

Aprendizaje Supervisado

Fernando Lozano

Universidad de los Andes

6 de agosto de 2013



Ejemplos

- Reconocimiento de patrones o clasificación:

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.
 - ▶ Identificación de sistemas.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.
 - ▶ Identificación de sistemas.
 - ▶ Aproximación de funciones.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.
 - ▶ Identificación de sistemas.
 - ▶ Aproximación de funciones.
- Ranking

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.
 - ▶ Identificación de sistemas.
 - ▶ Aproximación de funciones.
- Ranking
 - ▶ Sistema de recomendación.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.
 - ▶ Identificación de sistemas.
 - ▶ Aproximación de funciones.
- Ranking
 - ▶ Sistema de recomendación.
 - ▶ Information retrieval.

Ejemplos

- Reconocimiento de patrones o clasificación:
 - ▶ Diagnóstico médico.
 - ▶ Reconocimiento de caracteres.
 - ▶ Categorización de texto.
- Regresión:
 - ▶ Predicción de series de tiempo.
 - ▶ Identificación de sistemas.
 - ▶ Aproximación de funciones.
- Ranking
 - ▶ Sistema de recomendación.
 - ▶ Information retrieval.
- Otros.

Aprendizaje Supervisado

- Entrada \mathbf{x} , salida y .

Aprendizaje Supervisado

- Entrada \mathbf{x} , salida y .
- Queremos un sistema que **prediga** el valor de y a partir de \mathbf{x} .

Aprendizaje Supervisado

- Entrada \mathbf{x} , salida y .
- Queremos un sistema que **prediga** el valor de y a partir de \mathbf{x} .
- Existe un **supervisor** o **maestro** que conoce la respuesta correcta para patrones de entrada.

Aprendizaje Supervisado

- Entrada \mathbf{x} , salida y .
- Queremos un sistema que **prediga** el valor de y a partir de \mathbf{x} .
- Existe un **supervisor** o **maestro** que conoce la respuesta correcta para patrones de entrada.
- Conjunto de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^n$

Aprendizaje Supervisado

- Entrada \mathbf{x} , salida y .
- Queremos un sistema que **prediga** el valor de y a partir de \mathbf{x} .
- Existe un **supervisor** o **maestro** que conoce la respuesta correcta para patrones de entrada.
- Conjunto de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^n$

Visión Conceptual



- Queremos modelar S .

Visión Conceptual



- Queremos modelar S .
- No es fácil obtener un modelo analítico.

Visión Conceptual

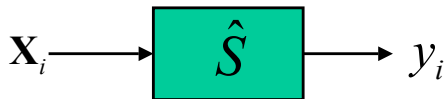


- Queremos modelar S .
- No es fácil obtener un modelo analítico.
- Usar modelo para predecir valores de la salida para nuevas entradas.

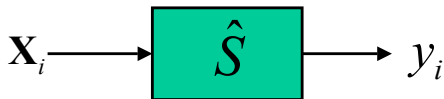
Visión Conceptual



- Queremos modelar S .
- No es fácil obtener un modelo analítico.
- Usar modelo para predecir valores de la salida para nuevas entradas.



- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.

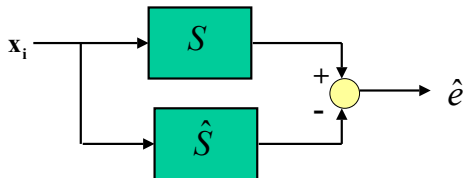


- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Conjunto de modelos a utilizar.



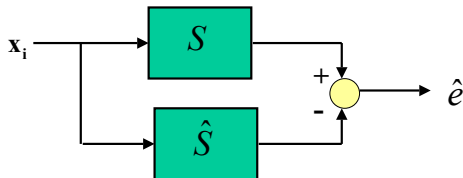
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Conjunto de modelos a utilizar.
- Conjunto de datos de prueba $\{\mathbf{x}_i, y_i\}_{i=1}^q$.

Aprendizaje=Construir modelo



- El objetivo es **aproximar** S .

Aprendizaje=Construir modelo



- El objetivo es **aproximar** S .
- Cuál es un criterio de error apropiado?

Marco Probabilístico

- $(\mathbf{x}, y) \sim \mathcal{D}$

Marco Probabilístico

- $(\mathbf{x}, y) \sim \mathcal{D}$
- $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) .

- $(\mathbf{x}, y) \sim \mathcal{D}$
- $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) .
- $\{\mathbf{x}_i, y_i\}_{i=1}^q$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) e independientes de $\{\mathbf{x}_i, y_i\}_{i=1}^n$.

- $(\mathbf{x}, y) \sim \mathcal{D}$
- $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) .
- $\{\mathbf{x}_i, y_i\}_{i=1}^q$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) e independientes de $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Criterio de error:

- $(\mathbf{x}, y) \sim \mathcal{D}$
- $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) .
- $\{\mathbf{x}_i, y_i\}_{i=1}^q$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) e independientes de $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Criterio de error:
 - ▶ Para clasificación binaria:

$$\mathbf{P}_{\mathcal{D}} \left[\hat{S}(\mathbf{x}) \neq y \right]$$

- $(\mathbf{x}, y) \sim \mathcal{D}$
- $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) .
- $\{\mathbf{x}_i, y_i\}_{i=1}^q$ copias independientes e idénticamente distribuidas de (\mathbf{x}, y) e independientes de $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Criterio de error:
 - ▶ Para clasificación binaria:

$$\mathbf{P}_{\mathcal{D}} \left[\hat{S}(\mathbf{x}) \neq y \right]$$

- ▶ Para regresión:

$$\mathbf{E}_{\mathcal{D}} \left[\hat{S}(\mathbf{X}) - y \right]^2$$

Generalización

- Entrenamiento sobre un conjunto de datos.

Generalización

- Entrenamiento sobre un conjunto de datos.
- Es relativamente fácil construir un modelo que no se equivoque en datos de entrenamiento.

Generalización

- Entrenamiento sobre un conjunto de datos.
- Es relativamente fácil construir un modelo que no se equivoque en datos de entrenamiento.
- Queremos un modelo que tenga error pequeño en datos **nuevos**

Generalización

- Entrenamiento sobre un conjunto de datos.
- Es relativamente fácil construir un modelo que no se equivoque en datos de entrenamiento.
- Queremos un modelo que tenga error pequeño en datos **nuevos**
- Aprendizaje=generalización.

Problemas Fundamentales

Problemas Fundamentales

Existencia: Es posible solucionar **en principio** el problema usando **cualquier** modelo? (es decir, está el problema bien definido?).

Problemas Fundamentales

Existencia: Es posible solucionar **en principio** el problema usando **cualquier** modelo? (es decir, está el problema bien definido?).

Capacidad de Representación: Es posible solucionar el problema usando una clase de modelos dada?

Problemas Fundamentales

Existencia: Es posible solucionar **en principio** el problema usando **cualquier** modelo? (es decir, está el problema bien definido?).

Capacidad de Representación: Es posible solucionar el problema usando una clase de modelos dada?

Estimación: Es posible determinar el modelo a partir de un conjunto de datos?

Problemas Fundamentales

Existencia: Es posible solucionar **en principio** el problema usando **cualquier** modelo? (es decir, está el problema bien definido?).

Capacidad de Representación: Es posible solucionar el problema usando una clase de modelos dada?

Estimación: Es posible determinar el modelo a partir de un conjunto de datos?

Computación: Es posible determinar el modelo **eficientemente**?

Problemas Fundamentales

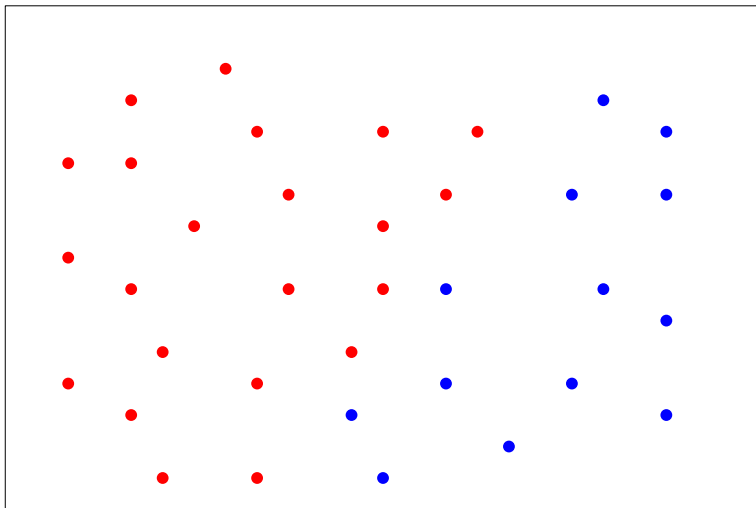
Existencia: Es posible solucionar **en principio** el problema usando **cualquier** modelo? (es decir, está el problema bien definido?).

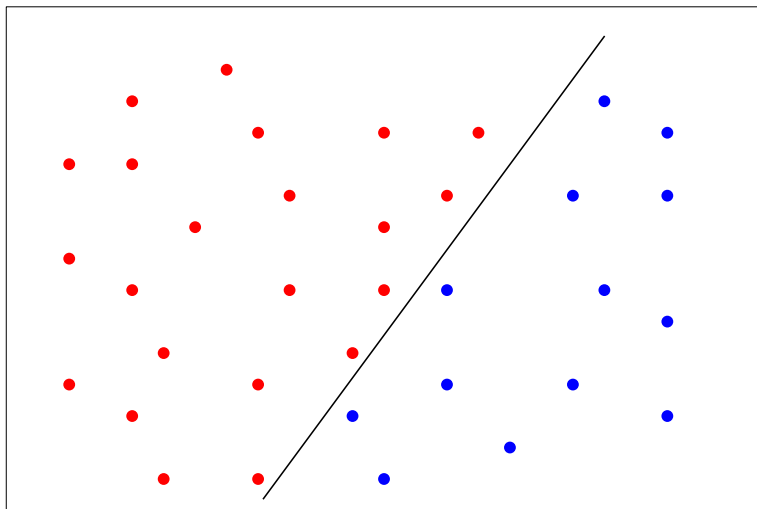
Capacidad de Representación: Es posible solucionar el problema usando una clase de modelos dada?

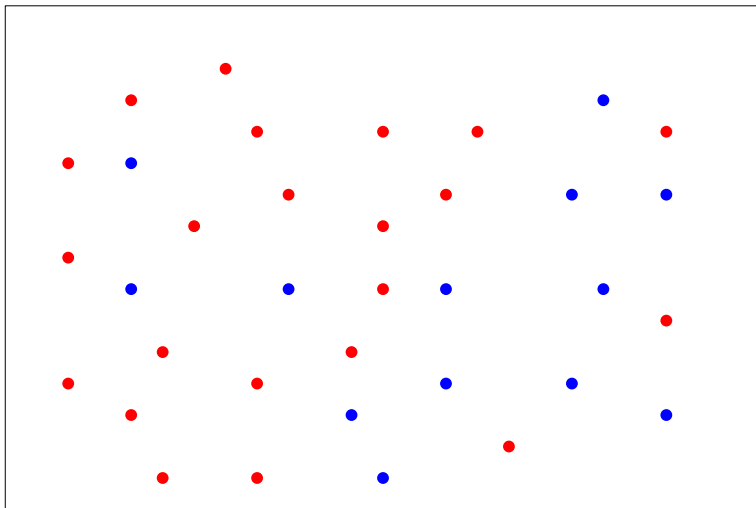
Estimación: Es posible determinar el modelo a partir de un conjunto de datos?

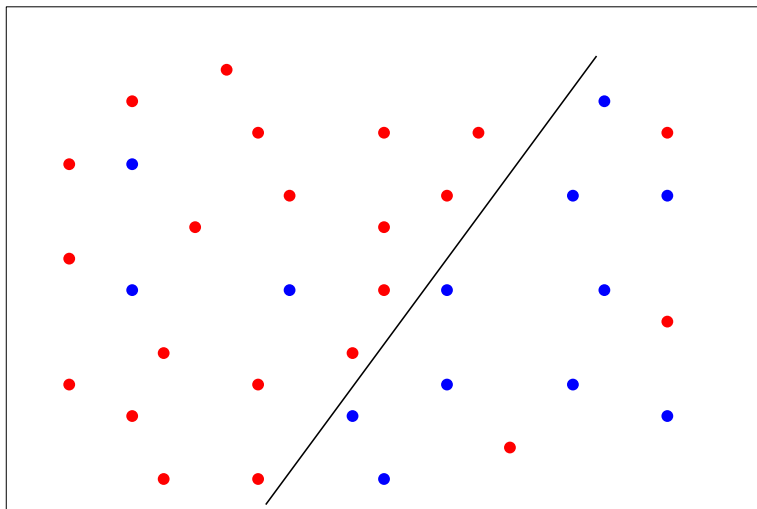
Computación: Es posible determinar el modelo **eficientemente**?

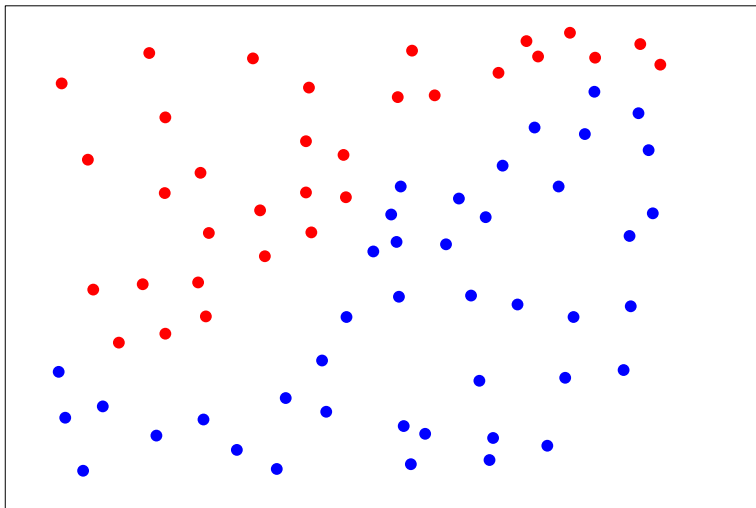
Implementación: Es posible diseñar e implementar el modelo usando precisión finita?

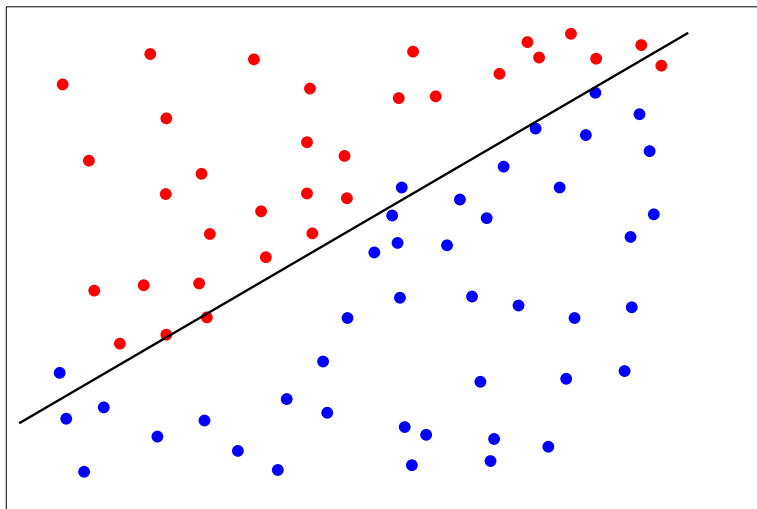


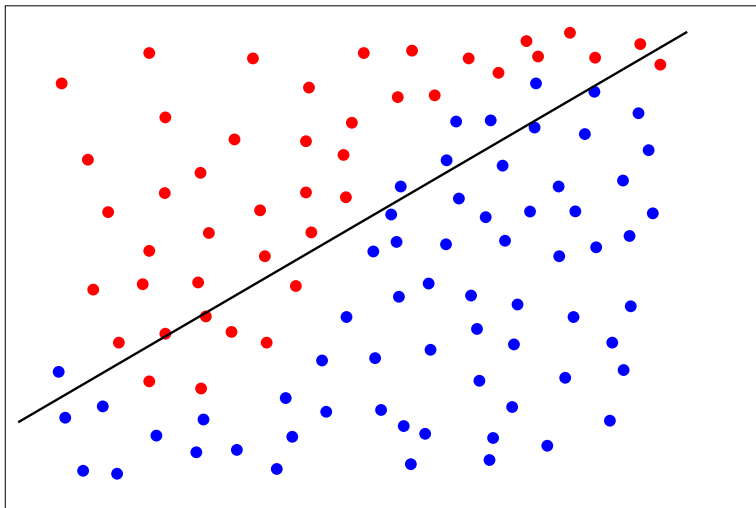


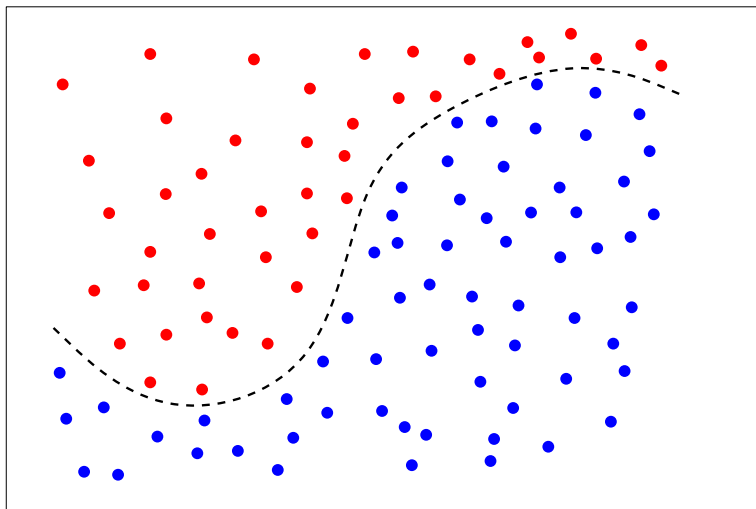


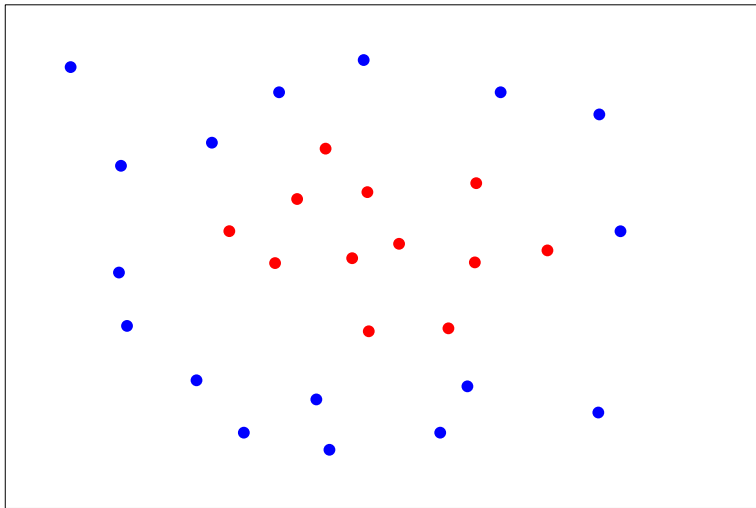


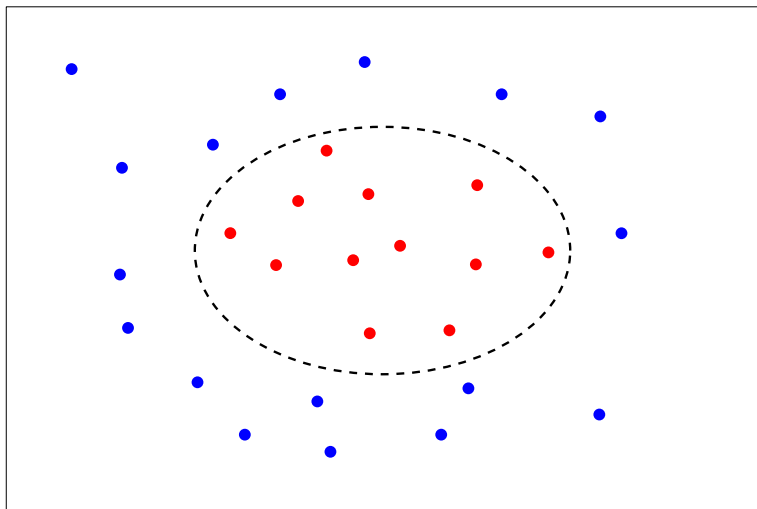


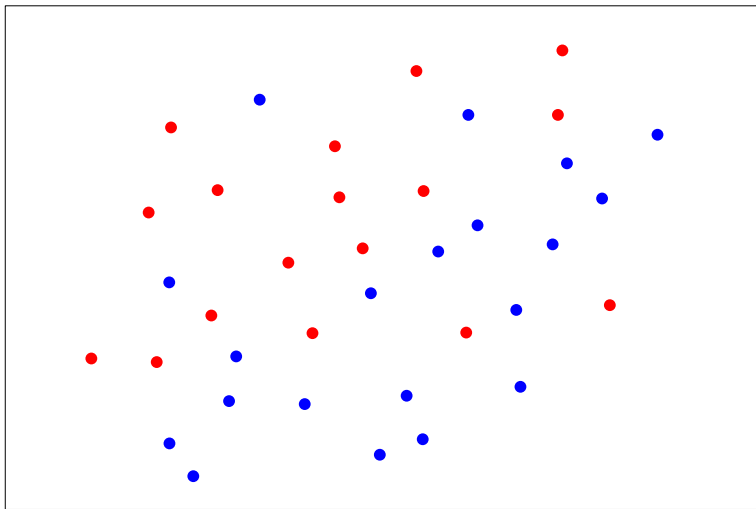


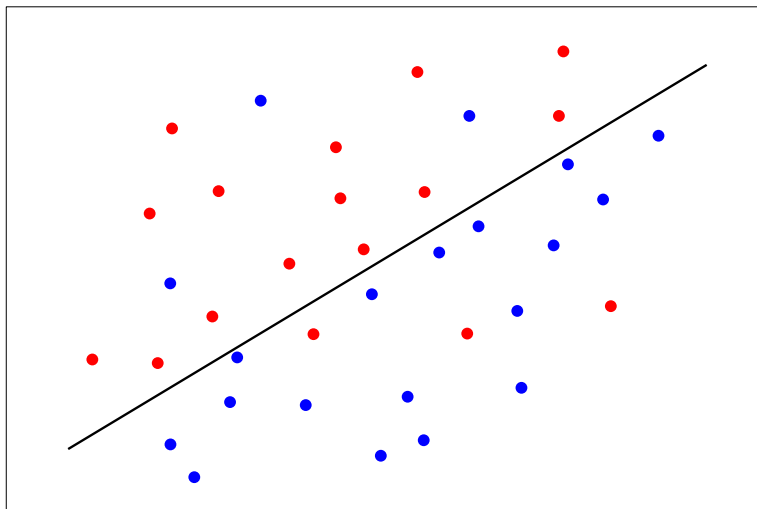


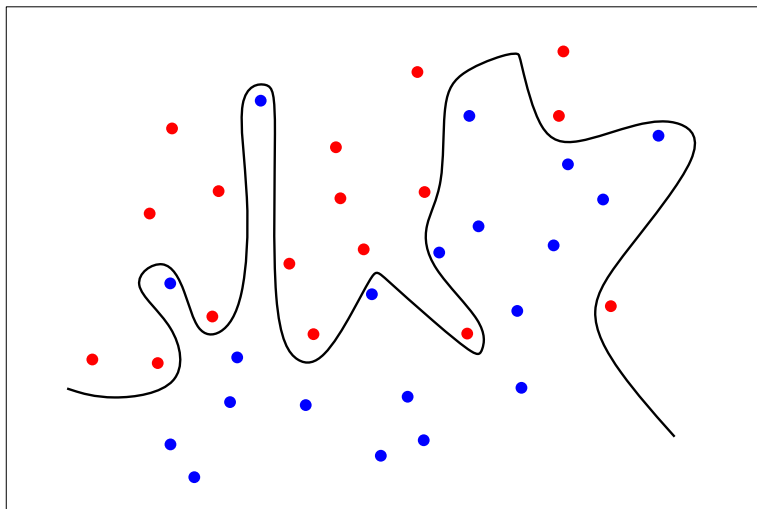


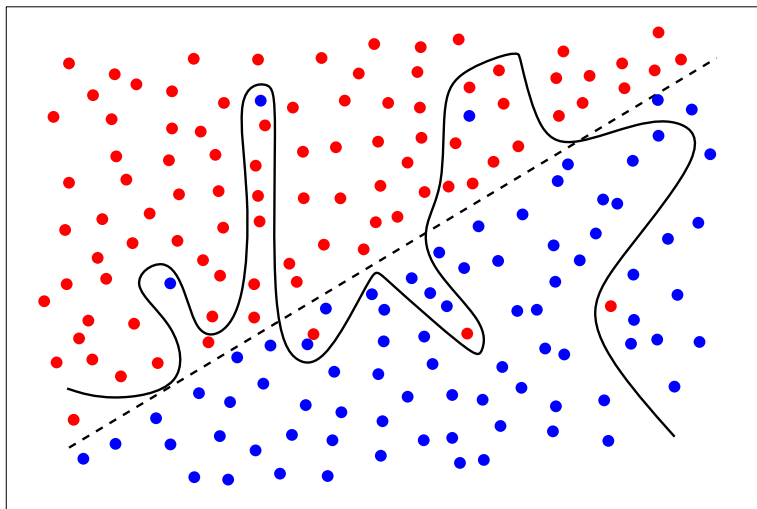


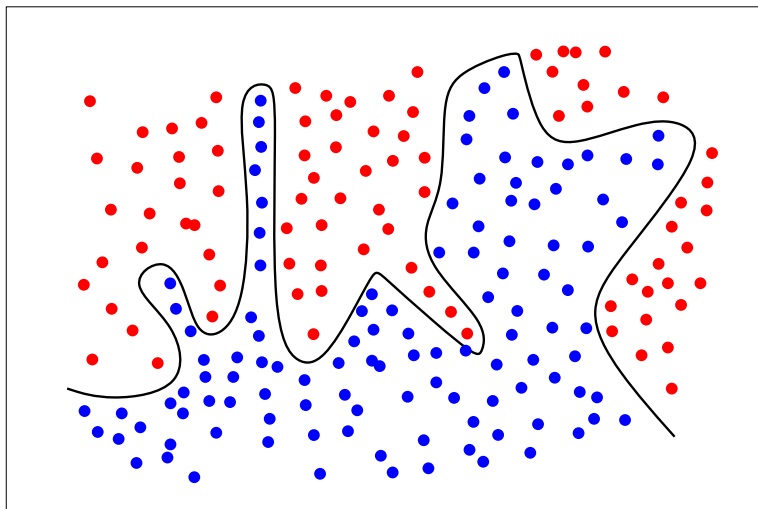


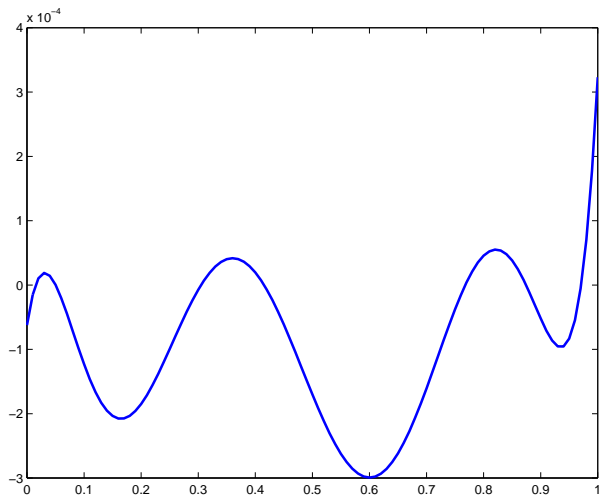


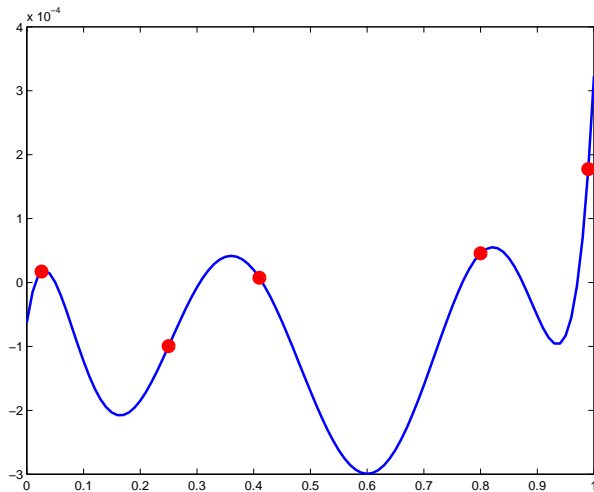


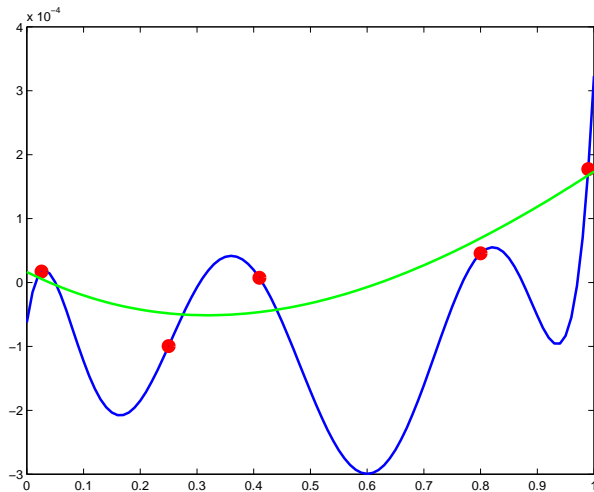


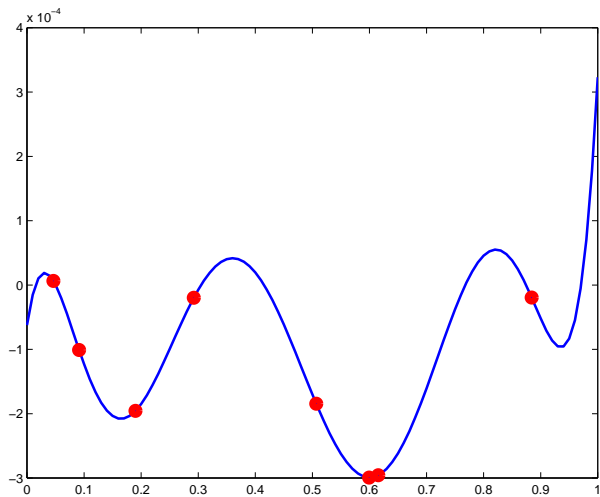


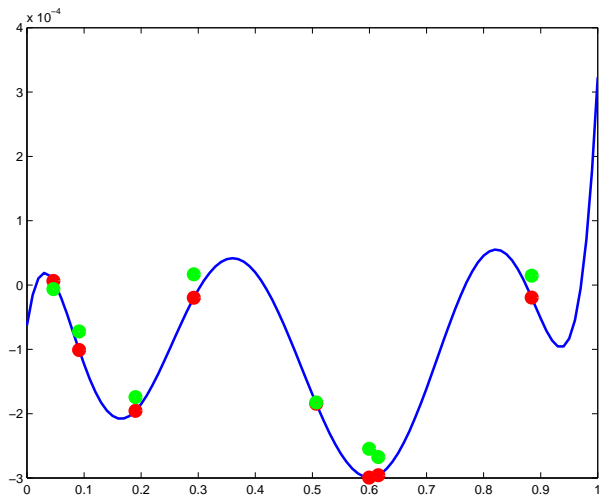


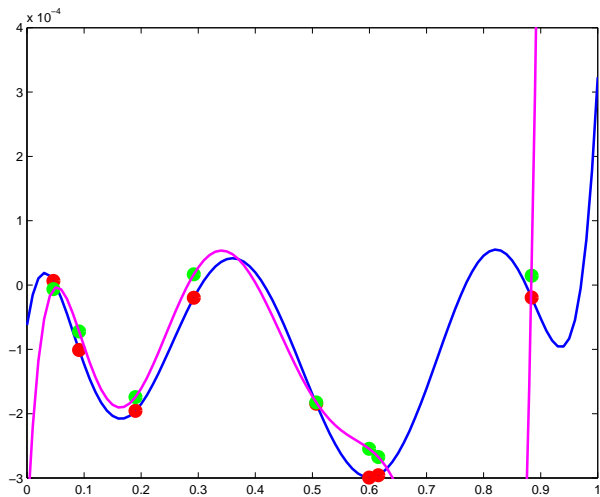


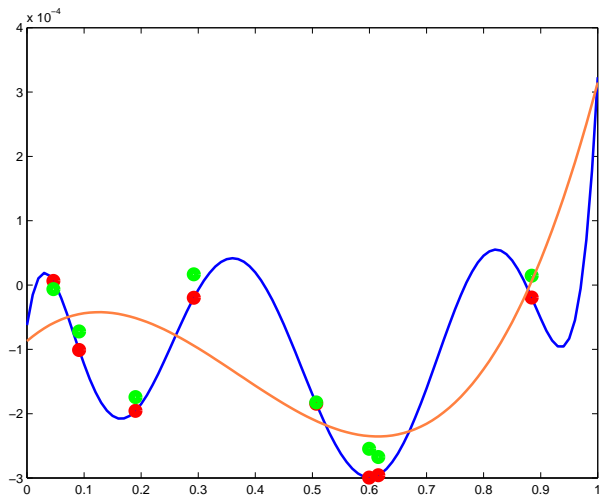




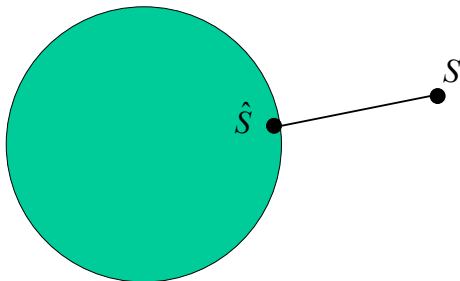




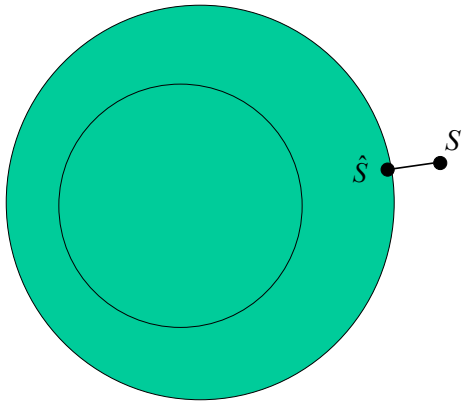




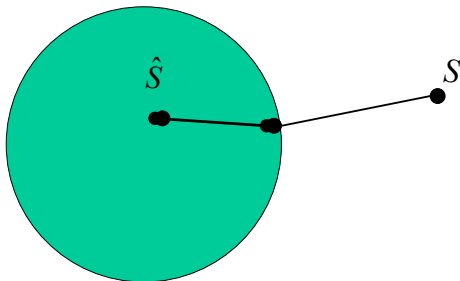
Error de Aproximación



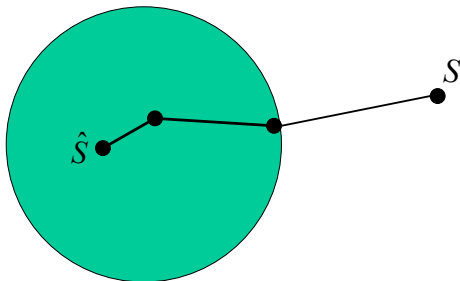
Incrementar complejidad



Error de Estimación



Error Computacional



Dificultades Adicionales

- Modelos altamente no lineales:

Dificultades Adicionales

- Modelos altamente no lineales:
 - ▶ Aprendizaje es lento (comparado por ejemplo con modelos lineales).

Dificultades Adicionales

- Modelos altamente no lineales:
 - ▶ Aprendizaje es lento (comparado por ejemplo con modelos lineales).
 - ▶ Algoritmos más complicados.

Dificultades Adicionales

- Modelos altamente no lineales:
 - ▶ Aprendizaje es lento (comparado por ejemplo con modelos lineales).
 - ▶ Algoritmos más complicados.
 - ▶ Difícil interpretación.

Dificultades Adicionales

- Modelos altamente no lineales:
 - ▶ Aprendizaje es lento (comparado por ejemplo con modelos lineales).
 - ▶ Algoritmos más complicados.
 - ▶ Difícil interpretación.
- Dimensionalidad alta de la entrada (“maldición” de la dimensionalidad).

Dificultades Adicionales

- Modelos altamente no lineales:
 - ▶ Aprendizaje es lento (comparado por ejemplo con modelos lineales).
 - ▶ Algoritmos más complicados.
 - ▶ Difícil interpretación.
- Dimensionalidad alta de la entrada (“maldición” de la dimensionalidad).
- Usualmente se tienen pocos datos, relativo a la dimensión de la entrada.

Con respecto a los datos

Con respecto a los datos

- Cuáles?

Con respecto a los datos

- Cuáles?
 - ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.

Con respecto a los datos

- Cuáles?

- ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
- ▶ Datos de prueba provienen de la misma distribución.

Con respecto a los datos

- Cuáles?
 - ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
 - ▶ Datos de prueba provienen de la misma distribución.
 - ▶ Otra opción: escoger los datos más “convenientes” (aprendizaje activo).

Con respecto a los datos

- Cuáles?
 - ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
 - ▶ Datos de prueba provienen de la misma distribución.
 - ▶ Otra opción: escoger los datos más “convenientes” (aprendizaje activo).
- Cuántos?

Con respecto a los datos

- Cuáles?
 - ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
 - ▶ Datos de prueba provienen de la misma distribución.
 - ▶ Otra opción: escoger los datos más “convenientes” (aprendizaje activo).
- Cuántos?
 - ▶ Los que se consigan.

Con respecto a los datos

- Cuáles?
 - ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
 - ▶ Datos de prueba provienen de la misma distribución.
 - ▶ Otra opción: escoger los datos más “convenientes” (aprendizaje activo).
- Cuántos?
 - ▶ Los que se consigan.
 - ▶ Tantos como sea posible.

Con respecto a los datos

- Cuáles?

- ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
- ▶ Datos de prueba provienen de la misma distribución.
- ▶ Otra opción: escoger los datos más “convenientes” (aprendizaje activo).

- Cuántos?

- ▶ Los que se consigan.
- ▶ Tantos como sea posible.
- ▶ Depende de la complejidad del modelo y de la función a aproximar.

Con respecto a los datos

- Cuáles?

- ▶ $\{x_i, y_i\}_{i=1}^n$ son i.i.d.
- ▶ Datos de prueba provienen de la misma distribución.
- ▶ Otra opción: escoger los datos más “convenientes” (aprendizaje activo).

- Cuántos?

- ▶ Los que se consigan.
- ▶ Tantos como sea posible.
- ▶ Depende de la complejidad del modelo y de la función a aproximar.
- ▶ Regla práctica: $n_{min} = 10 \times dim$

Preprocesamiento

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.
 - ▶ Sin una representación adecuada, es probable que **ningún** algoritmo funcione bien!

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.
 - ▶ Sin una representación adecuada, es probable que **ningún** algoritmo funcione bien!
- Selección de características útiles.

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.
 - ▶ Sin una representación adecuada, es probable que **ningún** algoritmo funcione bien!
- Selección de características útiles.
- Reducción de dimensionalidad.

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.
 - ▶ Sin una representación adecuada, es probable que **ningún** algoritmo funcione bien!
- Selección de características útiles.
- Reducción de dimensionalidad.
- Ejemplo: Datos MEG para detección de epilepsia:

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.
 - ▶ Sin una representación adecuada, es probable que **ningún** algoritmo funcione bien!
- Selección de características útiles.
- Reducción de dimensionalidad.
- Ejemplo: Datos MEG para detección de epilepsia:
 - ▶ 122 Canales.

Preprocesamiento

- Es necesario obtener una representación apropiada del problema.
 - ▶ Con una representación apropiada de las entradas, es probable que muchos algoritmos de aprendizaje funcionen bien.
 - ▶ Sin una representación adecuada, es probable que **ningún** algoritmo funcione bien!
- Selección de características útiles.
- Reducción de dimensionalidad.
- Ejemplo: Datos MEG para detección de epilepsia:
 - ▶ 122 Canales.
 - ▶ Cientos de miles de muestras por canal.

Solución de un problema de Aprendizaje Supervisado

Solución de un problema de Aprendizaje Supervisado

- Recolección de datos

Solución de un problema de Aprendizaje Supervisado

- Recolección de datos
- Preprocesamiento.

Solución de un problema de Aprendizaje Supervisado

- Recolección de datos
- Preprocesamiento.
- Seleccionar método (redes neuronales, SVM, ...)

Solución de un problema de Aprendizaje Supervisado

- Recolección de datos
- Preprocesamiento.
- Seleccionar método (redes neuronales, SVM, ...)
- Algoritmo de entrenamiento.

Solución de un problema de Aprendizaje Supervisado

- Recolección de datos
- Preprocesamiento.
- Seleccionar método (redes neuronales, SVM, ...)
- Algoritmo de entrenamiento.
- Selección de modelo.

Solución de un problema de Aprendizaje Supervisado

- Recolección de datos
- Preprocesamiento.
- Seleccionar método (redes neuronales, SVM, ...)
- Algoritmo de entrenamiento.
- Selección de modelo.
- Evaluación.

Cuál es el mejor clasificador?

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.
- Clasificador $C \subseteq \mathcal{S}$.

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.
- Clasificador $C \subseteq \mathcal{S}$.
- Función indicadora:

$$I_C(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} \in C \\ 0 & \text{si } \mathbf{x} \notin C \end{cases}$$

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.
- Clasificador $C \subseteq \mathcal{S}$.
- Función indicadora:

$$I_C(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} \in C \\ 0 & \text{si } \mathbf{x} \notin C \end{cases}$$

- Error de generalización a minimizar:

$$L(C) = \mathbb{P}\{y \neq I_C(\mathbf{x})\}$$

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.
- Clasificador $C \subseteq \mathcal{S}$.
- Función indicadora:

$$I_C(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} \in C \\ 0 & \text{si } \mathbf{x} \notin C \end{cases}$$

- Error de generalización a minimizar:

$$L(C) = \mathbb{P}\{y \neq I_C(\mathbf{x})\}$$

- Probabilidades a priori de cada clase:

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.
- Clasificador $C \subseteq \mathcal{S}$.
- Función indicadora:

$$I_C(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} \in C \\ 0 & \text{si } \mathbf{x} \notin C \end{cases}$$

- Error de generalización a minimizar:

$$L(C) = \mathbb{P}\{y \neq I_C(\mathbf{x})\}$$

- Probabilidades a priori de cada clase:

$$\mathbb{P}\{y = 1\} = \alpha, \quad \mathbb{P}\{y = 0\} = 1 - \alpha$$

- Probabilidades marginales

Cuál es el mejor clasificador?

- Par aleatorio $(\mathbf{x}, y) \in \mathcal{S} \times \{0, 1\}$.
- Clasificador $C \subseteq \mathcal{S}$.
- Función indicadora:

$$I_C(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{x} \in C \\ 0 & \text{si } \mathbf{x} \notin C \end{cases}$$

- Error de generalización a minimizar:

$$L(C) = \mathbb{P}\{y \neq I_C(\mathbf{x})\}$$

- Probabilidades a priori de cada clase:

$$\mathbb{P}\{y = 1\} = \alpha, \quad \mathbb{P}\{y = 0\} = 1 - \alpha$$

- Probabilidades marginales

$$\mathbf{P}[\mathbf{x}|y = 1] = p_1(\mathbf{x}), \quad \mathbf{P}[\mathbf{x}|y = 0] = p_0(\mathbf{x})$$

$$L(C) = \mathbb{P}\{y = 1, \mathbf{x} \notin C\} + \mathbb{P}\{y = 0, \mathbf{x} \in C\}$$

$$\begin{aligned} L(C) &= \mathbb{P}\{y = 1, \mathbf{x} \notin C\} + \mathbb{P}\{y = 0, \mathbf{x} \in C\} \\ &= \mathbb{P}\{\mathbf{x} \notin C | y = 1\} \mathbb{P}\{y = 1\} + \mathbb{P}\{\mathbf{x} \in C | y = 0\} \mathbb{P}\{y = 0\} \end{aligned}$$

$$\begin{aligned}
L(C) &= \mathbb{P}\{y = 1, \mathbf{x} \notin C\} + \mathbb{P}\{y = 0, \mathbf{x} \in C\} \\
&= \mathbb{P}\{\mathbf{x} \notin C | y = 1\} \mathbb{P}\{y = 1\} + \mathbb{P}\{\mathbf{x} \in C | y = 0\} \mathbb{P}\{y = 0\} \\
&= \alpha \int_{\mathcal{S}-C} p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x}
\end{aligned}$$

$$\begin{aligned}
L(C) &= \mathbb{P}\{y = 1, \mathbf{x} \notin C\} + \mathbb{P}\{y = 0, \mathbf{x} \in C\} \\
&= \mathbb{P}\{\mathbf{x} \notin C | y = 1\} \mathbb{P}\{y = 1\} + \mathbb{P}\{\mathbf{x} \in C | y = 0\} \mathbb{P}\{y = 0\} \\
&= \alpha \int_{\mathcal{S}-C} p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x} \\
&= \alpha \int_{\mathcal{S}} p_1(\mathbf{x}) d\mathbf{x} - \alpha \int_C p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x}
\end{aligned}$$

$$\begin{aligned}
L(C) &= \mathbb{P}\{y = 1, \mathbf{x} \notin C\} + \mathbb{P}\{y = 0, \mathbf{x} \in C\} \\
&= \mathbb{P}\{\mathbf{x} \notin C | y = 1\} \mathbb{P}\{y = 1\} + \mathbb{P}\{\mathbf{x} \in C | y = 0\} \mathbb{P}\{y = 0\} \\
&= \alpha \int_{S-C} p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x} \\
&= \alpha \int_S p_1(\mathbf{x}) d\mathbf{x} - \alpha \int_C p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x} \\
&= \alpha + \int_C [(1 - \alpha)p_0(\mathbf{x}) - \alpha p_1(\mathbf{x})] d\mathbf{x}
\end{aligned}$$

$$\begin{aligned}
L(C) &= \mathbb{P}\{y = 1, \mathbf{x} \notin C\} + \mathbb{P}\{y = 0, \mathbf{x} \in C\} \\
&= \mathbb{P}\{\mathbf{x} \notin C | y = 1\} \mathbb{P}\{y = 1\} + \mathbb{P}\{\mathbf{x} \in C | y = 0\} \mathbb{P}\{y = 0\} \\
&= \alpha \int_{S-C} p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x} \\
&= \alpha \int_S p_1(\mathbf{x}) d\mathbf{x} - \alpha \int_C p_1(\mathbf{x}) d\mathbf{x} + (1 - \alpha) \int_C p_0(\mathbf{x}) d\mathbf{x} \\
&= \alpha + \int_C [(1 - \alpha)p_0(\mathbf{x}) - \alpha p_1(\mathbf{x})] d\mathbf{x}
\end{aligned}$$

Cómo escogemos el C que minimiza $L(C)$?

Clasificador de Bayes

- El clasificador óptimo esta dado por la función indicadora del siguiente conjunto:

Clasificador de Bayes

- El clasificador óptimo esta dado por la función indicadora del siguiente conjunto:

$$C = \{\mathbf{x} : (1 - \alpha)p_0(\mathbf{x}) \leq \alpha p_1(\mathbf{x})\}$$

Clasificador de Bayes

- El clasificador óptimo está dado por la función indicadora del siguiente conjunto:

$$\begin{aligned} C &= \{ \mathbf{x} : (1 - \alpha)p_0(\mathbf{x}) \leq \alpha p_1(\mathbf{x}) \} \\ &= \left\{ \mathbf{x} : \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \leq \frac{\alpha}{1 - \alpha} \right\} \end{aligned}$$

Clasificador de Bayes

- El clasificador óptimo está dado por la función indicadora del siguiente conjunto:

$$\begin{aligned} C &= \{ \mathbf{x} : (1 - \alpha)p_0(\mathbf{x}) \leq \alpha p_1(\mathbf{x}) \} \\ &= \left\{ \mathbf{x} : \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \leq \frac{\alpha}{1 - \alpha} \right\} \\ &= \left\{ \mathbf{x} : l(\mathbf{x}) \leq \frac{\alpha}{1 - \alpha} \right\} \end{aligned}$$

Clasificador de Bayes

- El clasificador óptimo está dado por la función indicadora del siguiente conjunto:

$$\begin{aligned} C &= \{ \mathbf{x} : (1 - \alpha)p_0(\mathbf{x}) \leq \alpha p_1(\mathbf{x}) \} \\ &= \left\{ \mathbf{x} : \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \leq \frac{\alpha}{1 - \alpha} \right\} \\ &= \left\{ \mathbf{x} : l(\mathbf{x}) \leq \frac{\alpha}{1 - \alpha} \right\} \end{aligned}$$

- El clasificador óptimo recibe el nombre de **clasificador de Bayes**.

Clasificador de Bayes

- El clasificador óptimo está dado por la función indicadora del siguiente conjunto:

$$\begin{aligned} C &= \{ \mathbf{x} : (1 - \alpha)p_0(\mathbf{x}) \leq \alpha p_1(\mathbf{x}) \} \\ &= \left\{ \mathbf{x} : \frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} \leq \frac{\alpha}{1 - \alpha} \right\} \\ &= \left\{ \mathbf{x} : l(\mathbf{x}) \leq \frac{\alpha}{1 - \alpha} \right\} \end{aligned}$$

- El clasificador óptimo recibe el nombre de **clasificador de Bayes**.
- $l(\mathbf{x})$ es la razón de verosimilitud.

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son gaussianas:

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son gaussianas:

$$\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} = \frac{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_0|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_0)^T \Sigma_0^{-1} (\mathbf{x} - \mathbf{m}_0) \right\}}{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_1|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) \right\}} \leq \frac{\alpha}{1 - \alpha}$$

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son gaussianas:

$$\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} = \frac{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_0|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_0)^T \Sigma_0^{-1} (\mathbf{x} - \mathbf{m}_0) \right\}}{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_1|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) \right\}} \leq \frac{\alpha}{1 - \alpha}$$

- Tomando logaritmos:

Caso Especial

- Cuando $p_0(\mathbf{x})$ y $p_1(\mathbf{x})$ son gaussianas:

$$\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} = \frac{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_0|}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mathbf{m}_0)^T \Sigma_0^{-1} (\mathbf{x} - \mathbf{m}_0) \right\}}{\frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\Sigma_1|}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) \right\}} \leq \frac{\alpha}{1 - \alpha}$$

- Tomando logaritmos:

$$\begin{aligned} \frac{1}{2}(\mathbf{x} - \mathbf{m}_0)^T \Sigma_0^{-1} (\mathbf{x} - \mathbf{m}_0) - \frac{1}{2}(\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) \\ + \frac{1}{2} \ln \left(\frac{|\Sigma_0|}{|\Sigma_1|} \right) > \ln \left(\frac{1 - \alpha}{\alpha} \right) \end{aligned}$$

- Si además $\Sigma_0 = \Sigma_1 = \Sigma$:

$$\begin{aligned} \mathbf{x}^T \Sigma^{-1} \mathbf{x} - 2\mathbf{m}_0^T \Sigma^{-1} \mathbf{x} + \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_0 - \mathbf{x}^T \Sigma^{-1} \mathbf{x} \\ + 2\mathbf{m}_1^T \Sigma^{-1} \mathbf{x} - \mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 > 2 \ln \left(\frac{1 - \alpha}{\alpha} \right) \end{aligned}$$

- Si además $\Sigma_0 = \Sigma_1 = \Sigma$:

$$\begin{aligned} & \mathbf{x}^T \Sigma^{-1} \mathbf{x} - 2\mathbf{m}_0^T \Sigma^{-1} \mathbf{x} + \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_0 - \mathbf{x}^T \Sigma^{-1} \mathbf{x} \\ & + 2\mathbf{m}_1^T \Sigma^{-1} \mathbf{x} - \mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 > 2 \ln \left(\frac{1 - \alpha}{\alpha} \right) \end{aligned}$$

- entonces:

$$\underbrace{(\mathbf{m}_1 - \mathbf{m}_0)^T \Sigma^{-1} \mathbf{x}}_{\mathbf{w}^T} > \underbrace{2 \ln \left(\frac{1 - \alpha}{\alpha} \right) + \frac{1}{2} (\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 - \mathbf{m}_0^T \Sigma^{-1} \mathbf{m}_0)}_{-w_0}$$

