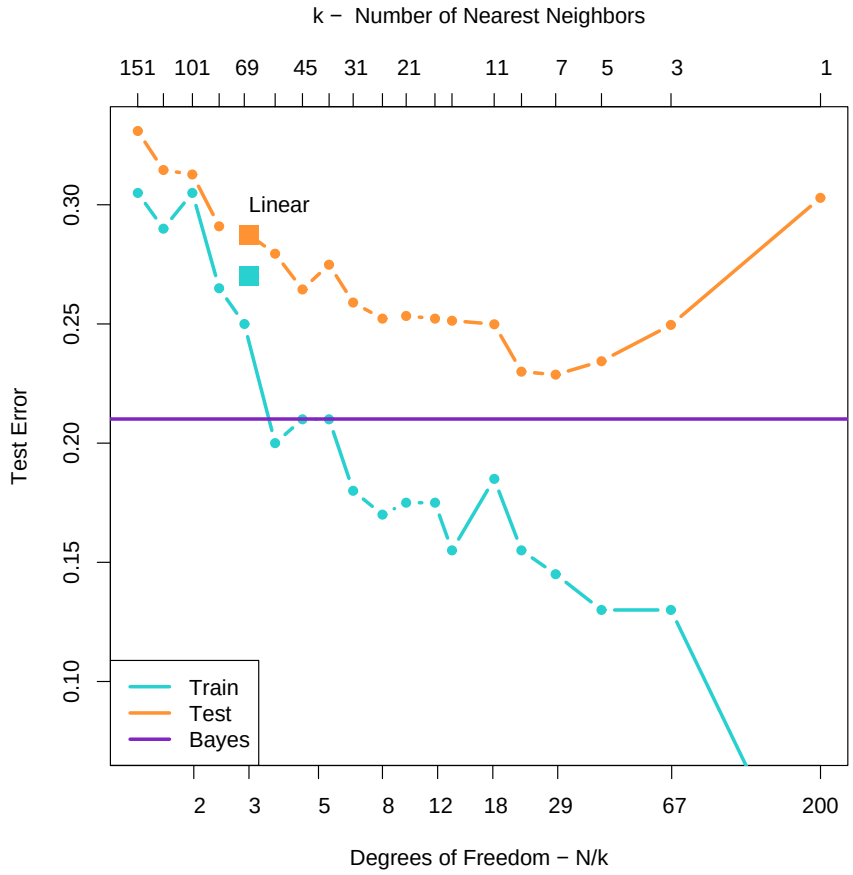
si $P(Y = 1 \,|X = x) \geq 0{,}5$ y cero caso contrario.

Aprendizaje Estadístico: El problema de clasificación formalmente

Example

Supongamos que $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$ y P es tal que X_i se distribuye normal $(0{,}5, 1)$ (truncada) en $[0,1]$ (no importa la distribución) y $P[Y = 1 \,|X = x] = 0{,}9$. En este caso el clasificador de Bayes es $f_{\text{Bayes}}(x) = 1$ para todo $x \in \Xi$. Por lo tanto $R(f_{\text{Bayes}}) = 0{,}1$

Aprendizaje Estadístico: Error de entrenamiento vs error de prueba



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

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:

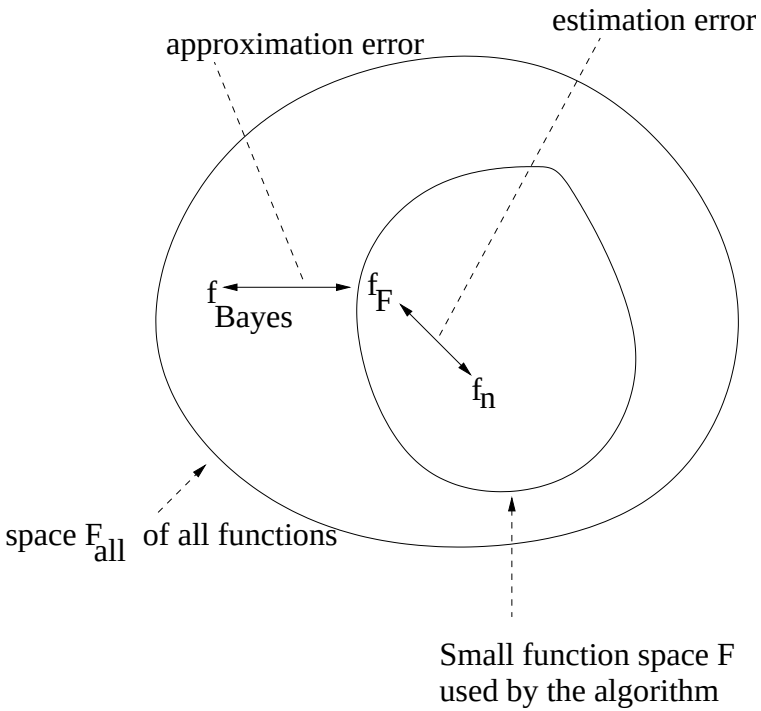
$$R(f_n) - R(f_{Bayes}) \tag{19}$$

que se puede reescribir como:

$$R(f_n) - R(f_{Bayes}) = R(f_{\mathbb{F}_0}) - R(f_{Bayes}) + R(f_n) - R(f_{\mathbb{F}_0}) \tag{20}$$

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

Error de estimación y aproximación



Sesgo y varianza

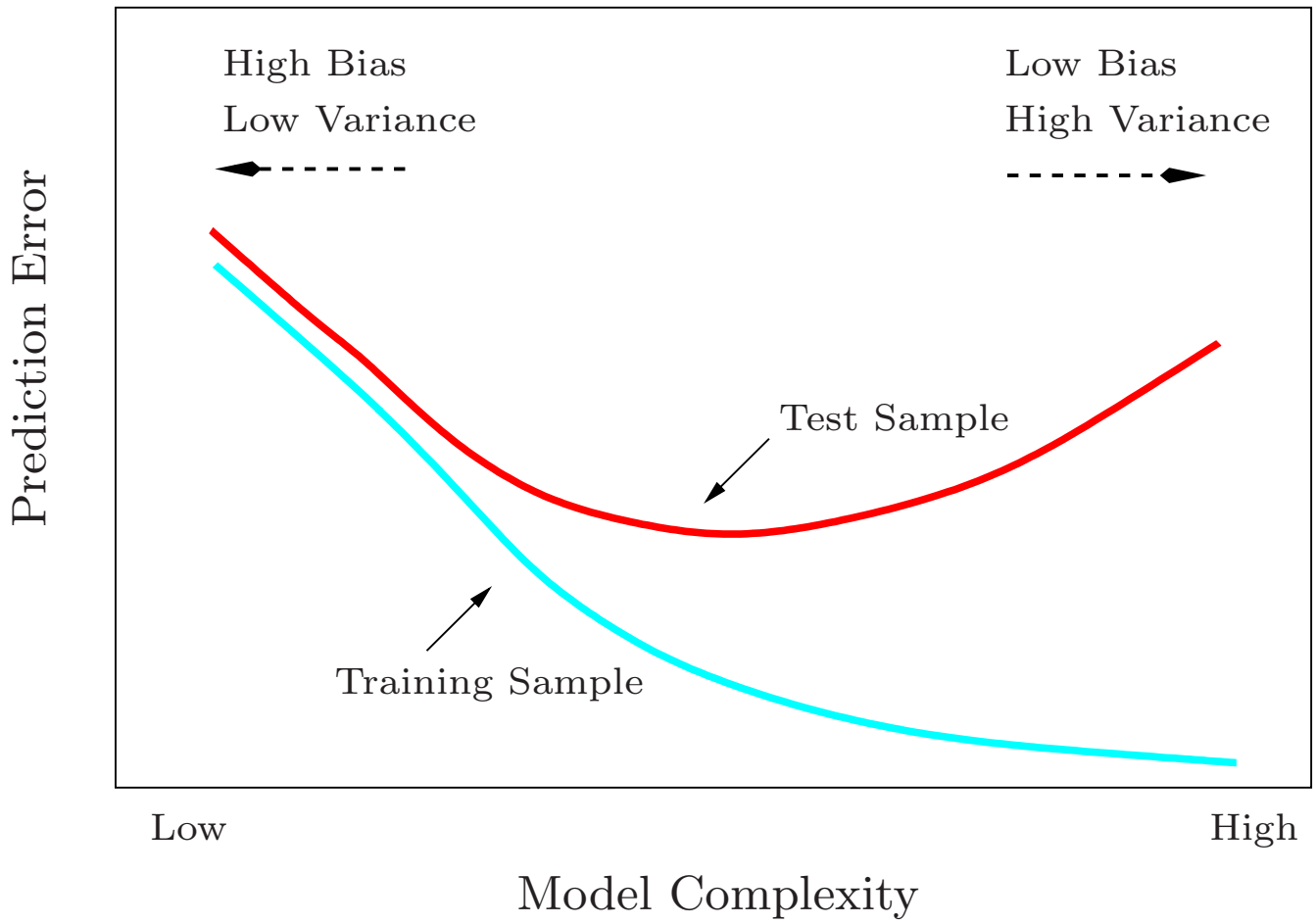
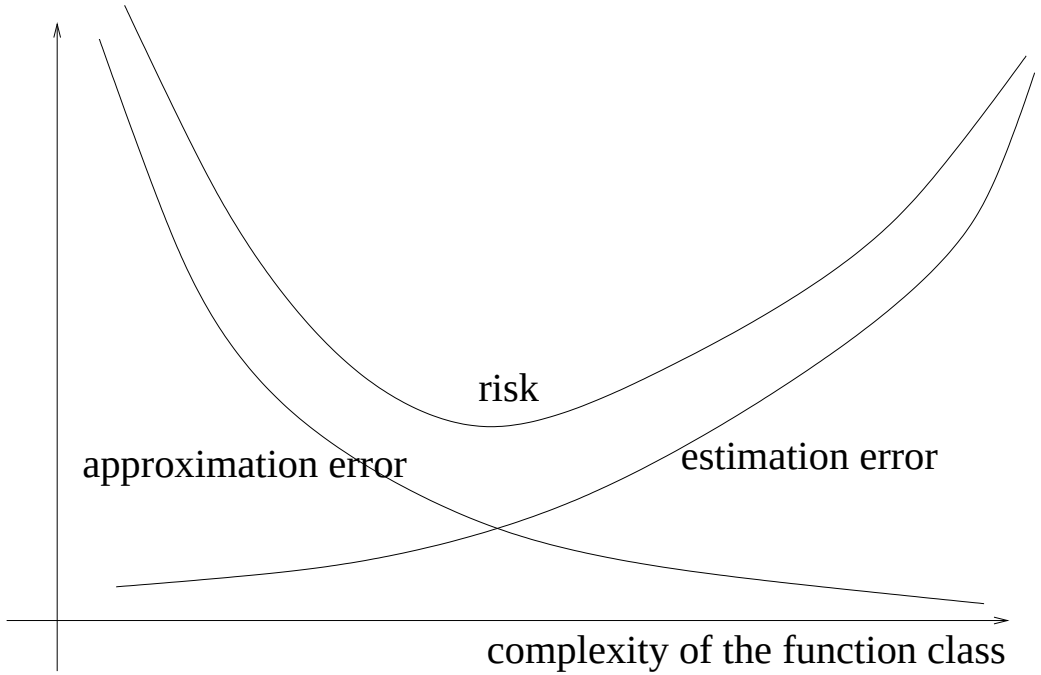


FIGURE 2.11. *Test and training error as a function of model complexity.*

Sesgo y varianza



Contenido

Aprendizaje Estadístico: El problema de regresión

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

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{21}$$

donde $R[f]$ involucra una función de pérdida apropiada para variable dependientes continuas. obsérvese que $\mathcal{F}_0$ 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 la la función de aprendizaje de Bayes y se denota por f_{Bayes}.

Aprendizaje Estadístico: El problema de regresión

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

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{21}$$

donde $R[f]$ involucra una función de pérdida apropiada para variable dependientes continuas. obsérvese que $\mathcal{F}_0$ 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 la la función de aprendizaje de Bayes y se denota por f_{Bayes}.

Aprendizaje Estadístico: El problema de regresión

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

$$f^* = argmin_{f \in \mathcal{F}_0} R[f] \tag{21}$$

donde $R[f]$ involucra una función de pérdida apropiada para variable dependientes continuas. obsérvese que $\mathcal{F}_0$ 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 la la función de aprendizaje de Bayes y se denota por f_{Bayes}.

Aprendizaje Estadístico: Resumen

- De la misma forma que en el problema de clasificación, el objetivo principal del apredizaje de máquinas para problemas de regresión es: dada una muestra finita τ_n y una función de pérdida, encontrar una espacio de funciones $\mathcal{F}_0$ y una función de aprendizaje óptima $f_{\mathcal{F}_0}$ tal que su riesgo sea lo más cercano posible a la función de aprendizaje de Bayes.
- El riesgo empírico de una regresión no es necesarimente un buen estimador del riesgo.

Aprendizaje Estadístico: Resumen

- De la misma forma que en el problema de clasificación, el objetivo principal del apredizaje de máquinas para problemas de regresión es: dada una muestra finita τ_n y una función de pérdida, encontrar una espacio de funciones $\mathcal{F}_0$ y una función de aprendizaje óptima $f_{\mathcal{F}_0}$ tal que su riesgo sea lo más cercano posible a la función de aprendizaje de Bayes.
- El riesgo empírico de una regresión no es necesarimente un buen estimador del riesgo.

Funciones de aprendizaje óptimas: Variables continuas

- Supongamos que Y es una variable continua que toma valores en los números reales.
- Definamos una función de pérdida L como el error cuadrático.
- Sea $\widehat{f}$ una función de aprendizaje.
- Entonces el riesgo de $\widehat{f}$ es:

$$E_X \int (\widehat{f}(X) - Y)^2 dF(Y \mid X) \tag{22}$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables continuas

- Supongamos que Y es una variable continua que toma valores en los números reales.
- Definamos una función de pérdida L como el error cuadrático.
- Sea $\widehat{f}$ una función de aprendizaje.
- Entonces el riesgo de $\widehat{f}$ es:

$$E_X \int (\widehat{f}(X) - Y)^2 dF(Y \mid X) \tag{22}$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables continuas

- Supongamos que Y es una variable continua que toma valores en los números reales.
- Definamos una función de pérdida L como el error cuadrático.
- Sea $\widehat{f}$ una función de aprendizaje.
- Entonces el riesgo de $\widehat{f}$ es:

$$E_X \int (\widehat{f}(X) - Y)^2 dF(Y \mid X) \tag{22}$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables continuas

- Supongamos que Y es una variable continua que toma valores en los números reales.
- Definamos una función de pérdida L como el error cuadrático.
- Sea $\widehat{f}$ una función de aprendizaje.
- Entonces el riesgo de $\widehat{f}$ es:

$$E_X \int (\widehat{f}(X) - Y)^2 dF(Y \mid X) \tag{22}$$

donde el valor esperado se toma con respecto a la distribución marginal X.

Funciones de aprendizaje óptimas: Variables continuas

- El problema de minimización lo podemos resolver si para cada $X = x$ minimizamos con respecto a y:

$$\int (y - Y)^2 dF(Y \mid X = x) \tag{23}$$

- Es fácil demostra que: $y = E[Y \mid X = x]$ o $\widehat{f}(X) = E[Y \mid X]$

Contenido

Descomposición Sesgo y Varianza

- El riesgo es definido como:

$$R(f_n) = \int L(x, y, f_n(x)) dP(x, y)$$

- El riesgo esperado:

$$E_n[R(f_n)] = \int R(f_n) dP^n(x_1, y_1, ..., x_n, y_n)$$

- El riesgo esperado en un punto (x_0, y_0):

$$E_n[L(X, Y, f_n(X)) \mid (X, Y) = (x_0, y_0)]$$

$$= \int L(x_0, y_0, f_n(x_0)) dP^n(x_1, y_1, ..., x_n, y_n)$$

Descomposición Sesgo y Varianza

- Supongamos que la función de pérdida es cuadrática:

$$R(f_n) = \int (f_n(x) - y)^2 dP(x, y)$$

$$\int (f_n(x) - y)^2 dP(y \mid x) dP(x)$$

- De otra parte, recuerde las siguientes definiciones:

$$E[Y \mid X] = \int y dP(y \mid x)$$

y

$$E[Y^2 \mid X] = \int y^2 dP(y \mid x)$$

Descomposición Sesgo y Varianza

- Entonces:

$$\begin{aligned}(f_n(X) - Y)^2 \\ = (f_n(X) - E[Y \mid X])^2 + \\ 2(f_n(X) - E[Y \mid X])(E[Y \mid X] - Y) + (E[Y \mid X] - Y)^2\end{aligned}$$

Luego:

$$\begin{aligned}R(f_n) = \int (f_n(x) - y)^2 dP(x, y) \\ = \int (f_n(x) - E[Y \mid x])^2 dP(y \mid x) dP(x) + \\ \int (E[y \mid x] - y)^2 dP(y \mid x) dP(x)\end{aligned}$$

Descomposición Sesgo y Varianza

- Obsérvese que el segundo término no depende de la muestra es un error irreducible que denotamos por $Irr(P)$.
- Ahora el primer término no depende de Y (es una variable aleatoria en X). Entonces:

$$R(f_n) = \int (f_n(x) - E[Y \mid x])^2 P(x) dx + Irr(P)$$

Descomposición Sesgo y Varianza

- Un argumento similar nos permite mostrar que el riesgo esperado en x_0 es:

$$\int (f_n(x_0 - E[Y \mid X = x_0])^2 dP^n(x_1, ..., x_n)$$
$$+Irr(P)$$
$$= E_{P^n}([f_n(x_0) - E[Y \mid X = x_0])^2]$$
$$+Irr(P)$$
$$= (E_{P^n}[f_n(x_0)] - E[Y \mid X = x_0])^2 + E_{P^n}[(f_n(x_0) - E_{P^n}[f_n(x_0)])^2]$$
$$+Irr(P)$$

El primer término es el sesgo en x_0. El segundo término es la varianza en x_0.

Descomposición Sesgo y Varianza

- Finalmente el riesgo esperado se puedo obtener de la anterior expresión integrando con respecto a x_0.

Vecinos más cercanos

- Considere un proceso generador de datos de la forma $y = f(x) + \epsilon$, donde $E[\epsilon] = 0$:

$$E_n[L(X, Y, f_n(X)) \mid (X, Y) = (x_0, y_0)] =$$
$$E_n[R(f_n)](x_0) = E[(Y - f_n(x_0))^2 \mid X = x_0]$$
$$= \sigma_\varepsilon^2 + (E_{P^n} f_n(x_0) - f(x_0))^2 + E_{P^n}[(f_n(x_0) - E[f_n(x_0)])^2]$$
$$= \sigma_\varepsilon^2 + Bias^2(f_n(x_0)) + Var(f_n(x_0))$$

- Estos errores son de diferentes naturalezas.

Vecinos más cercanos

- El primero es irreducible y corresponde a la desviación de los datos alrededor del verdadero valor $f(x_0)$. El segundo es qué tan lejos está el modelo del verdadero del verdadero valor. El tercero es qué tanto afecta la muestra le estimación del modelo en el punto de interés.

Vecinos más cercanos

- Esta ecuación se puede aproximar por:

$$\sigma_\varepsilon^2 + (f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)}))^2 + \frac{\sigma_\varepsilon^2}{k}$$

- A medida que k crece disminuye la complejidad del modelo
- Para k pequeños $\hat{f}_k(x)$ puede acercarse más a $f(x)$
- A medida que crece k el sesgo generalmente aumenta mientras la varianza disminuye