Aprendizaje de máquinas
Selección de modelos (parte 1)

Felipe Tobar

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Otoño, 2021.

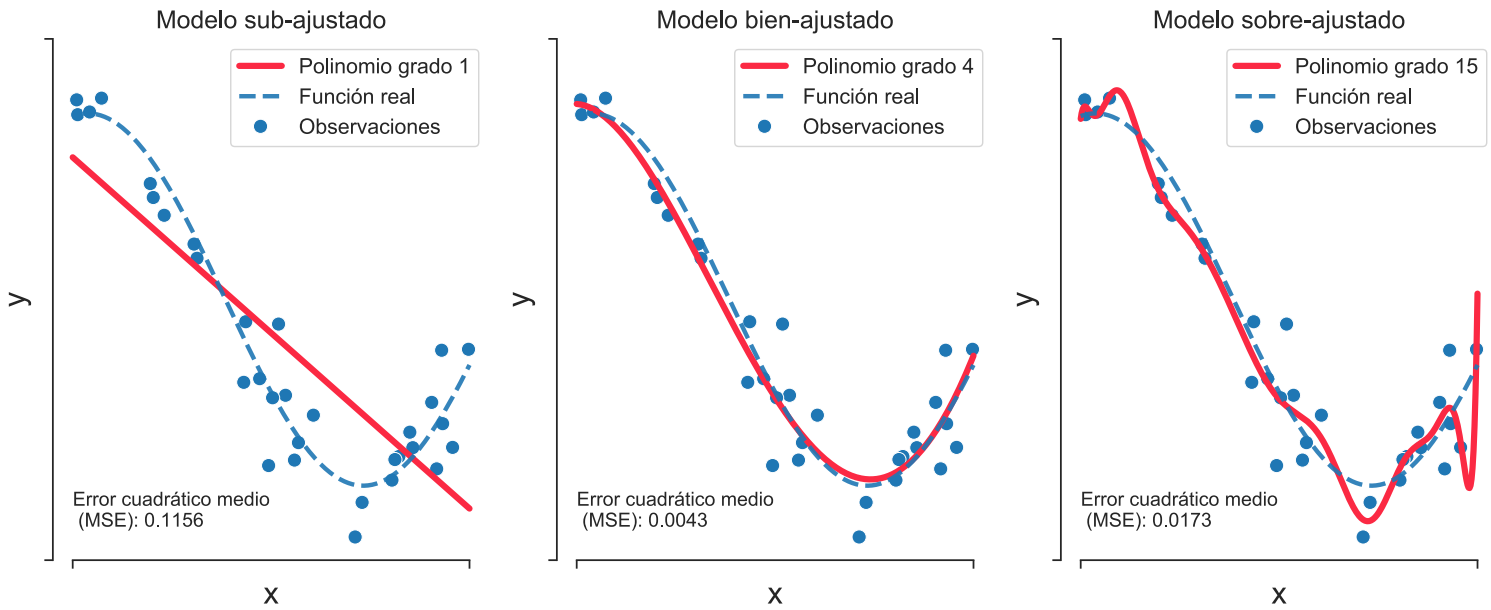
El problema de selección de modelos

Dado un conjunto de datos, existen muchos posibles modelos para poder realizar el aprendizaje, por lo que surge la pregunta natural de qué modelo elegir.

El problema de selección de modelos

Dado un conjunto de datos, existen muchos posibles modelos para poder realizar el aprendizaje, por lo que surge la pregunta natural de qué modelo elegir.

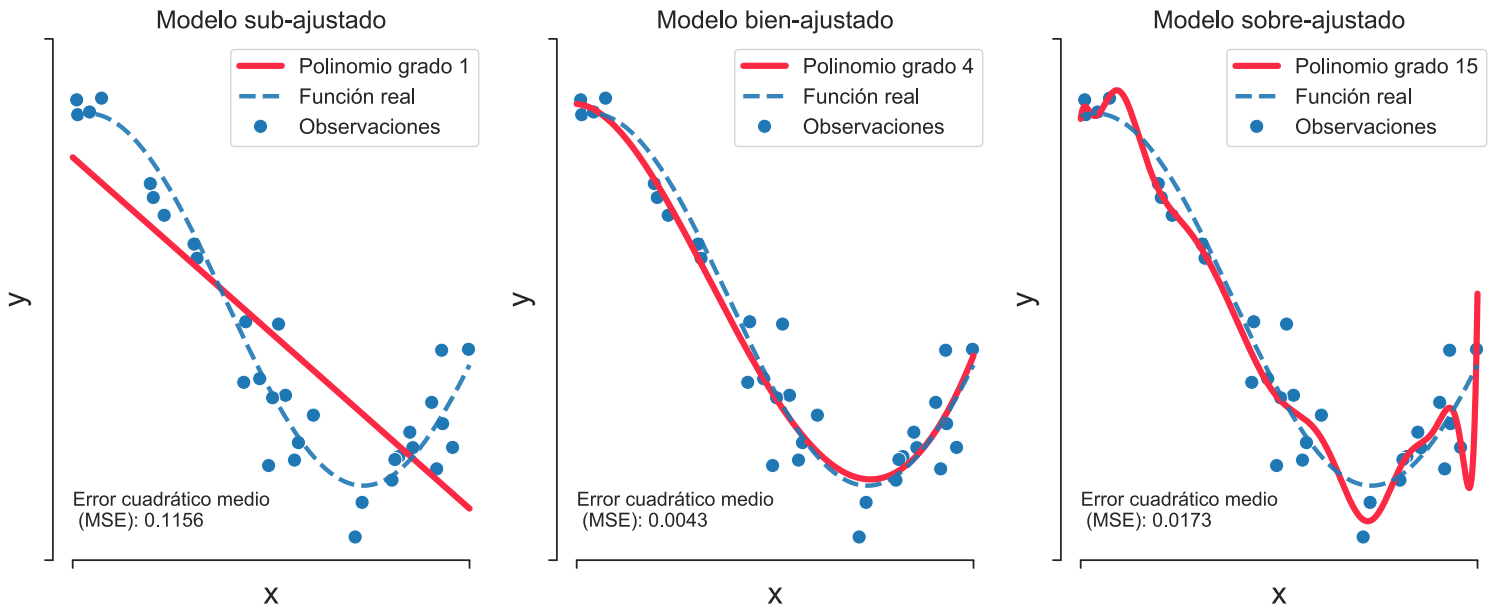
El problema de utilizar el modelo que mejor se ajuste a los datos en el entrenamiento es el sobreajuste, el cual puede ser observado en la siguiente figura:



El problema de selección de modelos

Dado un conjunto de datos, existen muchos posibles modelos para poder realizar el aprendizaje, por lo que surge la pregunta natural de qué modelo elegir.

El problema de utilizar el modelo que mejor se ajuste a los datos en el entrenamiento es el sobreajuste, el cual puede ser observado en la siguiente figura:



El problema de sobreajuste puede ser observado al elegir el grado en una regresión polinomial ya que para n puntos siempre existirá un polinomio de grado menor a n que pase exactamente por dichos puntos, por lo que es posible tener un error de ajuste nulo en los datos de entrenamiento, pero con un alto error de predicción en datos no vistos.

Descomposición sesgo-varianza

En el capítulo de regresión se probó que, si bien MCR introduce sesgo en el estimador, también disminuye la varianza del regresor, lo cual resultaba en una disminución del error esperado. El objetivo de esta parte será estudiar la descomposición sesgo-varianza para un modelo general.

Descomposición sesgo-varianza

En el capítulo de regresión se probó que, si bien MCR introduce sesgo en el estimador, también disminuye la varianza del regresor, lo cual resultaba en una disminución del error esperado. El objetivo de esta parte será estudiar la descomposición sesgo-varianza para un modelo general.

Definition

Sea $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^N \times \mathbb{R}$ un conjunto de observaciones generadas por una función desconocida $f : \mathbb{R}^N \to \mathbb{R}$ mediante $y = f(x) + \epsilon$ donde ϵ es una v.a. (ruido) con $\mathbb{E}(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Sea $\hat{f}(\cdot|\mathcal{D})$ un estimador de f determinado a partir de $\mathcal{D}$, entonces, se tienen las siguientes definiciones:

- Error (cuadrático) esperado: $\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right)$.

Descomposición sesgo-varianza

En el capítulo de regresión se probó que, si bien MCR introduce sesgo en el estimador, también disminuye la varianza del regresor, lo cual resultaba en una disminución del error esperado. El objetivo de esta parte será estudiar la descomposición sesgo-varianza para un modelo general.

Definition

Sea $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^N \times \mathbb{R}$ un conjunto de observaciones generadas por una función desconocida $f : \mathbb{R}^N \to \mathbb{R}$ mediante $y = f(x) + \epsilon$ donde ϵ es una v.a. (ruido) con $\mathbb{E}(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Sea $\hat{f}(\cdot|\mathcal{D})$ un estimador de f determinado a partir de $\mathcal{D}$, entonces, se tienen las siguientes definiciones:

- Error (cuadrático) esperado: $\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right)$.
- Sesgo del estimador: $\text{Bias}(\hat{f}(x|D)) := \mathbb{E}(\hat{f}\,(x|\mathcal{D}) - y) = \mathbb{E}(\hat{f}(x|\mathcal{D})) - f(x)$.

Descomposición sesgo-varianza

En el capítulo de regresión se probó que, si bien MCR introduce sesgo en el estimador, también disminuye la varianza del regresor, lo cual resultaba en una disminución del error esperado. El objetivo de esta parte será estudiar la descomposición sesgo-varianza para un modelo general.

Definition

Sea $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^N \times \mathbb{R}$ un conjunto de observaciones generadas por una función desconocida $f : \mathbb{R}^N \to \mathbb{R}$ mediante $y = f(x) + \epsilon$ donde ϵ es una v.a. (ruido) con $\mathbb{E}(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Sea $\hat{f}(\cdot|\mathcal{D})$ un estimador de f determinado a partir de $\mathcal{D}$, entonces, se tienen las siguientes definiciones:

- Error (cuadrático) esperado: $\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right)$.
- Sesgo del estimador: $\text{Bias}(\hat{f}(x|D)) := \mathbb{E}(\hat{f}\,(x|\mathcal{D}) - y) = \mathbb{E}(\hat{f}(x|\mathcal{D})) - f(x)$.
- Varianza del estimador: $\text{Var}(\hat{f}(x|D)) = \mathbb{E}\left(\left(\hat{f}(x|\mathcal{D}) - \mathbb{E}(\hat{f}(x|\mathcal{D}))\right)^2\right)$.

Descomposición sesgo-varianza

Bajo las definiciones anteriores, se tiene el siguiente teorema:

Theorem (descomposición sesgo-varianza)

Sea $\hat{f}(x|\mathcal{D})$ un estimador de f, entonces:

$$\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right) = Bias^2(\hat{f}(x|\mathcal{D})) + Var(\hat{f}(x|\mathcal{D})) + \sigma^2$$

Descomposición sesgo-varianza

Bajo las definiciones anteriores, se tiene el siguiente teorema:

Theorem (descomposición sesgo-varianza)

Sea $\hat{f}(x|\mathcal{D})$ un estimador de f, entonces:

$$\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right) = Bias^2(\hat{f}(x|\mathcal{D})) + Var(\hat{f}(x|\mathcal{D})) + \sigma^2$$

- la varianza intrínseca del ruido imposibilita realizar predicciones exactas bajo cualquier modelo aleatorio.

Descomposición sesgo-varianza

Bajo las definiciones anteriores, se tiene el siguiente teorema:

Theorem (descomposición sesgo-varianza)

Sea $\hat{f}(x|\mathcal{D})$ un estimador de f, entonces:

$$\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right) = Bias^2(\hat{f}(x|\mathcal{D})) + Var(\hat{f}(x|\mathcal{D})) + \sigma^2$$

- la varianza intrínseca del ruido imposibilita realizar predicciones exactas bajo cualquier modelo aleatorio.
- se puede introducir sesgo en el modelo con el fin de disminuir la varianza y viceversa.

Descomposición sesgo-varianza

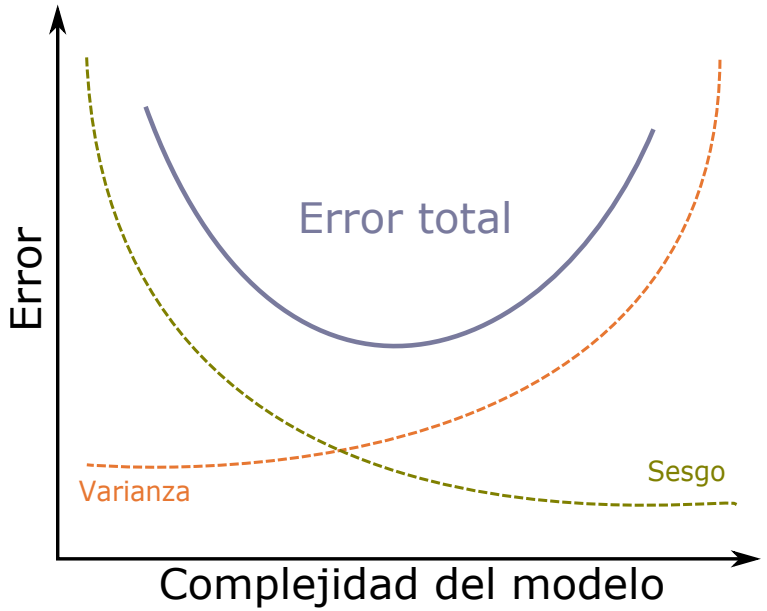
Bajo las definiciones anteriores, se tiene el siguiente teorema:

Theorem (descomposición sesgo-varianza)

Sea $\hat{f}(x|\mathcal{D})$ un estimador de f, entonces:

$$\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right) = Bias^2(\hat{f}(x|\mathcal{D})) + Var(\hat{f}(x|\mathcal{D})) + \sigma^2$$

- la varianza intrínseca del ruido imposibilita realizar predicciones exactas bajo cualquier modelo aleatorio.
- se puede introducir sesgo en el modelo con el fin de disminuir la varianza y viceversa.
- La combinación sesgo-varianza crea un error total convexo tal como se puede observar en la figura.



Descomposición sesgo-varianza

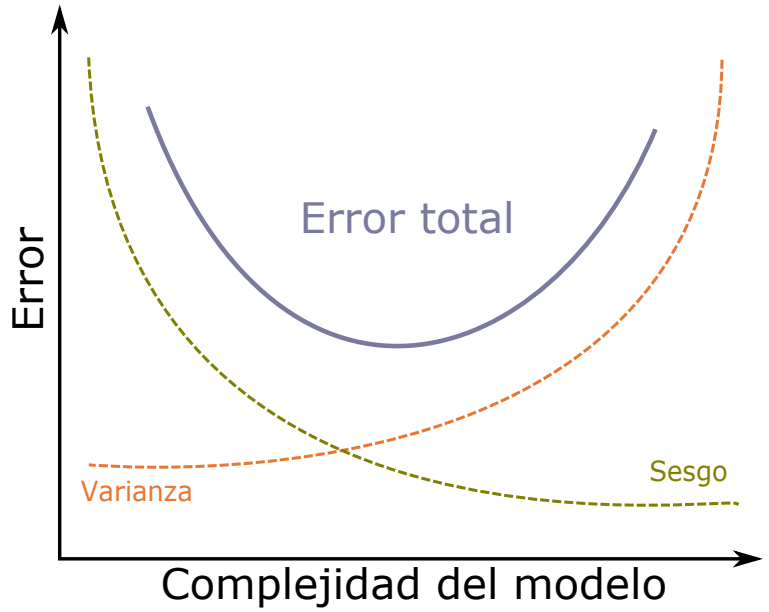
Bajo las definiciones anteriores, se tiene el siguiente teorema:

Theorem (descomposición sesgo-varianza)

Sea $\hat{f}(x|\mathcal{D})$ un estimador de f, entonces:

$$\mathbb{E}\left((y - \hat{f}(x|\mathcal{D}))^2\right) = Bias^2(\hat{f}(x|\mathcal{D})) + Var(\hat{f}(x|\mathcal{D})) + \sigma^2$$

- la varianza intrínseca del ruido imposibilita realizar predicciones exactas bajo cualquier modelo aleatorio.
- se puede introducir sesgo en el modelo con el fin de disminuir la varianza y viceversa.
- La combinación sesgo-varianza crea un error total convexo tal como se puede observar en la figura.



Esto crea la pregunta acerca de cuál es el par sesgo-varianza óptimo que minimiza el error total (dilema sesgo-varianza).

Validación cruzada

Por lo general no se cuenta con la forma analítica del error cuadrático esperado como para poder elegir el mejor modelo. En estos casos, se vuelve necesario poder comparar modelos de forma relativa, utilizando el conjunto de datos $\mathcal{D}$.

Validación cruzada

Por lo general no se cuenta con la forma analítica del error cuadrático esperado como para poder elegir el mejor modelo. En estos casos, se vuelve necesario poder comparar modelos de forma relativa, utilizando el conjunto de datos $\mathcal{D}$.

Una primera forma de elegir y evaluar un modelo fuera de muestra, consiste en particionar el conjunto de datos $\mathcal{D}$ en dos:

- Conjunto de entrenamiento: se utilizará para entrenar el modelo.

Validación cruzada

Por lo general no se cuenta con la forma analítica del error cuadrático esperado como para poder elegir el mejor modelo. En estos casos, se vuelve necesario poder comparar modelos de forma relativa, utilizando el conjunto de datos $\mathcal{D}$.

Una primera forma de elegir y evaluar un modelo fuera de muestra, consiste en particionar el conjunto de datos $\mathcal{D}$ en dos:

- Conjunto de entrenamiento: se utilizará para entrenar el modelo.
- Conjunto de validación: medirá el rendimiento del modelo de acuerdo a algún criterio predefinido (por ejemplo, ECM).

Validación cruzada

Por lo general no se cuenta con la forma analítica del error cuadrático esperado como para poder elegir el mejor modelo. En estos casos, se vuelve necesario poder comparar modelos de forma relativa, utilizando el conjunto de datos $\mathcal{D}$.

Una primera forma de elegir y evaluar un modelo fuera de muestra, consiste en particionar el conjunto de datos $\mathcal{D}$ en dos:

- Conjunto de entrenamiento: se utilizará para entrenar el modelo.
- Conjunto de validación: medirá el rendimiento del modelo de acuerdo a algún criterio predefinido (por ejemplo, ECM).

Con el fin de evitar posibles sesgos provocados por una partición en específico, la evaluación de desempeño se debe realizar varias veces sobre conjuntos de validación distintos.

Validación cruzada

Por lo general no se cuenta con la forma analítica del error cuadrático esperado como para poder elegir el mejor modelo. En estos casos, se vuelve necesario poder comparar modelos de forma relativa, utilizando el conjunto de datos $\mathcal{D}$.

Una primera forma de elegir y evaluar un modelo fuera de muestra, consiste en particionar el conjunto de datos $\mathcal{D}$ en dos:

- Conjunto de entrenamiento: se utilizará para entrenar el modelo.
- Conjunto de validación: medirá el rendimiento del modelo de acuerdo a algún criterio predefinido (por ejemplo, ECM).

Con el fin de evitar posibles sesgos provocados por una partición en específico, la evaluación de desempeño se debe realizar varias veces sobre conjuntos de validación distintos.

De esta forma, al promediar los rendimientos de cada partición se obtiene un rendimiento estimado fuera de muestra, lo cual permite finalmente elegir un modelo, quedándose con aquel que reporte el menor error out-sample.

Validación cruzada

Por lo general no se cuenta con la forma analítica del error cuadrático esperado como para poder elegir el mejor modelo. En estos casos, se vuelve necesario poder comparar modelos de forma relativa, utilizando el conjunto de datos $\mathcal{D}$.

Una primera forma de elegir y evaluar un modelo fuera de muestra, consiste en particionar el conjunto de datos $\mathcal{D}$ en dos:

- Conjunto de entrenamiento: se utilizará para entrenar el modelo.
- Conjunto de validación: medirá el rendimiento del modelo de acuerdo a algún criterio predefinido (por ejemplo, ECM).

Con el fin de evitar posibles sesgos provocados por una partición en específico, la evaluación de desempeño se debe realizar varias veces sobre conjuntos de validación distintos.

De esta forma, al promediar los rendimientos de cada partición se obtiene un rendimiento estimado fuera de muestra, lo cual permite finalmente elegir un modelo, quedándose con aquel que reporte el menor error out-sample.

Las distintas formas de mezclar y particionar los datos se conocen como *validación cruzada*.

Validación cruzada exhaustiva

Un primer tipo de validación cruzada corresponde a CV exhaustiva. En este tipo de validación cruzada, se prueban todas las posibles permutaciones de los datos al particionar el conjunto $\mathcal{D}$.

Validación cruzada exhaustiva

Un primer tipo de validación cruzada corresponde a CV exhaustiva. En este tipo de validación cruzada, se prueban todas las posibles permutaciones de los datos al particionar el conjunto $\mathcal{D}$. Se tienen 2 técnicas exhaustivas:

- **leave p out (LpOCV):** el conjunto $\mathcal{D}$ se particiona dejando p elementos para validación y los $N - p$ elementos restantes se utilizan para entrenar el modelo. Este entrenamiento y cálculo de desempeño se repite $C_p^N = \frac{N!}{(N-p)!p!}$ veces, pasando por todos los posibles conjuntos de validación de tamaño p.

Validación cruzada exhaustiva

Un primer tipo de validación cruzada corresponde a CV exhaustiva. En este tipo de validación cruzada, se prueban todas las posibles permutaciones de los datos al particionar el conjunto $\mathcal{D}$. Se tienen 2 técnicas exhaustivas:

- **leave p out (LpOCV):** el conjunto $\mathcal{D}$ se particiona dejando p elementos para validación y los $N - p$ elementos restantes se utilizan para entrenar el modelo. Este entrenamiento y cálculo de desempeño se repite $C_p^N = \frac{N!}{(N-p)!p!}$ veces, pasando por todos los posibles conjuntos de validación de tamaño p.
- **leave one out (LOOCV):** corresponde al caso anterior con $p = 1$. En este caso cada dato de $\mathcal{D}$ es utilizado como único elemento de validación mientras el resto de los datos se utiliza para entrenar.

Validación cruzada no exhaustiva

Por otra parte, existen dos tipos de validación cruzada no exhaustiva:

Validación cruzada no exhaustiva

Por otra parte, existen dos tipos de validación cruzada no exhaustiva:

- k-**fold:** el conjunto $\mathcal{D}$ es dividido en k grupos de igual tamaño. Luego, uno de esos grupos es utilizado como validador y el resto como entranemiento. Esto se repite k veces de forma de que todos los grupos sean validadores una y solo una vez.

Validación cruzada no exhaustiva

Por otra parte, existen dos tipos de validación cruzada no exhaustiva:

- k-**fold:** el conjunto $\mathcal{D}$ es dividido en k grupos de igual tamaño. Luego, uno de esos grupos es utilizado como validador y el resto como entranemiento. Esto se repite k veces de forma de que todos los grupos sean validadores una y solo una vez.
- **Monte Carlo CV:** se realizan particiones binarias aleatorias de $\mathcal{D}$. Se entrena y evalúa usando el par de conjuntos creados en cada partición.

Variante de la validación cruzada

Una variante de la validación cruzada es dividir el conjunto $\mathcal{D}$ en 3:

Variante de la validación cruzada

Una variante de la validación cruzada es dividir el conjunto $\mathcal{D}$ en 3:

- Los primeros dos conjuntos son utilizados para entrenamiento y validación.

Variante de la validación cruzada

Una variante de la validación cruzada es dividir el conjunto $\mathcal{D}$ en 3:

- Los primeros dos conjuntos son utilizados para entrenamiento y validación.
- El tercer conjunto (conocido como test set) es utilizado para obtener una estimación real del desempeño fuera de muestra del modelo elegido a partir de los dos conjuntos anteriores.

Variante de la validación cruzada

Una variante de la validación cruzada es dividir el conjunto $\mathcal{D}$ en 3:

- Los primeros dos conjuntos son utilizados para entrenamiento y validación.
- El tercer conjunto (conocido como test set) es utilizado para obtener una estimación real del desempeño fuera de muestra del modelo elegido a partir de los dos conjuntos anteriores.

Esto se realiza ya que al considerar únicamente el desempeño en el conjunto de validación, por lo general se sobreestima el desempeño real fuera de muestra debido a que el modelo fue elegido precisamente tomando el que reporta el menor error dentro del conjunto de validación.

Variante de la validación cruzada

Una variante de la validación cruzada es dividir el conjunto $\mathcal{D}$ en 3:

- Los primeros dos conjuntos son utilizados para entrenamiento y validación.
- El tercer conjunto (conocido como test set) es utilizado para obtener una estimación real del desempeño fuera de muestra del modelo elegido a partir de los dos conjuntos anteriores.

Esto se realiza ya que al considerar únicamente el desempeño en el conjunto de validación, por lo general se sobreestima el desempeño real fuera de muestra debido a que el modelo fue elegido precisamente tomando el que reporta el menor error dentro del conjunto de validación.

Observación: Si bien no hay una regla estándar que indique cómo particionar el conjunto, una división usual es utilizar el 50 % para entrenamiento y 25 % para validación y test.

Desventajas de la validación cruzada

Si bien la técnica de validación cruzada es bastante efectiva, tiene la limitación de requerir una gran cantidad de datos para poder realizar la partición de $\mathcal{D}$.

Para los casos que en los que no se cuenta con una cantidad considerable de observaciones, se requieren herramientas más sofisticadas para poder tomar una decisión acerca de qué modelo elegir.

Desventajas de la validación cruzada

Si bien la técnica de validación cruzada es bastante efectiva, tiene la limitación de requerir una gran cantidad de datos para poder realizar la partición de $\mathcal{D}$.

Para los casos que en los que no se cuenta con una cantidad considerable de observaciones, se requieren herramientas más sofisticadas para poder tomar una decisión acerca de qué modelo elegir.

La próxima clase se estudiarán dos criterios para selección de modelos:

- Criterio de información de Akaike (AIC).
- Criterio de información bayesiano (BIC).