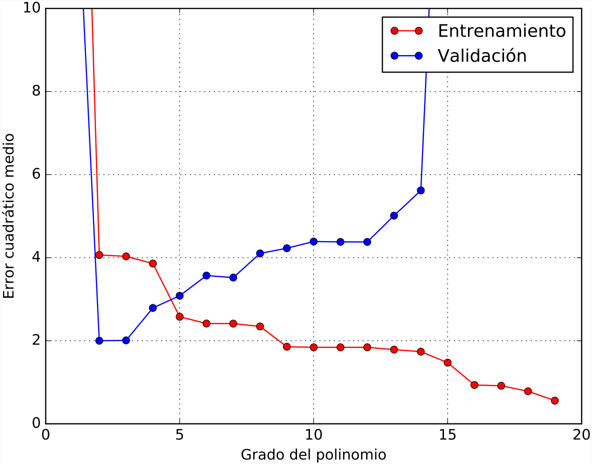
Aprendizaje automatizado

Selección de modelos

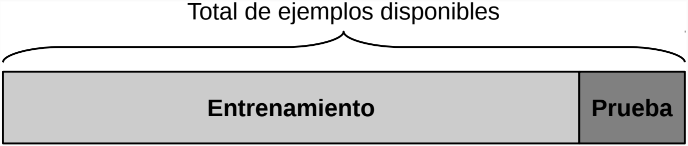
Gibran Fuentes-Pineda

Marzo 2020

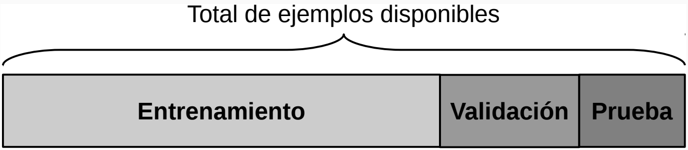
El problema de la generalización revisitado



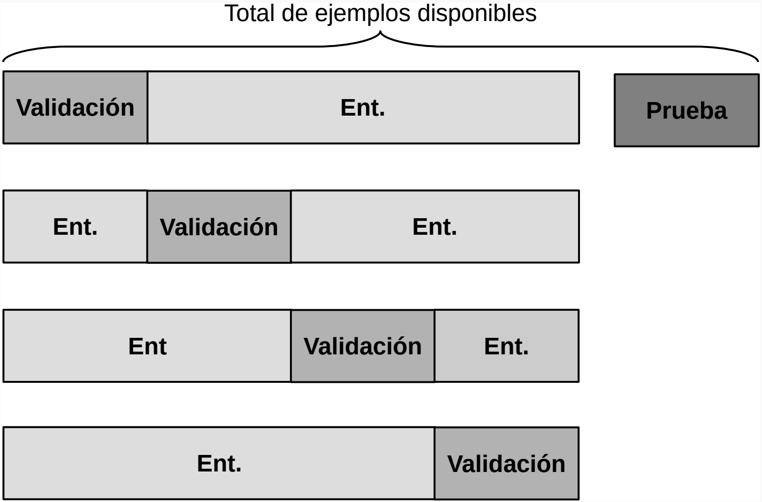
Partición de los datos en entrenamiento y prueba



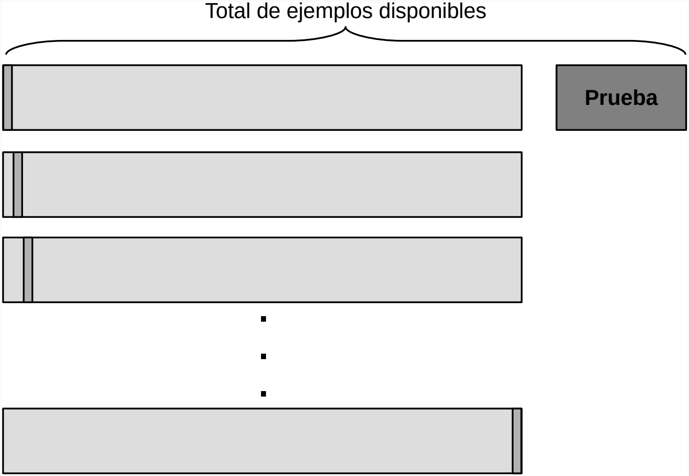
Dividiendo los datos en entrenamiento, validación y prueba



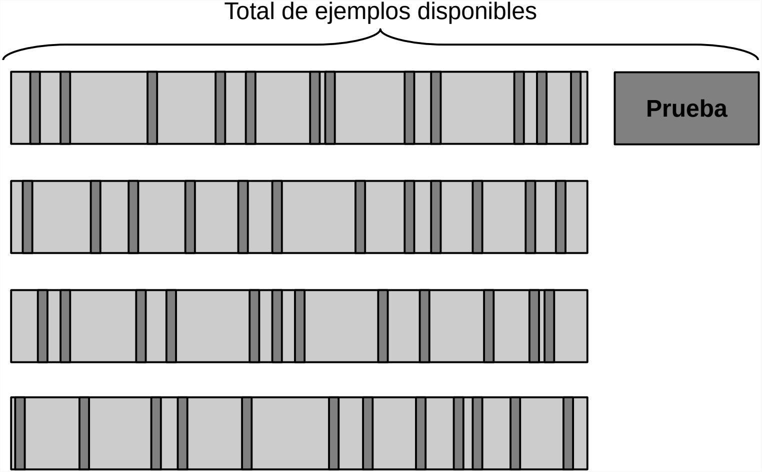
Validación cruzada con K particiones



Validación cruzada dejando uno fuera (LOOCV)



Validación cruzada aleatoria



Cálculo del error en validación cruzada

- Promedio de los errores en cada partición

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

- En el caso de LOOCV

$$E = \frac{1}{n}\sum_{i=1}^{n} E_i$$

Medidas de rendimiento de clasificadores binarios

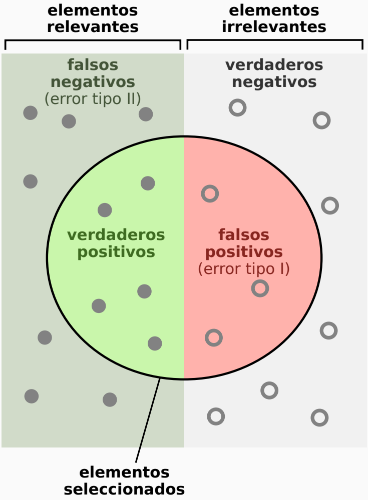


Figura traducida de Wikipedia (entrada de *Precision and Recall*)

Medidas de rendimiento de clasificadores binarios

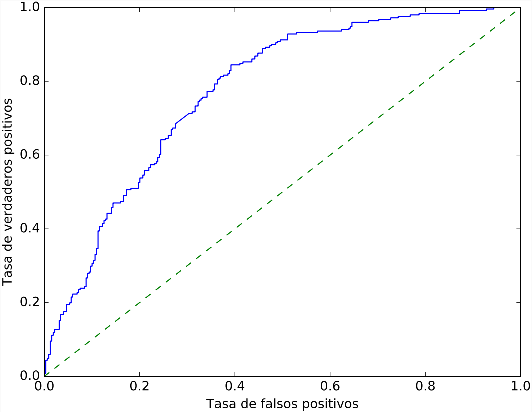
$$\text{precisión} = \frac{|\text{verdaderos positivos}|}{|\text{elementos seleccionados}|}$$

$$\text{exhaustividad} = \frac{|\text{verdaderos positivos}|}{|\text{elementos relevantes}|}$$

$$\text{tasa de verdaderos positivos} = \text{exhaustividad}$$

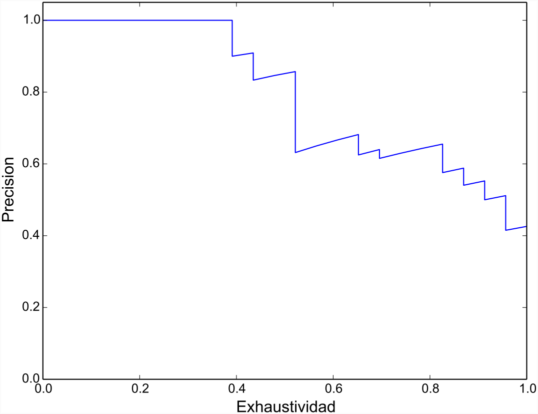
$$\text{tasa de falsos positivos} = \frac{|\text{falsos positivos}|}{|\text{elementos irrelevantes}|}$$

Curva ROC



Tasa de verdaderos positivos
Tasa de falsos positivos
0.0
0.2
0.4
0.6
0.8
1.0

Curva de precisión-exhaustividad



Matriz de confusión

		Clase Verdadera	
		Cáncer	No Cáncer
Clase	Cáncer	5 VP	3 FP
Predicha	No Cáncer	10 FN	6 VN

Medidas de rendimiento para regresión

- Error cuadrático medio

$$ECM(\mathbf{X}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2$$

- Erro absoluto medio

$$EAM(\mathbf{X}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \mid \hat{y}^{(i)} - y^{(i)} \mid$$

Métricas internas para agrupamiento

- **Compacidad**: Mide qué tan cerca están los elementos del mismo clústeres
- **Separación**: Mide qué tan separados están los elementos de diferentes clústeres

- **Pureza**: Mide la proporción de la clase con mayor número de elementos en el clúster con respecto al tamaño del mismo

- **Jaccard**:

$$J(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} = \frac{VP}{VP + FP + FN}$$

Métricas con clases desbalanceadas

- Considera la tarea de clasificación de correo no deseado.

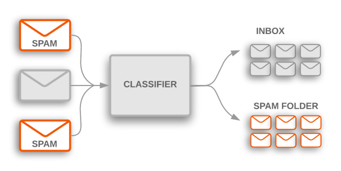


Figura reproducida de `https://developers.google.com/machine-learning/guides/text-classification`

Métricas con clases desbalanceadas

- Considera la tarea de clasificación de correo no deseado.

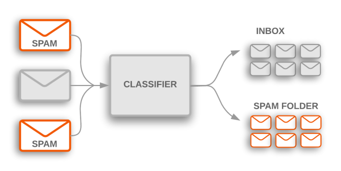


Figura reproducida de `https://developers.google.com/machine-learning/guides/text-classification`

- Nuestro conjunto de datos disponible contiene 96 % de correo normal y tan sólo 4 % de correo no deseado.

Métricas con clases desbalanceadas

- Considera la tarea de clasificación de correo no deseado.

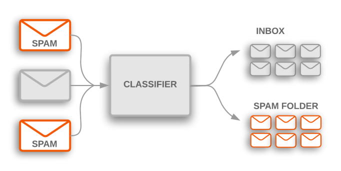


Figura reproducida de https://developers.google.com/machine-learning/guides/text-classification

- Nuestro conjunto de datos disponible contiene 96 % de correo normal y tan sólo 4 % de correo no deseado.
- Entrenamos un clasificador con un subconjunto de estos datos y evaluamos su exactitud con el restante, obteniendo un 96 % de exactitud. ¿Es este un buen modelo?

Impacto del desbalance en el aprendizaje

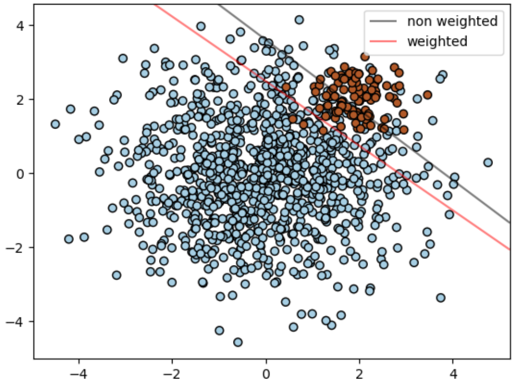


Figura reproducida de `https://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane_unbalanced.html`

Estrategias de aprendizaje para clases desbalanceadas

- Generar ejemplos artificiales de clase más escasa (*oversampling*)
- Elegir un subconjunto más pequeño de las clases más comunes (*undersampling*)
- Usar función de pérdida pesada

Aprendizaje en la práctica: demasiados atributos

- Demasiadas características pueden degradar el rendimiento de modelos
    - Maldición de la dimensionalidad
    - Atributos redundantes
    - Atributos irrelevantes

Extracción de características vs selección de atributos

- **Extracción de características**: mapea los atributos a un espacio de dimensiones menores
- **Selección de atributos**: elige un subconjunto de los atributos existentes
  - *Filtros*: evalúan el contenido de los atributos (por ej. distancia entre clases)
  - *Envolventes*: usan el clasificador para evaluar subconjuntos de atributos
  - *Híbridos*: tratan de combinar las ventajas de los filtros y los envolventes

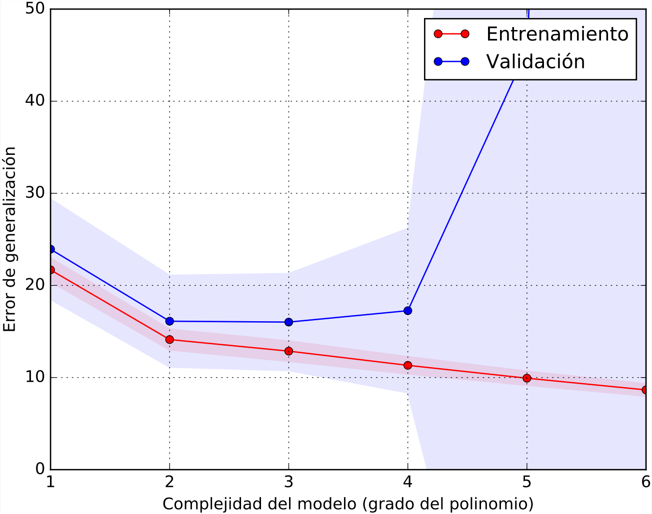
¿Por qué reducir el número de atributos?

- Menos efectos de la maldición de la dimensionalidad
- Menos espacio y mediciones
- Más rápido de entrenar y ejecutar
- Más fácil de interpretar y visualizar

¿Cómo elegimos los atributos más adecuados?

- Búsqueda óptima de subconjuntos es intratable
- **Selección hacia adelante**: se va añadiendo incrementalemnte el atributo que disminuya más el error
- **Selección hacia atrás**: se va eliminando decrementalmente el atributo que aumente más el error

Sesgo vs varianza



Criterio de información bayesiana (BIC)

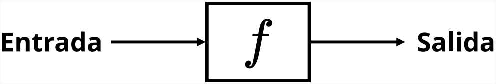
- Es posible incrementar la verosimilitud de cualquier modelo haciéndolo más complejo a costo de posible sobre-ajuste

- BIC es un criterio que penaliza modelos con muchos parámetros

$$BIC = -2 \cdot \log\left(\text{máx}\,\text{likelihood}\right) + \log\left(n\right) \cdot d$$

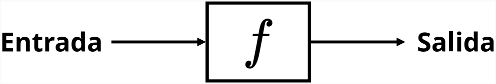
Aprendizaje universal

- ¿Es posible aprender cualquier tarea (función *f*)? ¿Es necesario el conocimiento a priori?



Aprendizaje universal

- ¿Es posible aprender cualquier tarea (función *f*)? ¿Es necesario el conocimiento a priori?



- *No existe la comida gratis*
    - Sólo es posible aprender de forma eficiente un pequeño subconjunto de todas las tareas posibles
    - El *sesgo inductivo* ayuda a aprender ciertas tareas

Toma de decisiones

- Nuestros modelos nos ofrecen una descripción de la incertidumbre de cierta situación resumidas en probabilidades
- Esta descripción nos sirve para tomar decisiones, es decir, saber qué acciones tomar (e.g. si es un tumor maligno realizar una operación)
- La *teoría de decisión* trata de cómo tomar decisiones óptimas a partir de nuestros modelos

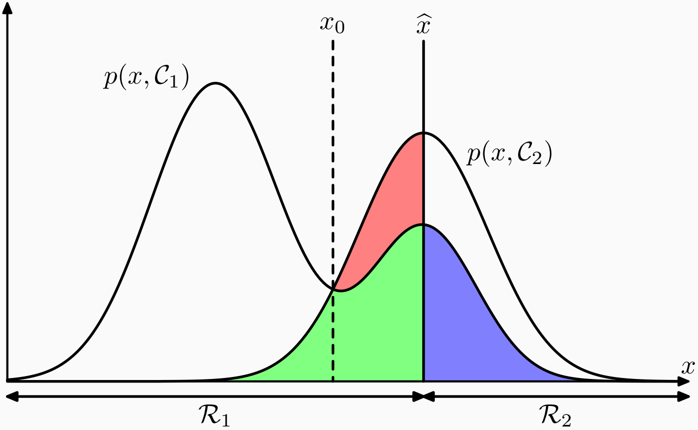
Decisión por minimización de equivocaciones

- Para clasificación binaria dividimos el espacio de entrada en **regiones de decisión** $\mathcal{R}_0$ y $\mathcal{R}_1$ (para clase 0 y 1)

$$
\begin{aligned}
P(\text{equivocación}) &= P(\mathbf{x} \in \mathcal{R}_0, y = 0) + P(\mathbf{x} \in \mathcal{R}_1, y = 1) \\
&= \int_{\mathcal{R}_0} P(\mathbf{x}, y = 1) + \int_{\mathcal{R}_1} P(\mathbf{x}, y = 0)
\end{aligned}
$$

- $P(\text{equivocación})$ es mínima cuando a cada $\mathbf{x}$ se le asigna la clase k con $P(y = k|\mathbf{x})$ más alta.
- A la frontera entre las dos regiones se le conoce como **frontera de decisión**

Regiones de decisión para clasificación binaria



Tomado de Bishop 2009

Decisión por maximización de aciertos

- Para K clases, es más fácil calcular la probabilidad de acierto

$$
\begin{aligned}
P(\text{acierto}) &= \sum_{k=1}^{K} P(\mathbf{x} \in \mathcal{R}_k, y = k) \\
&= \sum_{k=1}^{K} \int_{\mathcal{R}_k} P(\mathbf{x}, y = k)
\end{aligned}
$$

- La $P(\text{acierto})$ máxima se logra al asignar cada $\mathbf{x}$ a la clase k con mayor $P(y = k|\mathbf{x})$.

Decisión por minimización de pérdida esperada

- La **función de pérdida** $\mathcal{L}$ (función de utilidad en otros contextos), cuantifica el costo de equivocación y acierto
- Por ejemplo, una matriz de pérdida diagnóstico de cáncer sería

$$\mathcal{L} = \begin{array}{c c} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

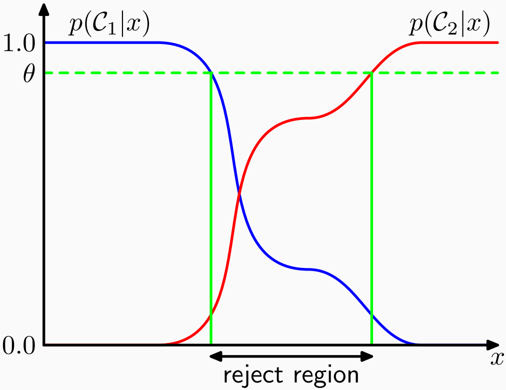
- Tomamos la decisión que minimice la pérdida esperada

$$\mathbb{E}[\mathcal{L}] = \sum_{k=1}^{K} \sum_{i=1}^{n} \int_{\mathcal{R}_i} \mathcal{L}_{ki} P(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

Decisión por opción de rechazo (1)

- Hay regiones en el espacio de entrada donde es incierto tomar decisiones
- En algunas aplicaciones es posible evitar la toma de decisiones en esas regiones (se conoce como **opción de rechazo**)
- Definimos un **región de rechazo** como aquellos valores en los que la probabilidad es menor que cierto umbral

Decisión por región de rechazo (2)

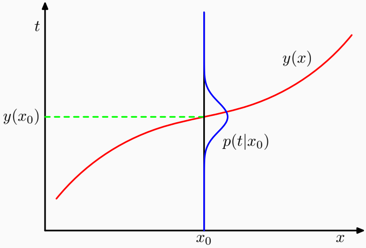


Tomado de Bishop 2009

Teoría de decisión para regresión

- La función de regresión que minimiza la pérdida esperada cuadrática está dada por la media de $P(y|\mathbf{x})$.

$$\begin{aligned}\mathbb{E}[\mathcal{L}] &= \int\int \mathcal{L}(y, \hat{y})P(\mathbf{x}, y)d\mathbf{x}dy \\ &= \int\int (y-\hat{y})^2 P(\mathbf{x}, y)d\mathbf{x}dy = \mathbb{E}_y[y|\mathbf{x}]\end{aligned}$$



Tomado de Bishop 2009