Evaluación, validación y sobre-ajuste

Eduardo Morales, Hugo Jair Escalante

Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica

Agosto, 2015

Introducción

Evaluación de métodos de aprendizaje

- Al desarrollar/implementar un clasificador como parte de algún sistema de toma de decisiones, es crítico evaluar su desempeño.
- La evaluación nos dará evidencia necesaria para anticipar el correcto funcionamiento del sistema.
- Una evaluación sistemática es imprescindible para publicar resultados y avanzar el estado del arte.

Introducción

Evaluación de métodos de aprendizaje

- Diferentes métodos hacen diferentes suposiciones, tienen sesgos y características.
- Con tantas variantes de algoritmos de aprendizaje es crítico evaluar objetivamente su desempeño.
- Tal evaluación también es imprescindible para seleccionar el mejor modelo (optimización de parámetros).

Introducción

Lo que veremos en esta sesión.

- Evaluación y comparación del desempeño de claificadores.
- Selección de parámetros y clasificadores, sobre-ajuste.

Evaluación de métodos de aprendizaje

- ¿Cómo evaluar el desempeño de un clasificador en una tarea dada?.
  - Siguiendo una metodología adecuada.
- ¿Cómo escoger el mejor método para un problema dado?:
  - Usando conocimiento del dominio.
  - Usando conocimiento del aprendizaje computacional.
    - Métodos *informados*.
    - Métodos agnósticos.

Evaluación de métodos de aprendizaje

- ¿Cómo evaluar el desempeño de un clasificador en una tarea dada?.
  - Siguiendo una metodología adecuada.
- ¿Cómo escoger el mejor método para un problema dado?:
  - Usando conocimiento del dominio.
  - Usando conocimiento del aprendizaje computacional.
    - Métodos *informados*.
    - Métodos agnósticos.

Cómo evaluar el desempeño de un clasificador

- ¿Qué evaluar?
- ¿Cómo realizar esta evaluación?
- ¿Qué información se requiere?
- ¿Cómo saber cuál es el mejor clasificador para una tarea dada?
- ¿Qué aspectos son importantes para realizar una validación *justa*?

Cómo evaluar el desempeño de un clasificador

Receta:

- Obtener datos.
- Dividir datos.
- Elegir medida de evaluación.
- Diseñar experimentos.
- Realizar evaluación.
- Ejecutar pruebas estadísticas.
- Reporte y análisis de resultados.

Cómo evaluar el desempeño de un clasificador

Receta:

- Obtener datos.
- Dividir datos.
- Elegir medida de evaluación.
- Diseñar experimentos.
- Realizar evaluación.
- Ejecutar pruebas estadísticas.
- Reporte y análisis de resultados.

Conseguir datos

Convenciones. contamos con un conjunto de datos etiquetado: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{1,\ldots,N}$, $\mathbf{x}_i \in \mathcal{R}^d$, $y_i \in \{-1, 1\}$.

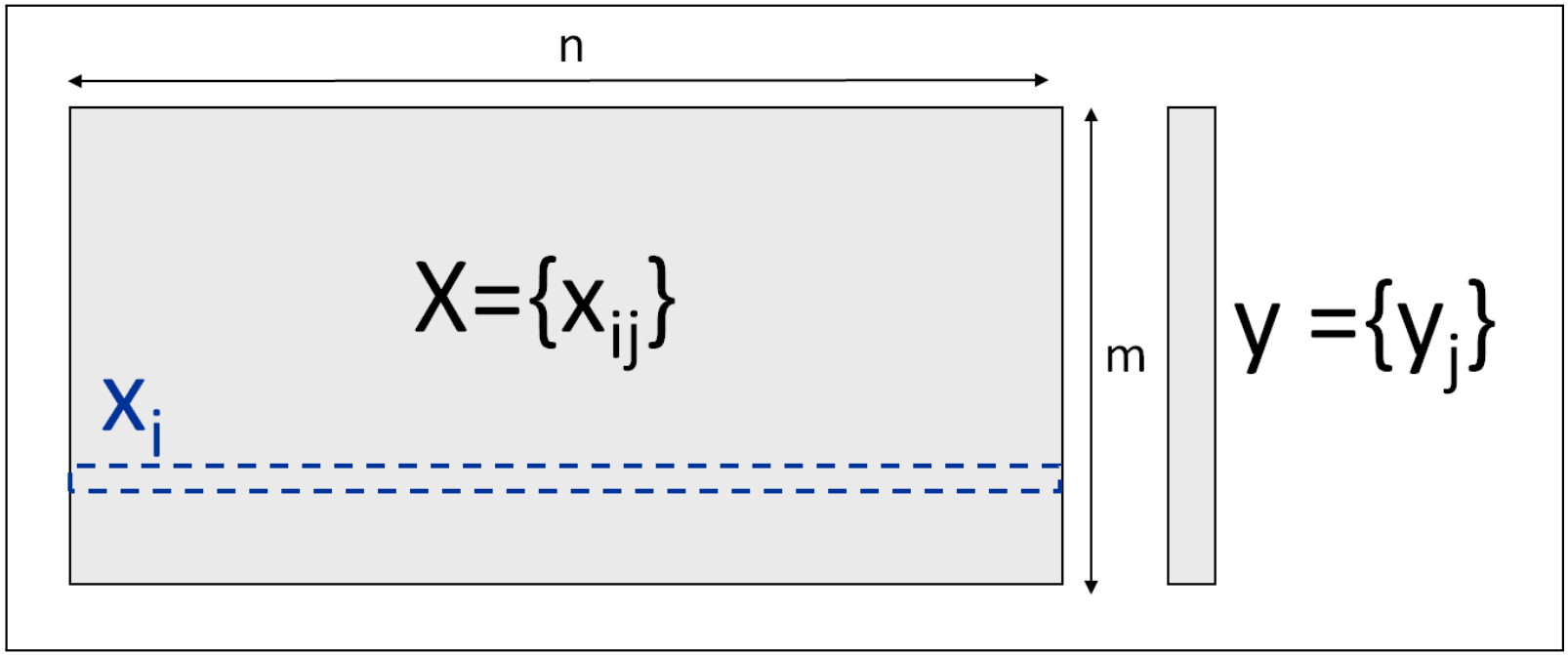


Figura: Datos para aprendizaje supervisado.

Dividir datos

Los datos disponibles se dividen en 3 subconjuntos:

- **Entrenamiento.** Construcción del clasificador.
- **Validación.** Optimización de parámetros.
- **Prueba** Evaluación del clasificador.

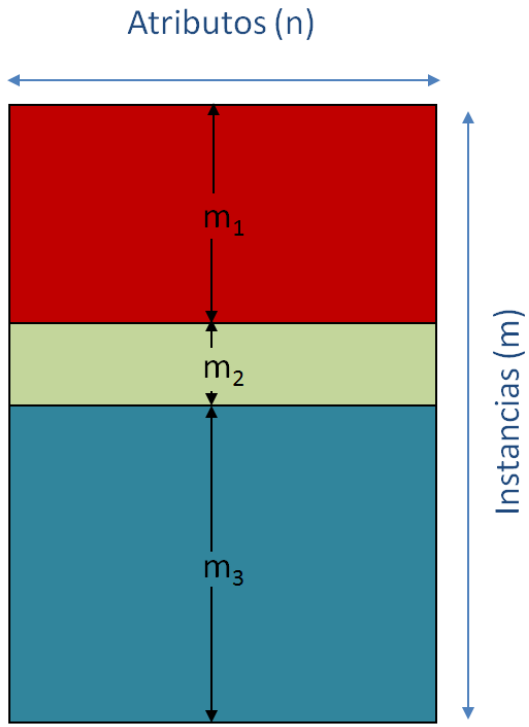


Figura: Partición sugerida.

Dividir datos

Caso de estudio 1: $k-$NN.

- **Entrenamiento.** Las instancias para clasificación se toman de m_1.
- **Validación.** Elegir el mejor valor de k para el clasificador, evaluar el desempeño en m_2
- **Prueba** El clasificador con el valor de k seleccionado se evalúa en m_3

Dividir datos

Caso de estudio 2: Redes neuronales.

- **Entrenamiento.** Una red neuronal con parámetros fijos se entrena en m_1.
- **Validación.** Los parámetros de la red (e.g., números neuronas, tasa de aprendizaje) se escogen evaluando el desempeño en m_2
- **Prueba** La red con los mejores parámetros se evalúa en m_3

¿Qué evaluar?

- Generalmente nos interesa maximizar la exactitud o minimizar el error de predicción.
- Sea $\mathcal{T} = \{(\mathbf{x}_i^T, y_i^T)\}_{1,\ldots,T}$ el conjunto de instancias en la partición m_3 (i.e., de prueba).
- Sea $\hat{y}_i^T = \hat{f}(\mathbf{x}_i^T)$, la predicción del modelo bajo estudio (i.e., $\hat{f}$) en la instancia de prueba i, con $i = 1, \ldots, T$.
- ¿Cómo deben ser las predicciones de $\hat{f}$ con respecto a y_i^T?

¿Qué evaluar?

- Idealmente $\hat{y}_i^{\mathcal{T}}$ debería ser igual a $y_i^{\mathcal{T}}$ para cualquier $\mathbf{x}$ (ojo: no solo para $\mathcal{T}$).
- ¿Cuando la salida es real?
- ¿Cuando la salida es categórica?

¿Qué evaluar?

- Idealmente $\hat{y}_i^{\mathcal{T}}$ debería ser igual a $y_i^{\mathcal{T}}$ para cualquier $\mathbf{x}$ (ojo: no solo para $\mathcal{T}$).
- **Cuando la salida es real:** *Minimizar la "distancia" entre* $\hat{y}_i^{\mathcal{T}}$, $y_i^{\mathcal{T}}$

¿Qué evaluar?

Medidas comúnmente usadas para evaluación en salidas continuas.

- Root Mean-Squared Error:

$$RMSE(\hat{f}) = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i^T - \hat{y}_i^T)^2}$$

- Mean Absolute Error:

$$MAE(\hat{f}) = \frac{1}{T}\sum_{i=1}^{T}|y_i^T - \hat{y}_i^T|$$

¿Qué evaluar?

- Idealmente $\hat{y}_i^T$ debería ser igual a y_i^T para cualquier **x** (ojo: no solo para $\mathcal{T}$).
- **Cuando la salida es categórica:** *Maximizar el número de instancias para las cuales* $\hat{y}_i^T = y_i^T$

¿Qué evaluar?

Medidas comúnmente usadas para evaluación en salidas categóricas.

- Exactitud (*accuracy*).

$$ACC(\hat{f}) = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{1}_{y_i^T = \hat{y}_i^T})$$

- Error (0-1 loss)

$$ERR(\hat{f}) = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{1}_{y_i^T \neq \hat{y}_i^T})$$

¿Qué evaluar?

¿Qué pasa con las medidas anteriores cuando se tienen datos deslabanceados?

- Del total de instancias T, 90 % son de la clase 1 y 10 % de la clase -1.
- Un clasificador trivial que siempre predice la clase 1 tendrá una efectividad del 90 %

Muchas veces es necesario analizar con mayor detalle los resultados.

¿Qué evaluar?

Una medida comúnmente usada para evaluación en salidas categóricas en datos desbalanceados:

- Balanced error rate

$$BER(\hat{f}) = \frac{E_- + E_+}{2}$$

donde E_- / E_+ es la tasa de error en instancias de la clase negativa/positiva, respectivamente.

¿Qué evaluar?

Tabla de contingencia y/o matriz de confusión: útil para estimar medidas de evaluación de la clasificación.

	$y_i^T = 1$	$y_i^T = -1$
$\hat{y}_i^T = 1$	***TP***	***FP***
$\hat{y}_i^T = -1$	***FN***	***TN***

Figura: Matriz de confusión 2-clases.

- **TP:** ciertos positivos.
- **FP:** falsos positivos.
- **TN:** ciertos negativos.
- **FN:** falsos negativos.

¿Qué evaluar?

- **Sensitividad (recall).** Tasa de ciertos positivos.

$$Sens(\hat{f}) = \frac{TP}{TP + FN}$$

- **Especifidad.** Tasa de ciertos negativos.

$$Esp(\hat{f}) = \frac{TN}{TN + FP}$$

	$y_i^T = 1$	$y_i^T = -1$
$\hat{y}_i^T = 1$	***TP***	***FP***
$\hat{y}_i^T = -1$	***FN***	***TN***

Figura: Matriz de confusión 2-clases.

¿Qué evaluar?

- **Recall.** Del total de positivos (resp. negativos) cuántas clasifico correctamente.

$$Rec_+(\hat{f}) = \frac{TP}{TP + FN} \vee Rec_-(\hat{f}) = \frac{TN}{TN + FP}$$

- **Precisión.** Del total de predicciones positivas (resp. negativas) cuántas clasifico correctamente.

$$Prec_+(\hat{f}) = \frac{TP}{TP + FP} \vee Prec_-(\hat{f}) = \frac{TN}{TN + FN}$$

- **Medida** f_β. Compromiso entre precisión y cobertura, usualmente $\beta = 1$.

$$f_\beta(\hat{f}) = \frac{2 \times Prec \times Rec}{Prec + Rec}$$

¿Qué evaluar?

Generalización de la medida f_1 para más de dos clases.

- **Macro-promedio.** Se calcula la medida f_1 para cada una de las clases del problema, y se promedian los resultados. Mismo peso a todas las clases.
- **Micro-promedio.** Calcula TP, FP, TN, FN para todas las categorías y se calcula la medida f_1. Mismo peso a todos las instancias.

Tabla de contingencia y/o matriz de confusión: útil para estimar medidas de evaluación de la clasificación.

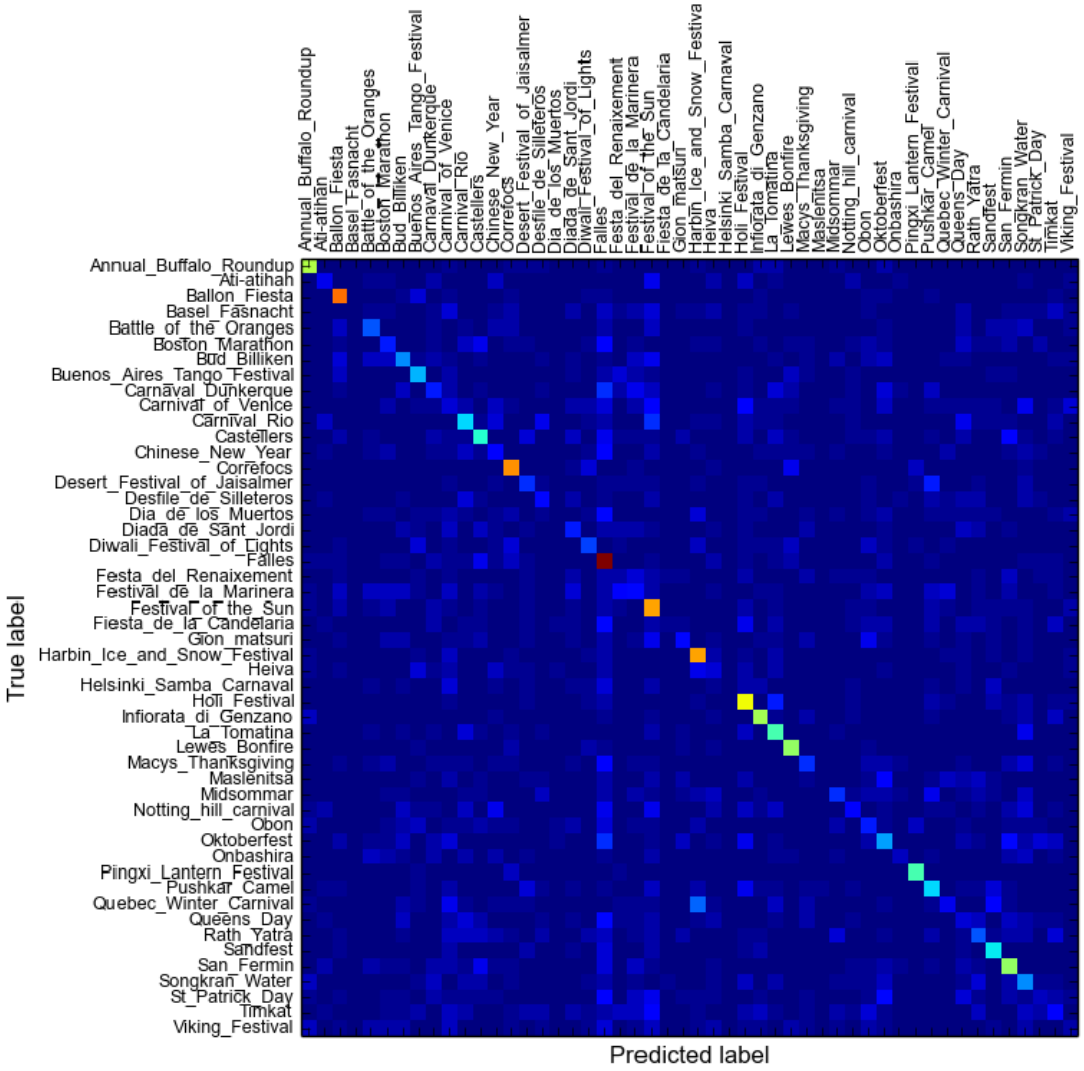


Figura: Matriz de confusión k-clases.

¿Qué evaluar?: salidas reales

- Todas las medidas anteriores asumen una clasificación *dura* (cada instancia se asocia a una clase).
- *Algunos clasificadores*, además de proveer una clasificación dura, pueden proveer un valor de confianza de la predicción.
  - Ejemplo: Clasificadores probabilistas. Por cada clase tenemos una probabilidad.
  - Ejemplo: $k-$NN. La confianza de predicción para cada una de las clases puede ser la distancia de la instancia hacia la instancia más cercana de cada clase.

¿Qué evaluar?: salidas reales

Vizualización de las confianzas de predicción de un clasificador Random-Forest:

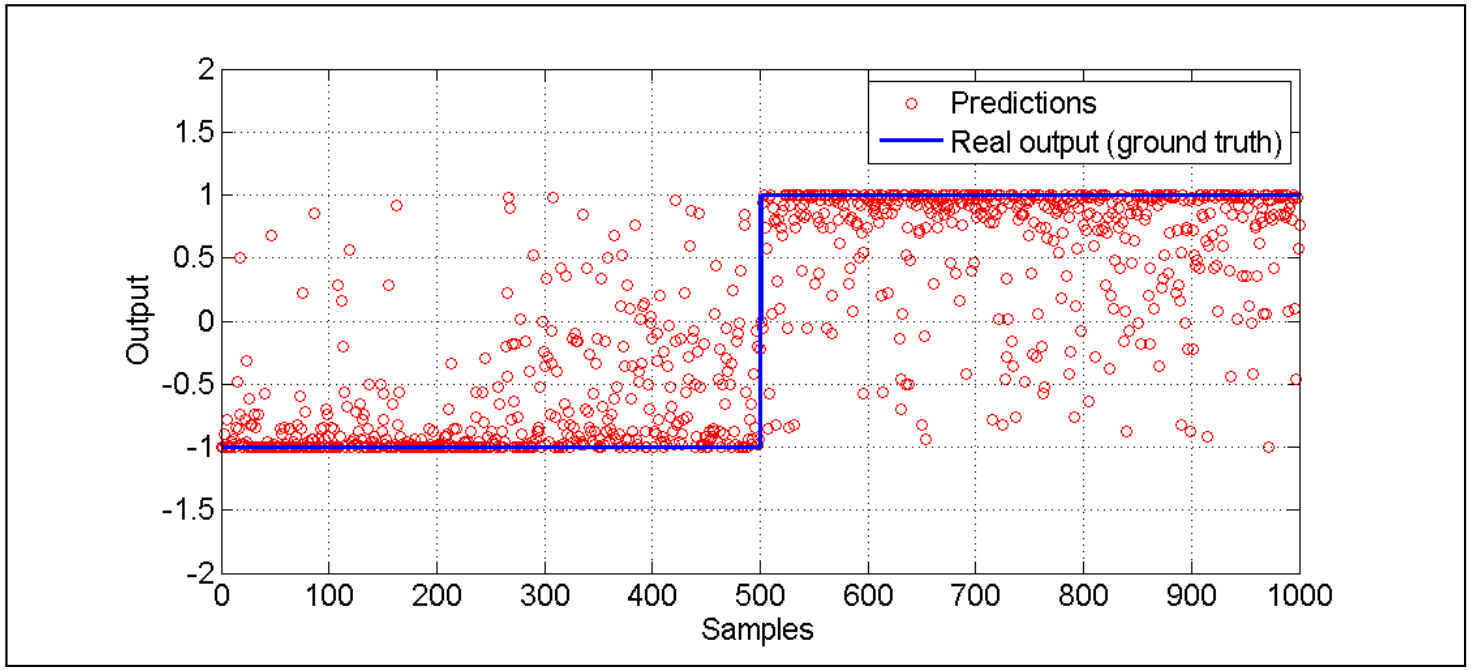


Figura: Predicciones de un clasificador RF.

¿Qué evaluar?: salidas reales

Vizualización de las confianzas de predicción de un clasificador Random-Forest: **clasificación dura**.

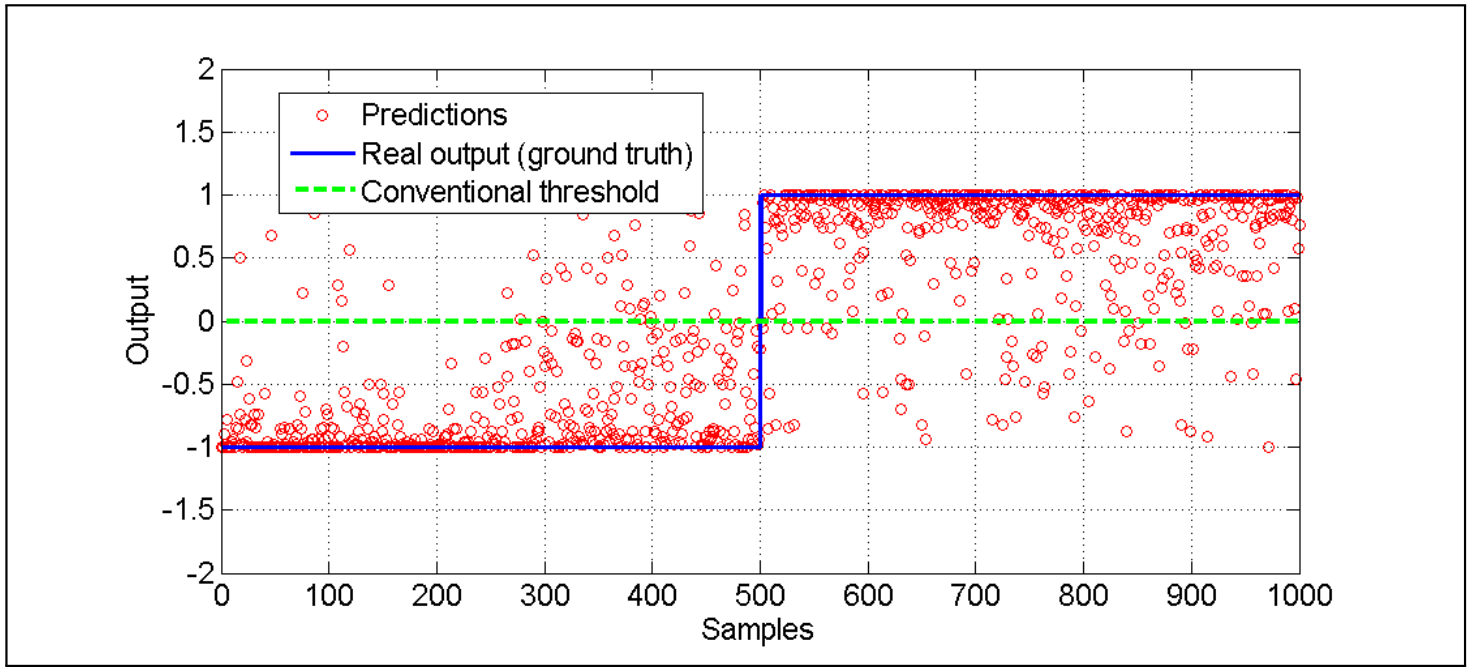


Figura: Predicciones de un clasificador RF.

¿Qué evaluar?: salidas reales

Vizualización de las confianzas de predicción de un clasificador naïve Bayes:

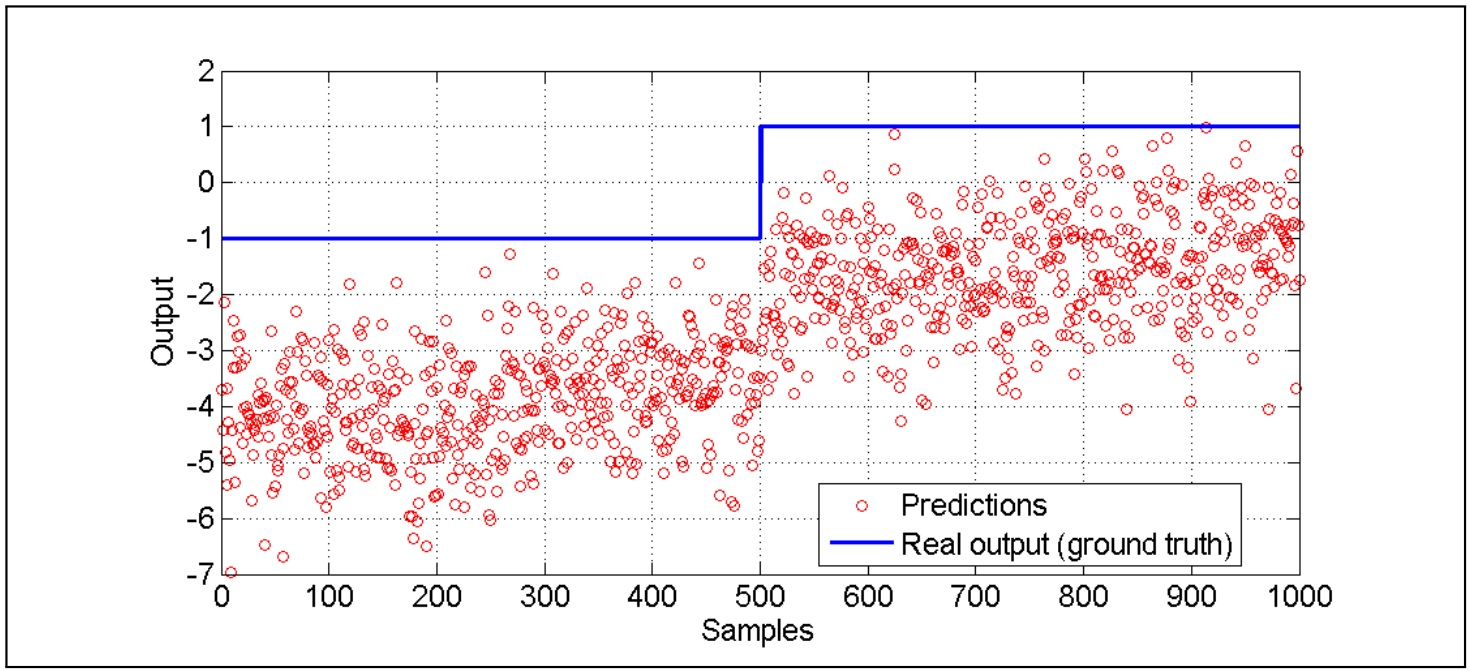


Figura: Predicciones de un clasificador.

¿Qué evaluar?: salidas reales

Vizualización de las confianzas de predicción de un clasificador naïve Bayes: **Clasificación dura**

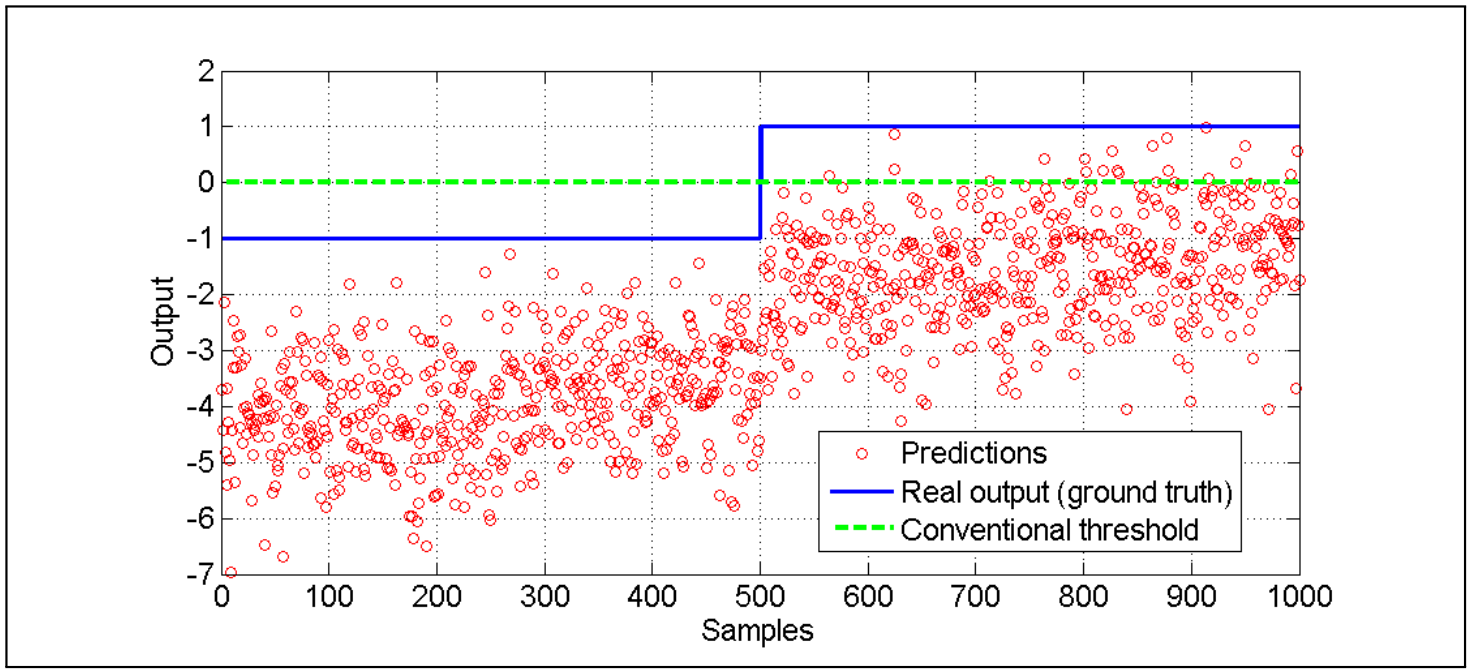


Figura: Predicciones de un clasificador.

¿Qué evaluar?: salidas reales

Realmente, ¿qué tan malo es este clasificador?

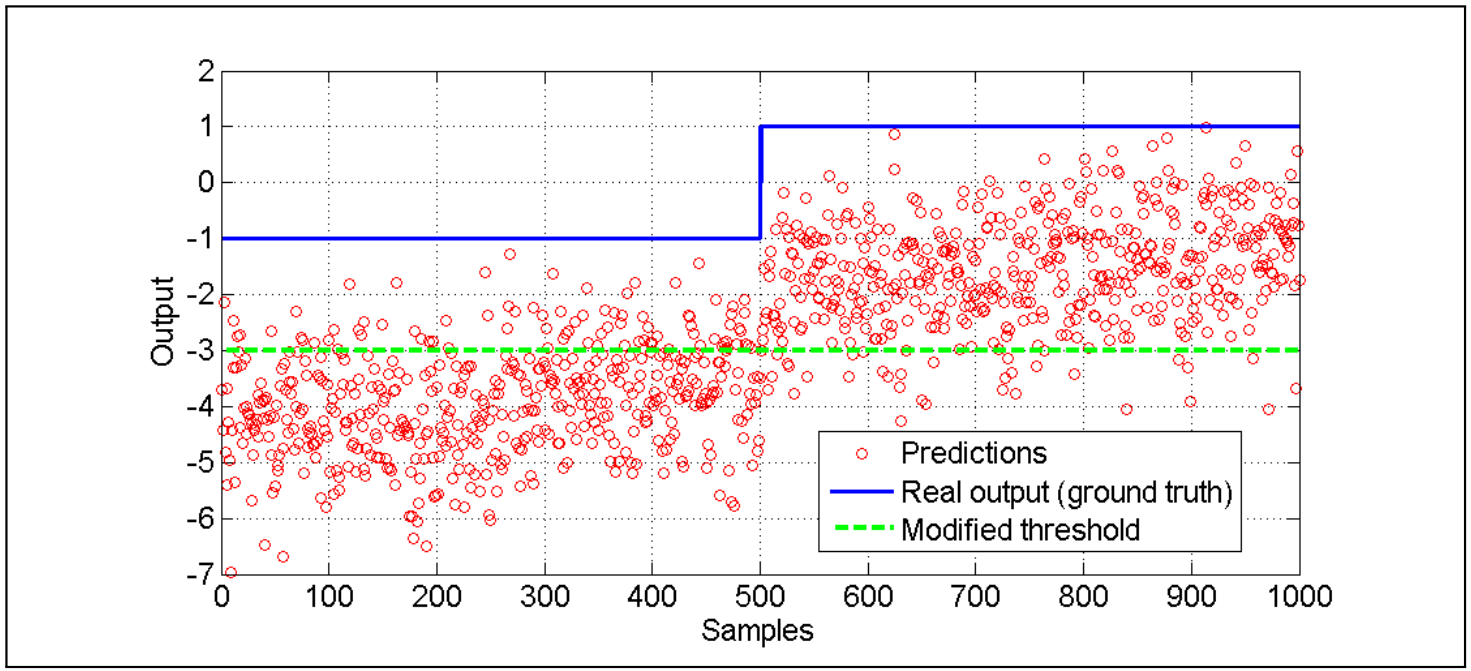


Figura: Predicciones de un clasificador.

¿Qué evaluar?: salidas reales

Vizualización de las confianzas de predicción de un clasificador naïve Bayes: **Clasificación dura**

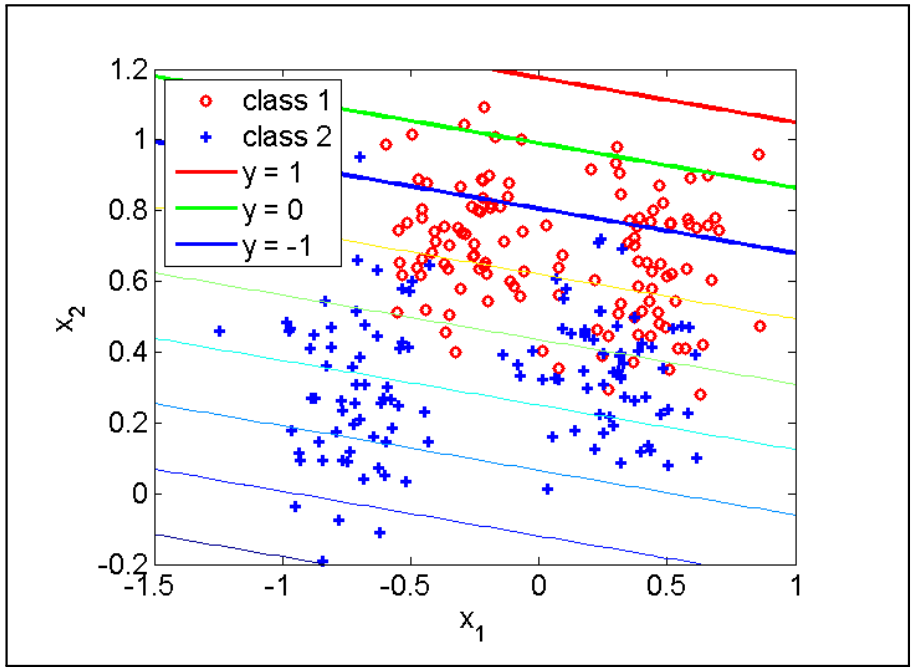


Figura: Predicciones de un clasificador.

¿Qué evaluar?: salidas reales

Realmente, ¿qué tan malo es este clasificador?

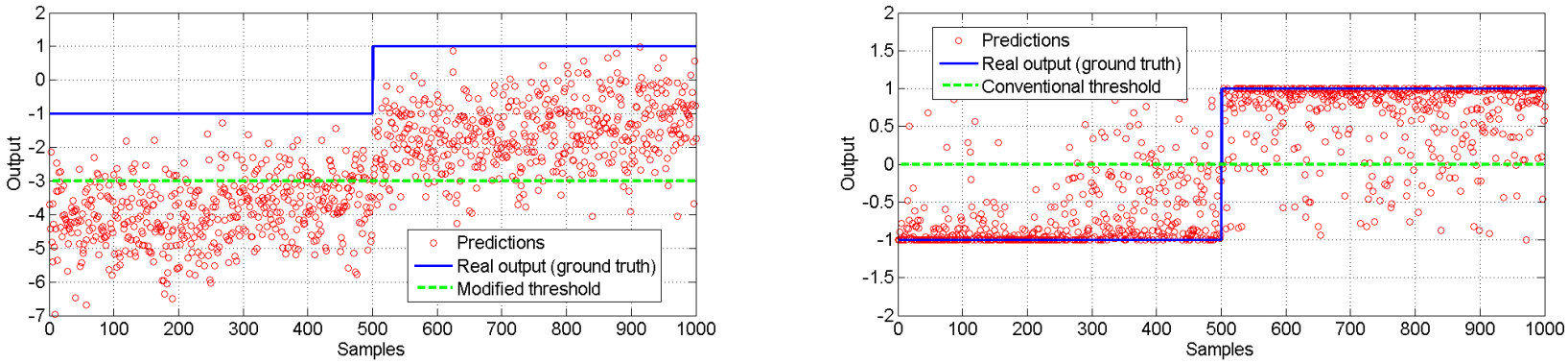


Figura: Comparación predicciones clasificador.

¿Qué evaluar?: salidas reales

- Es importante evaluar las predicciones de un clasificador independientemente del umbral de predicción.
- ¿Cómo hacerlo?

¿Qué evaluar?: salidas reales

La curva ROC (Receiving Operator Characteristic).
Para un umbral dado sobre $\hat{f}(\mathbf{x})$ se obtiene un punto de la curva ROC:

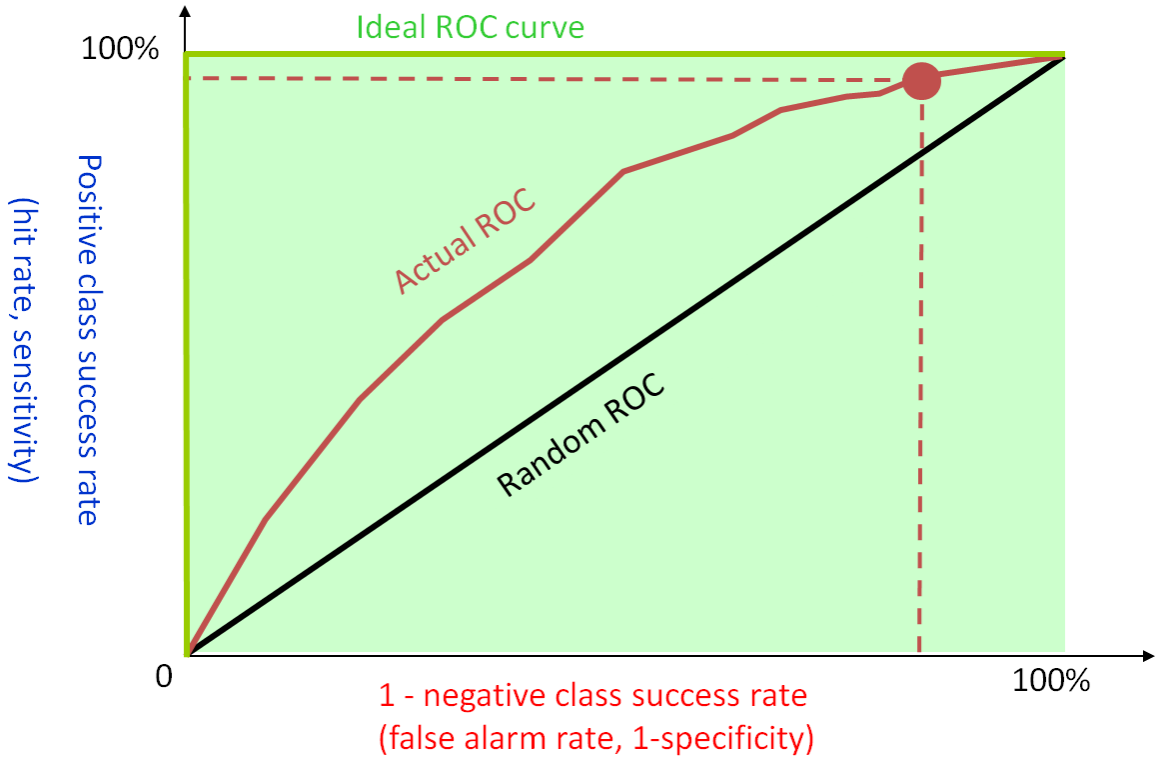


Figura: Curva ROC

¿Qué evaluar?: salidas reales

A menudo es complicado/subjetivo comparar curvas, ¿puede un solo número resumir una curva?

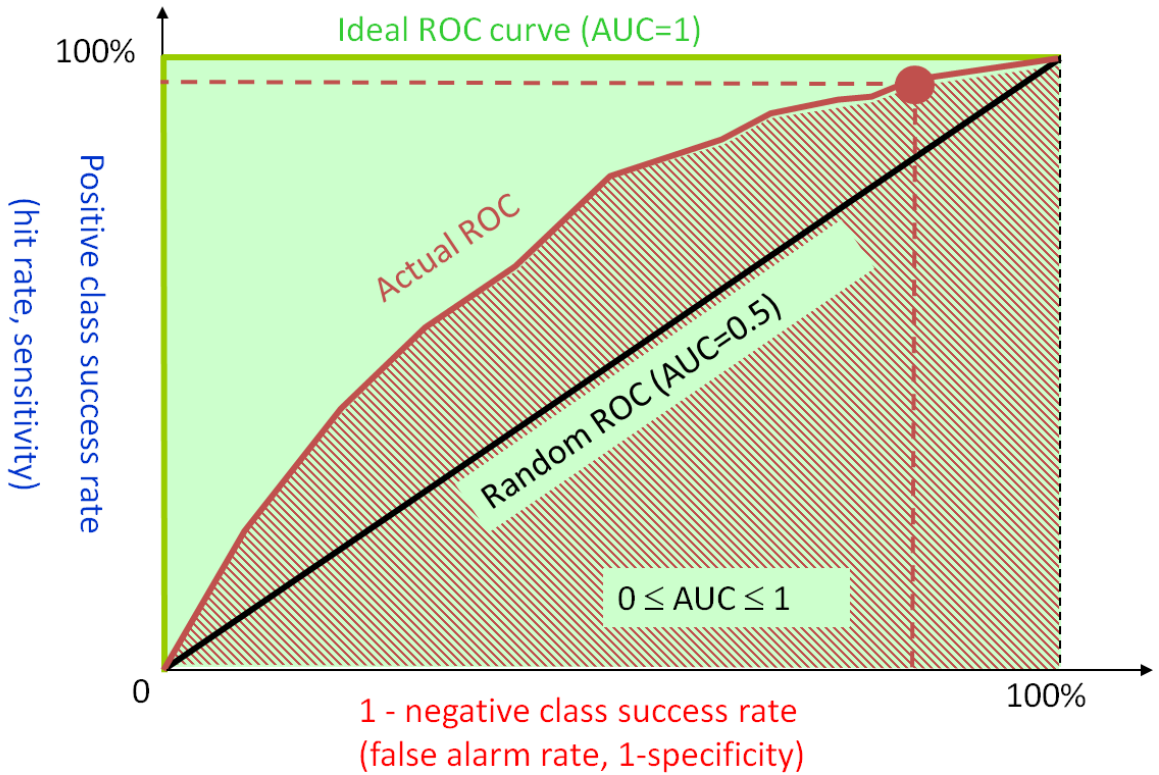


Figura: Curva ROC y AUC.

Comparación de clasificadores

Sean $\hat{f}_a$ y $\hat{f}_b$ dos clasificadores a comparar:

- Evaluar ambos clasificadores usando exactamente el mismo marco de evaluación (mismas división de datos, mismos recursos computacionales, medida de evaluación, etc.).
- Es común realizar la evaluación en varios conjuntos de datos, con diferentes características.
- Realizar pruebas de hipótesis para determinar la significancia estadísticas de la diferencia en efectividad.

Comparación de clasificadores

Se suele recurrir a la estadística inferencial para dar soporte a resultados experimentales:

- Sean $\mathcal{D}_1, \ldots, \mathcal{D}_k$, $k-$ conjuntos de datos que habrán de utilizarse para la comparación.
- Sean $r_1^a, \ldots, r_k^a$ y $r_1^b, \ldots, r_k^b$ los valores[1] de la medida de evaluación en los $k-$conjuntos de datos obtenidos por los clasificadores $\hat{f}_a$ y $\hat{f}_b$, respectivamente.
- Queremos, evaluar si los resultados obtenidos por $\hat{f}^a$ son *estadísticamente* diferentes a los obtenidos por $\hat{f}^b$

[1] Es común comparar promedios de medidas y no el resultado de una única medición por conjunto de datos.

Comparación de clasificadores

Data set	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$
Breast-cancer	$36.98^{+}_{-}0.08$	$33.59^{+}_{-}0.12$
Diabetes	$26.07^{+}_{-}0.03$	$25.37^{+}_{-}0.02$
Flare-solar	$32.87^{+}_{-}0.02$	$32.65^{+}_{-}0.01$
German	$28.65^{+}_{-}0.02$	$28.28^{+}_{-}0.02$
Heart	$19.50^{+}_{-}0.19$	$17.35^{+}_{-}0.06$
Image	$3.58^{+}_{-}0.01$	$2.50^{+}_{-}0.01$
Splice	$13.94^{+}_{-}0.99$	$9.46^{+}_{-}0.25$
Thyroid	$10.84^{+}_{-}0.39$	$5.98^{+}_{-}0.06$
Titanic	$29.94^{+}_{-}0.00$	$29.60^{+}_{-}0.00$

Cuadro: Ejemplo, resultado de dos métodos a comparar.

Comparación de clasificadores

Preguntas típicas:

- Supera significativamente $\hat{\mathbf{f}}^{\mathbf{b}}$ a $\hat{\mathbf{f}}^{\mathbf{a}}$ en el conjunto de datos X?
- En cuantos conjuntos de datos la diferencia de desempeño es significativa?
- Sobre todos los conjuntos de datos, qué clasificador obtiene el menor error?, es significativa la diferencia?

Data set	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$
Breast-cancer	$36.98^{+}_{-}0.08$	$33.59^{+}_{-}0.12$
Diabetes	$26.07^{+}_{-}0.03$	$25.37^{+}_{-}0.02$
Flare-solar	$32.87^{+}_{-}0.02$	$32.65^{+}_{-}0.01$
German	$28.65^{+}_{-}0.02$	$28.28^{+}_{-}0.02$
Heart	$19.50^{+}_{-}0.19$	$17.35^{+}_{-}0.06$
Image	$3.58^{+}_{-}0.01$	$2.50^{+}_{-}0.01$
Splice	$13.94^{+}_{-}0.99$	$9.46^{+}_{-}0.25$
Thyroid	$10.84^{+}_{-}0.39$	$5.98^{+}_{-}0.06$
Titanic	$29.94^{+}_{-}0.00$	$29.60^{+}_{-}0.00$

Cuadro: Ejemplo, resultado de dos métodos a comparar.

Comparación de clasificadores

Preguntas típicas:

- Supera significativamente $\hat{\mathbf{f}}^{\mathbf{b}}$ a $\hat{\mathbf{f}}^{\mathbf{a}}$ en el conjunto de datos X?
- En cuantos conjuntos de datos la diferencia de desempeño es significativa?
- Sobre todos los conjuntos de datos, qué clasificador obtiene el menor error?, es significativa la diferencia?

Data set	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$
Breast-cancer	36.98 ± 0.08	33.59 ± 0.12
Diabetes	26.07 ± 0.03	25.37 ± 0.02
Flare-solar	32.87 ± 0.02	32.65 ± 0.01
German	28.65 ± 0.02	28.28 ± 0.02
Heart	19.50 ± 0.19	17.35 ± 0.06
Image	3.58 ± 0.01	2.50 ± 0.01
Splice	13.94 ± 0.99	9.46 ± 0.25
Thyroid	10.84 ± 0.39	5.98 ± 0.06
Titanic	29.94 ± 0.00	29.60 ± 0.00

Cuadro: Ejemplo, resultado de dos métodos a comparar.

Comparación de clasificadores

Significancia estadística. En estadística, se dice que una diferencia es estadísticamente significante/significativa, cuando no es posible que se presente por azar.

- No se refiere a que se trata de algo “importante”

Comparación de clasificadores

Comparación de dos clasificadores ($\hat{\mathbf{f}}^{\mathbf{a}}$ y $\hat{\mathbf{f}}^{\mathbf{b}}$) en el conjunto de datos X.

- Se tienen múltiples resultados en diferentes particiones del mismo conjunto de datos X.
- Generalmente, se quiere determinar si el promedio de las diferencias proviene de una distribución normal con media cero, o no.
- ¿Cómo hacerlo?

Data set	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$	dif.
r_1	11.28	10.31	0.97
r_2	11.98	11.50	0.48
r_3	11.18	9.32	1.86
r_4	10.25	9.99	0.26
r_5	11.22	9.12	2.10
r_6	10.56	9.87	0.69
r_7	11.23	10.54	0.69
r_8	10.43	10.01	0.42
r_9	11.22	10.45	0.77
r_10	10.76	10.12	0.64
avg.	11.01 ± 0.511	10.12 ± 0.663	0.88 ± 0.61

Cuadro: Ejemplo, resultado de dos métodos a comparar.

Comparación de clasificadores

Comparación de dos clasificadores ($\hat{\mathbf{f}}^{\mathbf{a}}$ y $\hat{\mathbf{f}}^{\mathbf{b}}$) en N conjuntos de datos.

- Por cada conjunto de datos se tienen resultados de ambos métodos.
- Generalmente, se quiere determinar si el promedio de las diferencias proviene de una distribución normal con media cero, o no.
- ¿Cómo hacerlo?

Data set	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$	$BER(\hat{\mathbf{f}}^{\mathbf{a}})$
Breast-cancer	$36.98^{+}_{-}0.08$	$33.59^{+}_{-}0.12$
Diabetes	$26.07^{+}_{-}0.03$	$25.37^{+}_{-}0.02$
Flare-solar	$32.87^{+}_{-}0.02$	$32.65^{+}_{-}0.01$
German	$28.65^{+}_{-}0.02$	$28.28^{+}_{-}0.02$
Heart	$19.50^{+}_{-}0.19$	$17.35^{+}_{-}0.06$
Image	$3.58^{+}_{-}0.01$	$2.50^{+}_{-}0.01$
Splice	$13.94^{+}_{-}0.99$	$9.46^{+}_{-}0.25$
Thyroid	$10.84^{+}_{-}0.39$	$5.98^{+}_{-}0.06$
Titanic	$29.94^{+}_{-}0.00$	$29.60^{+}_{-}0.00$

Cuadro: Ejemplo, resultado de dos métodos a comparar.

Comparación de clasificadores

Lecturas *Altamente* recomendadas:

- T.G. Dietterich. **Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms** *Neural Computation*, Vol. 10:1895–1924, 1998.
- J. Demsar. **Statistical Comparisons of Classifiers over Multiple Data sets.** *Journal of Machine Learning Research*, Vol. 7:1–30, 2006.
- S. García, F. Herrera. **An Extension to "Statistical Comparisons of Classifiers over Multiple Data sets" for all Pairwise Comparisons.** *Journal of Machine Learning Research*, Vol. 9:2677–2694, 2008.

Evaluación, validación y sobre-ajuste

Evaluación de métodos de aprendizaje

- ¿Cómo evaluar el desempeño de un clasificador en una tarea dada?.
  - Siguiendo una metodología adecuada.
- ¿Cómo escoger el mejor método para un problema dado?:
  - Usando conocimiento del dominio.
  - Usando conocimiento del aprendizaje computacional.
    - Métodos *informados*.
    - Métodos agnósticos.

Evaluación, validación y sobre-ajuste

Evaluación de métodos de aprendizaje

- ¿Cómo evaluar el desempeño de un clasificador en una tarea dada?.
  - Siguiendo una metodología adecuada.
- ¿Cómo escoger el mejor método para un problema dado?:
  - Usando conocimiento del dominio.
  - Usando conocimiento del aprendizaje computacional.
    - Métodos *informados*.
    - Métodos agnósticos.

Evaluación, validación y sobre-ajuste

Usando el conocimiento del dominio.
Categorización de textos.

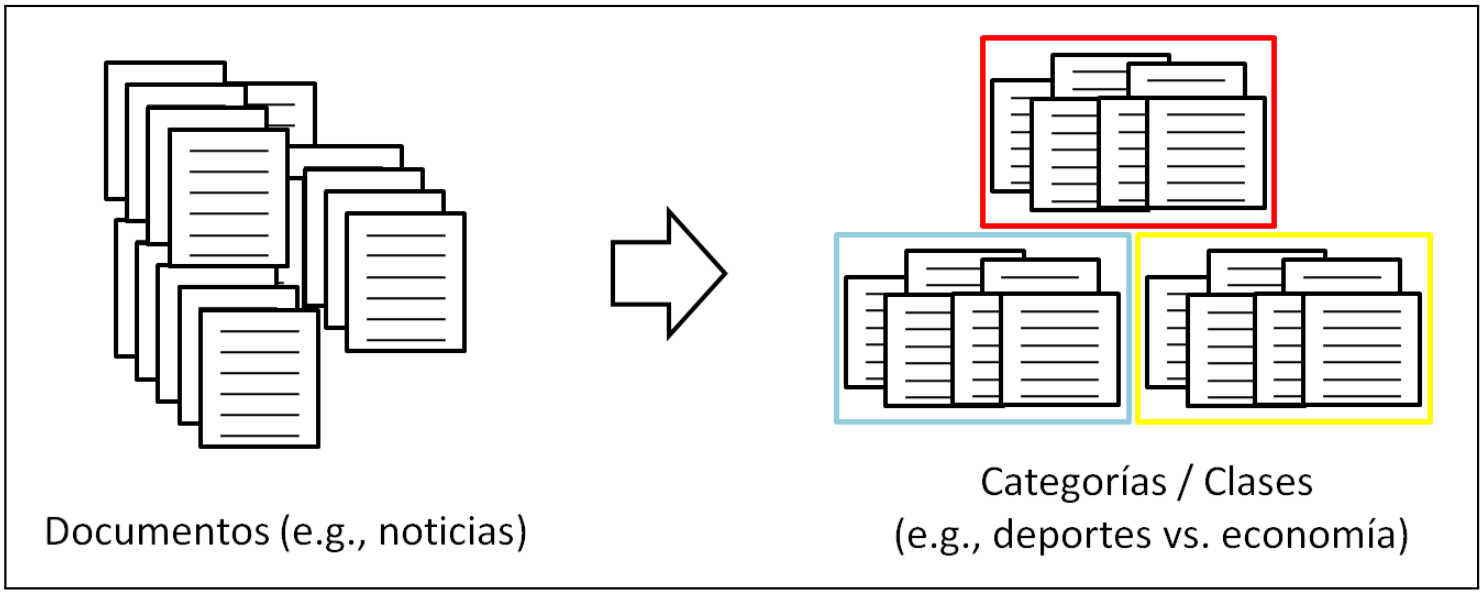


Figura: Clasificación de textos.

Evaluación, validación y sobre-ajuste

Usando el conocimiento del dominio.
Categorización de textos.

- *Qué:* Preprocesamiento a datos, generación de representaciones apropiadas, reducción de atributos, clasificadores recomendados: naïve Bayes, SVM.
- *Por qué:* Abundancia de información irrelevante, muchos datos faltantes (*sparse representation*), muchas dimensiones, representaciones mixtas, generalmente linealmente separable.

Evaluación, validación y sobre-ajuste

Usando el conocimiento del dominio.
Clasificación de acciones en video.

Figura: Reconocimiento de acciones.

Evaluación, validación y sobre-ajuste

Usando el conocimiento del dominio.
Clasificación de acciones en video.

- *Qué:* Transformación a escala de grises, trayectorias densas, descriptores TRJ, HOG, HOF, MBH, representación de vectores de Fisher, clasificadores recomendados: SVM, con kernel de intersección de bines.
- *Por qué:* IDT captura información altamente discriminativa (espacio-temporal), FVs modelan la incertidumbre en descriptores, muchas dimensiones, kernel apropiado para histogramas.

Evaluación, validación y sobre-ajuste

Estrategias de aprendizaje computacional.

- **Métodos informados.** Explotan características de los métodos de clasificación y aprendizaje (e.g., KRR). Permiten seleccionar parámetros para modelos específicos de clasificación.
- **Métodos agnósticos.** Métodos de caja negra, se define un criterio de *efectividad* y se intenta optimizar. Útiles para selección de entre variantes de diferente naturaleza.

Selección de modelo

Las medidas de evaluación vistas en la sesión anterior proveen un estimado de la efectividad de un modelo/clasificador $\hat{f}$.

- ¿Cómo seleccionar los mejores (híper-) parámetros para un clasificador dado?
- ¿Cómo seleccionar un clasificador de un conjunto de opciones?

Selección de modelo

Las medidas de evaluación vistas en la sesión anterior proveen un estimado de la efectividad de un modelo/clasificador $\hat{f}$.

- ¿Cómo seleccionar los mejores (híper-) parámetros para un clasificador dado?
- ¿Cómo seleccionar un clasificador de un conjunto de opciones?

Diferencia entre parámetro e híper-parámetro

En aprendizaje computacional:

- **Parámetro.** Parámetros son las variables (y/o sus valores) que se "aprenden" a partir de los datos. E.g., parámetros en $k-$NN?
- **Híper-parámetro.** Son las variables (y/o sus valores) de un modelo, clasificador, función, que deben especificarse antes de aprender los parámetros. E.g., híper-parámetros en $k-$NN?

Selección de modelo

Grid-search. Método básico para selección de híper-parámetros.
Si el modelo f tiene híper-parámetros $\Theta = \{\theta_1, \ldots, \theta_p\}$:

1. Discretizar el rango de valores que pueden tomar $\theta_{1,\ldots,p}$.
2. a cada posible combinación de híper-parámetros Θ':
   - Entrenar $f_{\Theta'}$ en m_1 (datos de entrenamiento)
   - Evaluar el desempeño de $\hat{f}_{\Theta'}$ en m_2 (datos de validación)
3. Seleccionar la mejor configuración de híper-parámetros Θ^* para f.

Selección de modelo

Grid-search. Método básico para selección de híper-parámetros.

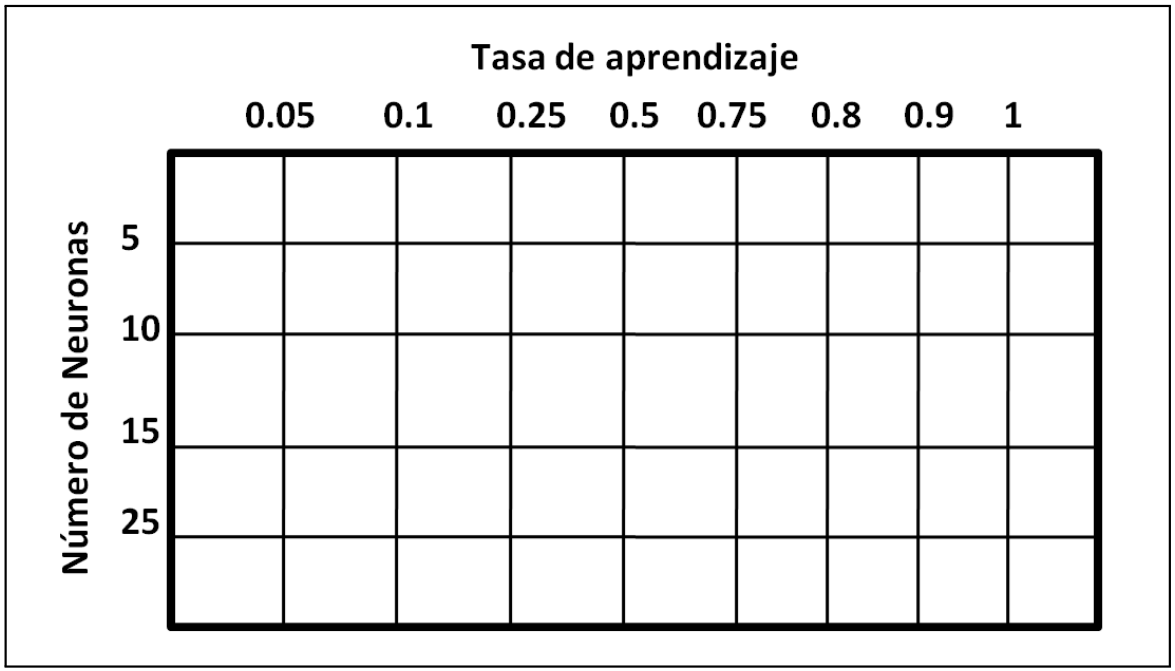


Figura: Grid search.

Selección de modelo

Grid-search. Método básico para selección de híper-parámetros.

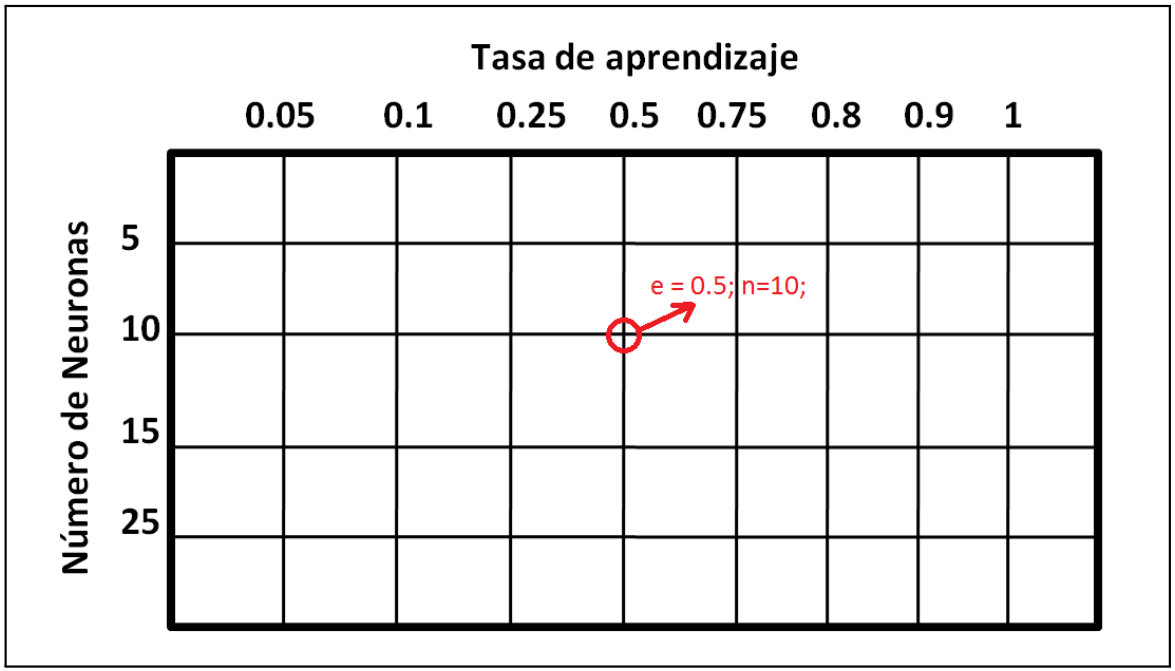


Figura: Grid search.

Selección de modelo

Grid-search. Método básico para selección de híper-parámetros.

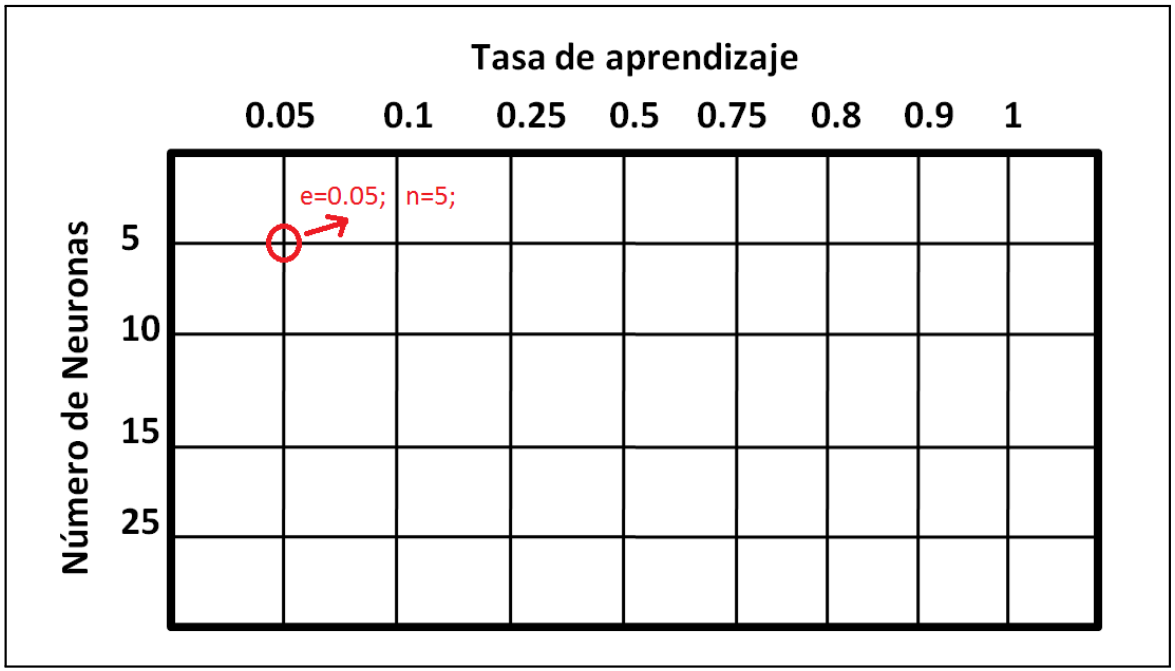


Figura: Grid search.

Selección de modelo

Grid-search. Método básico para selección de híper-parámetros.

- ¿Qué tan fina debe ser la discretización?
- ¿Discretización uniforme?
- ¿Qué pasa cuand p es muy grande?

Alternativa: Usar otra estrategia de búsqueda/optimización.

Selección de modelo

Las medidas de evaluación vistas en la sesión anterior proveen un estimado de la efectividad de un modelo/clasificador $\hat{f}$.

- ¿Cómo seleccionar los mejores (híper-) parámetros para un clasificador dado?
- ¿Cómo seleccionar un clasificador de un conjunto de opciones?

Selección de modelo

Diferentes niveles en selección de modelo:

- **Optimización de parámetros.** Dado un clasificador, optimizar sus híper-parámetros.
- **Selección de clasificador.** Dado un conjunto de clasificadores, seleccionar el mejor para un problema.
- **Selección de modelo completo.** Dado un toolbox de aprendizaje computacional, selecciona el mejor modelo posible que se pueda generar.

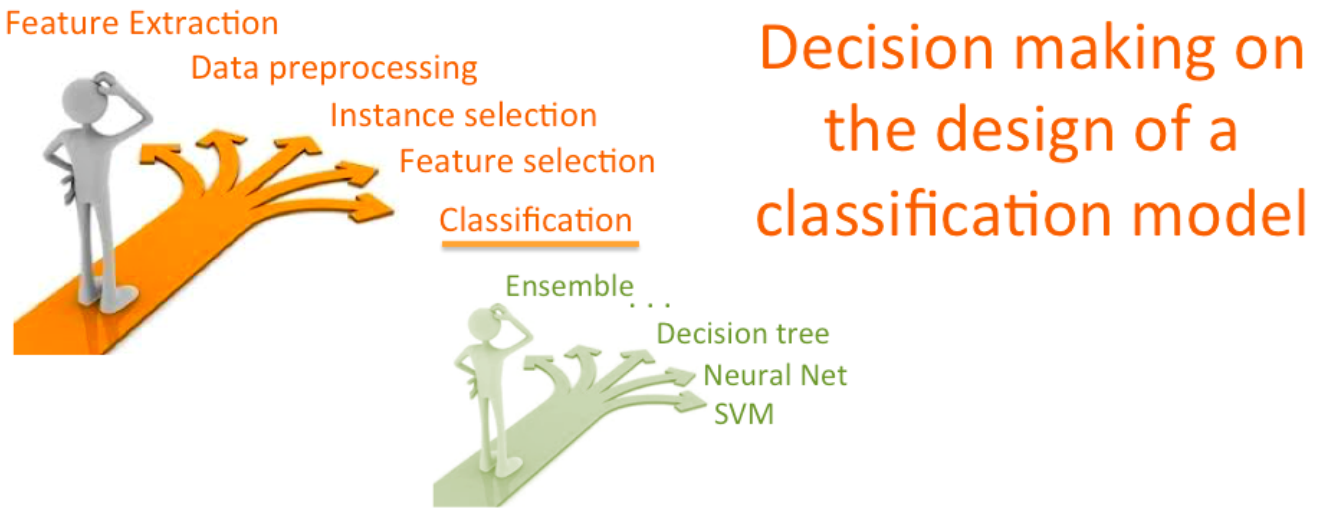
Selección de modelo

Diferentes niveles en selección de modelo:

Decision making on the design of a classification model

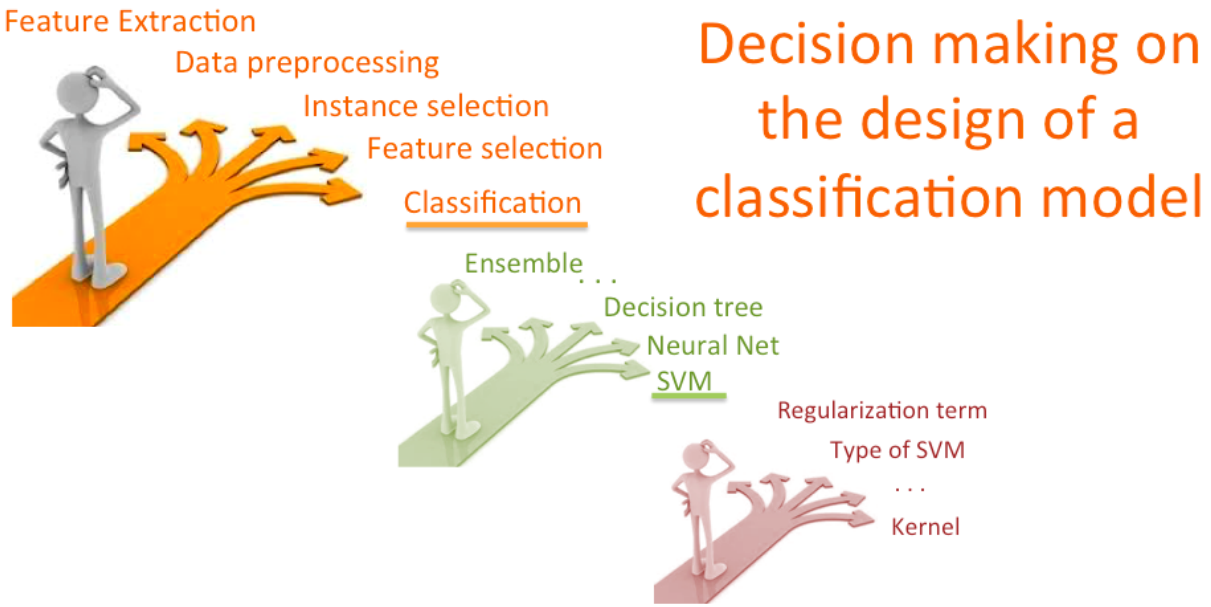
Selección de modelo

Diferentes niveles en selección de modelo:



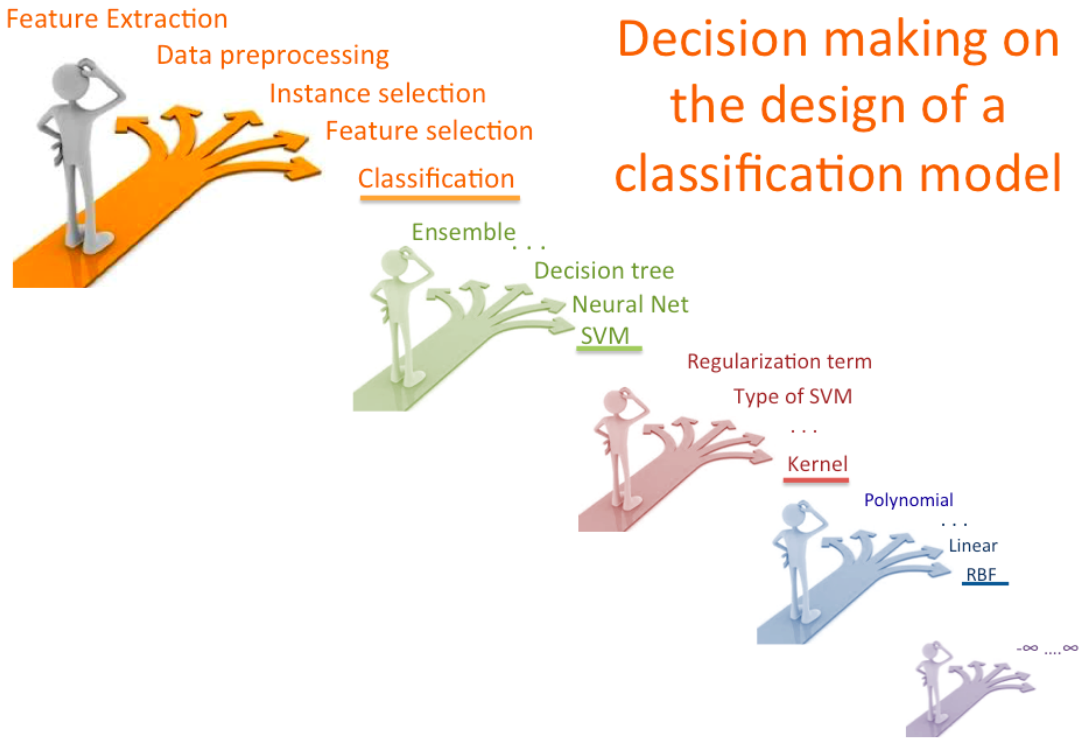
Selección de modelo

Diferentes niveles en selección de modelo:



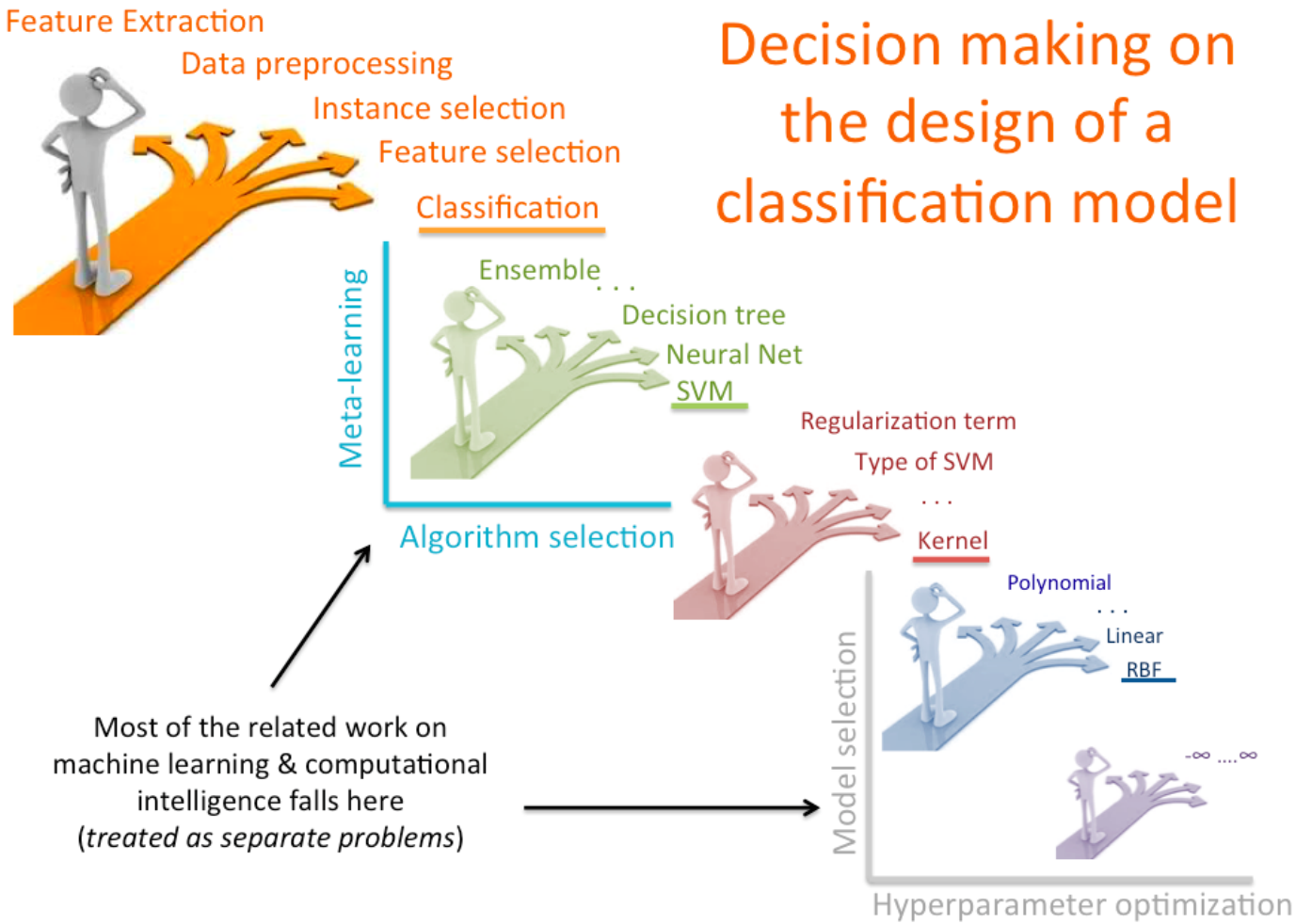
Selección de modelo

Diferentes niveles en selección de modelo:



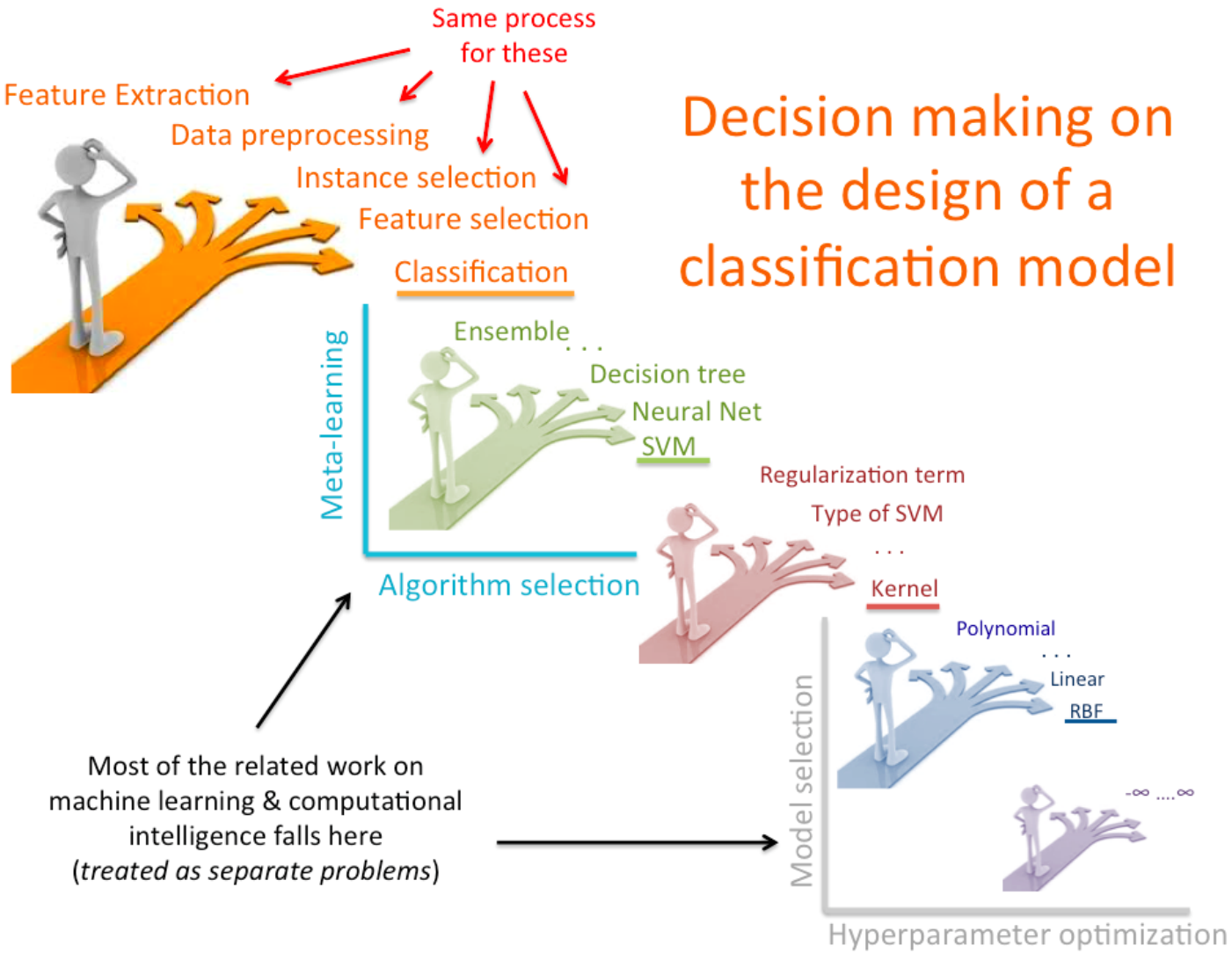
Selección de modelo

Diferentes niveles en selección de modelo:



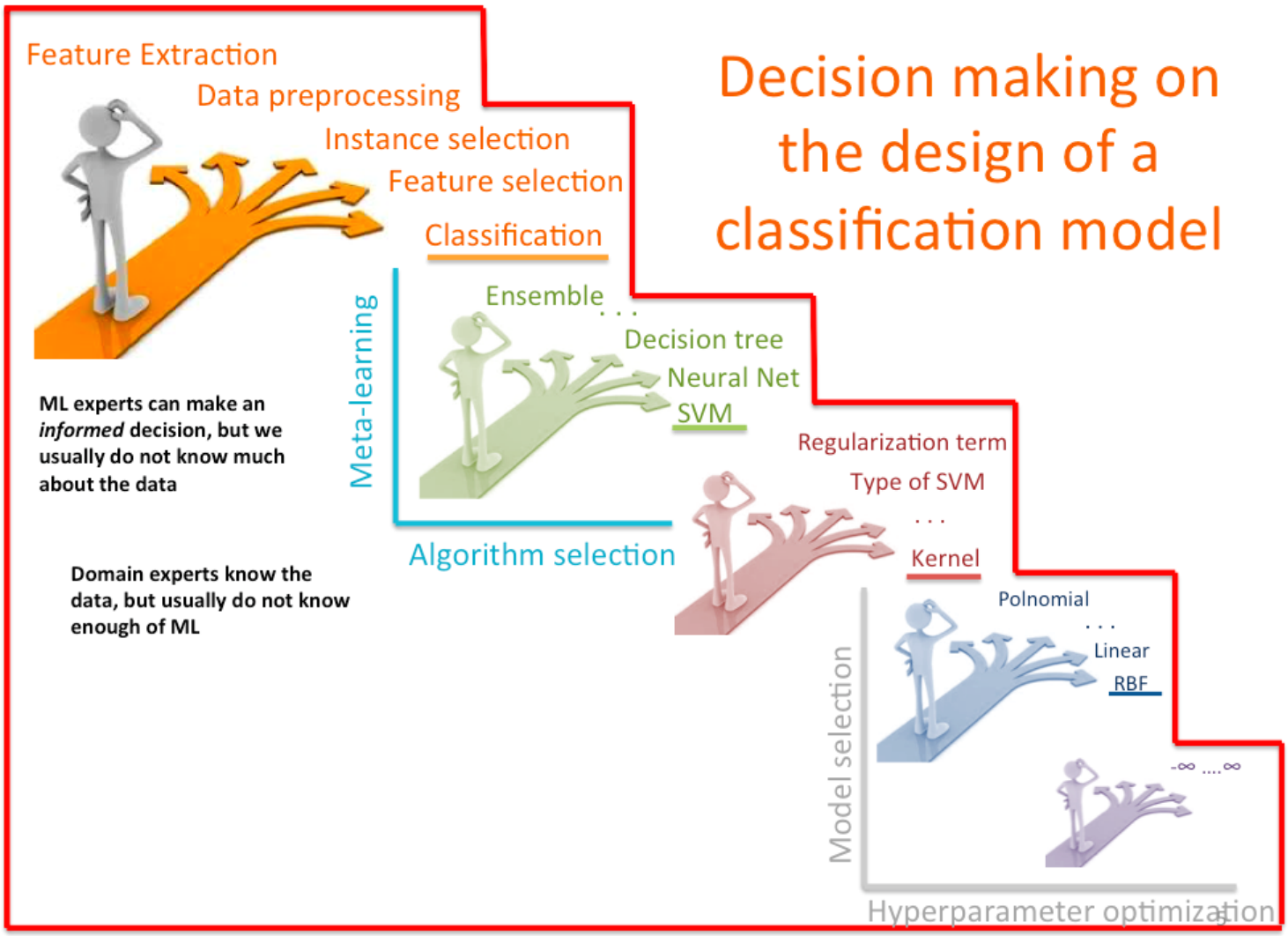
Selección de modelo

Diferentes niveles en selección de modelo:



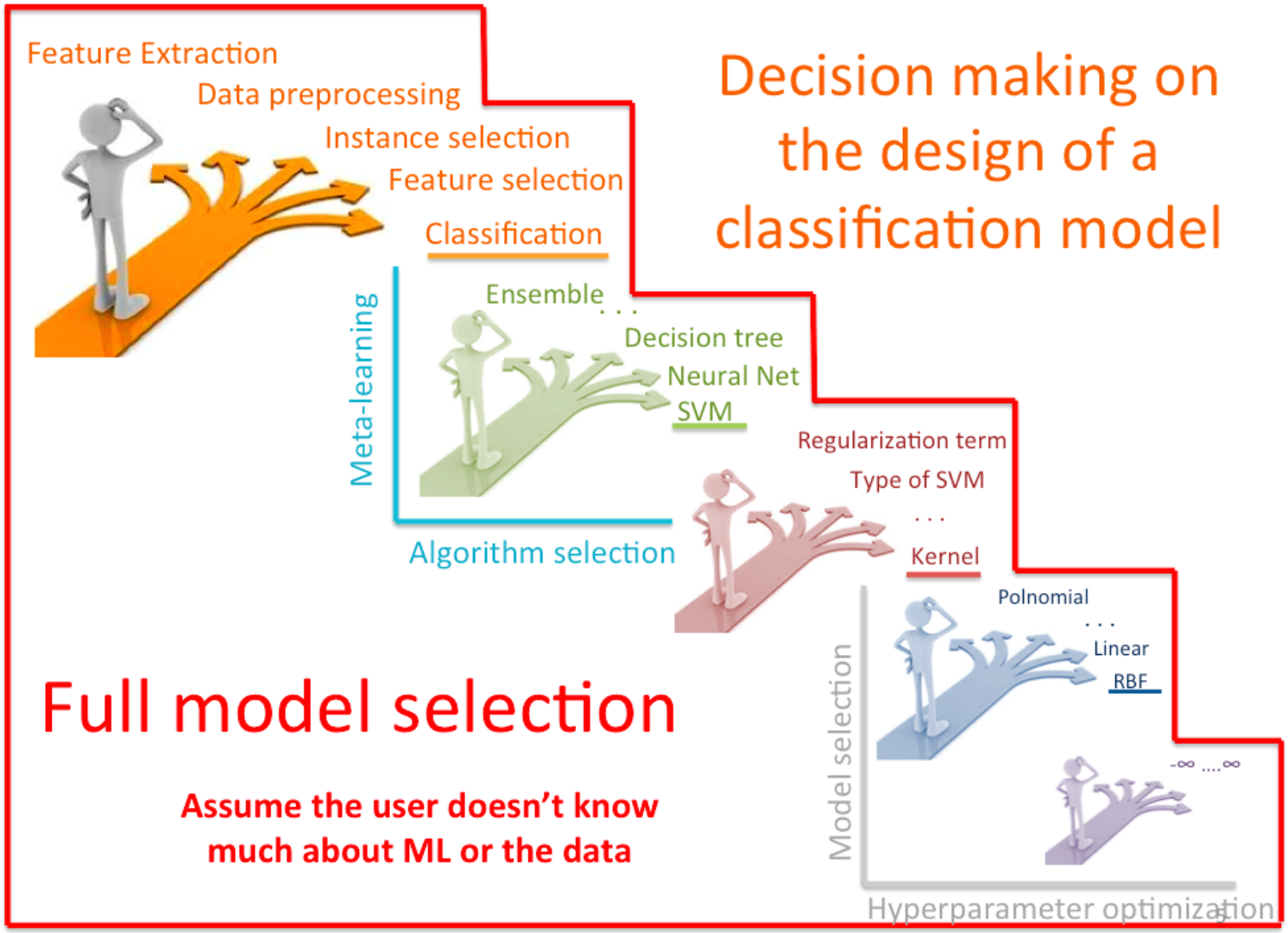
Selección de modelo

Diferentes niveles en selección de modelo:



Selección de modelo

Diferentes niveles en selección de modelo:



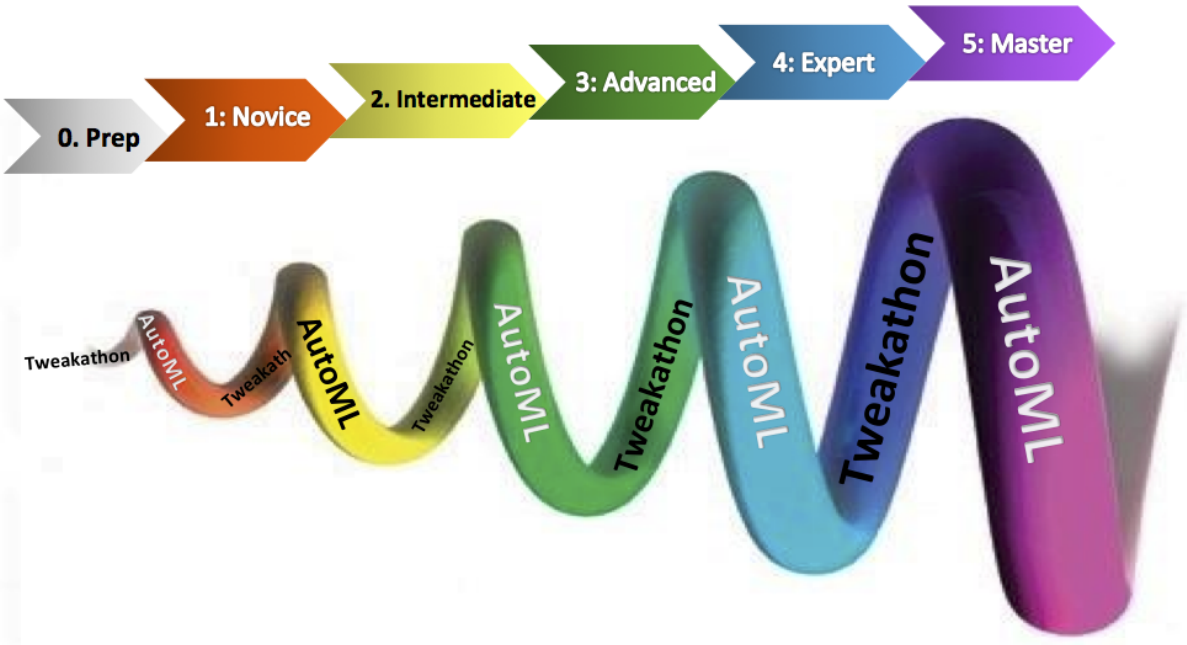
Selección de modelo

Métodos de caja negra.

- Seleccionar un criterio a optimizar (e.g., exactitud, error, AUC).
- Seleccionar una estrategia de evaluación (e.g., $k - fold$ CV).
- Seleccionar método de optimización.

Selección de modelo completo

Tendencias: Automatic Machine Learning



https://www.codalab.org/competitions/2321

Selección de modelo

Problemas en selección de modelo:

- Riesgo de sobre-ajustar el modelo a los datos.
- Problema de optimización computacionalmente costoso.
- Problema altamente complejo con muchos factores de aleatoriedad.

Error de generalización

Hasta ahora, hemos asumido que: se entrena en m_1, se valida en m_2 y se evalúa en m_3, ¿por qué?

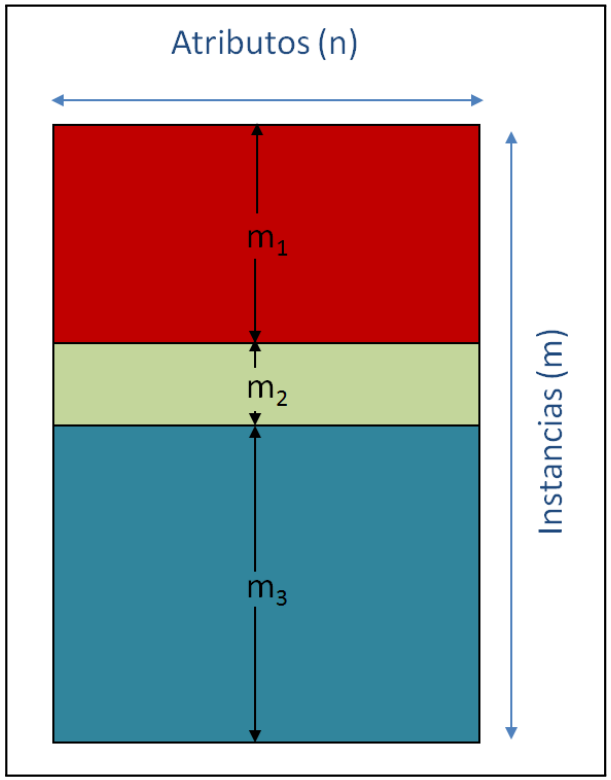


Figura: Datos para aprendizaje supervisado.

- **Error de generalización.** Dada una muestra finita de datos (i.e., $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{1,\ldots,N}$), queremos obtener el clasificador $\hat{f}$ que minimiza el error de clasificación en cualquier muestra de datos que son i.i.d. i.e., minimizar

$$E[L(Y, \hat{f}(X))]$$

donde L es una función de perdida, y X, Y muestreos aleatoriamente de su distribución conjunta.
¿Problema?

El error de entrenamiento no es un buen estimado del error de prueba:

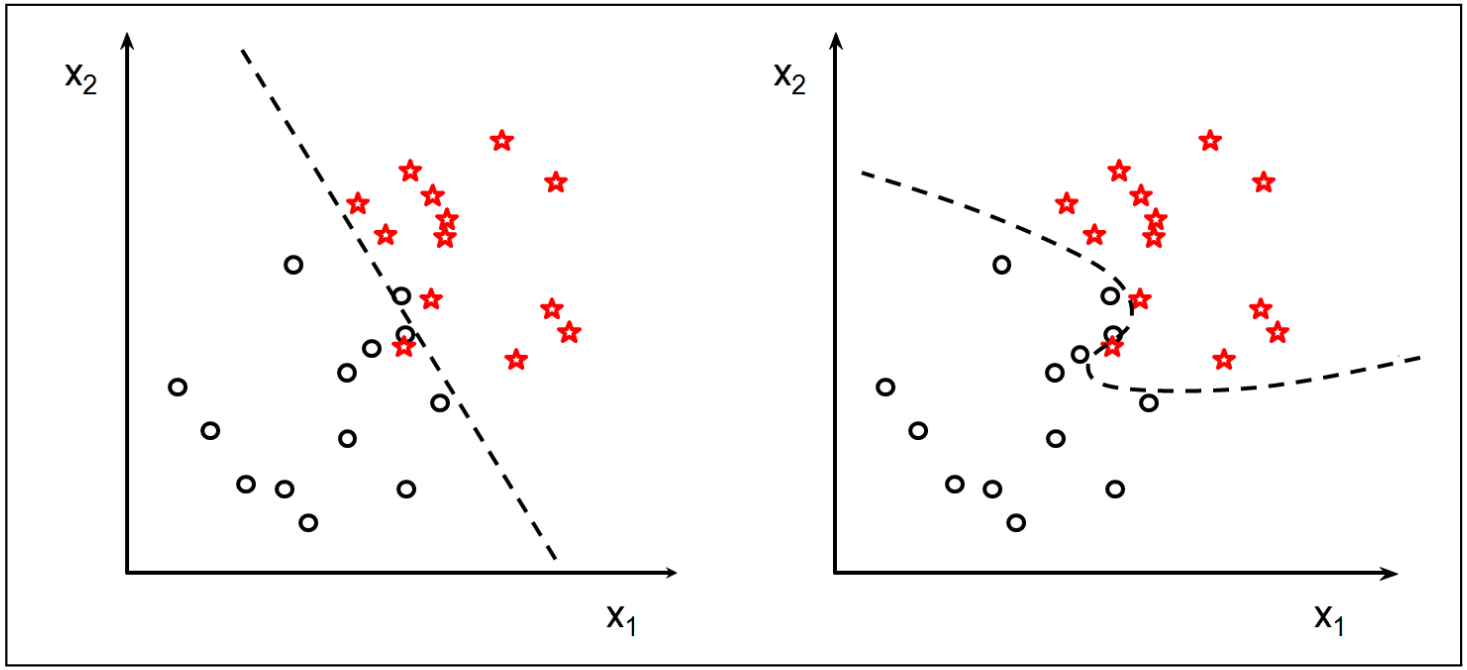


Figura: Capacidad de generalización.

El error de entrenamiento no es un buen estimado del error de prueba:

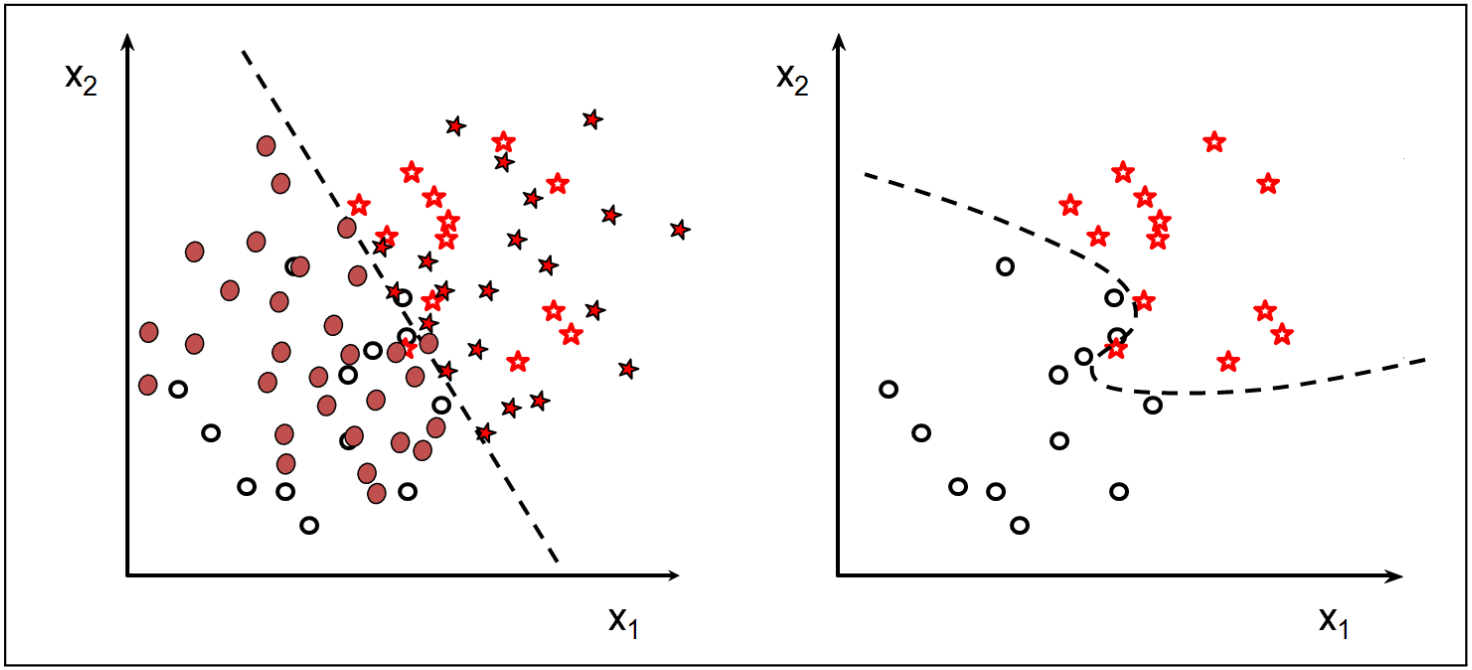


Figura: Capacidad de generalización.

El error de entrenamiento no es un buen estimado del error de prueba:

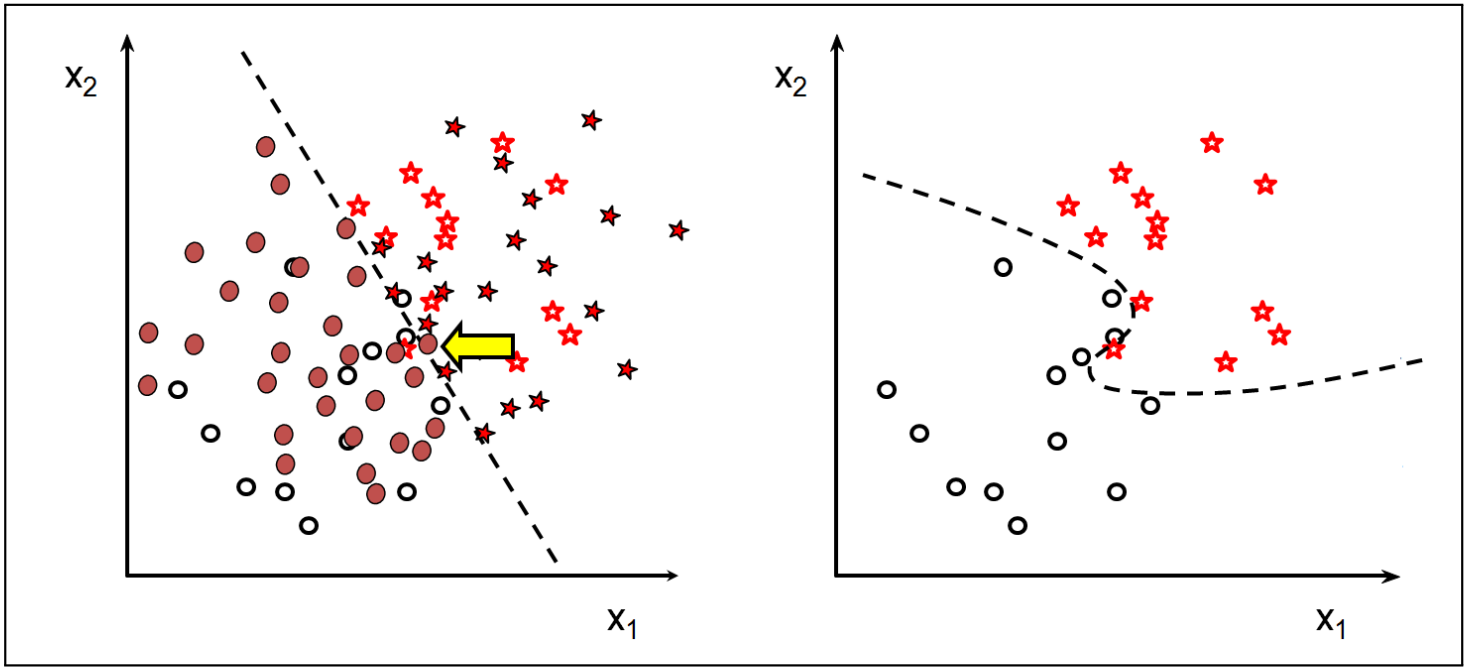


Figura: Capacidad de generalización.

El error de entrenamiento no es un buen estimado del error de prueba:

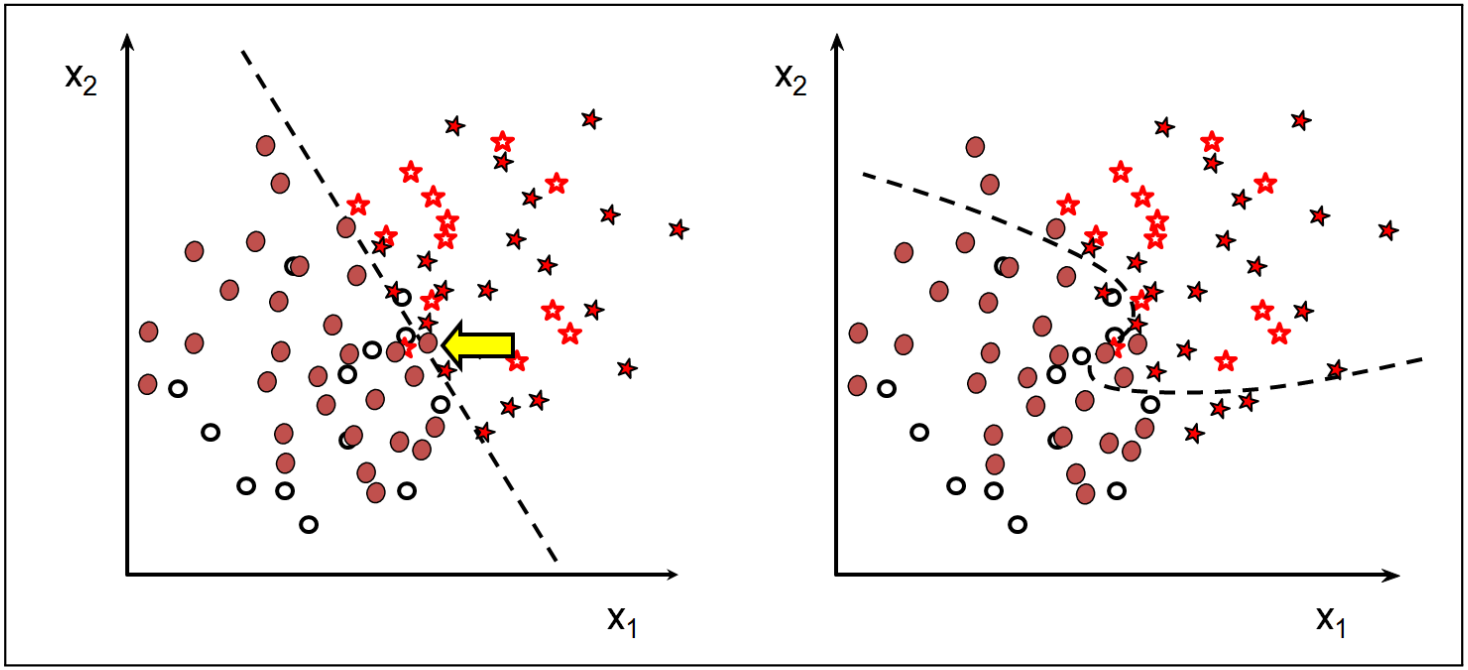


Figura: Capacidad de generalización.

El error de entrenamiento no es un buen estimado del error de prueba:

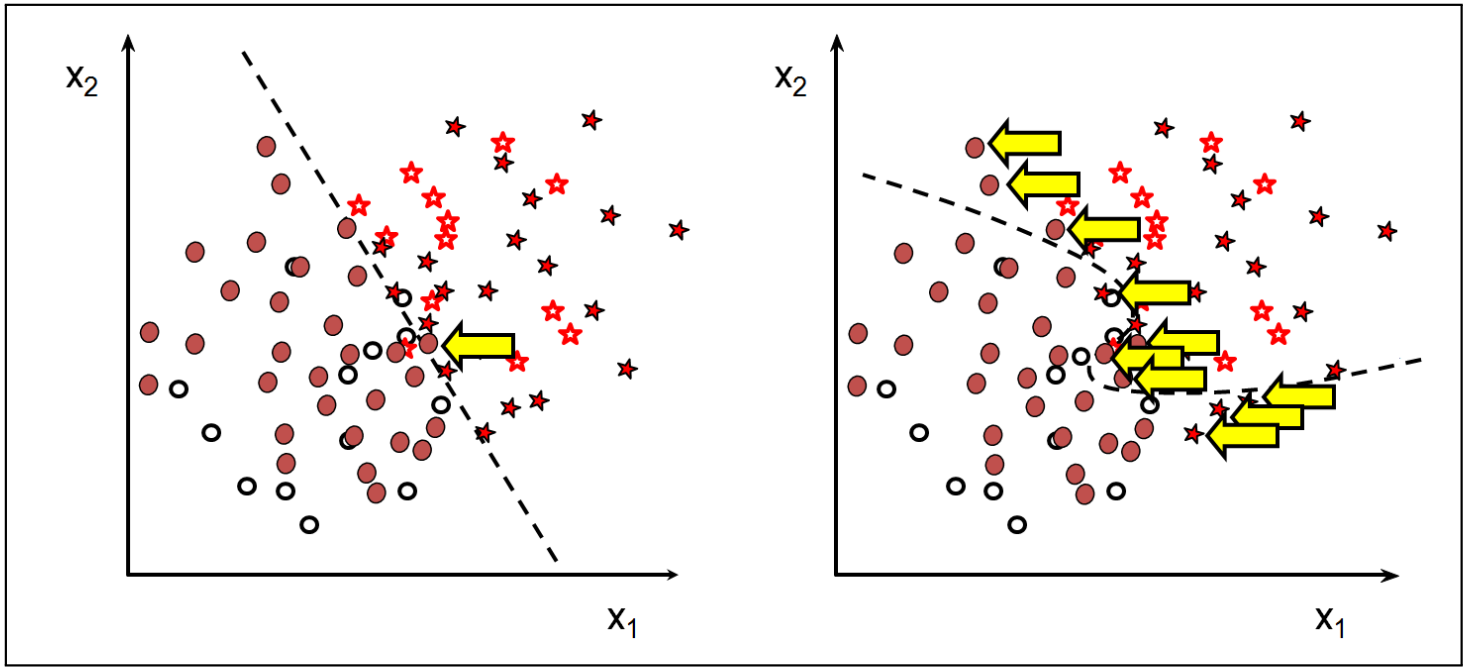
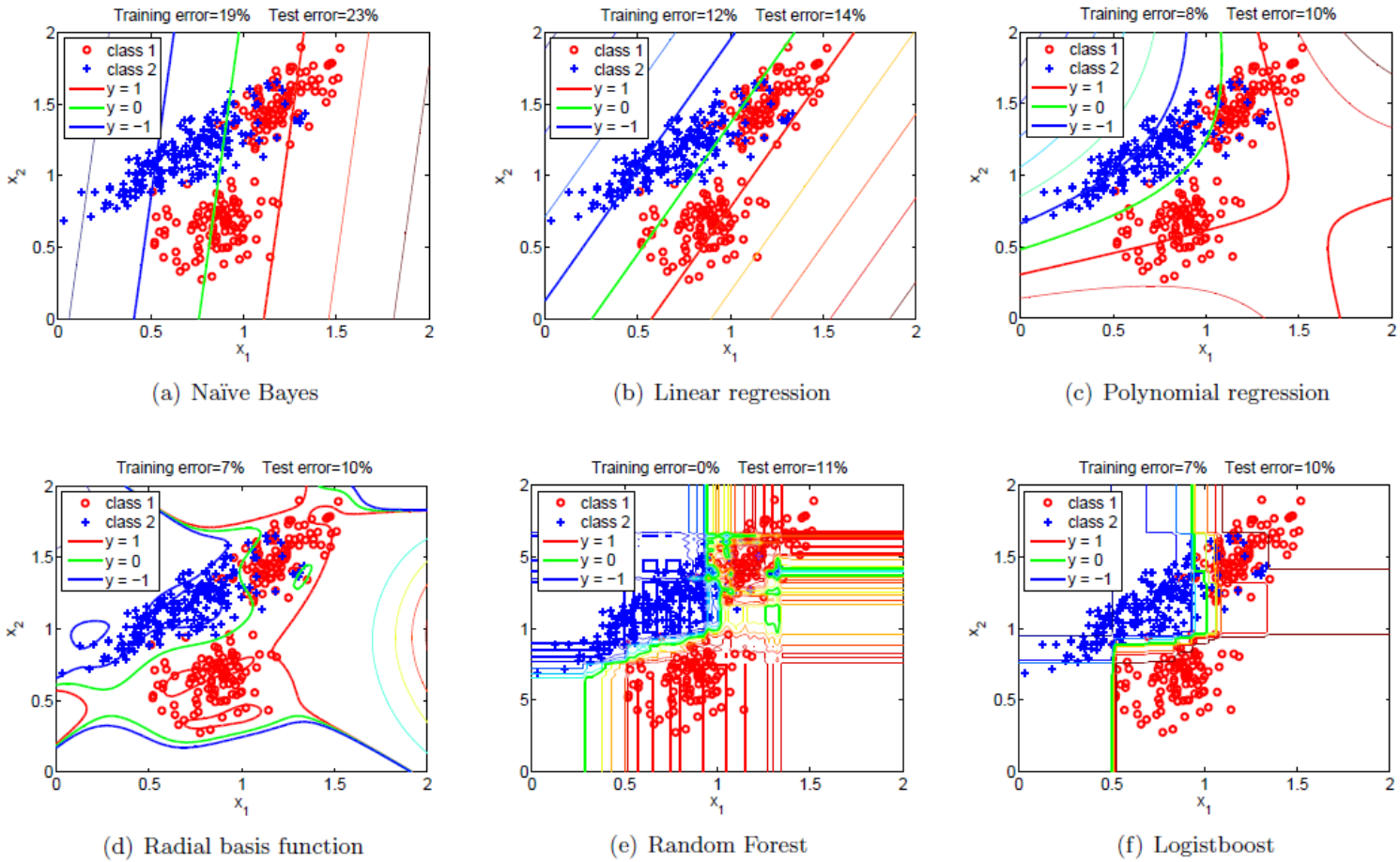


Figura: Capacidad de generalización.

Error de entrenamiento vs. error de generalización



(a) Naïve Bayes (b) Linear regression (c) Polynomial regression

(d) Radial basis function (e) Random Forest (f) Logistboost

Figura: Diferentes superficies de clasificación generadas por diferentes clasificadores.

- **Problema.** Generalmente la muestra $\mathcal{D}$ es finita y pequeña, entonces, ¿cómo podemos estimar el error en datos no vistos?
  - Hold-out.
  - Cross-validation.
  - Bootstrapping.
  - Jacknife.
  - ...
- La estimación aplica para evaluar un clasificador, seleccionar parámetros, o comparar técnicas.

Hold-out

El esquema visto hasta el momento:

- Dejar fuera una partición de datos para evaluación.
- Selección aleatoria de particiones.
- Generalmente se hacen varias repeticiones.

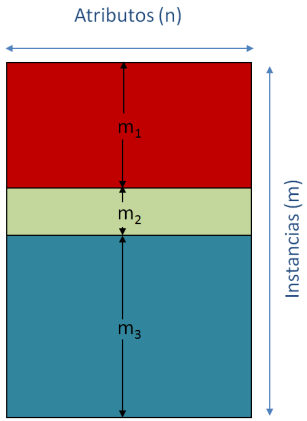


Figura: Partición sugerida esquema hold out.

$k-$fold Cross validation

1. Dividir el conjunto de datos $\mathcal{D}$ en $k-$particiones: $\mathcal{D}_1, \ldots, \mathcal{D}_k$.
2. Por cada subconjunto $\mathcal{D}_i$:
   - Entrenar clasificador usando $\bigcup \mathcal{D}_{j:j \neq i}$ ($k-1$ subconjuntos)
   - Evaluar el clasificador entrenado en $\mathcal{D}_i$, $Err_i(\hat{f})$
3. Reportar el promedio del desempeño obtenido: $CV_{Err} = \frac{1}{k} \sum_{i=1}^{k} Err_i(\hat{f})$

$k-$fold Cross validation

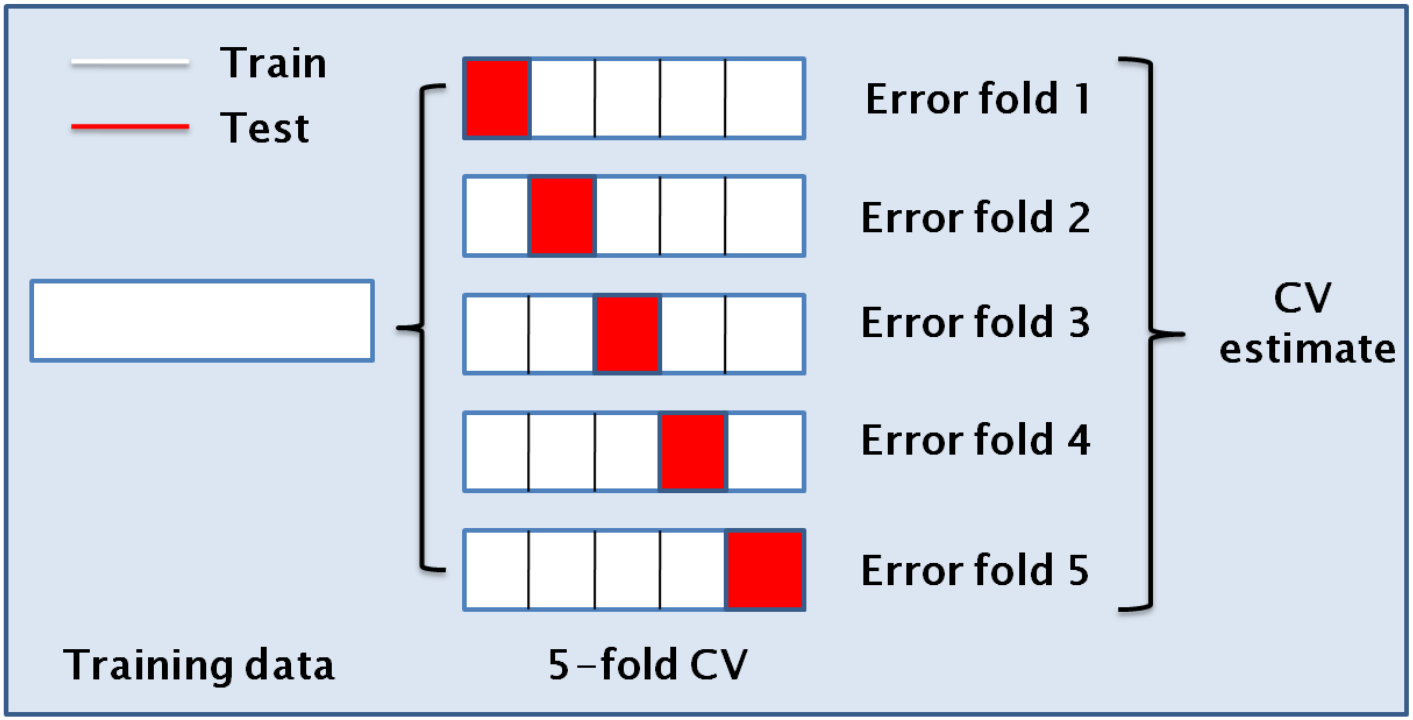


Figura: Validación cruzada.

Otras técnicas

- **Bootstraping.** Repetir Hold-out muestreando con remplazo.
- **Jacknife.** $N-$fold Cross Validation, con $N = |\mathcal{D}|$.
- **$5 \times 2-$fold CV**. 5 times $2-$fold Cross Validation.
- **Stratified CV.** CV manteniendo la distribución de las clases.
- ...

- Ojo: Al realizar una búsqueda intensiva para optimizar parámetros también es posible sobre-ajustar la estrategia de evaluación.
- Por qué?

El dilema sesgo-varianza

- Un modelo predictivo f puede verse como un estimador de $E(Y|X)$, el valor esperado de Y dado X
- Sean
  - $\hat{f}(X;\mathcal{D})$: el modelo entrenado en un conjunto de datos $\mathcal{D}$ de tamaño t
  - $E_{\mathcal{D}}[\cdot]$: el valor esperado tomado sobre todos los conjuntos de datos de tamaño t de acuerdo a $P(X, Y)$
- Se puede mostrar que:

$$E_{\mathcal{D}}[(\hat{f}(X;\mathcal{D}) - E[Y|X])^2] = (E_{\mathcal{D}}[\hat{f}(X;\mathcal{D})] - E(Y|X))^2 \ldots$$

$$\ldots + E_{\mathcal{D}}[(\hat{f}(X;\mathcal{D}) - E[\hat{f}(X;\mathcal{D})])^2]$$

El dilema sesgo-varianza

Sesgo. Qué tanto se desvia el promedio de $\hat{f}(X;\mathcal{D})$ sobre todos los conjuntos de datos de tamaño t de $E(Y|X)$ (*la media verdadera*)
Qué tanto se aleja el modelo bajo análisis al modelo que generó los datos

$$E_{\mathcal{D}}[(\hat{f}(X;\mathcal{D}) - E[Y|X])^2] = (E_{\mathcal{D}}[\hat{f}(X;\mathcal{D})] - E(Y|X))^2 \ldots$$

$$\ldots + E_{\mathcal{D}}[(\hat{f}(X;\mathcal{D}) - E[\hat{f}(X;\mathcal{D})])^2]$$

El dilema sesgo-varianza

Varianza. La desviación promedio de $\hat{f}(X;\mathcal{D})$ con respecto al promedio de $\hat{f}(X;\mathcal{D})$ sobre todos los conjuntos de datos de tamaño t.

Qué tanto depende el modelo del conjunto de datos que se usó para entrenarlo/generarlo. Qué tanto variarán las predicciones de tu modelo para distintos conjuntos de datos?

$$E_{\mathcal{D}}[(\hat{f}(X;\mathcal{D}) - E[Y|X])^2] = (E_{\mathcal{D}}[\hat{f}(X;\mathcal{D})] - E(Y|X))^2 \ldots$$

$$\ldots + E_{\mathcal{D}}[(\hat{f}(X;\mathcal{D}) - E[\hat{f}(X;\mathcal{D})])^2]$$

Dilema sesgo-varianza

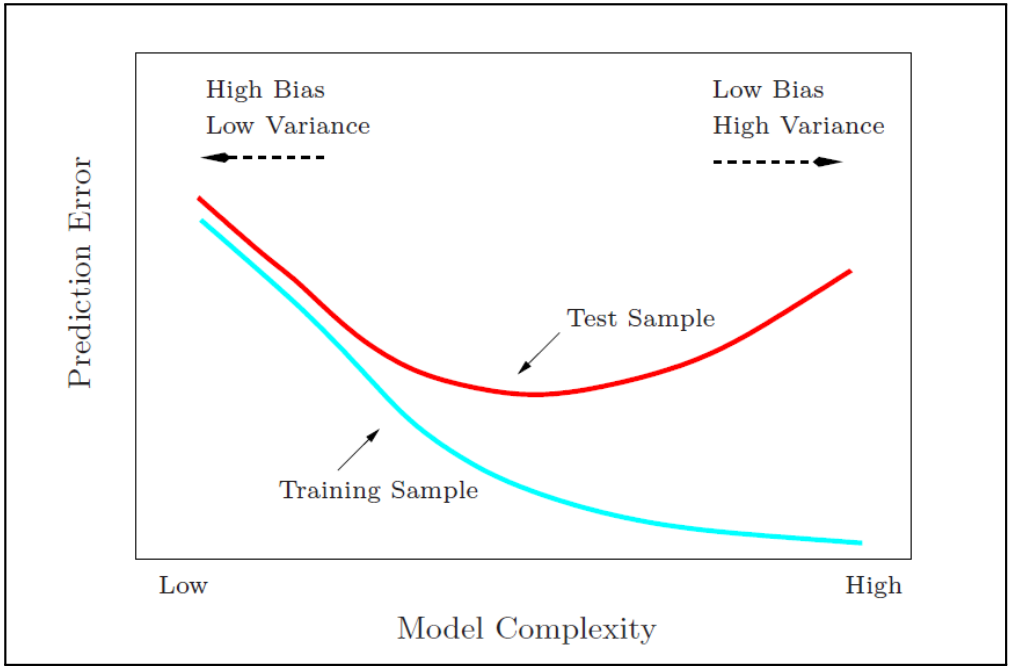


Figura: Dilema sesgo-varianza.

Dilema sesgo-varianza

Superficie de decisión $k-$NN.

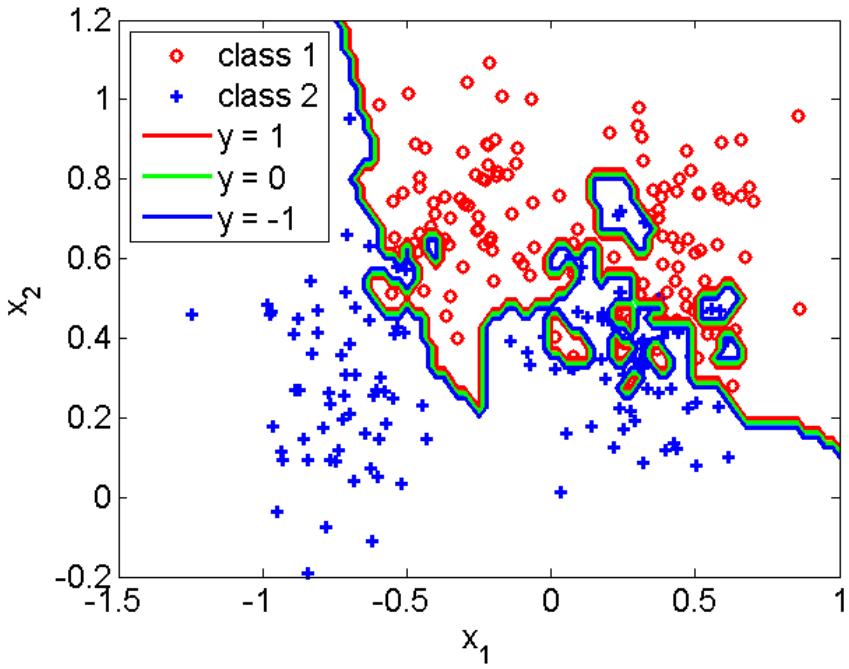


Figura: Superficie de clasificación inducida por 1-NN.

Dilema sesgo-varianza

Superficie de decisión $k-$NN.

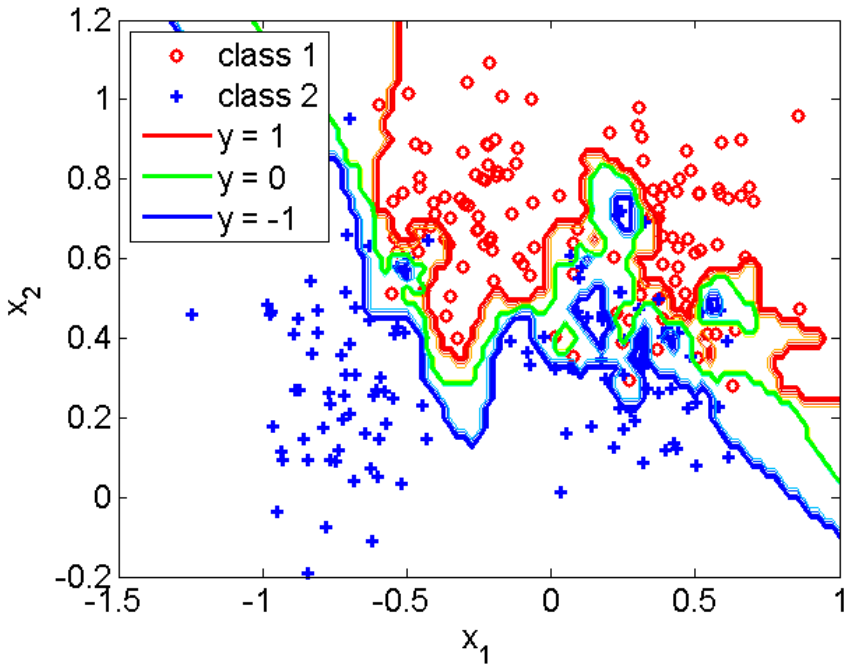


Figura: Superficie de clasificación inducida por 3-NN.

Dilema sesgo-varianza

Superficie de decisión $k-$NN.

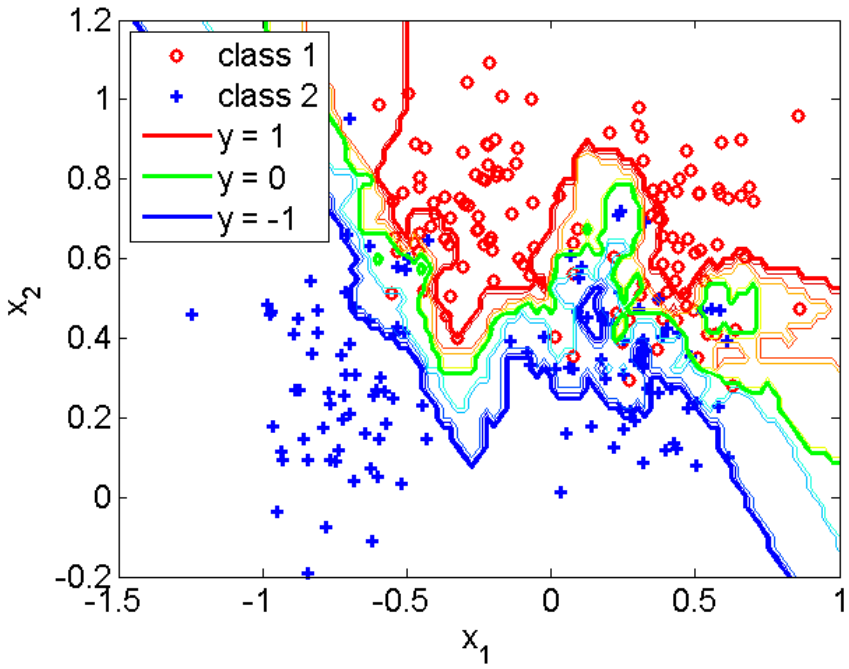


Figura: Superficie de clasificación inducida por 5-NN.

Dilema sesgo-varianza

Superficie de decisión $k-$NN.

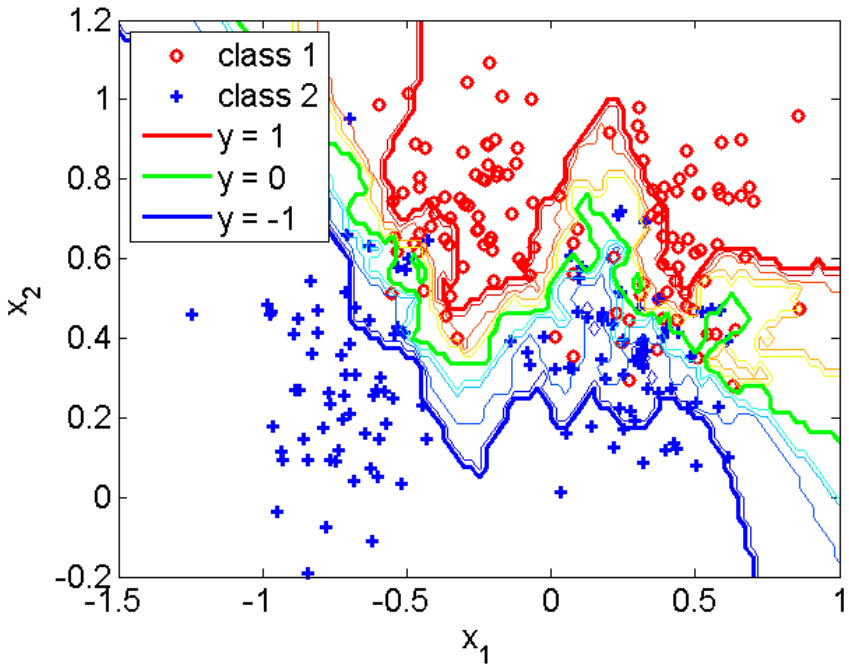


Figura: Superficie de clasificación inducida por 7-NN.

Dilema sesgo-varianza

Superficie de decisión $k-$NN.

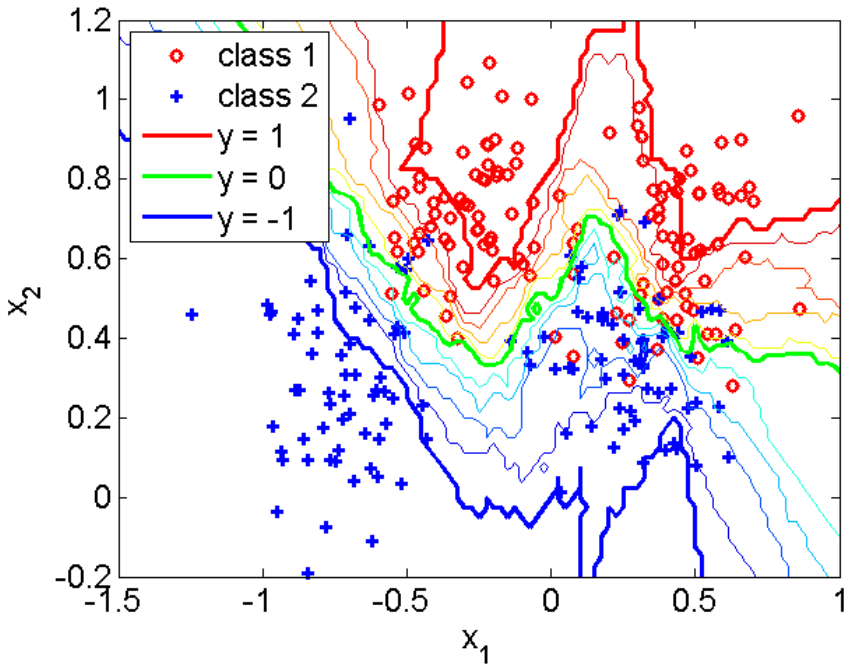


Figura: Superficie de clasificación inducida por 15-NN.

Dilema sesgo-varianza

Superficie de decisión $k-$NN.

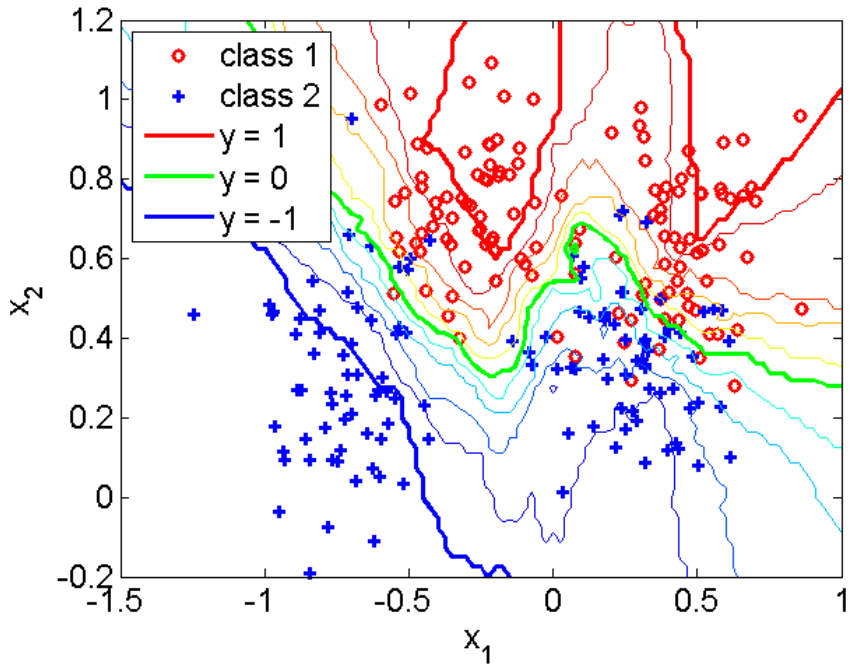


Figura: Superficie de clasificación inducida por 25-NN.

Dilema sesgo-varianza

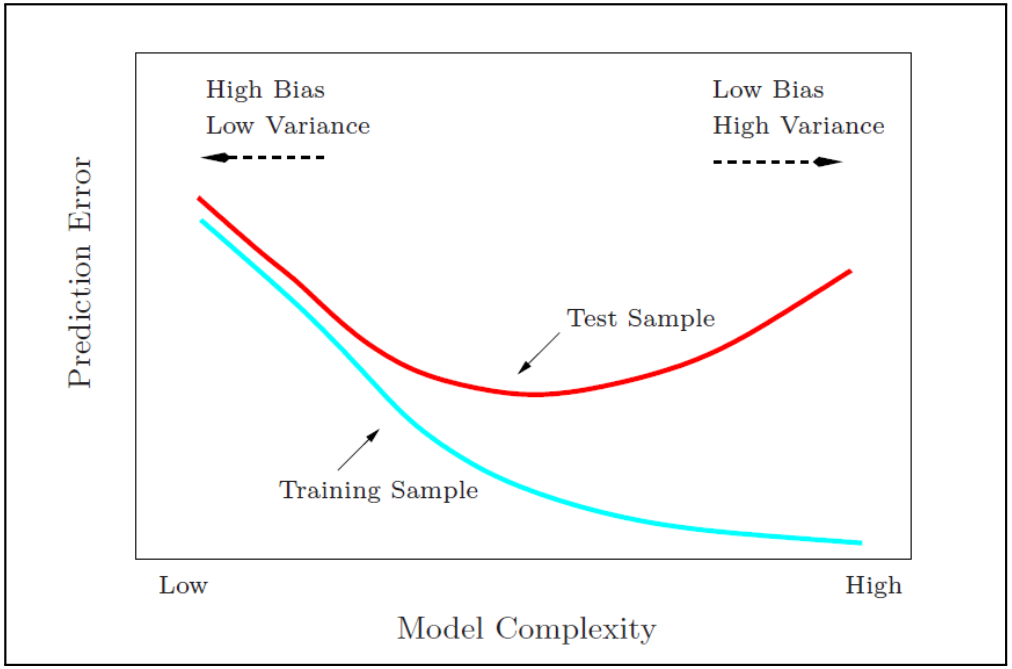


Figura: Dilema sesgo-varianza.

Discusión

Take-home messages.

- En problemas reales, no es posible estimar exactamente el error de generalización.
- El error de entrenamiento siempre será optimista.
- Estrategias de validación nos dan idea de cómo será el error de generalización.
- Intentos por mejorar el desempeño de un clasificador en datos de entrenamiento, mediante el incremento de la complejidad del modelo puede llevarnos a sobre-ajustar los datos: *el error de entrenamiento es engañoso!*