Auxiliar 1: selección de modelos

MA5204 Aprendizaje de Máquinas

Arie Wortsman, Nelson Moreno, Víctor Faragi,
Francisco Vásquez, Fernando Fêtis.

Departamento de ingeniería matemática
Universidad de Chile

7 de mayo de 2021

UNIVERSIDAD
DE CHILE

Descomposición sesgo-varianza

Se asumirá lo siguiente para un problema de regresión:

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^M \times \mathbb{R}$ conjunto de observaciones.
- cada observación es generada por $y = f(x) + \epsilon$ donde f es una función desconocida y ϵ es una v.a. de **media nula** y $\text{Var}(\epsilon) = \sigma^2$.

Definition

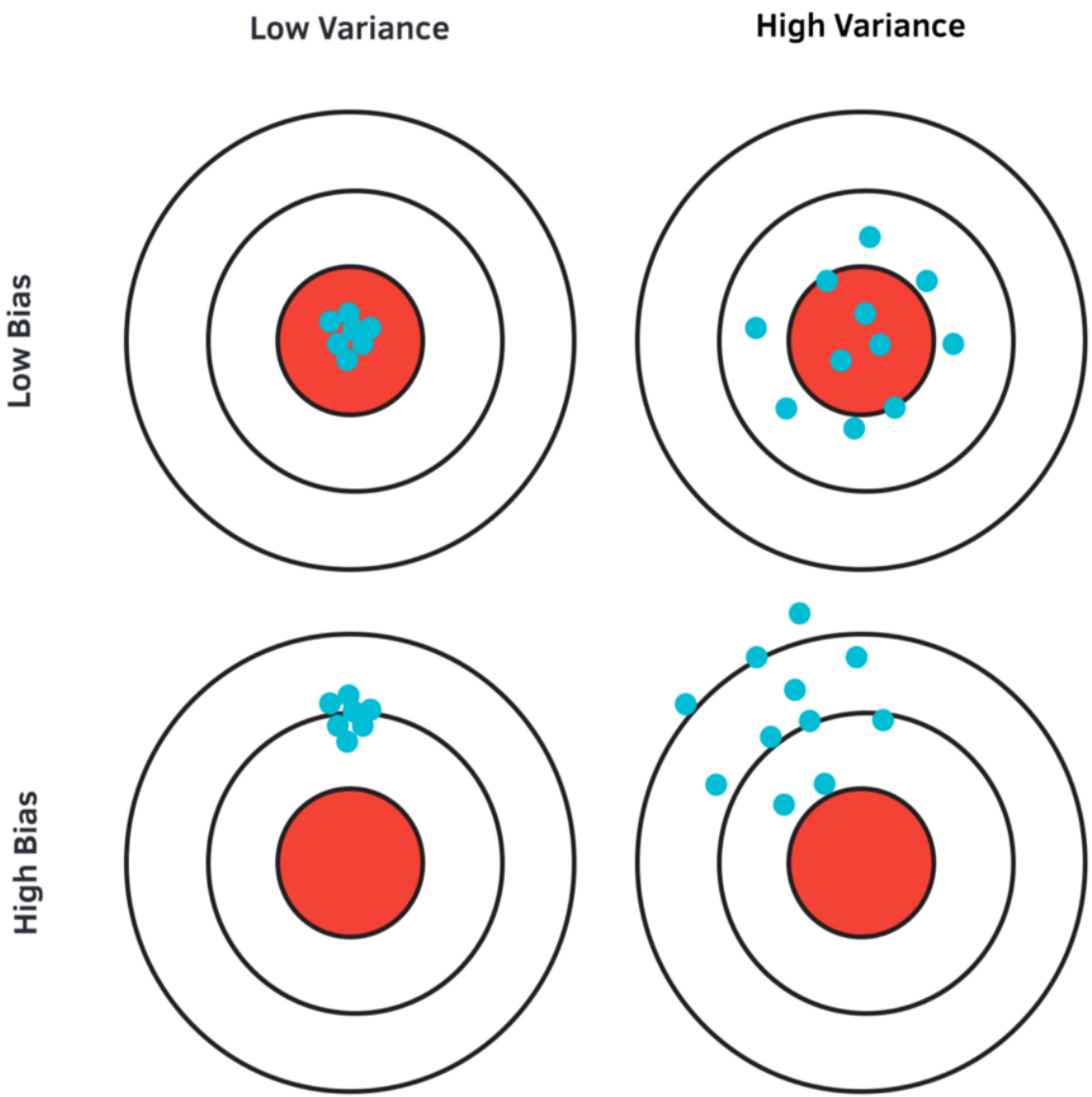
Sea $\hat{f}(\cdot|\mathcal{D})$ un **estimador de** f determinado a partir de $\mathcal{D}$, entonces, para una **nueva observación** (x_0, y_0) se tienen las siguientes definiciones:

- Error de generalización: $\mathbb{E}_{\mathcal{D}}\left((y_0 - \hat{f}(x_0|\mathcal{D}))^2\right)$.
- Sesgo del estimador de $f(x_0)$:

$$\text{Bias}(\hat{f}(x_0|\mathcal{D})) := \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0|\mathcal{D})) - f(x_0) = \mathbb{E}_{\mathcal{D}}(\hat{f}\,(x_0|\mathcal{D}) - f(x_0))$$

- Varianza del estimador: $\text{Var}(\hat{f}(x_0|\mathcal{D})) := \mathbb{E}_{\mathcal{D}}\left(\left(\hat{f}(x_0|\mathcal{D}) - \mathbb{E}_{\mathcal{D}}(\hat{f}(x_0|\mathcal{D}))\right)^2\right)$.

Descomposición sesgo-varianza



Descomposición sesgo-varianza

El error esperado de predicción se puede descomponer en 3 términos de acuerdo al siguiente teorema:

Theorem (descomposición sesgo-varianza)

Para una nueva muestra (x, y), el estimador $\hat{f}(x|\mathcal{D})$ de $f(x)$ cumple que:

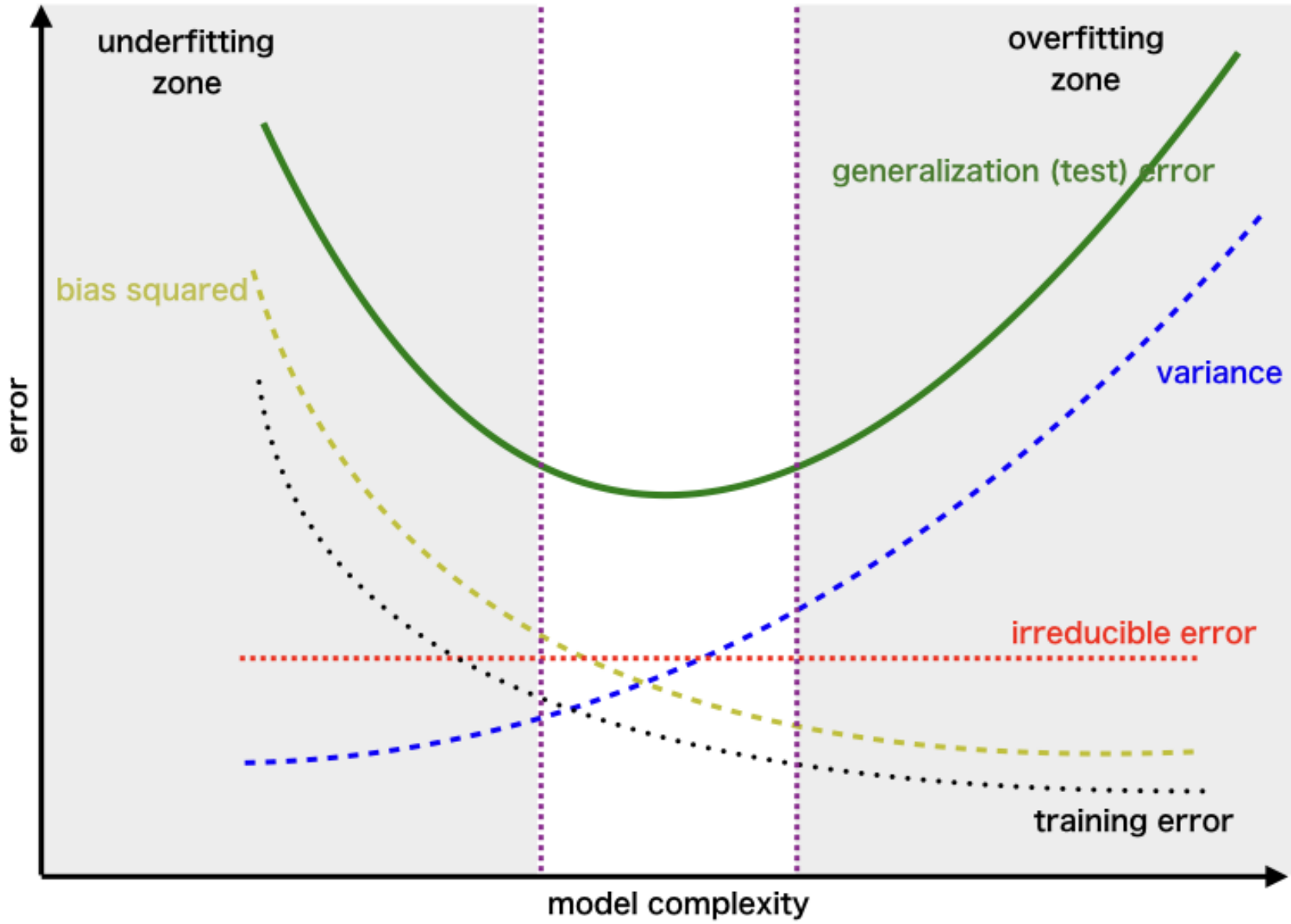
$$\mathbb{E}\left((y-\hat{f})^2\right) = \underbrace{\left(\mathbb{E}(\hat{f})-f(x)\right)^2}_{sesgo^2} + \underbrace{\mathbb{E}\left(\left(\hat{f}-\mathbb{E}(\hat{f})\right)^2\right)}_{varianza\ estimador} + \underbrace{\mathbb{E}\left((\epsilon-\mathbb{E}(\epsilon))^2\right)}_{\sigma^2\ (varianza\ ruido)}$$

Demostración.

$$\begin{aligned}
\mathbb{E}\left((y-\hat{f})^2\right) =& \mathbb{E}\left((f+\epsilon-\hat{f})^2\right) = \mathbb{E}\left(f^2+\epsilon^2+\hat{f}^2+2f\epsilon-2f\hat{f}-2\epsilon\hat{f}\right) \\
=& \left(\mathbb{E}^2(\hat{f})-2f\,\mathbb{E}(\hat{f})+f^2\right)+\mathbb{E}\left(\hat{f}^2-2\hat{f}\,\mathbb{E}(\hat{f})+\mathbb{E}^2(\hat{f})\right)+\mathbb{E}\left((\epsilon-\mathbb{E}(\epsilon))^2\right) \\
& -2\,\mathbb{E}(\epsilon\hat{f})-2\,\mathbb{E}^2(\hat{f})+2\,\mathbb{E}(\hat{f})\,\mathbb{E}(\hat{f}) \\
=& \left(\mathbb{E}(\hat{f})-f(x)\right)^2+\mathbb{E}\left(\left(\hat{f}-\mathbb{E}(\hat{f})\right)^2\right)+\mathbb{E}\left((\epsilon-\mathbb{E}(\epsilon))^2\right)-2\,\mathbb{E}(\epsilon)\,\mathbb{E}(\hat{f}) \\
=& \mathrm{Bias}^2(\hat{f}(x|\mathcal{D}))+\mathrm{Var}(\hat{f}(x|\mathcal{D}))+\sigma^2
\end{aligned}$$

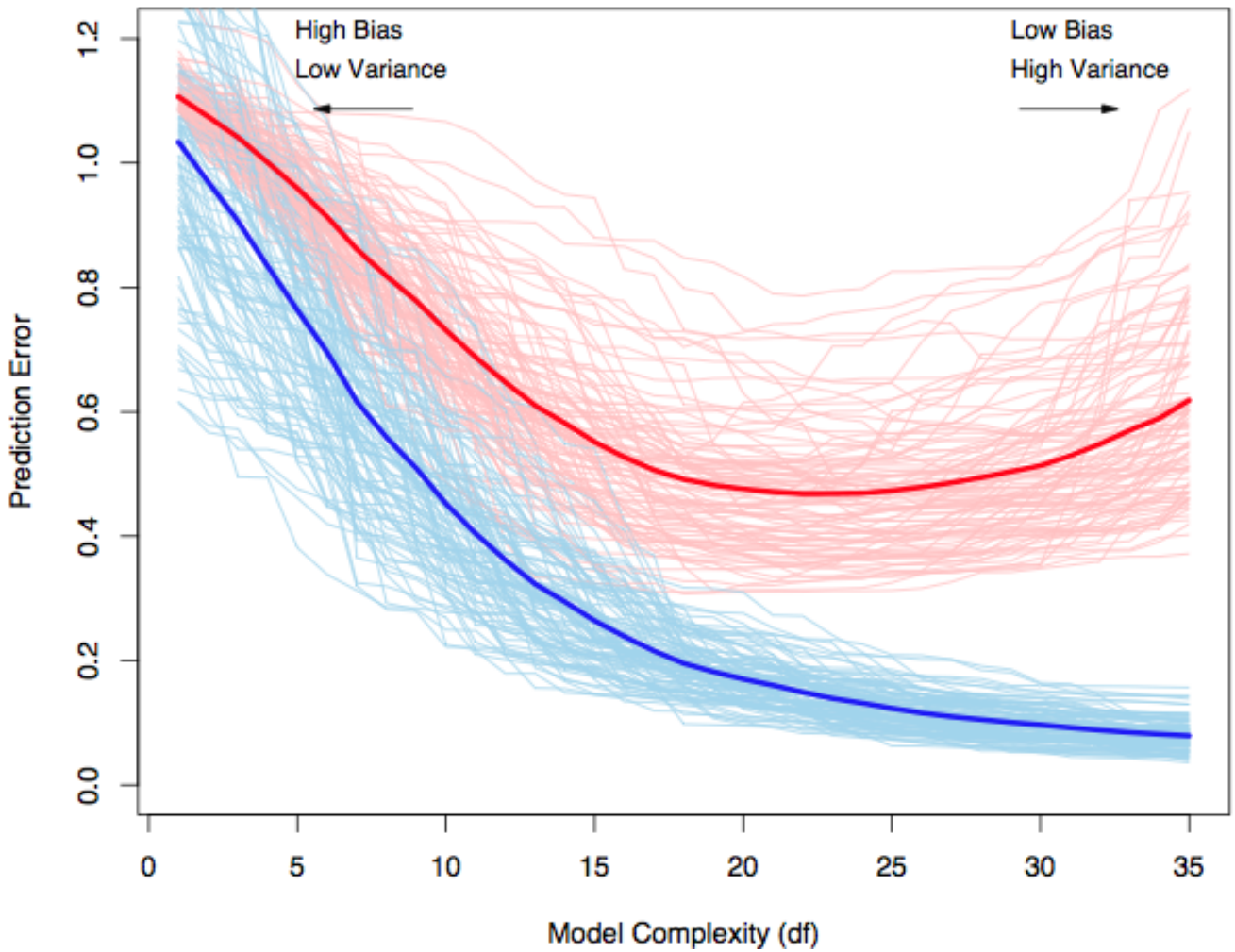
□

Descomposición sesgo-varianza



Descomposición sesgo-varianza

En la siguiente figura, la curva azul representa el error in-sample y la curva roja representa el error out-of-sample:



Validación cruzada

Es ideal poder encontrar los hiperparámetros que entreguen el par sesgo-varianza óptimo. Como esto no se puede hacer analíticamente, una forma intuitiva de comparar distintos hiperparámetros (y modelos en general) es la validación cruzada.

- Es utilizada para comparar modelos cuando hay una **cantidad considerable de datos**.
- $\mathcal{D}$ es dividido en 2:
  - **Conjunto de entrenamiento:** se utiliza para encontrar algún $\hat{\theta}$ (estimador de θ).
  - **Conjunto de validación:** se utiliza para evaluar el rendimiento out-of-sample.
- El conjunto $\mathcal{D}$ se particiona varias veces y luego se promedian los desempeños obtenidos en cada ciclo de entrenamiento-evaluación.

Los tipos de CV más comunes son:

- **leave p out (LpOCV):** se prueban todas las posibilidades donde se utilizan p elementos para evaluar el regresor.
- **leave one out (LOOCV):** caso anterior con $p = 1$.

También es posible particionar $\mathcal{D}$ en 3 conjuntos: entrenamiento, validación y testeo. Esto permite obtener una **mejor estimación del error de generalización**.

Criterio de información de Akaike (AIC)

Cuando no existen suficientes datos, no es buena aproximación utilizar CV para comparar el rendimiento de distintos modelos (indexados por $\theta \in \mathbb{R}^d$). Para estos casos, se puede realizar una **corrección a las verosimilitudes de los parámetros**.

- La verosimilitud $L(\theta) = p(\mathcal{D}|\theta)$ es la probabilidad de que el parámetro θ haya generado los datos $\mathcal{D}$.
- Entonces $\mathbb{E}_x(l(\theta|x)))$ representa la verosimilitud esperada sobre todo el espacio muestral Ω, es decir, la probabilidad de que el modelo θ genere Ω.
- Por lo tanto, lo ideal es encontrar el θ_0 que la maximice.
- Dado que solo se cuenta con un conjunto de entrenamiento $\mathcal{D} = \{x_i\}_{i=1}^N$ y no con Ω completo, no es posible calcular $\mathbb{E}_x$.
- Se puede aproximar $\mathbb{E}_x(l(\theta|x)))$ mediante la verosimilitud sobre $\mathcal{D}$.

Criterio de información de Akaike (AIC)

Una vez elegido el EMV $\hat{\theta} = \arg\max_{\theta \in \Theta} l(\theta|\mathcal{D})$ no se tiene la seguridad de que dicha verosimilitud se mantenga sobre otras observaciones (i.e., al generalizar sobre Ω). **AIC busca corregir la estimación $l(\theta)$ de la verosimilitud global**. Por lo anterior, se define lo siguiente:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV.
- **Riesgo real:** $R(\hat{\theta}) = -\mathbb{E}_{\mathcal{D}}(N \cdot \mathbb{E}_x(l(\hat{\theta}|x)))$, donde $\mathbb{E}_x(l(\hat{\theta}|x))$ corresponde a la log-verosimilitud de $\hat{\theta}$ sobre todo el espacio muestral Ω.

Lo que verdaderamente se busca minimizar es el riesgo real ya que se busca el modelo qué más probablemente genere todas las muestras (i.e., el EMV sobre Ω). Si el riesgo empírico se ve como un estimador del riesgo real, AIC busca corregir dicha estimación, por lo que **AIC es un estimador insesgado del riesgo real**.

Criterio de información de Akaike (AIC)

- Para que $\overline{\mathcal{R}}(\hat{\theta})$ sea un estimador insesgado del riesgo real $R(\hat{\theta})$ debe cumplirse que $\mathbb{E}(\overline{\mathcal{R}}(\hat{\theta})) = R(\hat{\theta})$, es decir, que en promedio el estadístico estime el riesgo real.
- El sesgo del riesgo empírico $R_{\mathcal{D}}(\hat{\theta})$ como estimador del riesgo real $R(\hat{\theta})$ viene dado por $\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta})$.
- Utilizando aproximaciones de Taylor de 2^{o} orden sobre el riesgo real y el riesgo empírico se prueba que $\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) \approx -d$ (d dimensión de θ).
- Por lo tanto, por linealidad de la esperanza, $R_{\mathcal{D}}(\hat{\theta}) + d$ es un **estimador insesgado del riesgo real**.

Definition (AIC)

Sea M un modelo d-paramétrico $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ y $\mathcal{D} = (x_i)_{i=1}^N$ observaciones. El AIC del modelo M (c/r a $\mathcal{D}$) se define como

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

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

Criterio de información de Akaike (AIC)

- Se busca elegir el modelo que entregue el menor riesgo real, por lo que se selecciona el modelo que presente la menor AIC.
- AIC penaliza la complejidad del modelo (dada por el número de parámetros).
- Para poder utilizar las aproximaciones de Taylor con igualdad, es necesario que $\mathcal{D}$ sea infinito. Dado que esto en general no se tiene, se puede nuevamente corregir el estimador:

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

Criterio de información bayesiano (BIC)

Para un problema de selección de modelos, supóngase lo siguiente:

- $\mathcal{M}$ es una familia de modelos.
- $p(m)$ es un prior sobre cada modelo $m \in \mathcal{M}$.

BIC elige al mejor modelo como aquel que maximice la posterior dada por

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

De forma análoga a AIC, se aproxima la verosimilitud del modelo $p(\mathcal{D}|m)$ probando que la posterior $p(m|\mathcal{D})$ es independiente del prior. Dicha aproximación lleva a la siguiente definición:

Definition (BIC)

Sea M un modelo d-paramétrico $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ y $\mathcal{D} = (x_i)_{i=1}^N$ observaciones. El BIC del modelo M (c/r a $\mathcal{D}$) 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 $\mathcal{D}$.

Criterio de información bayesiano (BIC)

- Se busca elegir el modelo que entregue la mayor posterior $p(m|\mathcal{D})$, por lo que se selecciona el modelo que presente el menor $BIC = d \cdot \log(N) - 2 \log \hat{L}(\mathcal{D})$.
- Al igual que $AIC = 2d - 2 \log \hat{L}(\mathcal{D})$, BIC penaliza la flexibilidad del modelo.
- **Teorema de Stone-Shao:** 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 el modelo lineal con ruido aditivo gaussiano

- Se considera el modelo lineal generativo usual $y = \theta^\top \tilde{x} + \epsilon$ con $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- Las observaciones (x, y) vienen dadas por la distribución $y|x \sim \mathcal{N}(y; c^\top x, \sigma^2)$.

Por lo tanto, si $\hat{\theta}$ y $\hat{\sigma}^2$ son los EMV asociados al conjunto de entrenamiento $\mathcal{D} = (x_i)_{i=1}^N$, la máxima log-verosimilitud viene dada por:

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

Donde $\mathrm{RSS}(\mathcal{D}) := \sum_{i=1}^N \left(y_i - c^\top x_i\right)^2$ y $C(N)$ es una constante independiente del modelo por lo que puede ser omitida en la comparación de modelos. Por lo tanto, para el modelo lineal gaussiano:

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