Aprendizaje de máquinas
Selección de modelos (parte 2)

Felipe Tobar

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Otoño, 2021.

Selección de modelos

Como se comentó la clase pasada, el método de validación cruzada tiene la limitación de requerir una gran cantidad de datos para poder realizar la partición de $\mathcal{D}$.

Selección de modelos

Como se comentó la clase pasada, el método de validación cruzada tiene la limitación de requerir una gran cantidad de datos para poder realizar la partición de $\mathcal{D}$.

En esta clase se estudiarán dos métodos más sofisticados para la elección de modelos:

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

Criterio de información de Akaike (AIC)

Sea $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones generadas por una distribución desconocida perteneciente a una familia paramétrica cuyos parámetros están en $\Theta \subset \mathbb{R}^d$.

Criterio de información de Akaike (AIC)

Sea $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones generadas por una distribución desconocida perteneciente a una familia paramétrica cuyos parámetros están en $\Theta \subset \mathbb{R}^d$. Bajo este modelo, se puede utilizar el estimador de máxima verosimilitud:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathcal{D}) = \arg\max_{\theta \in \Theta} l(\theta|\mathcal{D})$$

Una forma de evaluar el desempeño real de este estimador es mediante el **riesgo de predicción**, el cual se ve reflejado en la log-verosimilitud de $\hat{\theta}$ sobre todas las posibles observaciones: $\mathbb{E}(l(\hat{\theta}|x))$.

Criterio de información de Akaike (AIC)

Sea $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones generadas por una distribución desconocida perteneciente a una familia paramétrica cuyos parámetros están en $\Theta \subset \mathbb{R}^d$. Bajo este modelo, se puede utilizar el estimador de máxima verosimilitud:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathcal{D}) = \arg\max_{\theta \in \Theta} l(\theta|\mathcal{D})$$

Una forma de evaluar el desempeño real de este estimador es mediante el **riesgo de predicción**, el cual se ve reflejado en la log-verosimilitud de $\hat{\theta}$ sobre todas las posibles observaciones: $\mathbb{E}(l(\hat{\theta}|x))$.

Dado que solo se cuenta con una cantidad finita de muestras, solo es posible obtener un riesgo empírico.

AIC: derivación (riesgo empírico y real)

AIC busca ajustar este riesgo para obtener un **estimador asintóticamente insesgado del riesgo real**.

AIC: derivación (riesgo empírico y real)

AIC busca ajustar este riesgo para obtener un **estimador asintóticamente insesgado del riesgo real**.
Para eso, se tienen las siguientes definiciones para el estimador de máxima verosimilitud $\hat{\theta}$:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV empírico.

AIC: derivación (riesgo empírico y real)

AIC busca ajustar este riesgo para obtener un **estimador asintóticamente insesgado del riesgo real**.
Para eso, se tienen las siguientes definiciones para el estimador de máxima verosimilitud $\hat{\theta}$:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV empírico.
- **Riesgo real:** $R(\hat{\theta}) = -\mathbb{E}(N \cdot l_0(\hat{\theta}))$, donde $l_0(\theta) = \mathbb{E}(l(\theta|x))$ corresponde a la log-verosimilitud de θ sobre todo el espacio muestral. Notar que se multiplica por N ya que en el riesgo empírico no se normalizó por N.

AIC: derivación (riesgo empírico y real)

AIC busca ajustar este riesgo para obtener un **estimador asintóticamente insesgado del riesgo real**.
Para eso, se tienen las siguientes definiciones para el estimador de máxima verosimilitud $\hat{\theta}$:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV empírico.
- **Riesgo real:** $R(\hat{\theta}) = -\mathbb{E}(N \cdot l_0(\hat{\theta}))$, donde $l_0(\theta) = \mathbb{E}(l(\theta|x))$ corresponde a la log-verosimilitud de θ sobre todo el espacio muestral. Notar que se multiplica por N ya que en el riesgo empírico no se normalizó por N.

Para poder obtener el *AIC* se analizará el sesgo asintótico del riesgo empírico con respecto al riesgo real. Para esto:

- Se utilizarán aproximaciones sobre ambos riesgos.

AIC: derivación (riesgo empírico y real)

AIC busca ajustar este riesgo para obtener un **estimador asintóticamente insesgado del riesgo real**.
Para eso, se tienen las siguientes definiciones para el estimador de máxima verosimilitud $\hat{\theta}$:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV empírico.
- **Riesgo real:** $R(\hat{\theta}) = -\mathbb{E}(N \cdot l_0(\hat{\theta}))$, donde $l_0(\theta) = \mathbb{E}(l(\theta|x))$ corresponde a la log-verosimilitud de θ sobre todo el espacio muestral. Notar que se multiplica por N ya que en el riesgo empírico no se normalizó por N.

Para poder obtener el *AIC* se analizará el sesgo asintótico del riesgo empírico con respecto al riesgo real. Para esto:

- Se utilizarán aproximaciones sobre ambos riesgos.
- Se considerará que medida que N crece, el EMV empírico tiende al EMV global.

AIC: derivación (riesgo empírico y real)

AIC busca ajustar este riesgo para obtener un **estimador asintóticamente insesgado del riesgo real**.
Para eso, se tienen las siguientes definiciones para el estimador de máxima verosimilitud $\hat{\theta}$:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV empírico.
- **Riesgo real:** $R(\hat{\theta}) = -\mathbb{E}(N \cdot l_0(\hat{\theta}))$, donde $l_0(\theta) = \mathbb{E}(l(\theta|x))$ corresponde a la log-verosimilitud de θ sobre todo el espacio muestral. Notar que se multiplica por N ya que en el riesgo empírico no se normalizó por N.

Para poder obtener el *AIC* se analizará el sesgo asintótico del riesgo empírico con respecto al riesgo real. Para esto:

- Se utilizarán aproximaciones sobre ambos riesgos.
- Se considerará que medida que N crece, el EMV empírico tiende al EMV global.
- Lo anterior implica que el residuo de Taylor tenderá a 0.

AIC: derivación (aproximaciones para los riesgos)

Sea $\theta_0 = \arg\max_{\theta \in \Theta} l_0(\theta)$ el EMV sobre todo el espacio muestral. Utilizando una aproximación de Taylor de segundo orden sobre l_0 alrededor de θ_0 se prueba que:

$$l_0(\hat{\theta}) \approx l_0(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)$$

AIC: derivación (aproximaciones para los riesgos)

Sea $\theta_0 = \arg\max_{\theta \in \Theta} l_0(\theta)$ el EMV sobre todo el espacio muestral. Utilizando una aproximación de Taylor de segundo orden sobre l_0 alrededor de θ_0 se prueba que:

$$l_0(\hat{\theta}) \approx l_0(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)$$

De esta forma, se tiene una aproximación de segundo orden para el riesgo real:

$$R(\hat{\theta}) \approx -N \cdot l_0(\theta_0) - \frac{N}{2}\,\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

AIC: derivación (aproximaciones para los riesgos)

Sea $\theta_0 = \arg\max_{\theta \in \Theta} l_0(\theta)$ el EMV sobre todo el espacio muestral. Utilizando una aproximación de Taylor de segundo orden sobre l_0 alrededor de θ_0 se prueba que:

$$l_0(\hat{\theta}) \approx l_0(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)$$

De esta forma, se tiene una aproximación de segundo orden para el riesgo real:

$$R(\hat{\theta}) \approx -N \cdot l_0(\theta_0) - \frac{N}{2}\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

De forma análoga se prueba que:

$$\hat{l} \approx \sum_{i=1}^{N} l(\theta_0|x_i) + N(\hat{\theta} - \theta_0)^\top \mathbb{E}(H_l(\theta_0|x))(\theta_0 - \hat{\theta}) + \frac{N}{2}(\hat{\theta} - \theta_0)^\top \mathbb{E}(H_l(\theta_0|x))(\hat{\theta} - \theta_0)$$

AIC: derivación (aproximaciones para los riesgos)

Sea $\theta_0 = \arg\max_{\theta \in \Theta} l_0(\theta)$ el EMV sobre todo el espacio muestral. Utilizando una aproximación de Taylor de segundo orden sobre l_0 alrededor de θ_0 se prueba que:

$$l_0(\hat{\theta}) \approx l_0(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)$$

De esta forma, se tiene una aproximación de segundo orden para el riesgo real:

$$R(\hat{\theta}) \approx -N \cdot l_0(\theta_0) - \frac{N}{2}\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

De forma análoga se prueba que:

$$\hat{l} \approx \sum_{i=1}^{N} l(\theta_0|x_i) + N(\hat{\theta} - \theta_0)^\top \mathbb{E}(H_l(\theta_0|x))(\theta_0 - \hat{\theta}) + \frac{N}{2}(\hat{\theta} - \theta_0)^\top \mathbb{E}(H_l(\theta_0|x))(\hat{\theta} - \theta_0)$$

Obteniendo una aproximación de segundo orden para el riesgo empírico:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) = -N \cdot l_0(\theta_0) + \frac{N}{2}\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

AIC: derivación (corrección del sesgo)

De este modo, el sesgo del riesgo empírico como estimador del riesgo real es:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) = -N\,\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

AIC: derivación (corrección del sesgo)

De este modo, el sesgo del riesgo empírico como estimador del riesgo real es:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) = -N\,\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

Dado que $\sqrt{N}\left(\hat{\theta} - \theta_0\right) \approx \mathcal{N}\left(0, H_{l_0}(\theta_0)^{-1}\right)$, la forma cuadrática anterior puede ser aproximada por una distribución de Pearson: $N(\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0) \approx \mathcal{X}^2_d$, donde $\mathbb{E}(\mathcal{X}^2_d) = d$. De este modo:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) \approx -d$$

AIC: derivación (corrección del sesgo)

De este modo, el sesgo del riesgo empírico como estimador del riesgo real es:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) = -N\,\mathbb{E}\left((\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0)\right)$$

Dado que $\sqrt{N}\left(\hat{\theta} - \theta_0\right) \approx \mathcal{N}\left(0, H_{l_0}(\theta_0)^{-1}\right)$, la forma cuadrática anterior puede ser aproximada por una distribución de Pearson: $N(\hat{\theta} - \theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta} - \theta_0) \approx \mathcal{X}^2_d$, donde $\mathbb{E}(\mathcal{X}^2_d) = d$. De este modo:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) \approx -d$$

Por lo que corrigiendo $R_{\mathcal{D}}(\hat{\theta})$ se obtiene un estimador asintóticamente insesgado del riesgo real: $R_{\mathcal{D}}(\hat{\theta}) + d$.

AIC: definición

La corrección anterior motiva la siguiente definición:

Definition (AIC)

Sea M un modelo estadístico d-paramétrico y $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones. El AIC del modelo (aproximado por $\mathcal{D}$) se define como

$$AIC(M, \mathcal{D}) := 2d - \log\left(\hat{L}(\mathcal{D})\right)$$

Donde $\hat{L}(\mathcal{D})$ corresponde a la verosimilitud del EMV asociado a $\mathcal{D}$, es decir:

$$\hat{L}(\mathcal{D}) = \prod_{i=1}^{N} p(x_i|\hat{\theta}), \text{ para } \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathcal{D})$$

AIC: definición

La corrección anterior motiva la siguiente definición:

Definition (AIC)

Sea M un modelo estadístico d-paramétrico y $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones. El AIC del modelo (aproximado por $\mathcal{D}$) se define como

$$AIC(M, \mathcal{D}) := 2d - \log\left(\hat{L}(\mathcal{D})\right)$$

Donde $\hat{L}(\mathcal{D})$ corresponde a la verosimilitud del EMV asociado a $\mathcal{D}$, es decir:

$$\hat{L}(\mathcal{D}) = \prod_{i=1}^{N} p(x_i|\hat{\theta}), \text{ para } \hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|\mathcal{D})$$

Como se puede ver, el AIC corresponde al estimador asintóticamente insesgado del riesgo real multiplicado por 2. Esta ponderación es realizada por motivos históricos.

Observaciones acerca de AIC

- Para un conjunto de posibles modelos, se debe elegir el modelo que presente el menor valor AIC ya que será el que minimice el riesgo de predicción.

Observaciones acerca de AIC

- Para un conjunto de posibles modelos, se debe elegir el modelo que presente el menor valor AIC ya que será el que minimice el riesgo de predicción.
- AIC no se basa únicamente en la verosimilitud del modelo sino que agrega una penalización de acuerdo a la cantidad de parámetros, evitando elegir un modelo sobreajustado a los datos.

Observaciones acerca de AIC

- Para un conjunto de posibles modelos, se debe elegir el modelo que presente el menor valor AIC ya que será el que minimice el riesgo de predicción.
- AIC no se basa únicamente en la verosimilitud del modelo sino que agrega una penalización de acuerdo a la cantidad de parámetros, evitando elegir un modelo sobreajustado a los datos.
- Una de las hipótesis de AIC es que el espacio muestral es infinito ya que se asume que el error de Taylor es despreciable. Para una cantidad finita de datos, se puede realizar una corrección del estimador dada por:

$$AICc(M, \mathcal{D}) := AIC(M, \mathcal{D}) + \frac{2d(d+1)}{N-d-1}$$

Es importante notar que cuando $N \to \infty$ se recupera el AIC original.

Criterio de información bayesiano (BIC)

Otro enfoque para la selección de modelos corresponde al criterio de información bayesiano (o criterio de Schwarz).

Criterio de información bayesiano (BIC)

Otro enfoque para la selección de modelos corresponde al criterio de información bayesiano (o criterio de Schwarz).

- Dada una familia de modelos $\mathcal{M}$, se define un prior $p(m)$ para cada modelo $m \in \mathcal{M}$.

Criterio de información bayesiano (BIC)

Otro enfoque para la selección de modelos corresponde al criterio de información bayesiano (o criterio de Schwarz).

- Dada una familia de modelos $\mathcal{M}$, se define un prior $p(m)$ para cada modelo $m \in \mathcal{M}$.
- Además, se define un prior $p(\theta|m)$ sobre los parámetros de cada modelo.

Criterio de información bayesiano (BIC)

Otro enfoque para la selección de modelos corresponde al criterio de información bayesiano (o criterio de Schwarz).

- Dada una familia de modelos $\mathcal{M}$, se define un prior $p(m)$ para cada modelo $m \in \mathcal{M}$.
- Además, se define un prior $p(\theta|m)$ sobre los parámetros de cada modelo.

El criterio de información bayesiano (BIC) elige al mejor modelo de acuerdo a la posterior $p(m|\mathcal{D})$, la cual viene dada de acuerdo al teorema de Bayes:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} \propto p(\mathcal{D}|m)p(m)$$

Criterio de información bayesiano (BIC)

Otro enfoque para la selección de modelos corresponde al criterio de información bayesiano (o criterio de Schwarz).

- Dada una familia de modelos $\mathcal{M}$, se define un prior $p(m)$ para cada modelo $m \in \mathcal{M}$.
- Además, se define un prior $p(\theta|m)$ sobre los parámetros de cada modelo.

El criterio de información bayesiano (BIC) elige al mejor modelo de acuerdo a la posterior $p(m|\mathcal{D})$, la cual viene dada de acuerdo al teorema de Bayes:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} \propto p(\mathcal{D}|m)p(m)$$

De forma similar al criterio de Akaike, se puede calcular la verosimilitud del modelo $p(\mathcal{D}|m)$ mediante aproximaciones de Taylor, probando que es independiente del prior.

BIC: definición

La derivación de $p(\mathcal{D}|m)$ lleva a la siguiente definición:

Definition (BIC)

Sea M un modelo estadístico d-paramétrico y $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones. El BIC del modelo (aproximado por $\mathcal{D}$) se define como

$$BIC(M, \mathcal{D}) := d \cdot \log(N) - 2 \log\left(\hat{L}(\mathcal{D})\right)$$

Donde nuevamente $\hat{L}(\mathcal{D})$ corresponde a la verosimilitud del EMV asociado a $\mathcal{D}$.

BIC: definición

La derivación de $p(\mathcal{D}|m)$ lleva a la siguiente definición:

Definition (BIC)

Sea M un modelo estadístico d-paramétrico y $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones. El BIC del modelo (aproximado por $\mathcal{D}$) se define como

$$BIC(M, \mathcal{D}) := d \cdot \log(N) - 2 \log\left(\hat{L}(\mathcal{D})\right)$$

Donde nuevamente $\hat{L}(\mathcal{D})$ corresponde a la verosimilitud del EMV asociado a $\mathcal{D}$.

En este caso, se vuelve a elegir el modelo que presente el menor BIC. Se observa que, al igual que AIC, BIC contiene una penalización sobre el número de parámetros por lo que también evita el sobreajuste a los datos.

Teoremas de Stone y Shao

Al igual de validación cruzada, los criterios de Akaike y bayesiano buscan elegir el mejor modelo de acuerdo a su capacidad predictiva fuera de muestra. Existe una estrecha relación entre ambas técnicas, las cuales se resumen en el siguiente teorema:

Theorem (Stone (1977) - Shao (1997))

Para una familia de modelos, minimizar el AIC es asintóticamente equivalente a realizar LOOCV. Por otra parte, minimizar el BIC es asintóticamente equivalente a realizar leave p out cross validation para

$$p = \left\lfloor N \left(1 - \frac{1}{\log(N) - 1}\right) \right\rfloor$$

AIC y BIC para la regresión lineal

Para el modelo lineal gaussiano $y = c^\top x + \epsilon$, con $\epsilon \sim \mathcal{N}(0, \sigma^2)$ se tiene que $y|x \sim \mathcal{N}(y; c^\top x, \sigma^2)$. Sean $\hat{c}$ y $\hat{\sigma}^2$ los EMV del modelo, entonces la log-verosimilitud máxima viene dada por:

$$\hat{l}(\mathcal{D}) = \frac{-N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N}(y_i - \hat{c}^\top x_i) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}N\hat{\sigma}^2$$

$$= C(N) - \frac{N}{2}\log(\hat{\sigma}^2) = C(N) - \frac{N}{2}\log\left(\frac{1}{N}\text{RSS}(\mathcal{D})\right)$$

Donde $C(N) = -\frac{N}{2}\log(2\pi) - N$ y $\text{RSS}(\mathcal{D})$ corresponde a la suma de cuadrados residuales: $\text{RSS}(\mathcal{D}) := \sum_{i=1}^{N}\left(y_i - c^\top x_i\right)^2$.

AIC y BIC para la regresión lineal

Para el modelo lineal gaussiano $y = c^\top x + \epsilon$, con $\epsilon \sim \mathcal{N}(0, \sigma^2)$ se tiene que $y|x \sim \mathcal{N}(y; c^\top x, \sigma^2)$. Sean $\hat{c}$ y $\hat{\sigma}^2$ los EMV del modelo, entonces la log-verosimilitud máxima viene dada por:

$$\hat{l}(\mathcal{D}) = \frac{-N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N}(y_i - \hat{c}^\top x_i) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}N\hat{\sigma}^2$$

$$= C(N) - \frac{N}{2}\log(\hat{\sigma}^2) = C(N) - \frac{N}{2}\log\left(\frac{1}{N}\text{RSS}(\mathcal{D})\right)$$

Donde $C(N) = -\frac{N}{2}\log(2\pi) - N$ y $\text{RSS}(\mathcal{D})$ corresponde a la suma de cuadrados residuales: $\text{RSS}(\mathcal{D}) := \sum_{i=1}^{N}\left(y_i - c^\top x_i\right)^2$.

Dado que $C(N)$ es una constante independiente del modelo, puede ser omitida en la comparación de modelos, por lo tanto:

- $AIC = 2d - N\log(\frac{1}{N}\text{RSS}(\mathcal{D}))$
- $BIC = d\log(N) - N\log(\frac{1}{N}\text{RSS}(\mathcal{D}))$