Aprendizaje Supervisado

Fernando Lozano

Universidad de los Andes

29 de agosto de 2022

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.

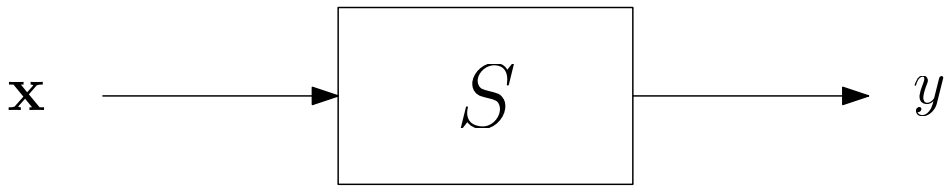
Aprendizaje Supervisado

- 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

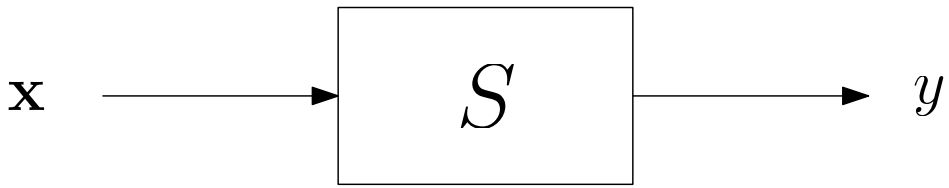
- Entrada $\mathbf{x}$, salida y.
- Queremos un sistema que prediga el valor de y a partir de $\mathbf{x}$.
- Existe un supervisor o maestro que conoce la respuesta correcta para patrones de entrada.
- Conjunto de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$

Visión Conceptual



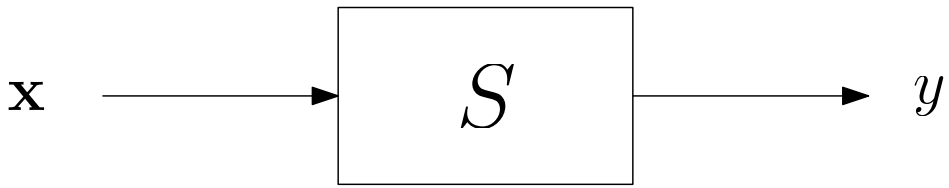
- Queremos modelar S.

Visión Conceptual



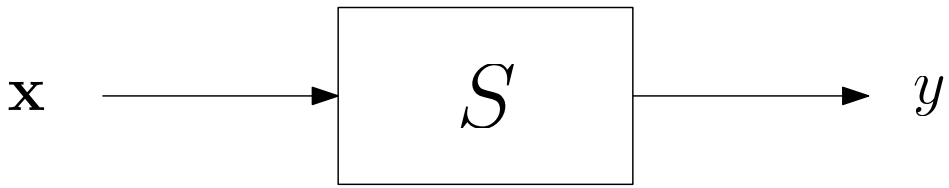
- Queremos modelar S.
- No es fácil obtener un modelo analítico.

Visión Conceptual



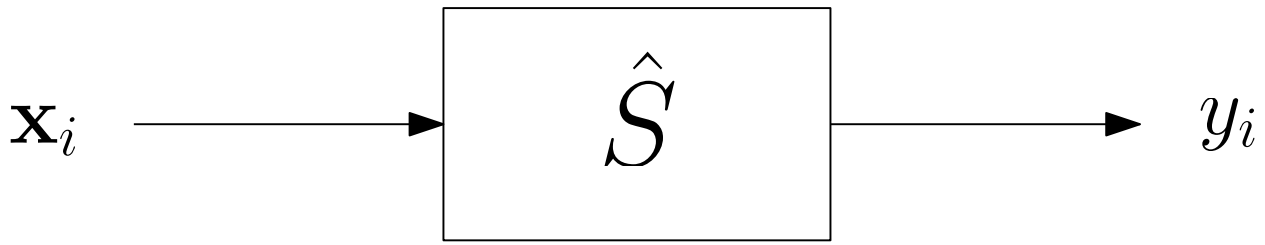
- Queremos modelar S.
- No es fácil obtener un modelo analítico.
- Usar modelo para predecir valores de la salida para nuevas entradas.

Visión Conceptual



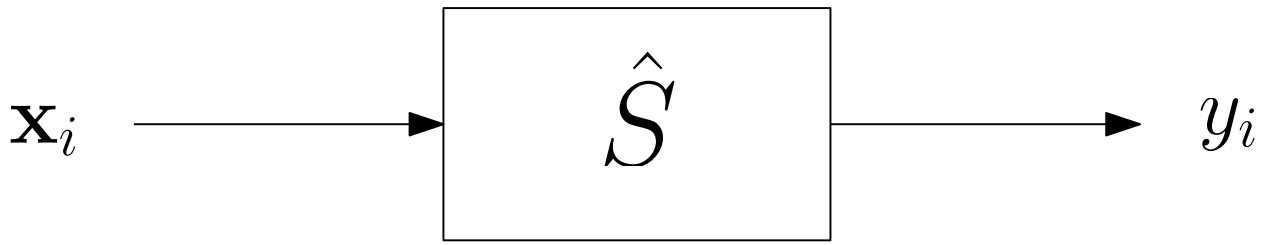
- Queremos modelar S.
- No es fácil obtener un modelo analítico.
- Usar modelo para predecir valores de la salida para nuevas entradas.

Elementos



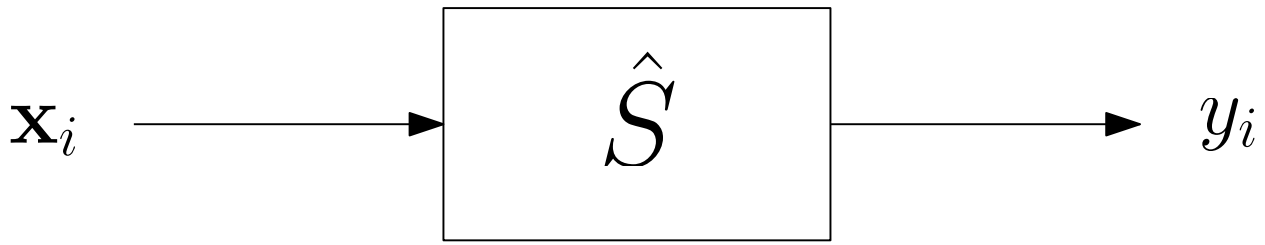
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.

Elementos



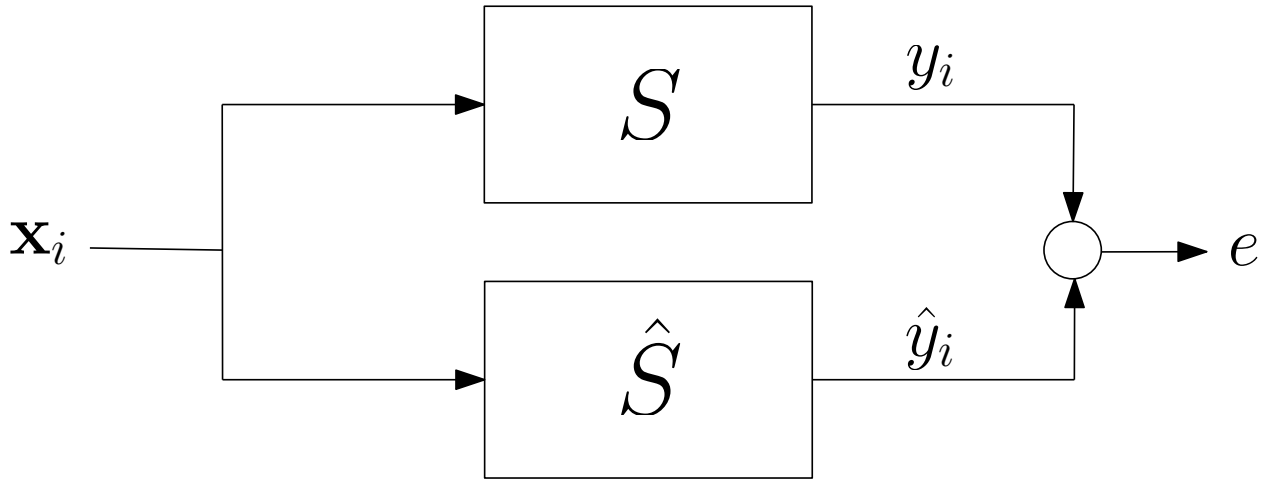
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Modelos a utilizar (NN,SVM,...)

Elementos



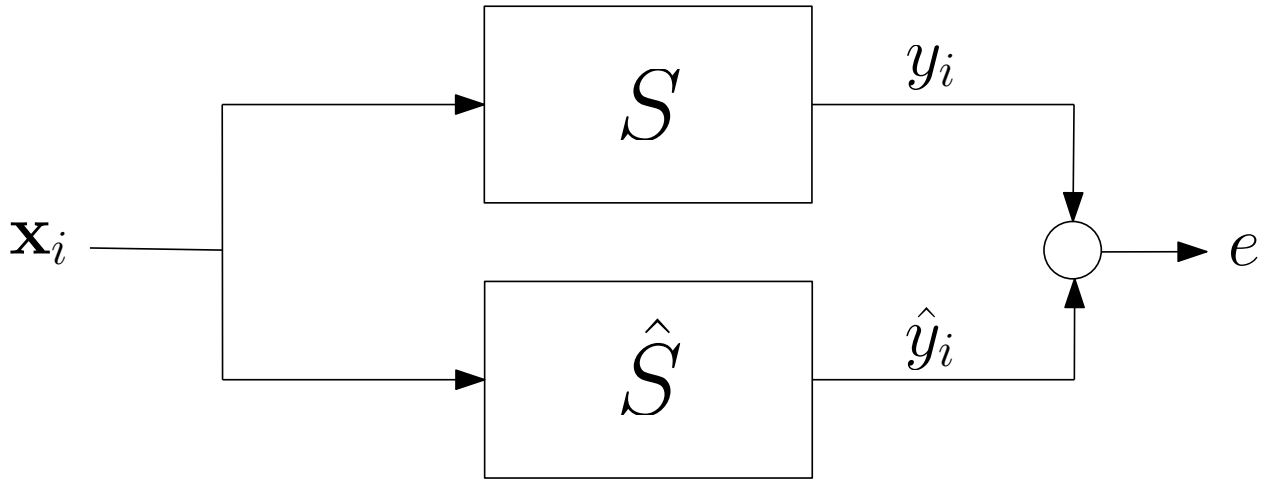
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$.
- Modelos a utilizar (NN,SVM,...)
- Conjunto de datos de prueba $\{\mathbf{x}_i, y_i\}_{i=1}^{q}$.

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

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

- Por ejemplo en clasificación:

Marco Probabilístico

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

- Por ejemplo en clasificación:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún clasificador determinístico c desconocido.

Marco Probabilístico

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

- Por ejemplo en clasificación:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún clasificador determinístico c desconocido.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbf{P}[\eta = 1] = p$.

Marco Probabilístico

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

- Por ejemplo en clasificación:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún clasificador determinístico c desconocido.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbf{P}[\eta = 1] = p$.
  - $\mathbf{x} \sim \mathcal{D}$ y $\mathbf{P}[y = 1 \mid \mathbf{x}] = \alpha(\mathbf{x})$.

Marco Probabilístico

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

- Por ejemplo en clasificación:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún clasificador determinístico c desconocido.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbf{P}[\eta = 1] = p$.
  - $\mathbf{x} \sim \mathcal{D}$ y $\mathbf{P}[y = 1 \mid \mathbf{x}] = \alpha(\mathbf{x})$.
- Por ejemplo en regresión:

Marco Probabilístico

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

- Por ejemplo en clasificación:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún clasificador determinístico c desconocido.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbf{P}[\eta = 1] = p$.
  - $\mathbf{x} \sim \mathcal{D}$ y $\mathbf{P}[y = 1 \mid \mathbf{x}] = \alpha(\mathbf{x})$.
- Por ejemplo en regresión:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = f(\mathbf{x})$ para alguna función determinística f desconocida.

Marco Probabilístico

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

- Por ejemplo en clasificación:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x})$ para algún clasificador determinístico c desconocido.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) \oplus \eta$ donde $\eta \in \{1, 0\}$ es ruido de clasificación con $\mathbf{P}[\eta = 1] = p$.
  - $\mathbf{x} \sim \mathcal{D}$ y $\mathbf{P}[y = 1 \mid \mathbf{x}] = \alpha(\mathbf{x})$.
- Por ejemplo en regresión:
  - $\mathbf{x} \sim \mathcal{D}$ y $y = f(\mathbf{x})$ para alguna función determinística f desconocida.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = f(\mathbf{x}) + \eta$ donde $\eta \sim \mathcal{D}_\eta$

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

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

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

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

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

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

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

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

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

Aprendizaje?

1. Datos:

Aprendizaje?

1. Datos: Entrenamiento/Prueba.
2. Modelo

Aprendizaje?

1. Datos: Entrenamiento/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos

Aprendizaje?

1. Datos: Entrenamiento/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:

Aprendizaje?

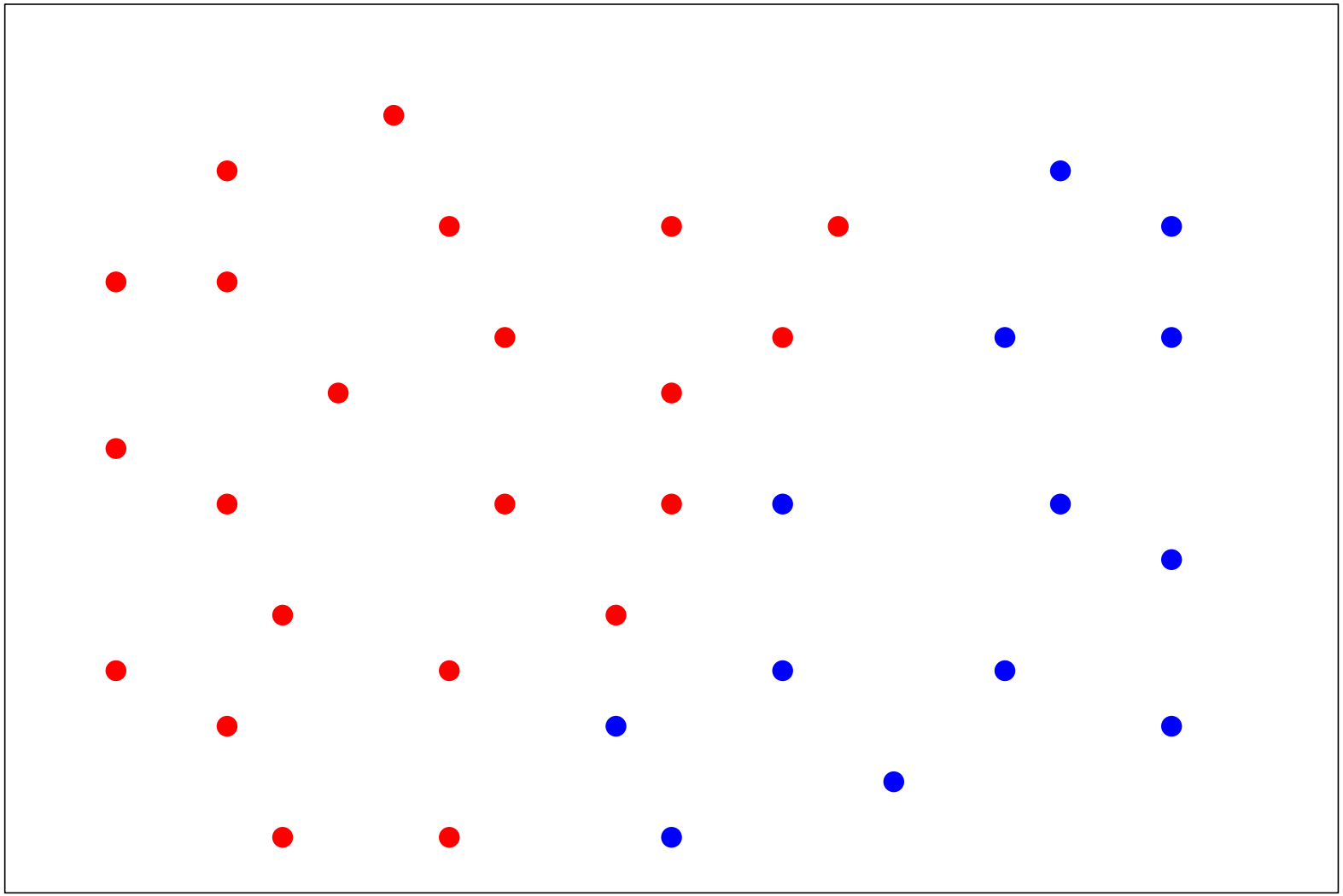
1. Datos: Entrenamiento/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error

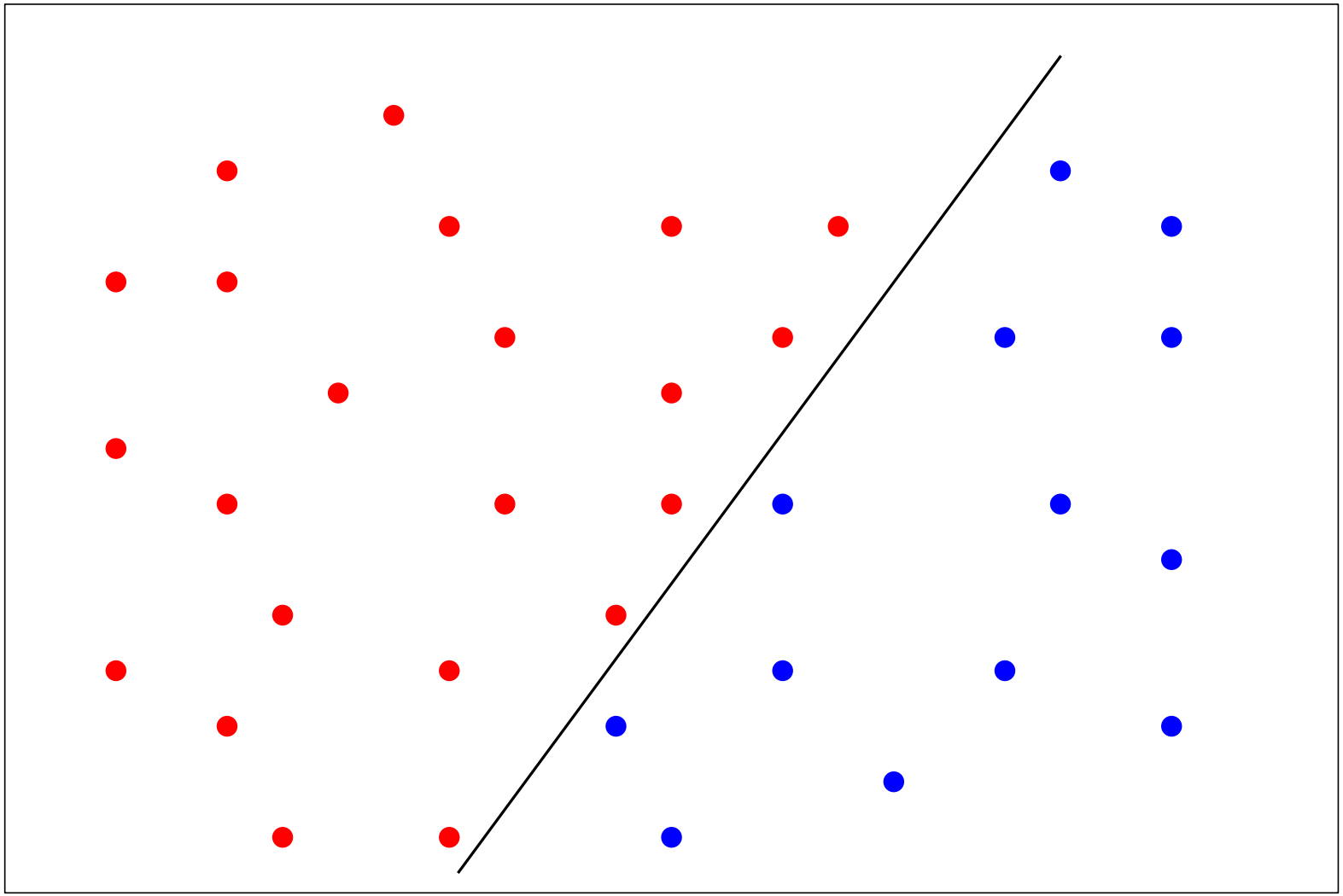
Aprendizaje?

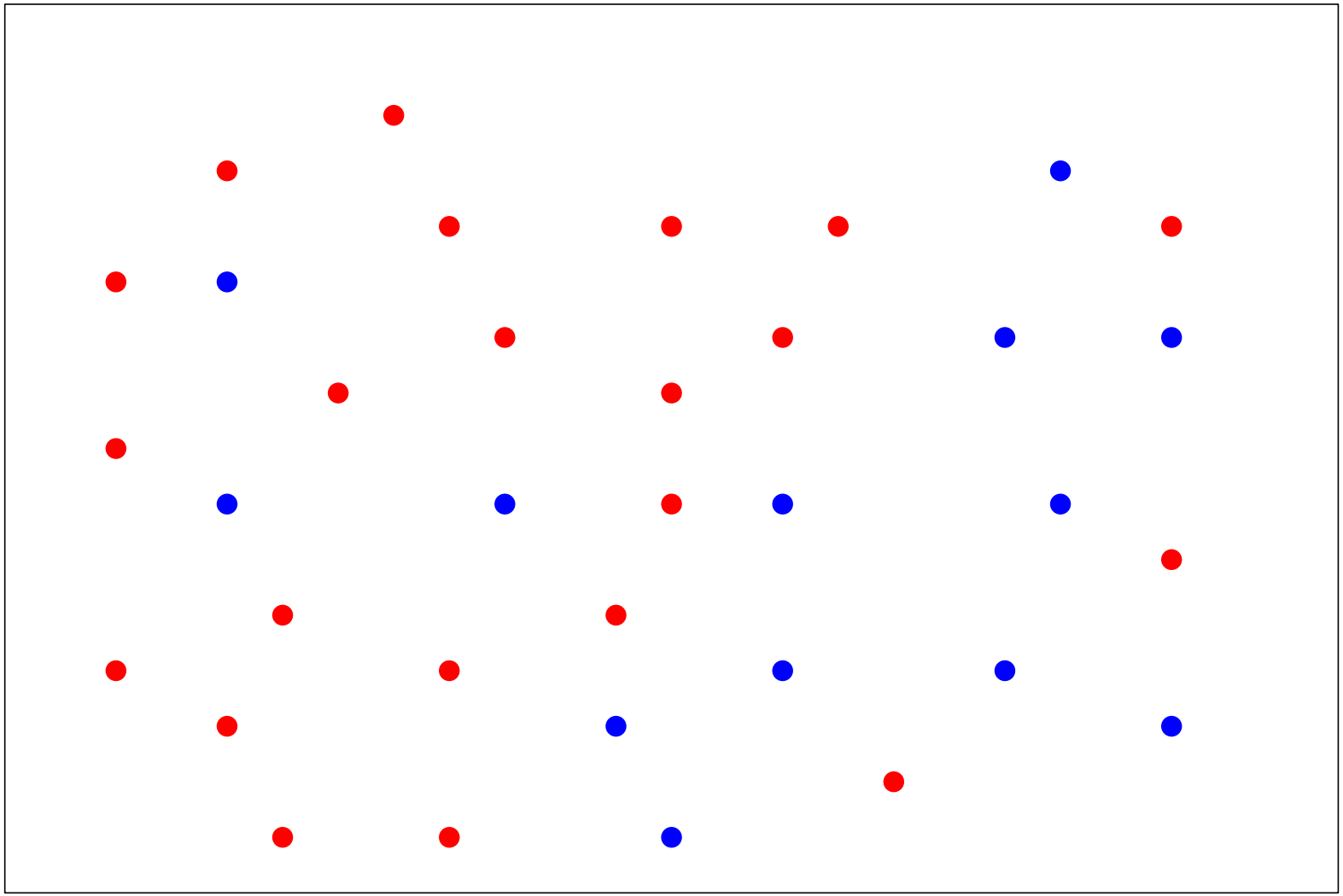
1. Datos: Entrenamiento/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Optimización de función de error en datos de entrenamiento.

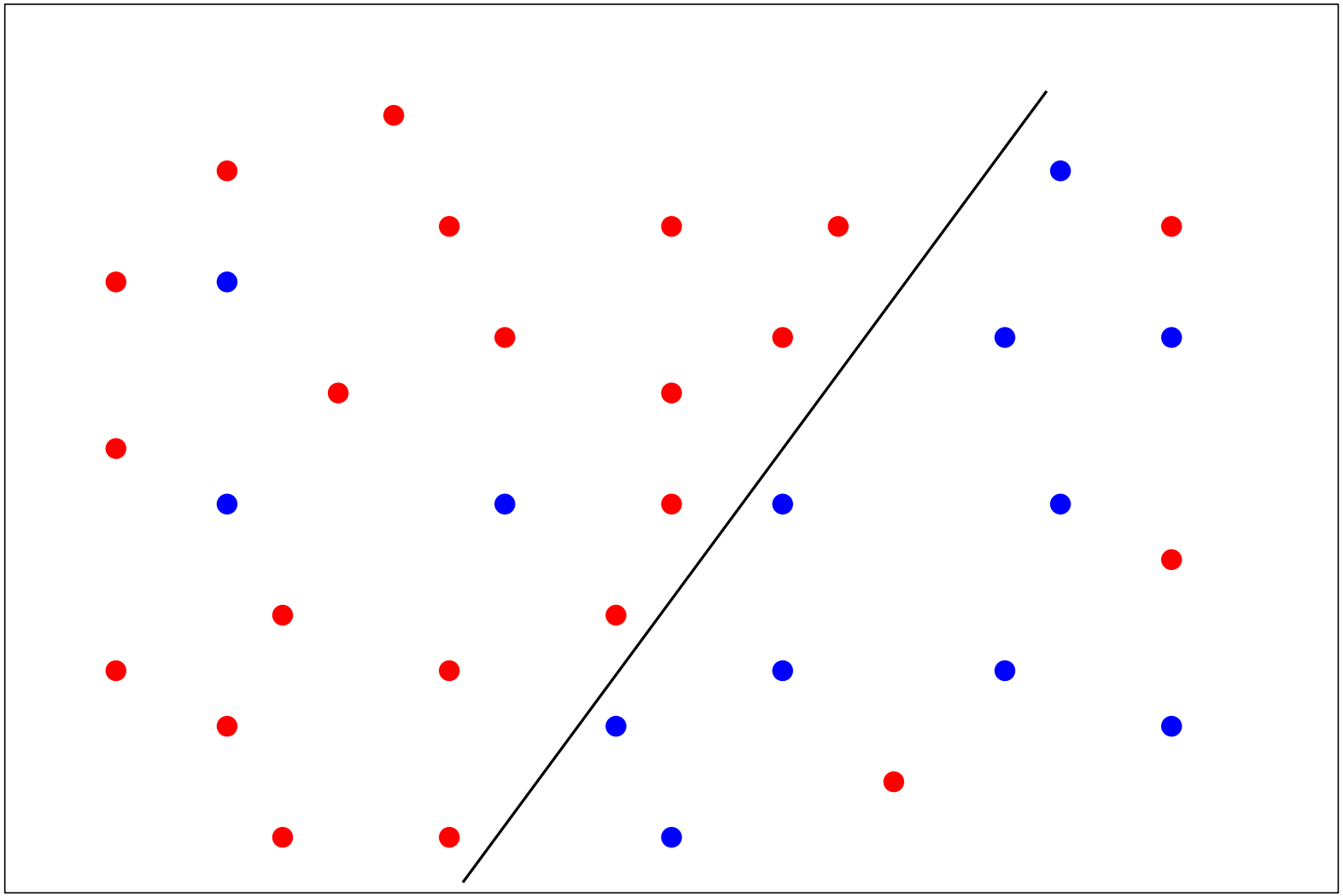
Aprendizaje?

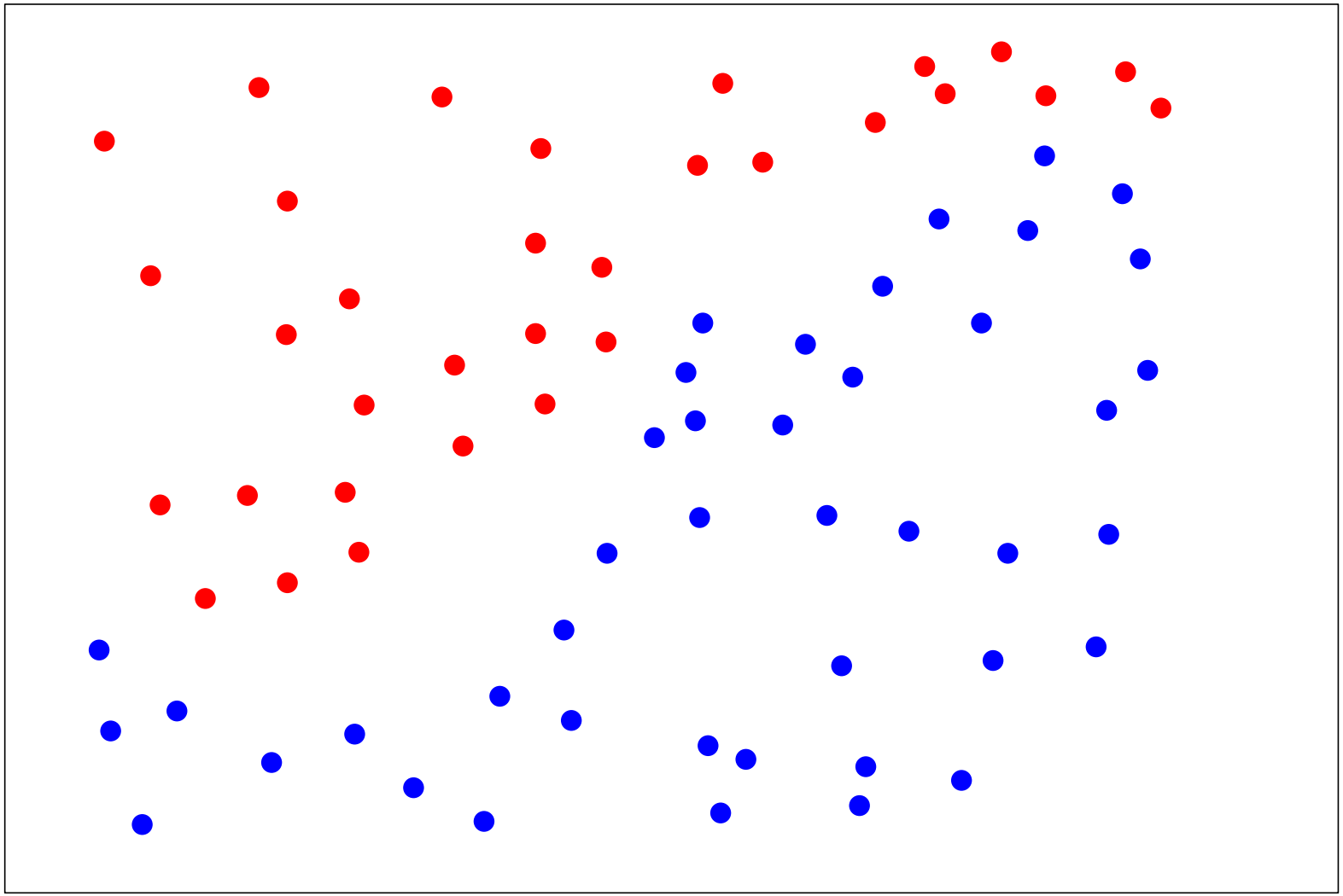
1. Datos: Entrenamiento/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Optimización de función de error en datos de entrenamiento.
5. Evaluar modelo en datos prueba.

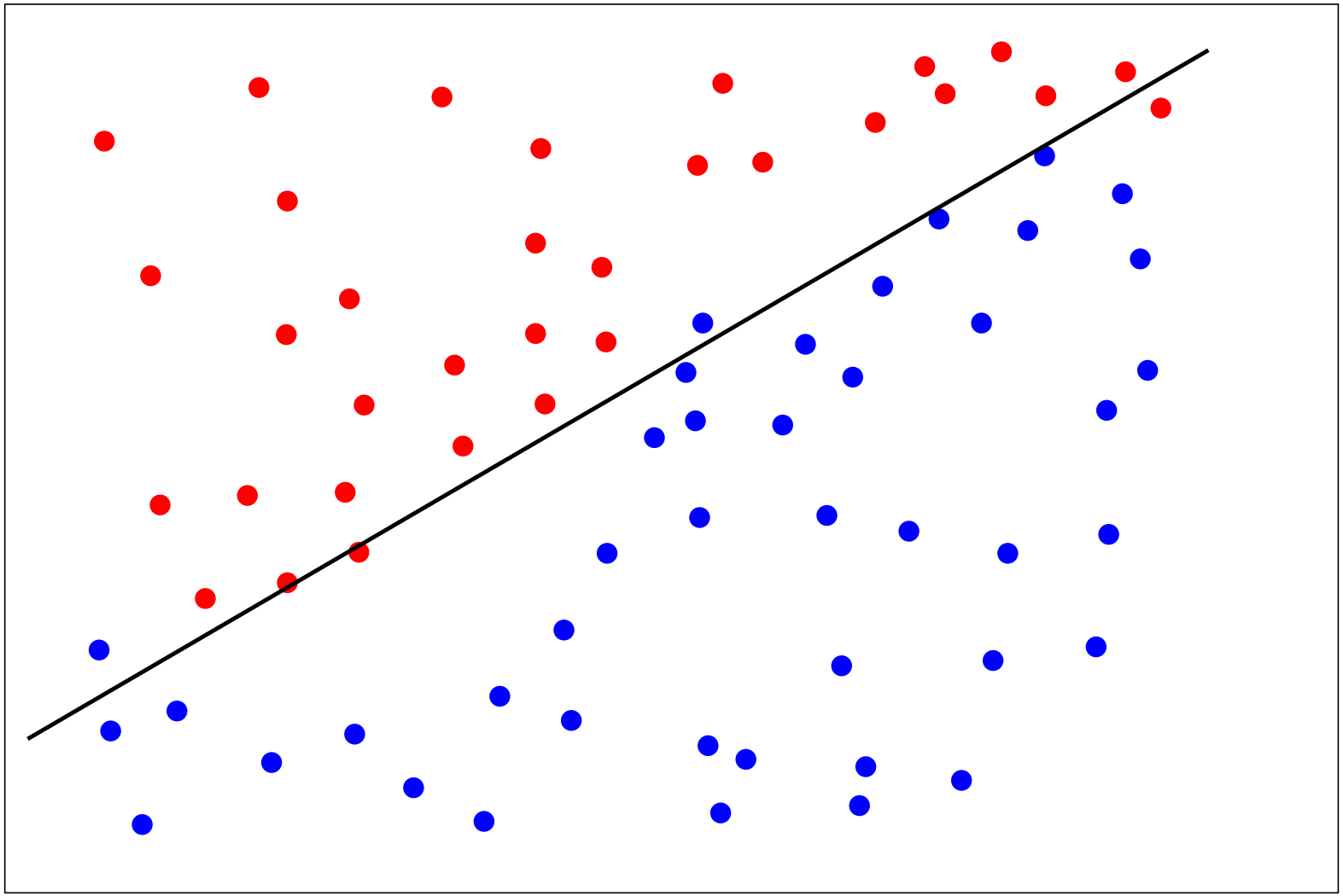


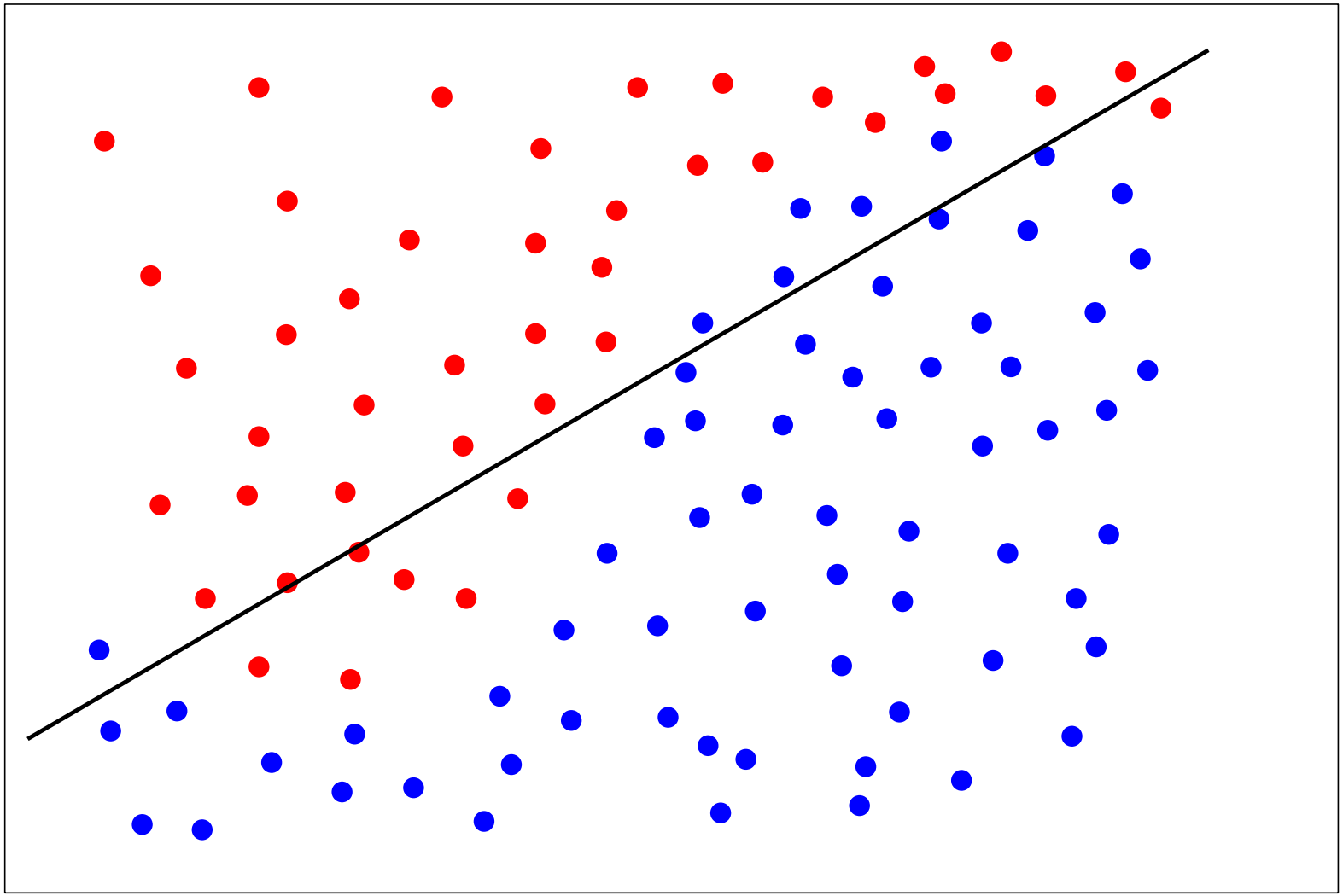


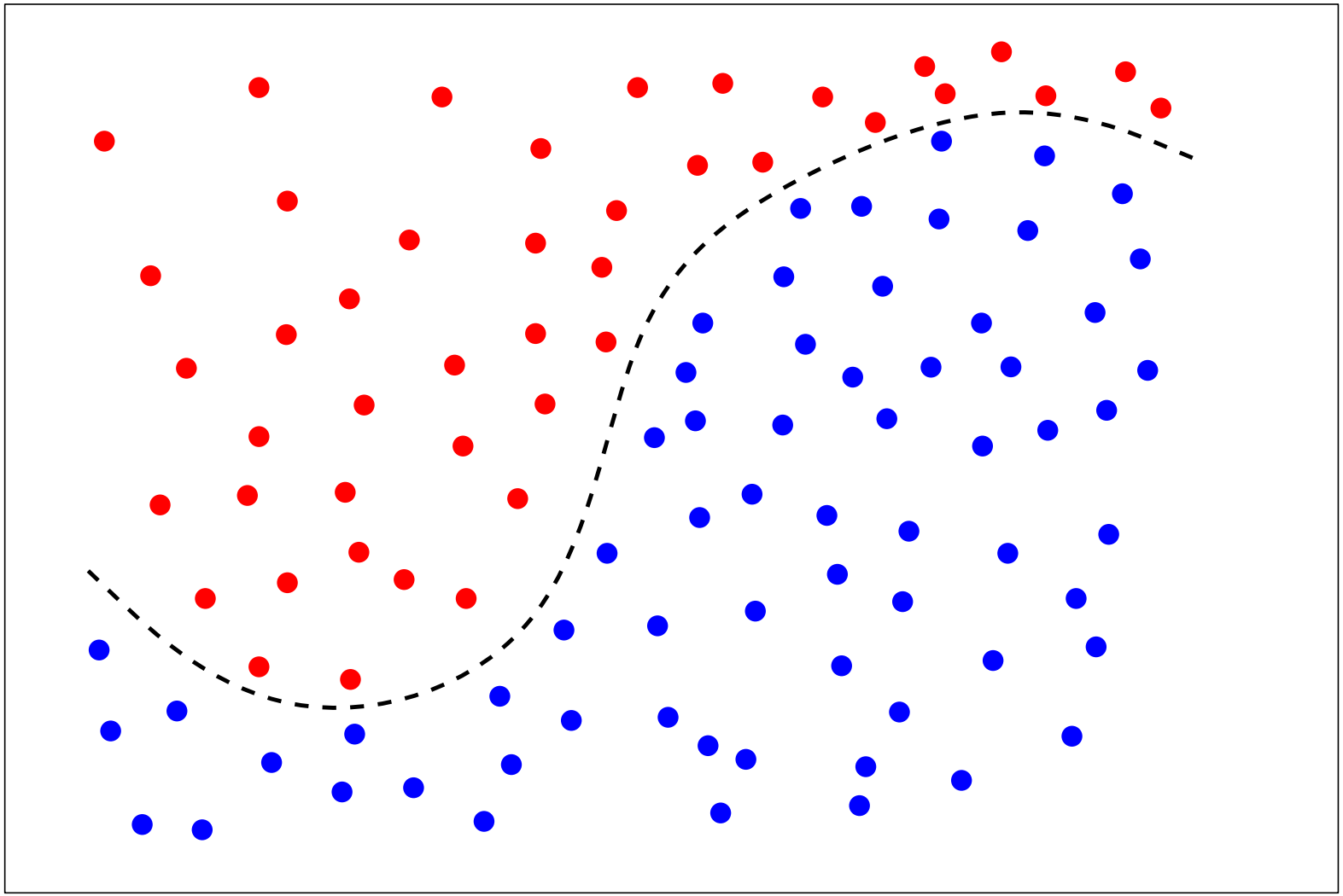


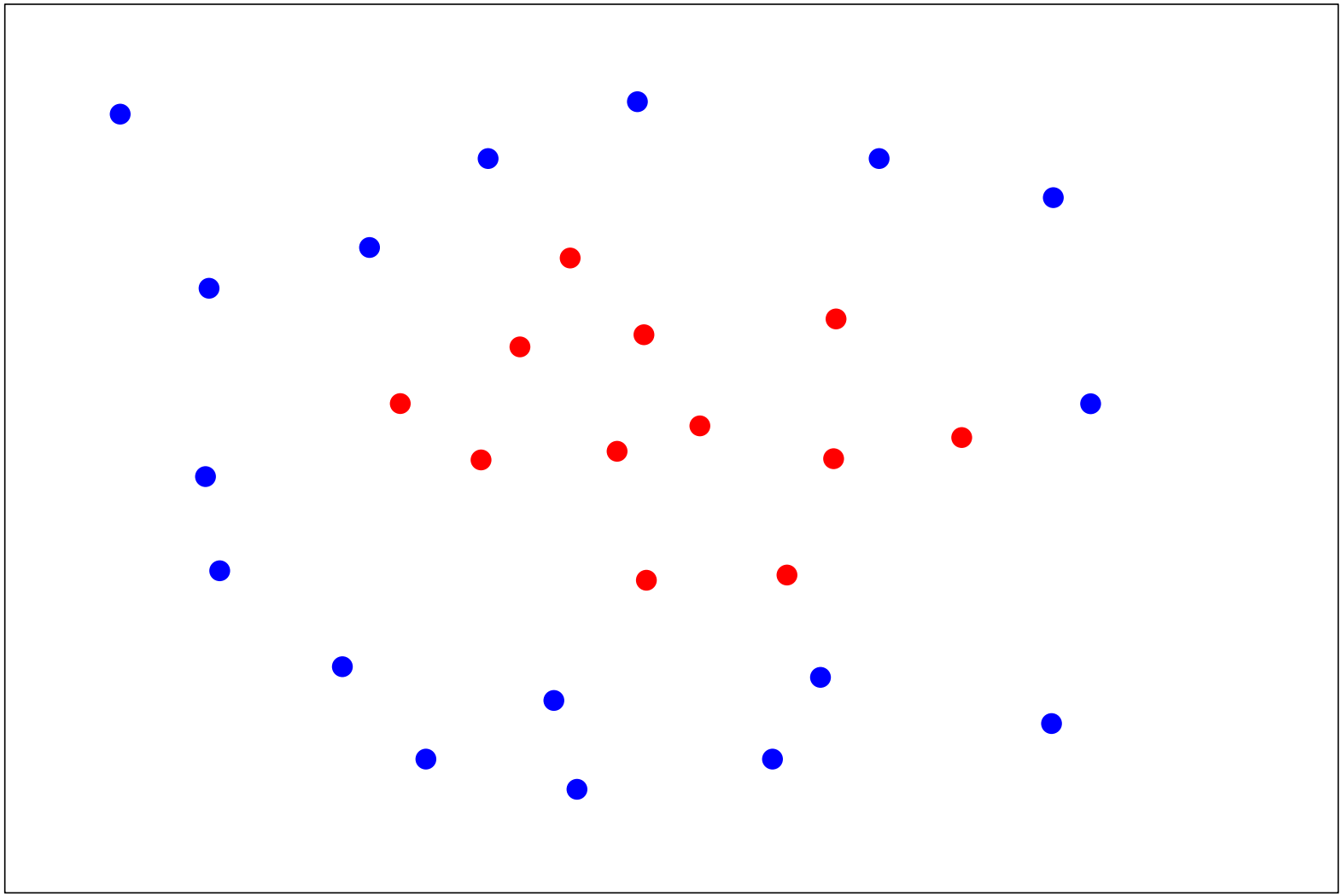


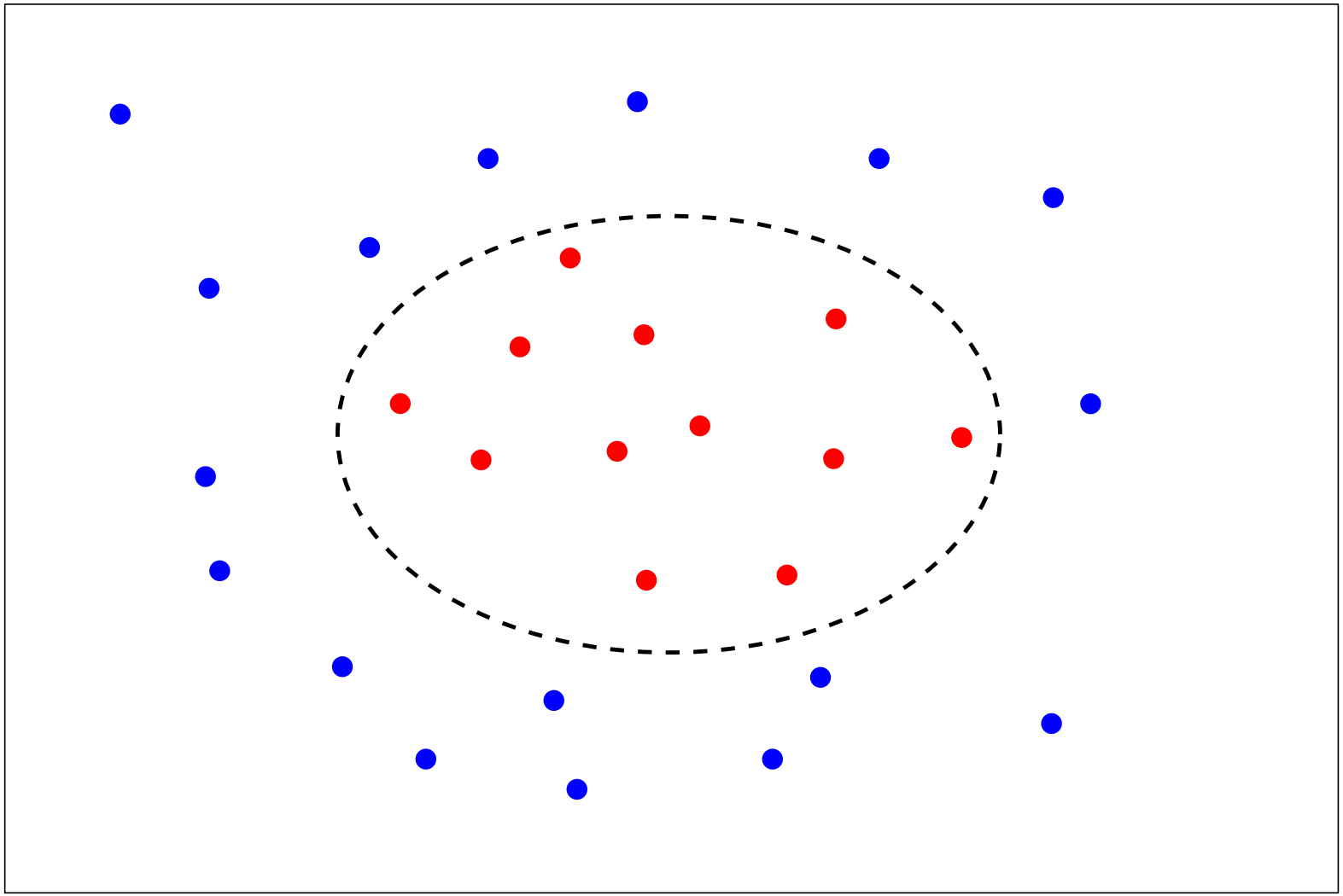


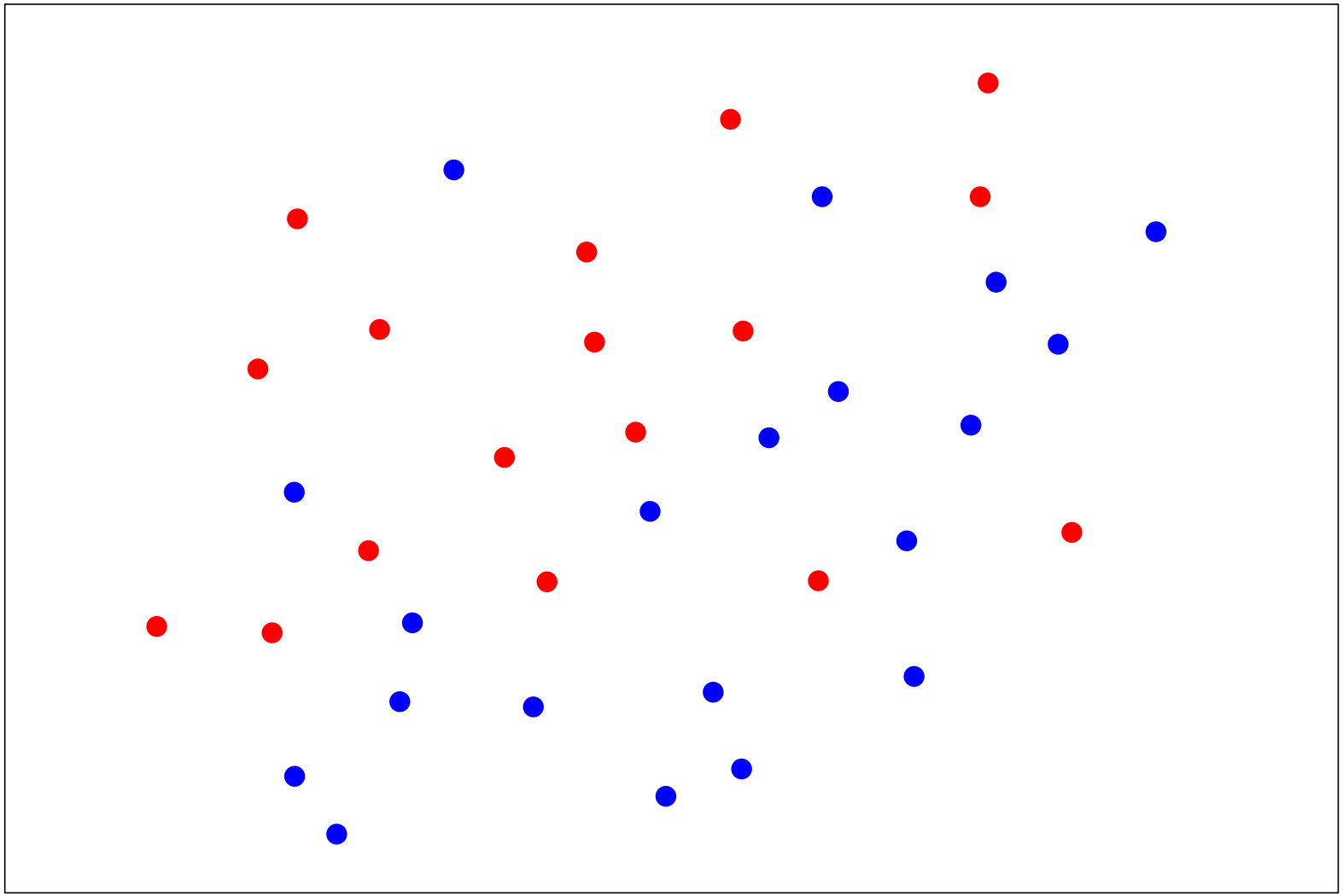


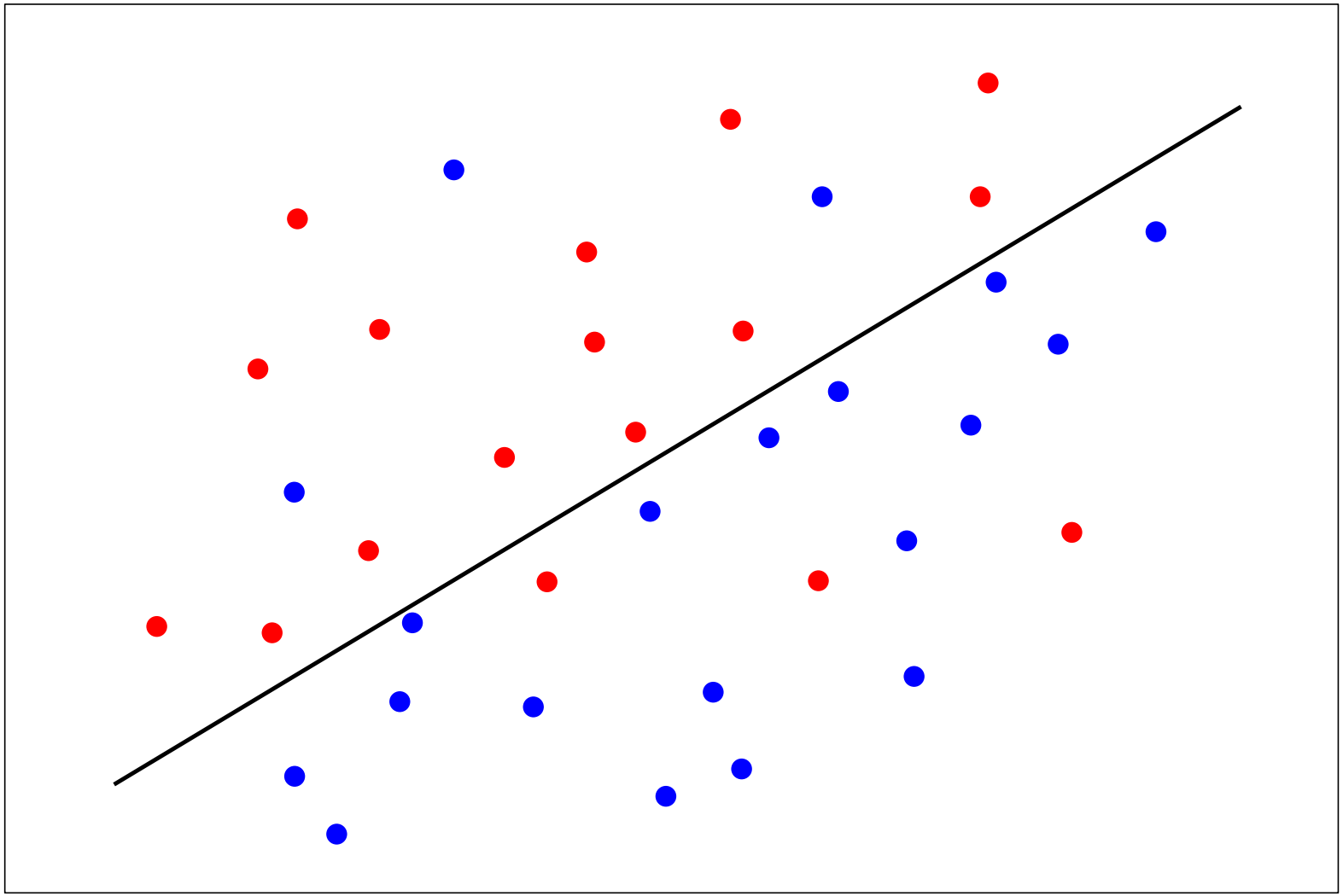


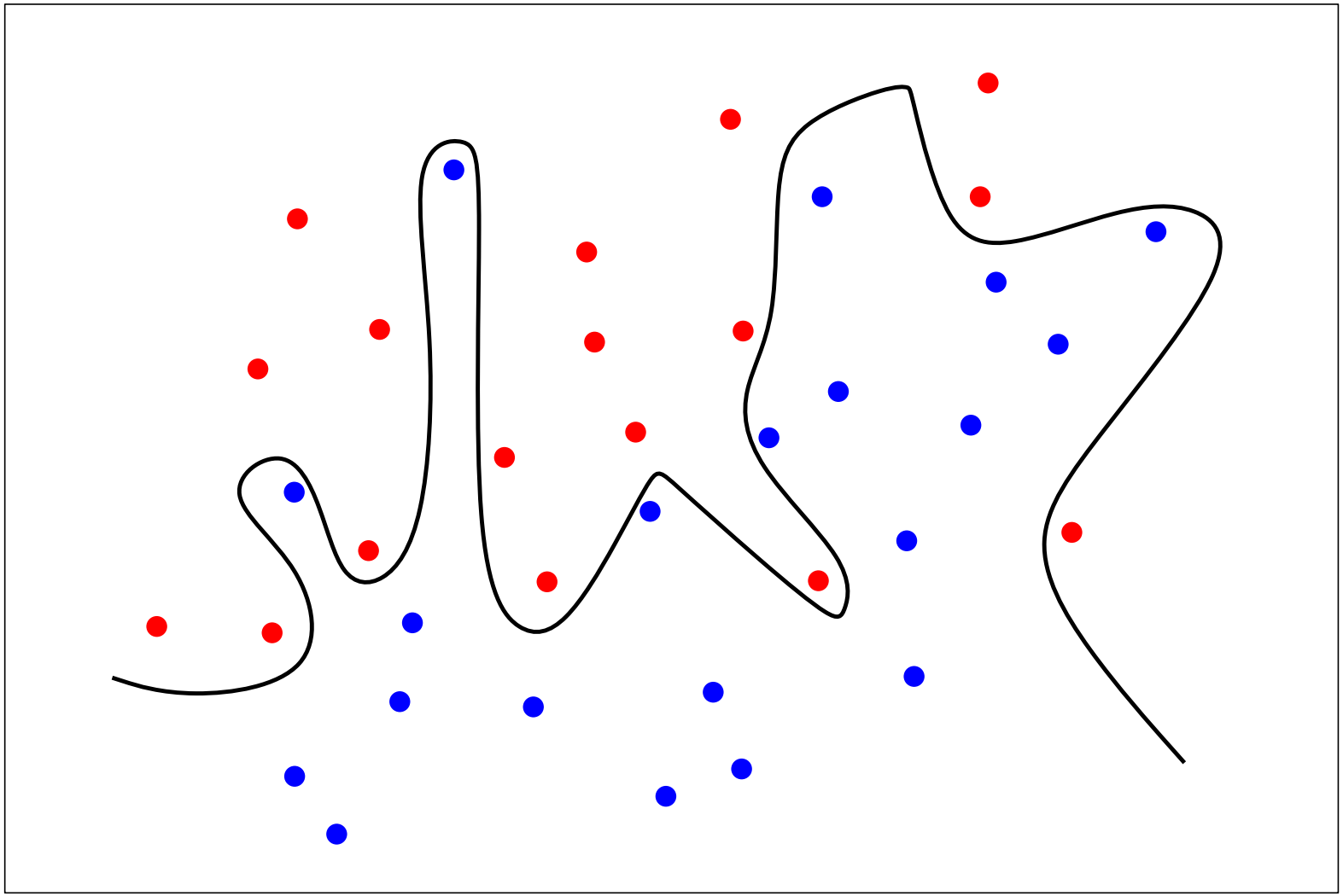


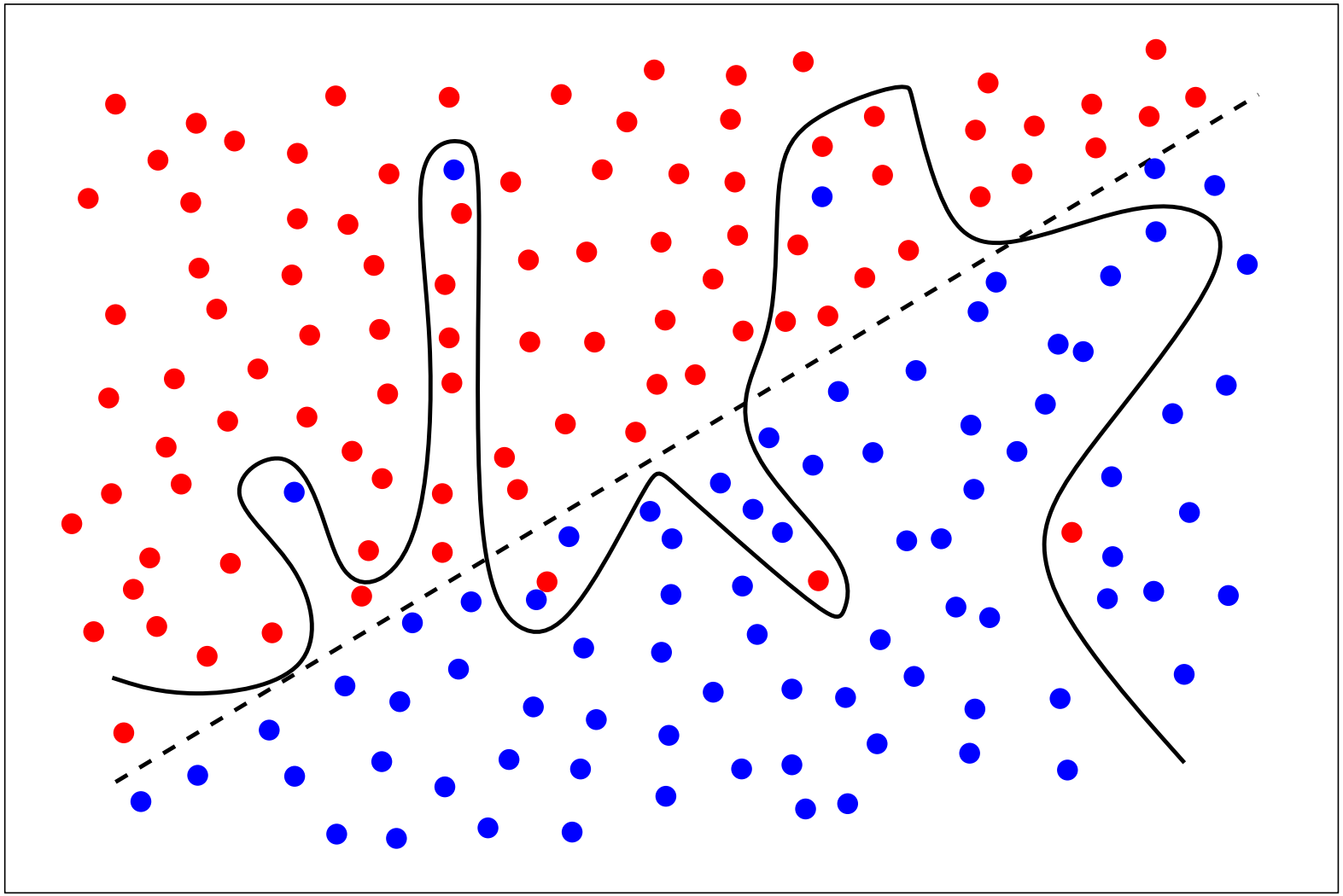


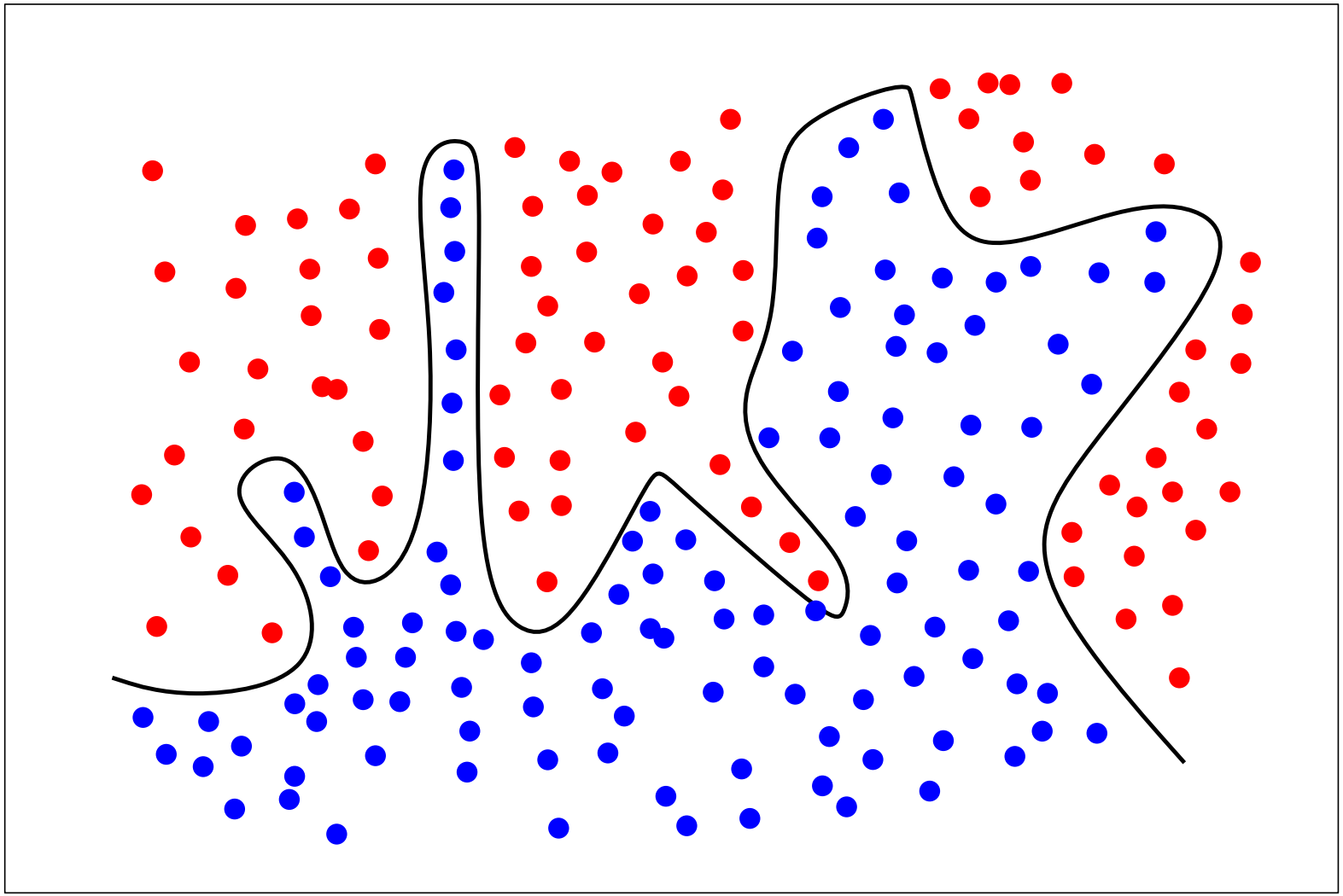


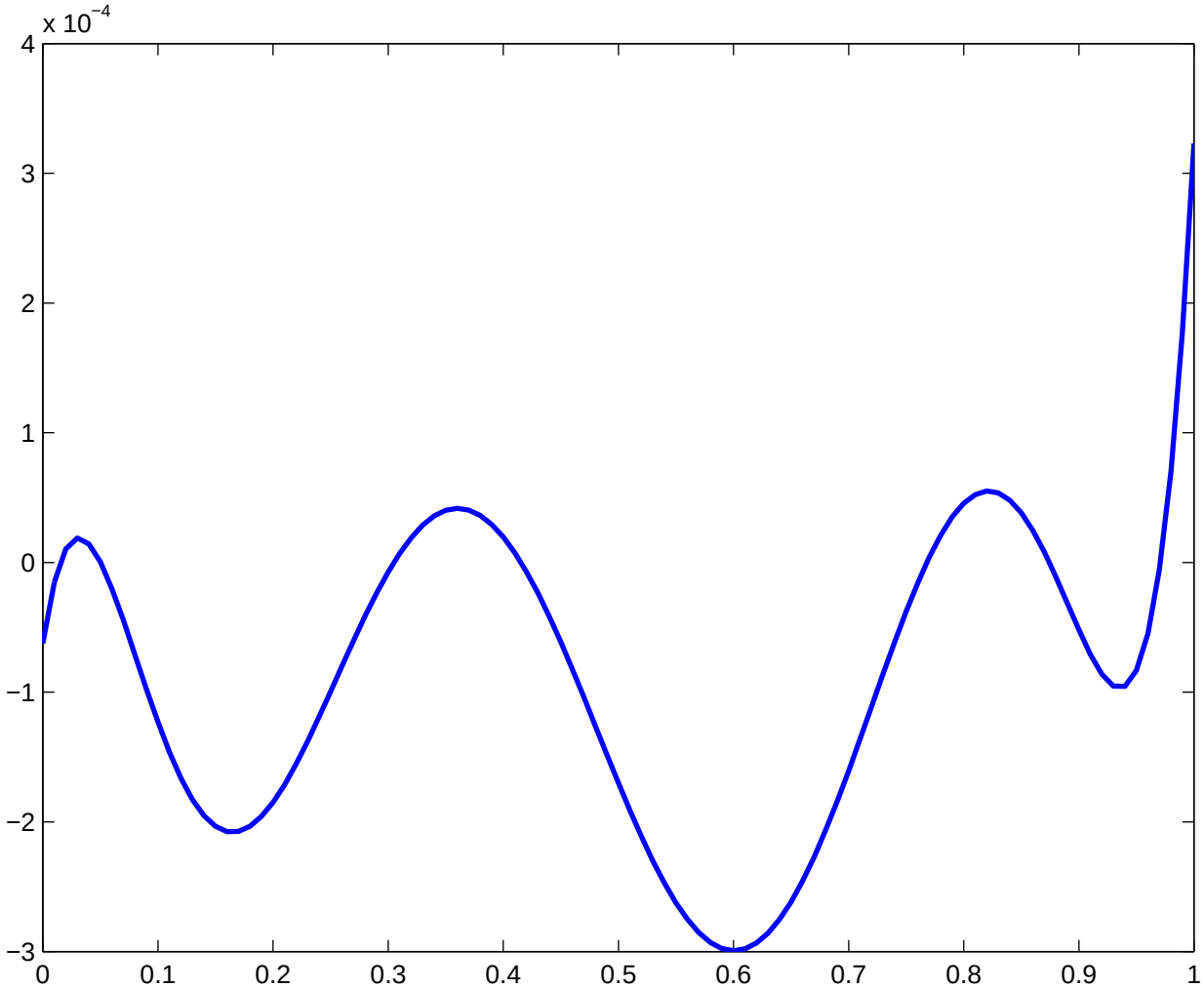


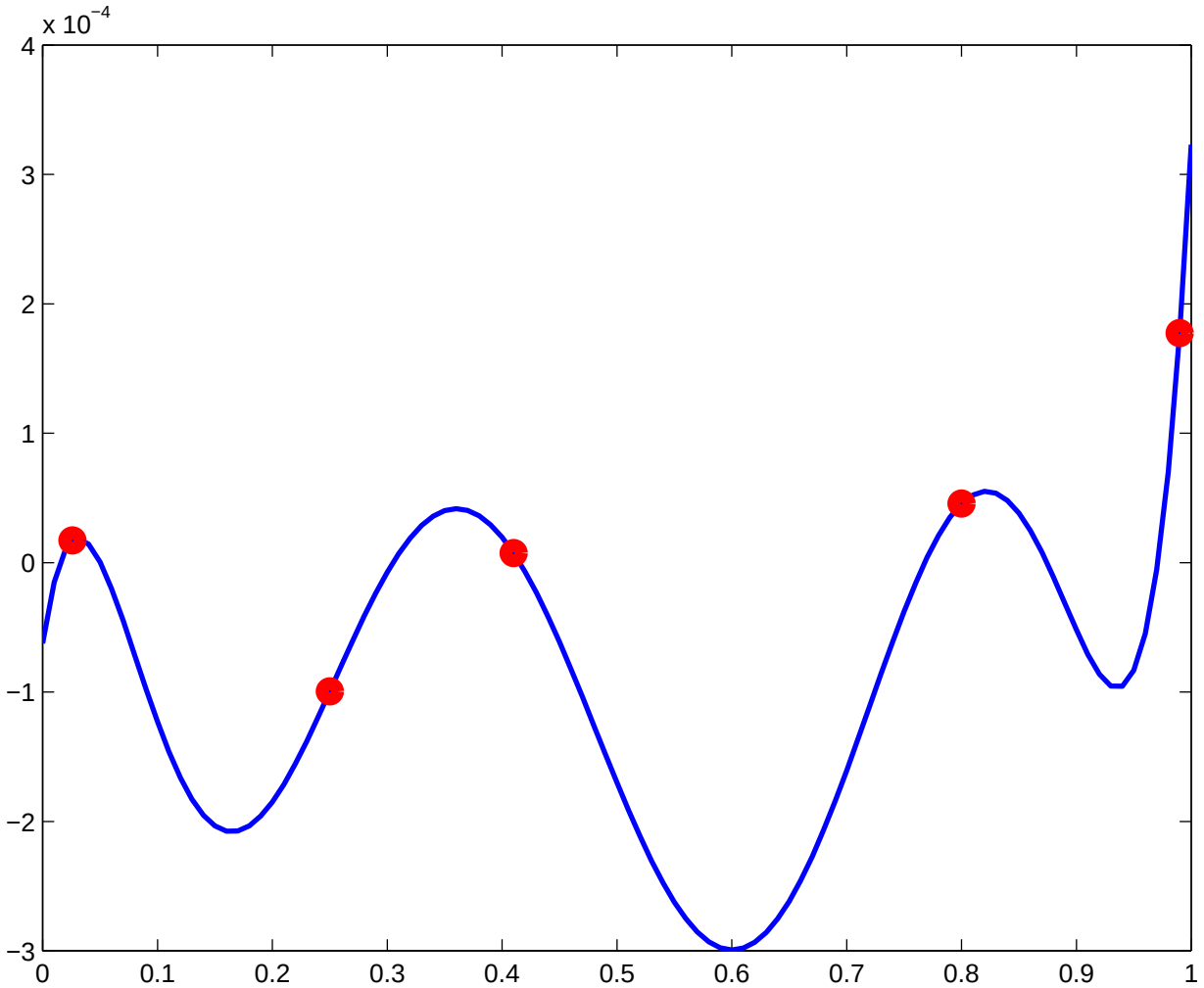


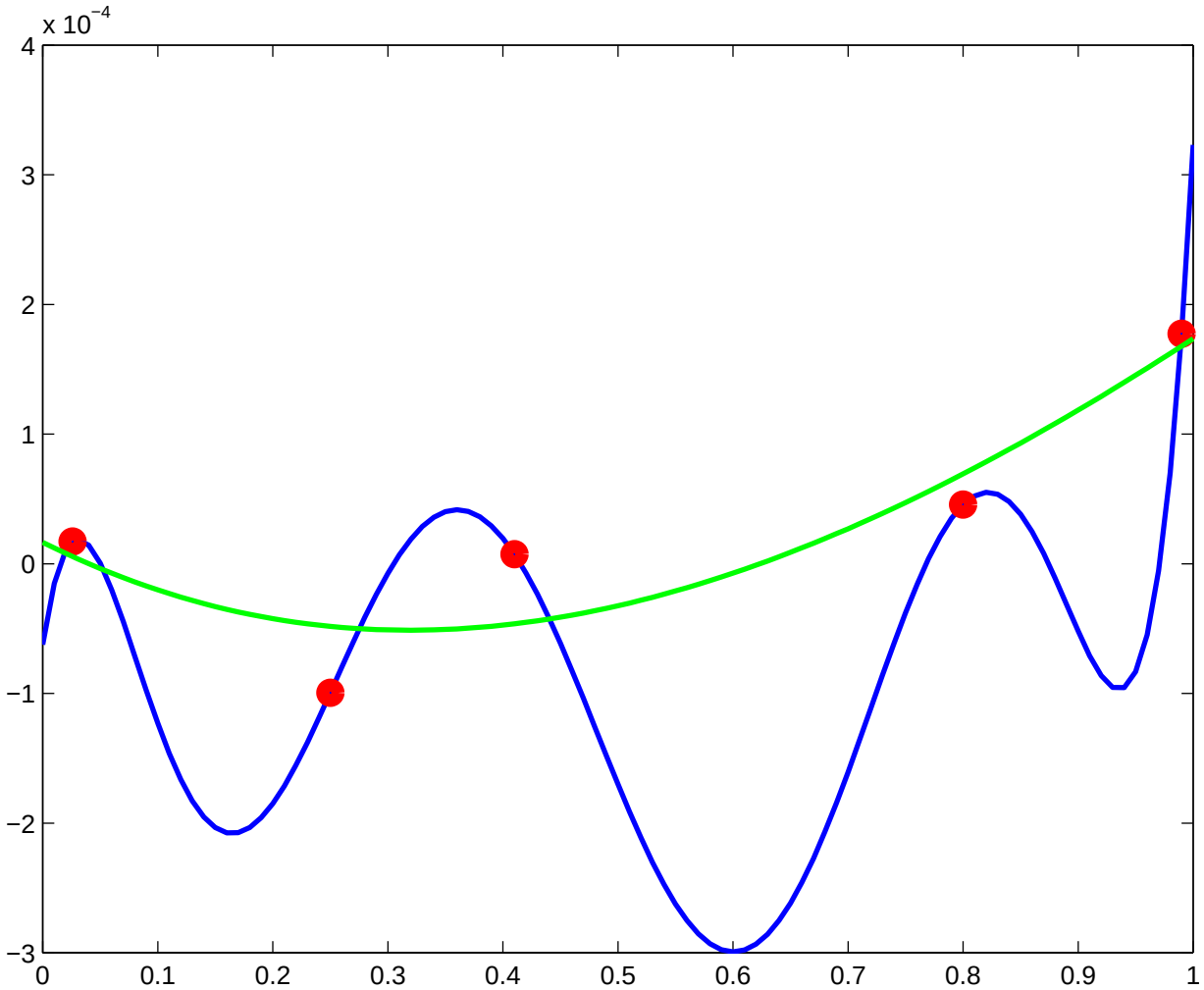


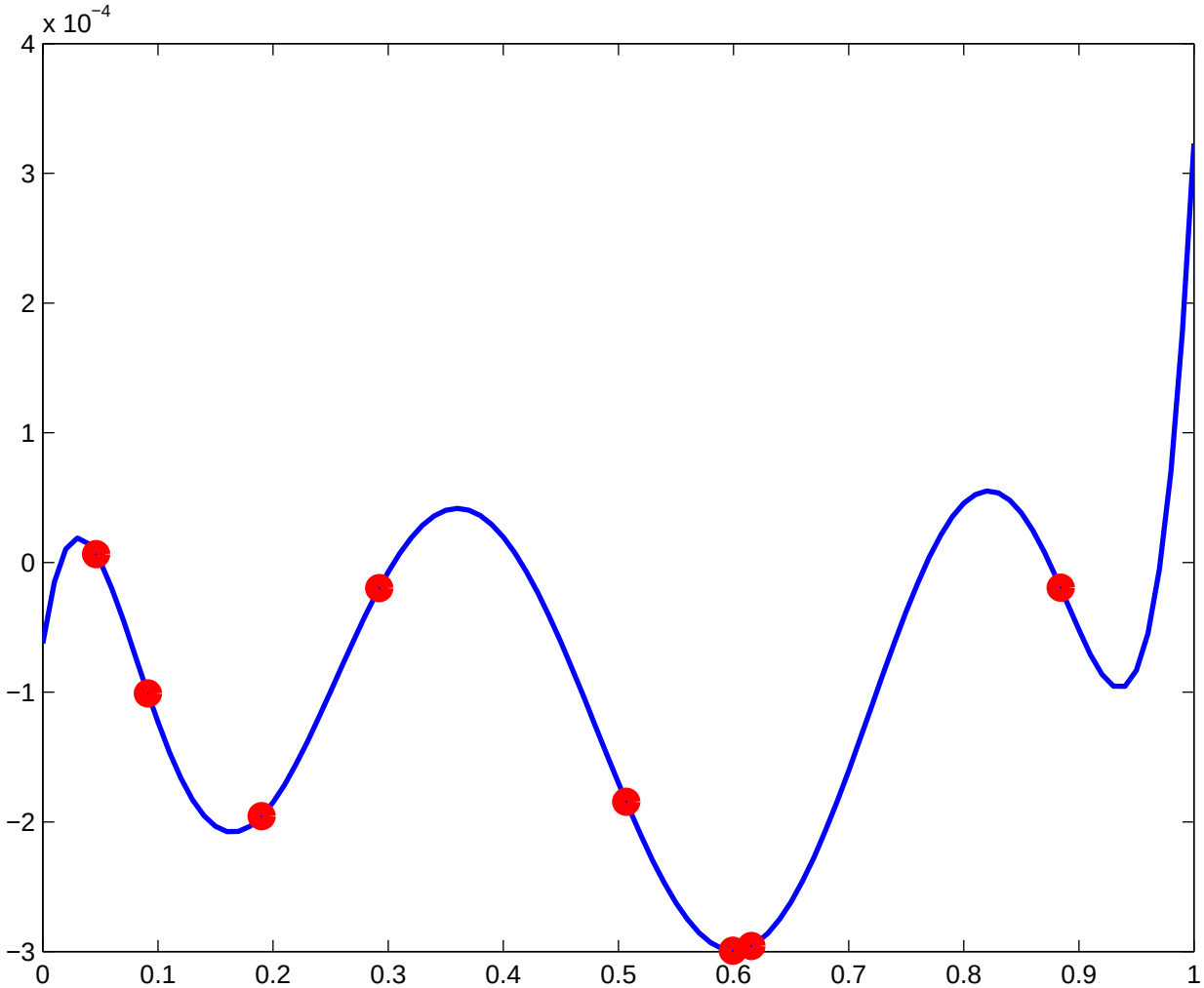


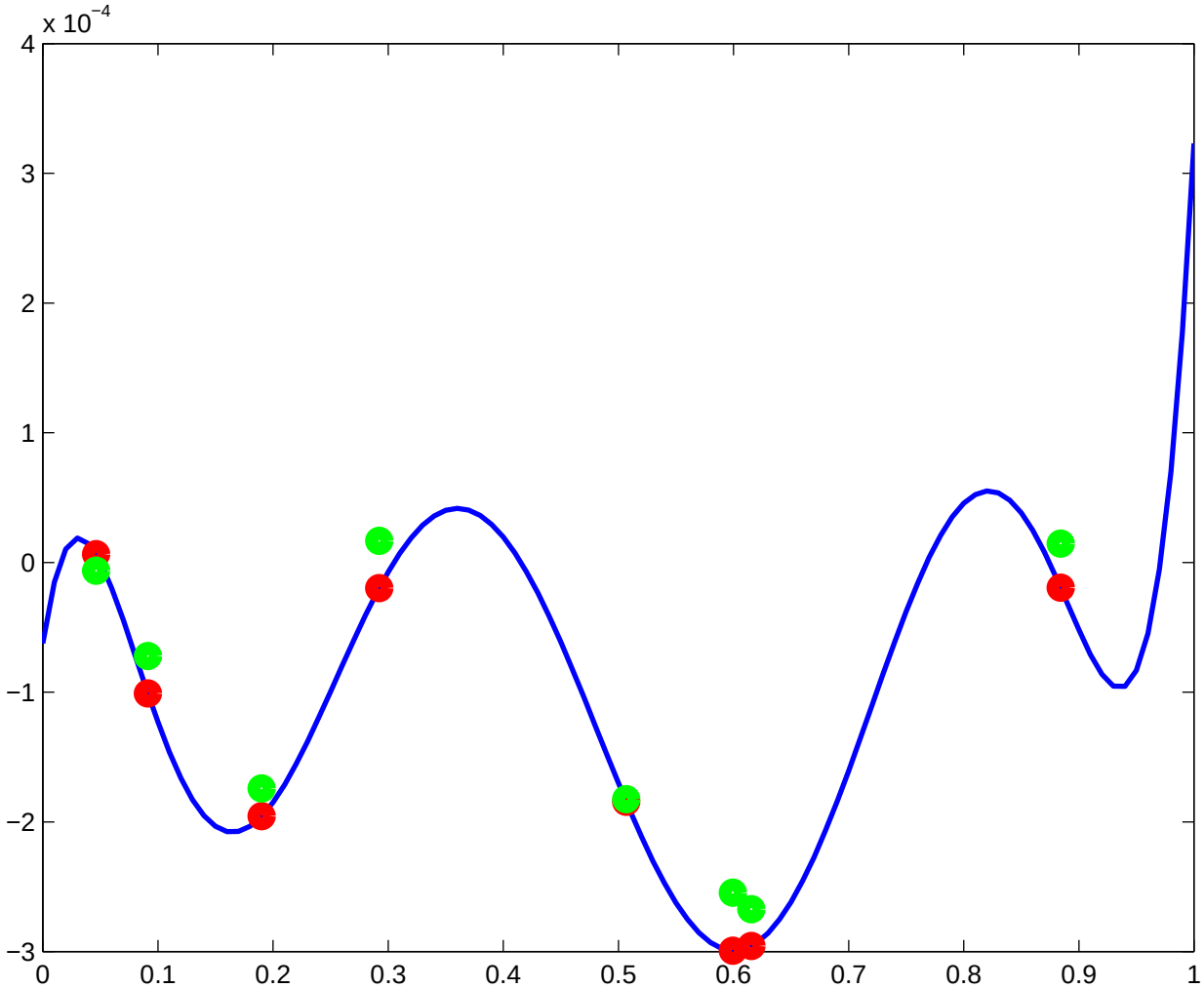


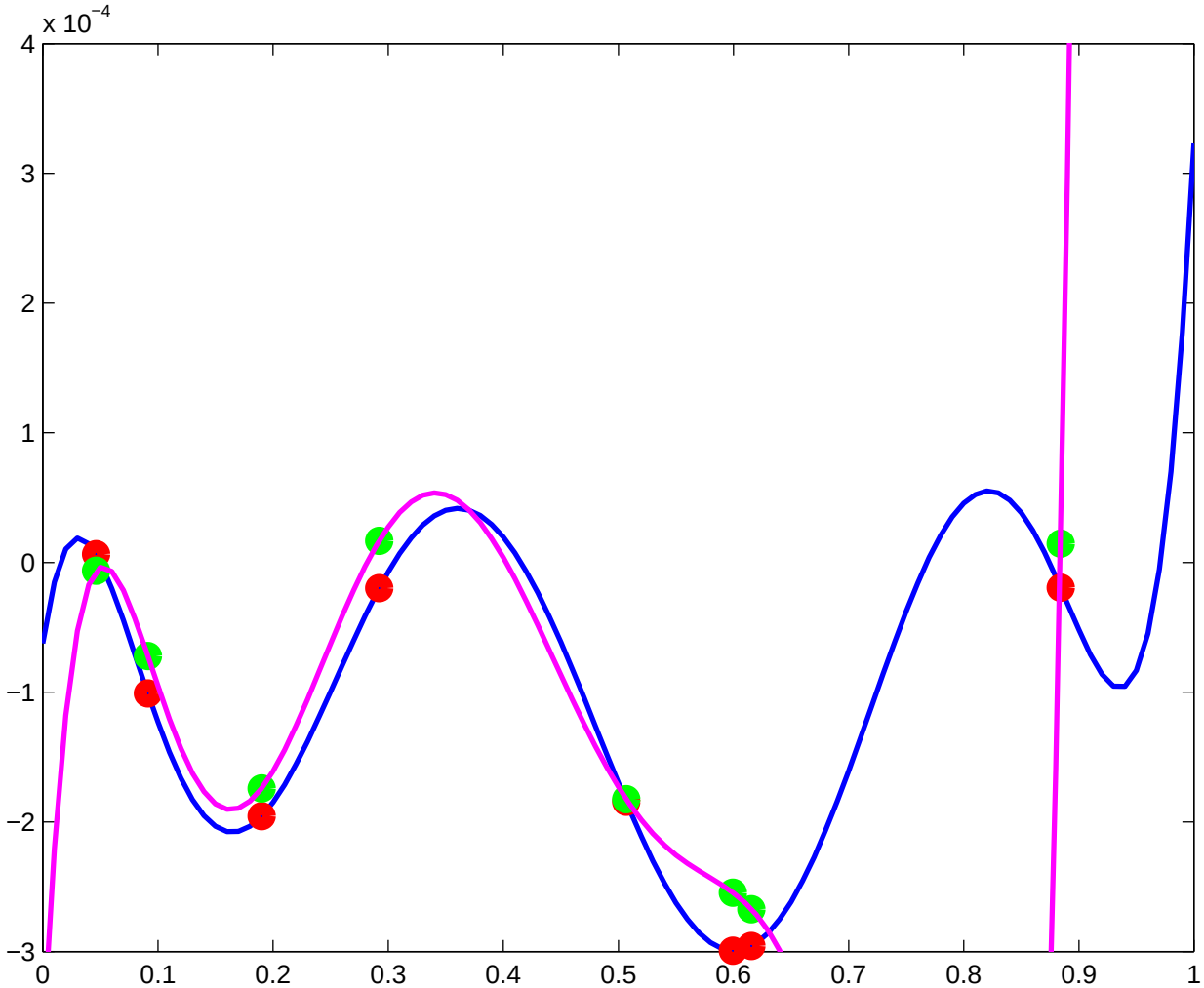

x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

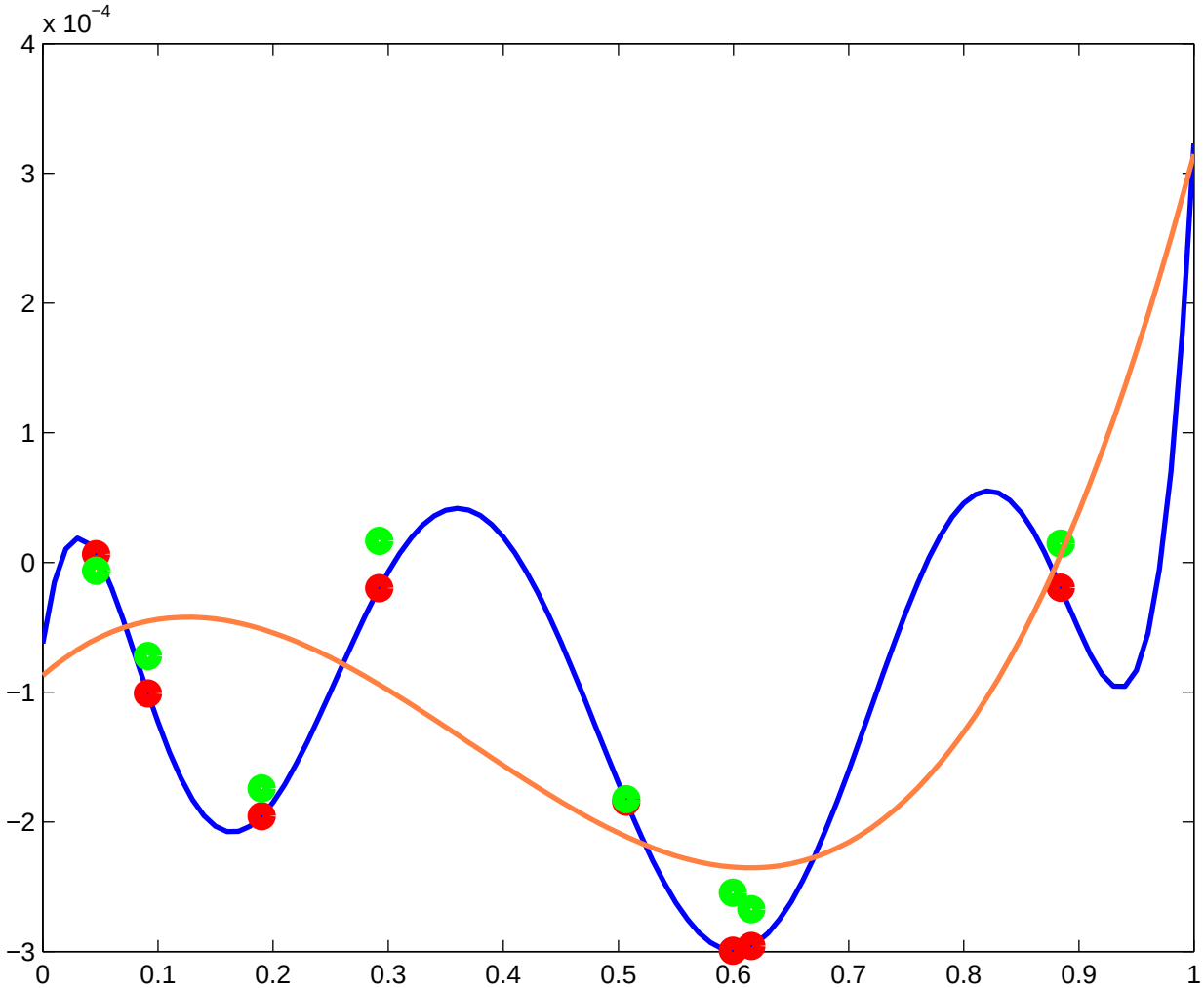

x 10^-4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


x 10^-4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Aprendizaje (selección de modelo)

1. Datos:

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Para modelos de diferentes “tamaños” o “complejidad”:

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Para modelos de diferentes “tamaños” o “complejidad”:
        1. Optimización de función de error en datos de entrenamiento.

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Para modelos de diferentes "tamaños" o "complejidad":
        1. Optimización de función de error en datos de entrenamiento.
        2. Evaluar modelo en datos de validación.

Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Para modelos de diferentes "tamaños" o "complejidad":
        1. Optimización de función de error en datos de entrenamiento.
        2. Evaluar modelo en datos de validación.
    3. Seleccionar modelo con menor error en datos de validación

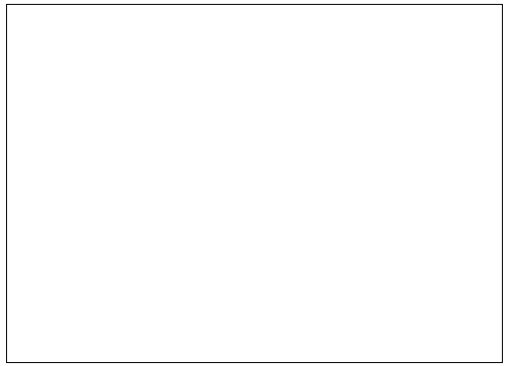
Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Para modelos de diferentes "tamaños" o "complejidad":
        1. Optimización de función de error en datos de entrenamiento.
        2. Evaluar modelo en datos de validación.
    3. Seleccionar modelo con menor error en datos de validación
5. Evaluar modelo en datos prueba.

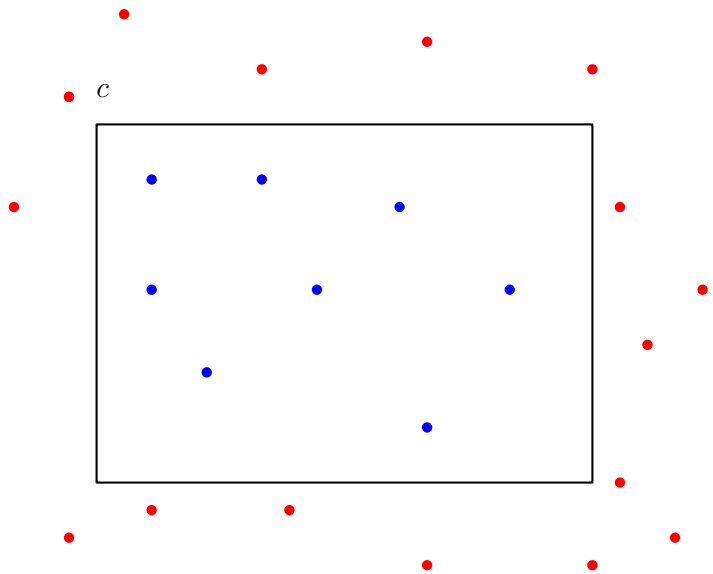
Aprendizaje (selección de modelo)

1. Datos: Entrenamiento/Validación/Prueba.
2. Modelo
3. Adecuación/ Preprocesamiento de los datos
4. Entrenamiento:
    1. Función de error
    2. Para modelos de diferentes "tamaños" o "complejidad":
        1. Optimización de función de error en datos de entrenamiento.
        2. Evaluar modelo en datos de validación.
    3. Seleccionar modelo con menor error en datos de validación
5. Evaluar modelo en datos prueba.

Ejemplo: Aprendizaje de rectángulos en $\mathbb{R}^2$

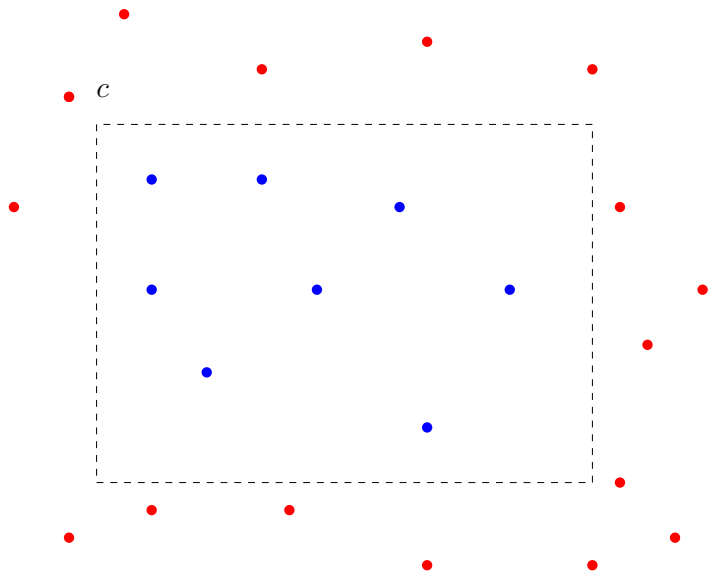


Ejemplo: Aprendizaje de rectángulos en $\mathbb{R}^2$



- $\mathbf{x}_i \sim \mathcal{D}$, independientes.
- $y_i = c(\mathbf{x}_i) \equiv I_c(\mathbf{x}_i)$

Ejemplo: Aprendizaje de rectángulos en $\mathbb{R}^2$



- $\mathbf{x}_i \sim \mathcal{D}$, independientes.
- $y_i = c(\mathbf{x}_i) \equiv I_c(\mathbf{x}_i)$

Algoritmo

- Encontrar a partir de los datos un rectángulo que minimice error:

Algoritmo

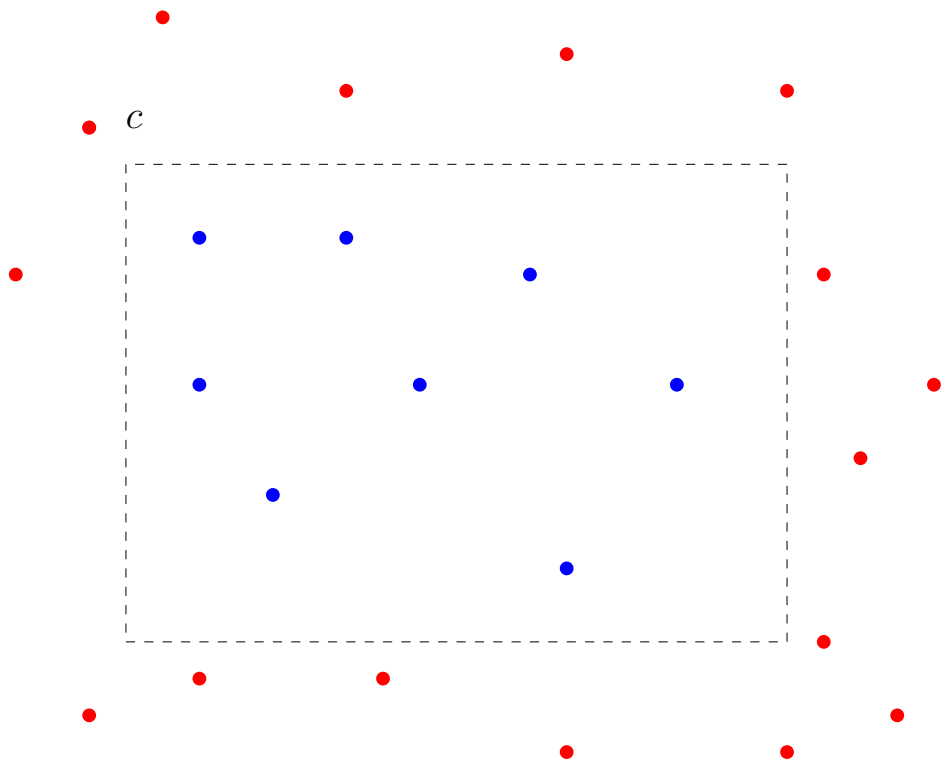
- Encontrar a partir de los datos un rectángulo que minimice error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[c(\mathbf{x}) \neq h(\mathbf{x})\right]$$

Algoritmo

- Encontrar a partir de los datos un rectángulo que minimice error:

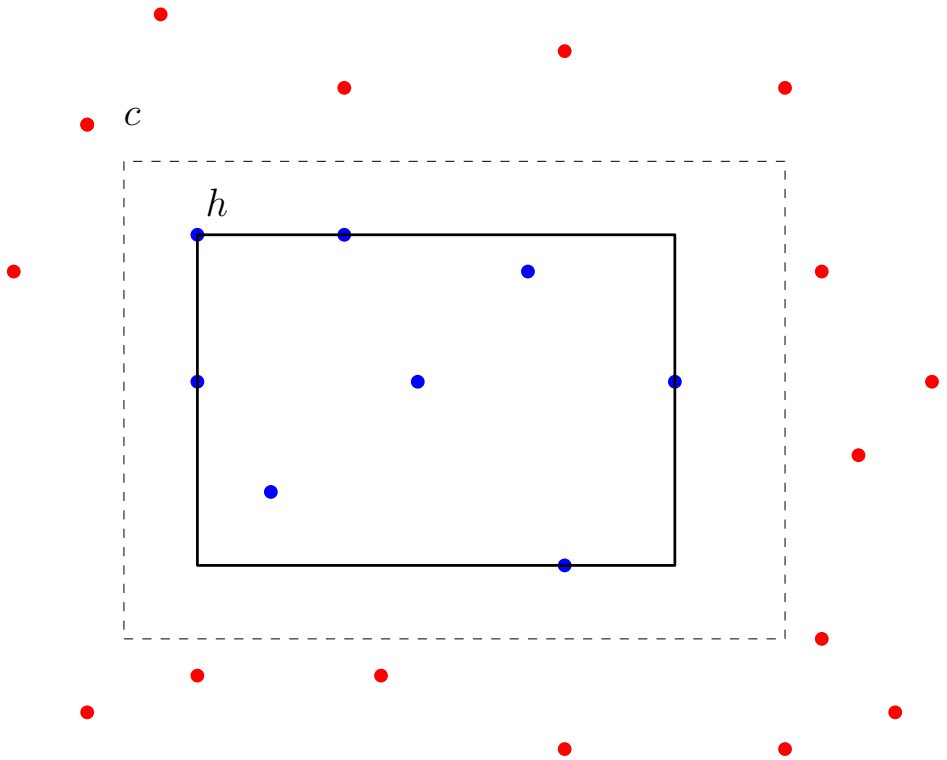
$$e(h) = \mathbf{P}_{\mathcal{D}}\left[c(\mathbf{x}) \neq h(\mathbf{x})\right]$$



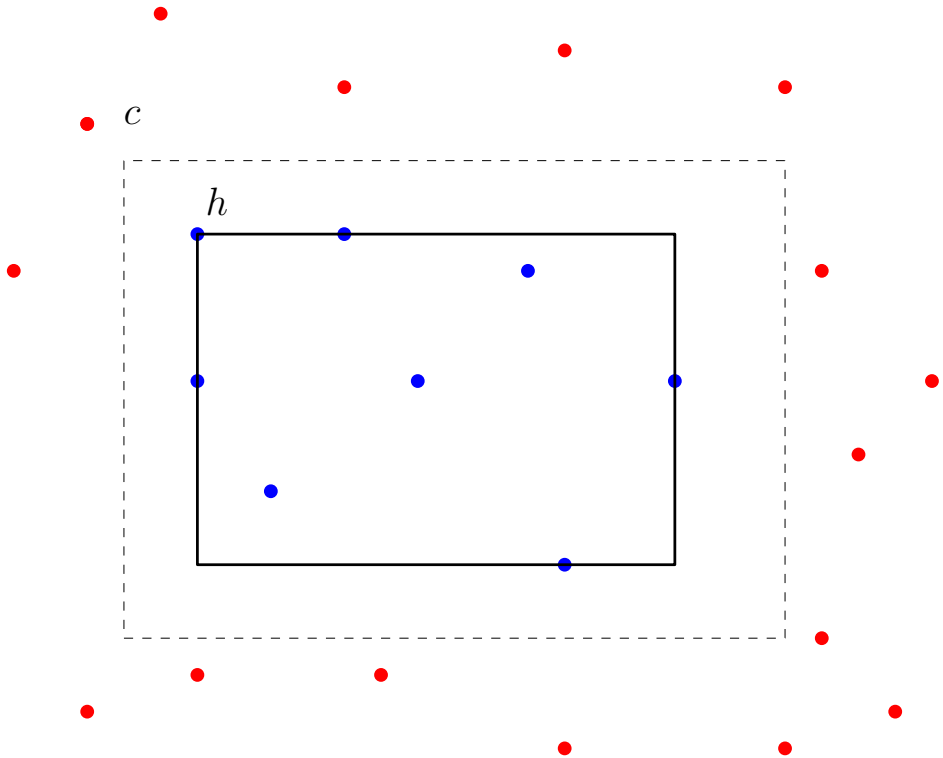
Algoritmo

- Encontrar a partir de los datos un rectángulo que minimice error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[c(\mathbf{x}) \neq h(\mathbf{x})\right]$$

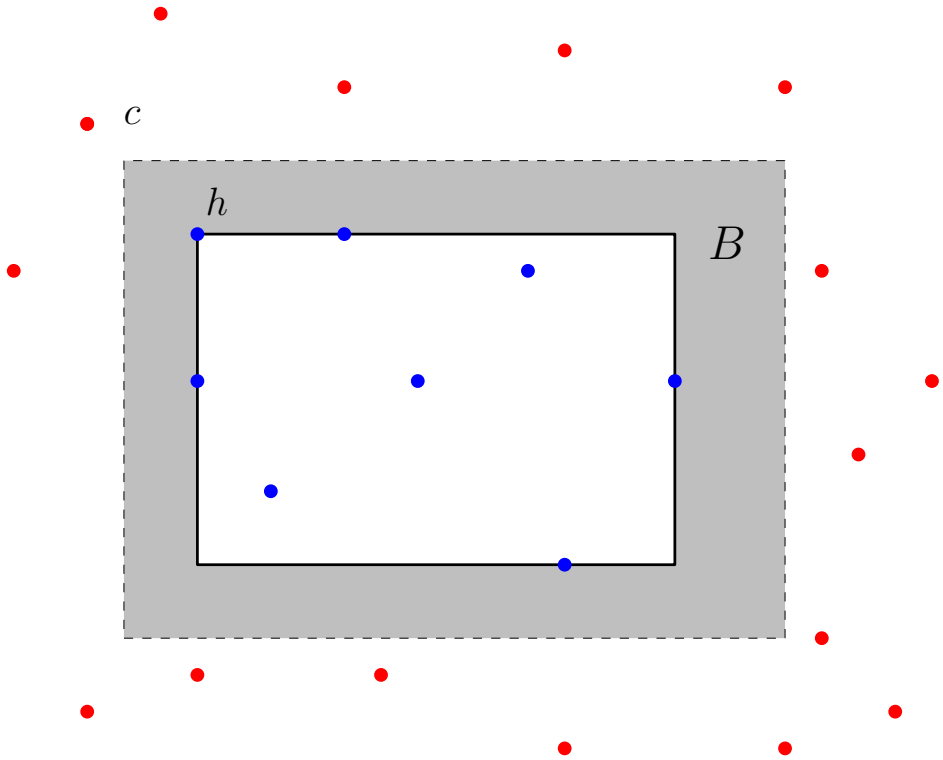


Análisis



Análisis

- Queremos $e(h) = \mathbf{P}_{\mathcal{D}}[B] \lll$

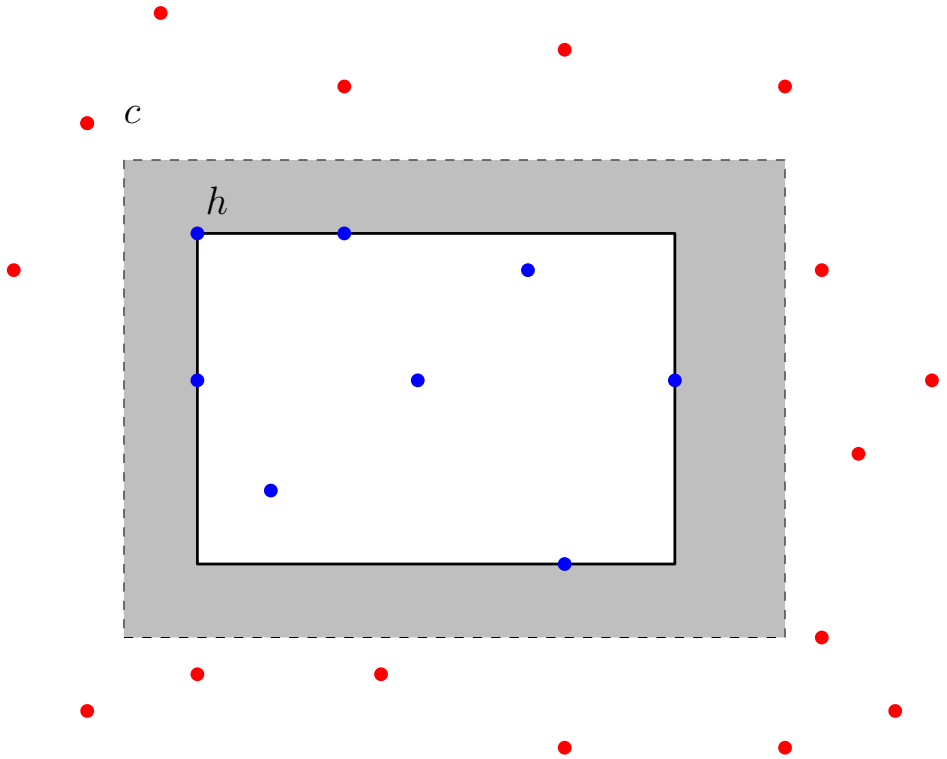


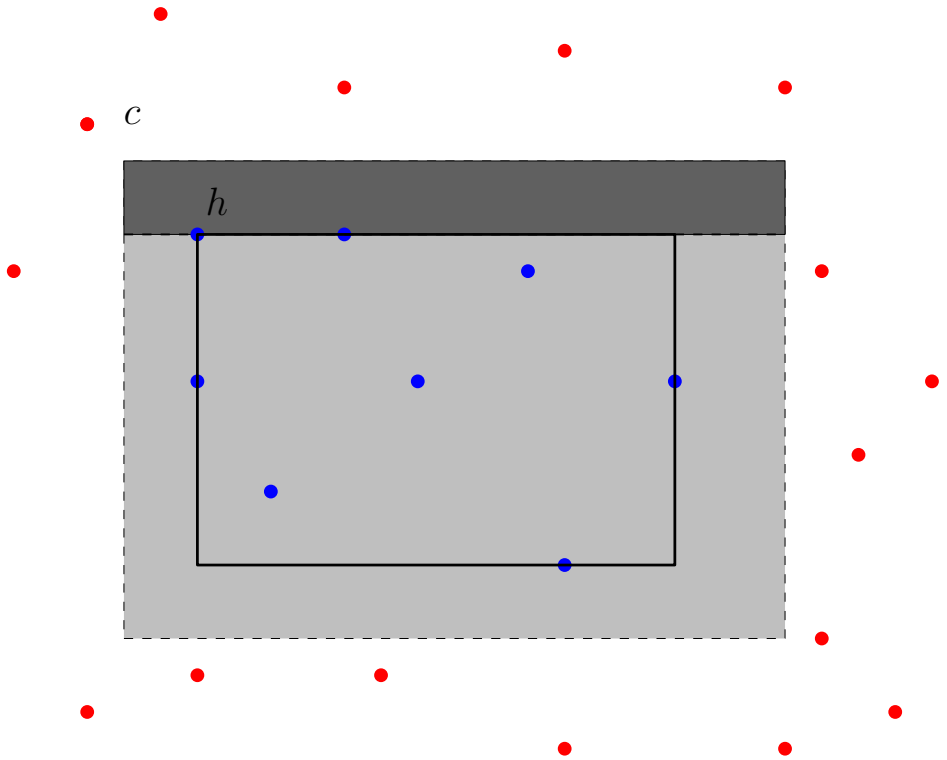
- $e(h)$ es una variable aleatoria porque h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.

- $e(h)$ es una variable aleatoria porque h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- Queremos que con alta probabilidad $\mathbf{P}_{\mathcal{D}}[B] \lll$

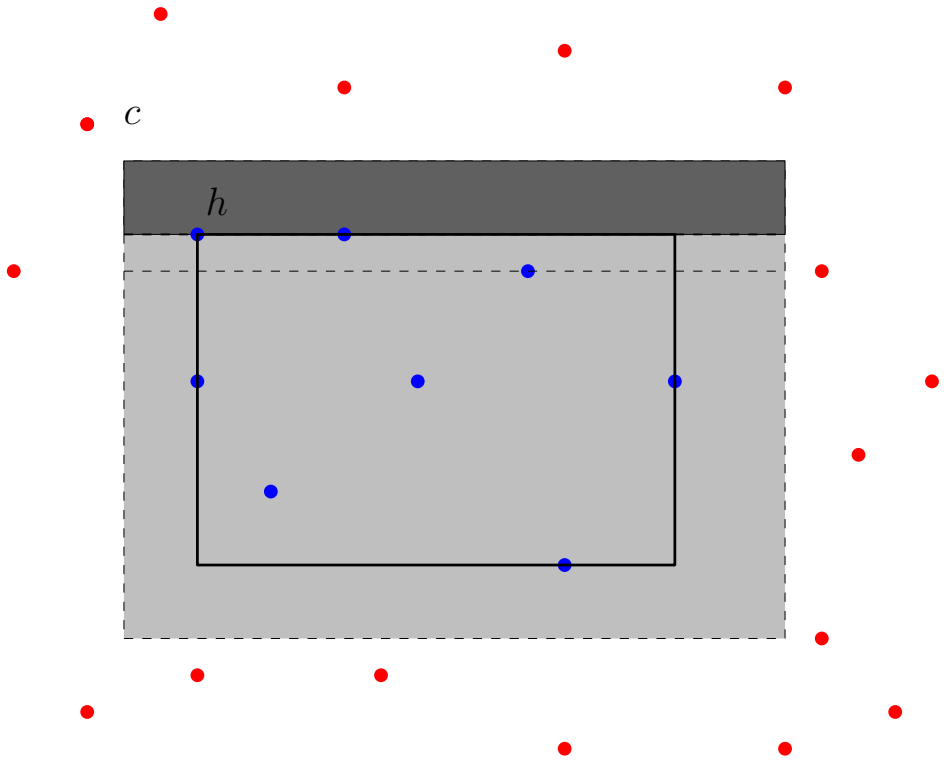
- $e(h)$ es una variable aleatoria porque h depende de los datos $\mathbf{x}_i \sim \mathcal{D}$.
- Queremos que con alta probabilidad $\mathbf{P}_{\mathcal{D}}[B] \lll$
- Dados $\varepsilon, \delta > 0$,

$$\mathbf{P}_{\mathcal{D}}[e(h) \geq \varepsilon] \leq \delta?$$

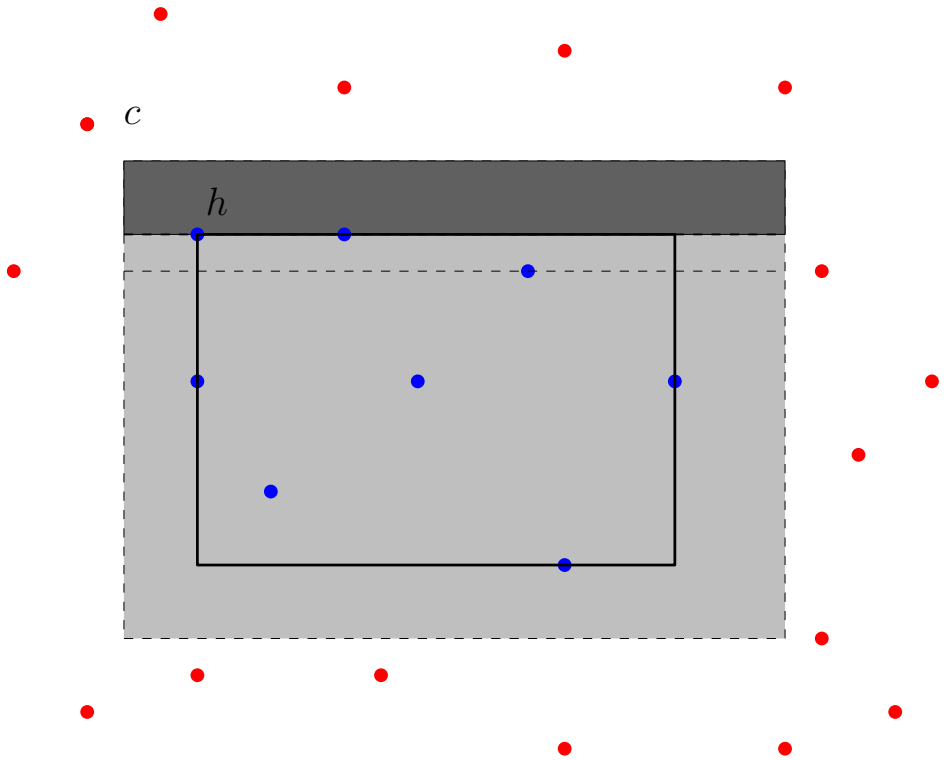

c
h



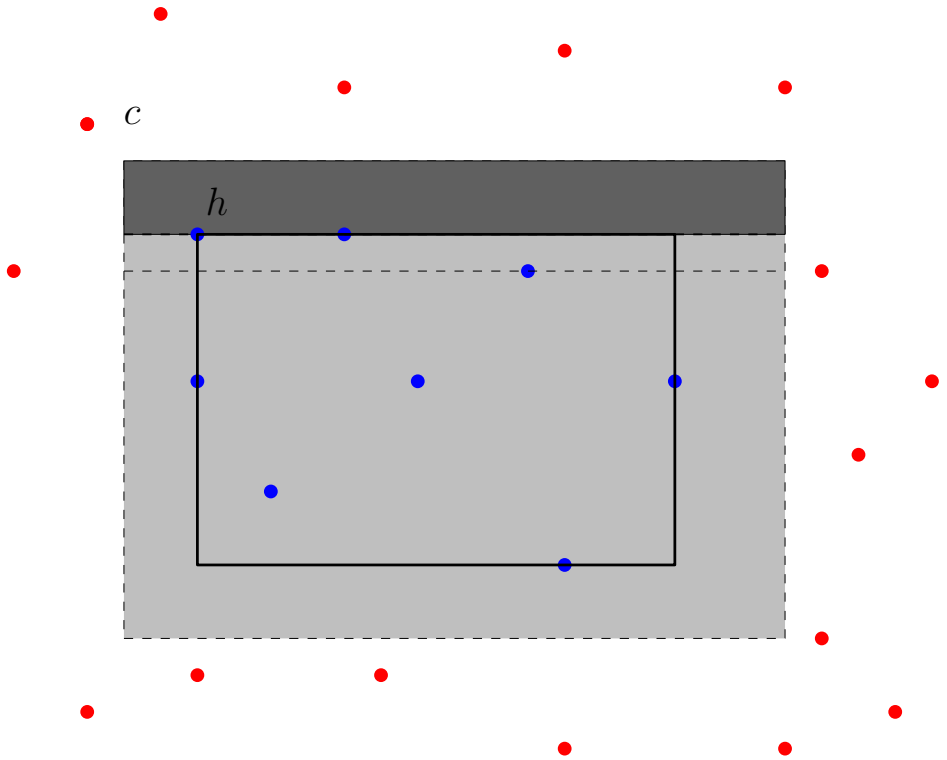
- Franja T'



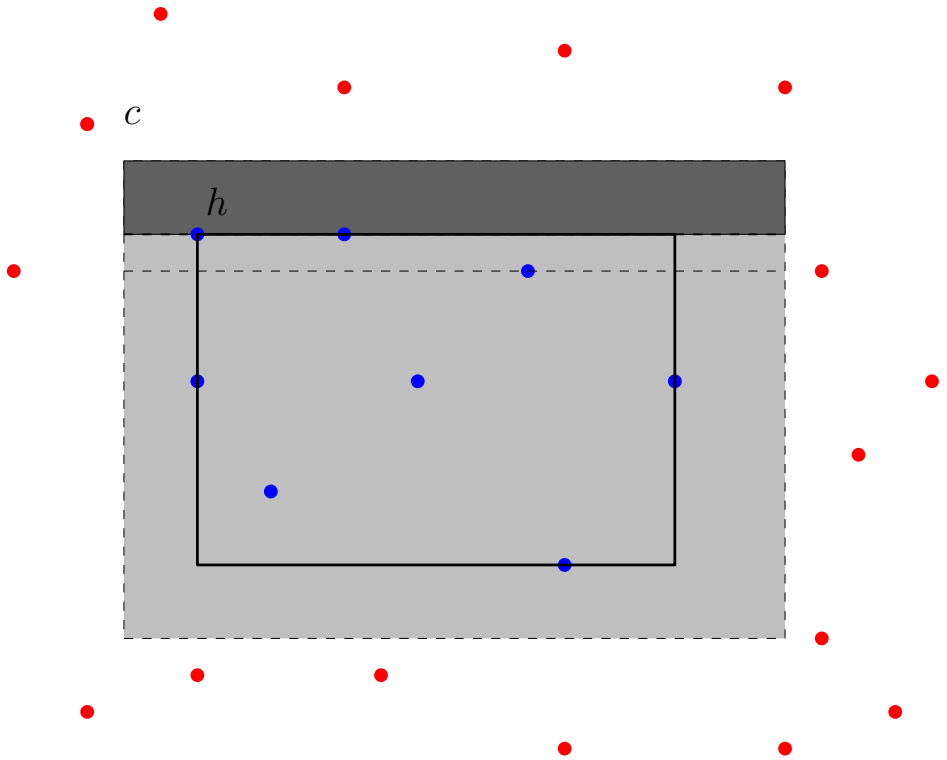
- Franja T'
- Franja T con $\mathbf{P}_{\mathcal{D}}[T] = \frac{\varepsilon}{4}$



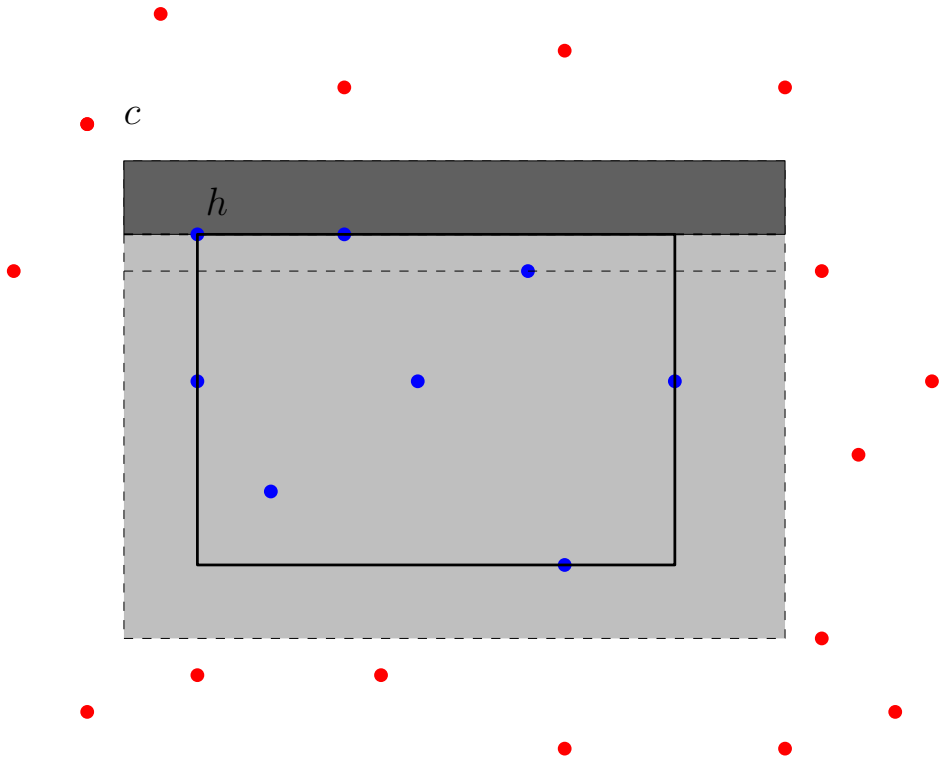
- Franja T'
- Franja T con $\mathbf{P}_{\mathcal{D}}[T] = \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}[T'] > \frac{\varepsilon}{4} \Leftrightarrow$ no hay puntos en T.



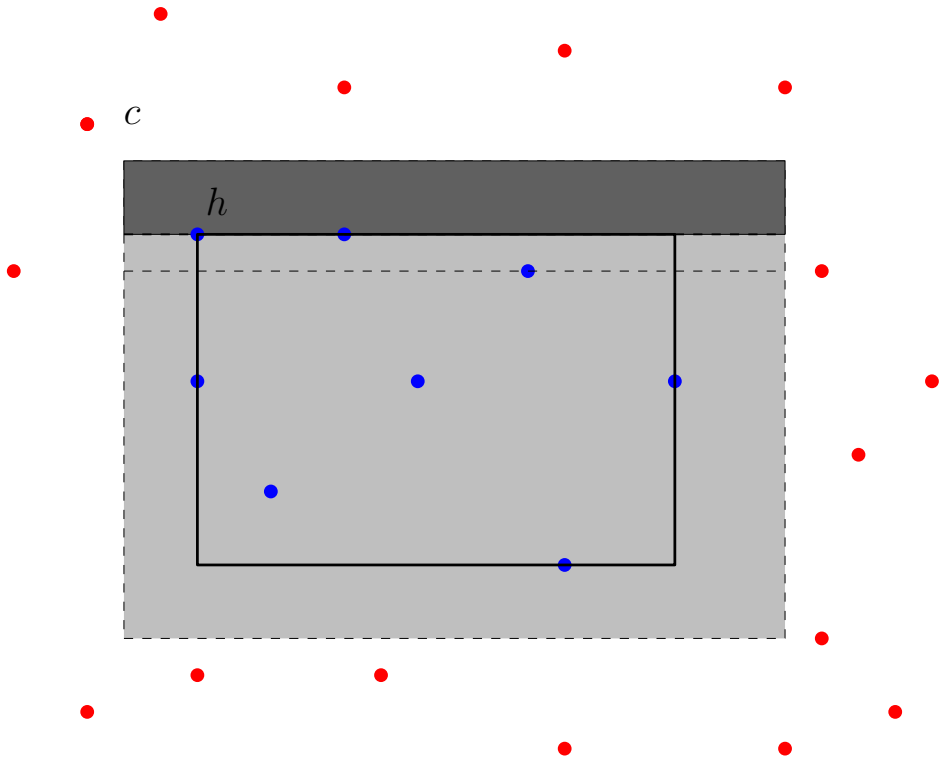
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$



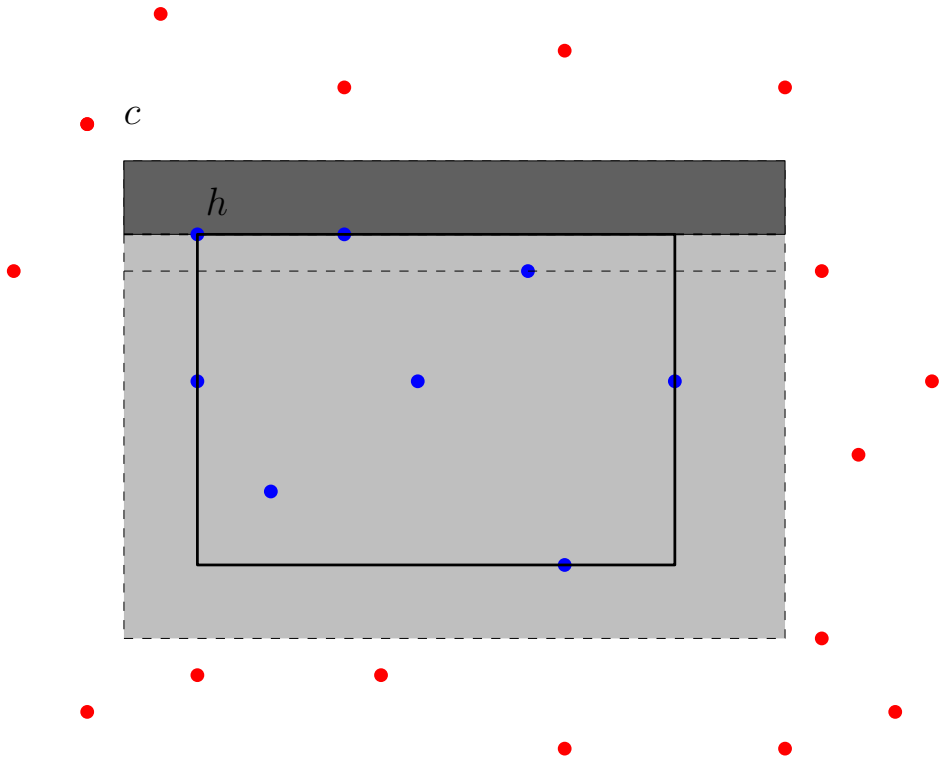
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] =$



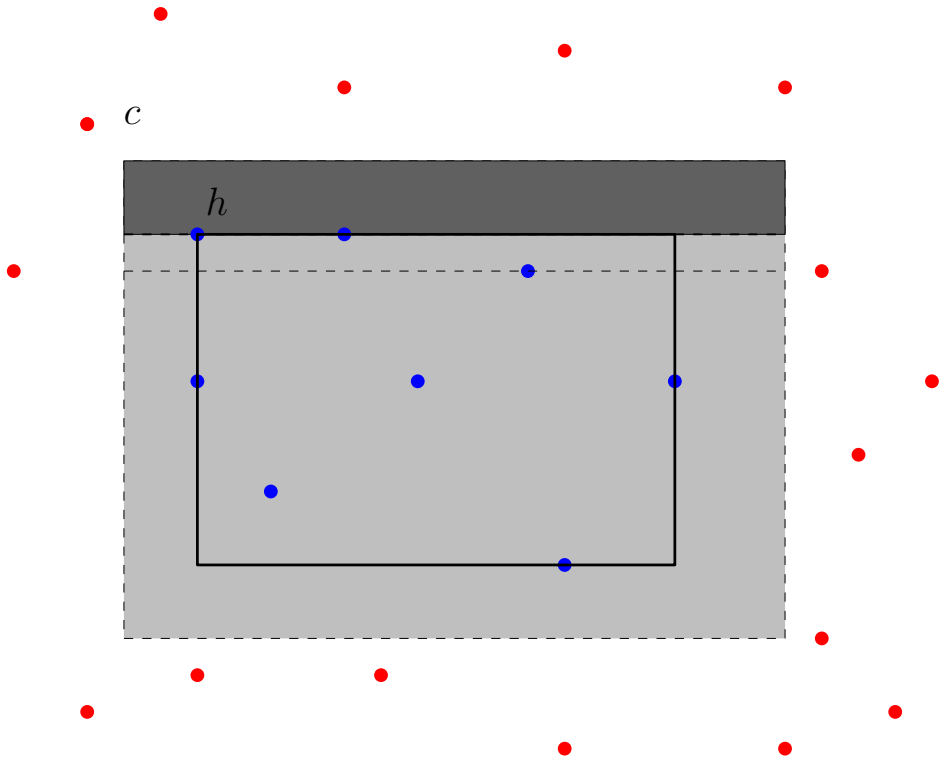
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$



- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin B\right]$



- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin B\right] \leq 4\left(1 - \frac{\varepsilon}{4}\right)^m$



- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x} \notin T\right] = 1 - \frac{\varepsilon}{4}$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin T\right] = \left(1 - \frac{\varepsilon}{4}\right)^m$
- $\mathbf{P}_{\mathcal{D}}\left[\mathbf{x}_1, \ldots, \mathbf{x}_m \notin B\right] \leq 4\left(1 - \frac{\varepsilon}{4}\right)^m$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

$$4e^{-\frac{\varepsilon m}{4}} \leq \delta$$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

$$4e^{-\frac{\varepsilon m}{4}} \leq \delta$$

o

$$m \geq \frac{4}{\varepsilon}\ln\frac{4}{\delta}$$

- Queremos escoger m que satisfaga

$$4\left(1-\frac{\varepsilon}{4}\right)^m \leq \delta$$

usando $1-x \leq e^{-x}$ tenemos

$$4e^{-\frac{\varepsilon m}{4}} \leq \delta$$

o

$$m \geq \frac{4}{\varepsilon}\ln\frac{4}{\delta}$$

- El algoritmo consistente con por lo menos $\frac{4}{\varepsilon}\ln\frac{4}{\delta}$ datos produce con probabilidad por lo menos $1-\delta$ una hipótesis que clasifica mal un nuevo dato con probabilidad máxima de ε.

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$

Clases de hipótesis finitas

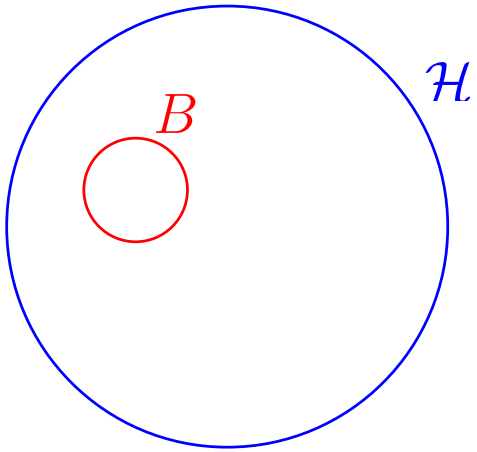
- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.

Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \ : \ e(h) > \varepsilon\}$,

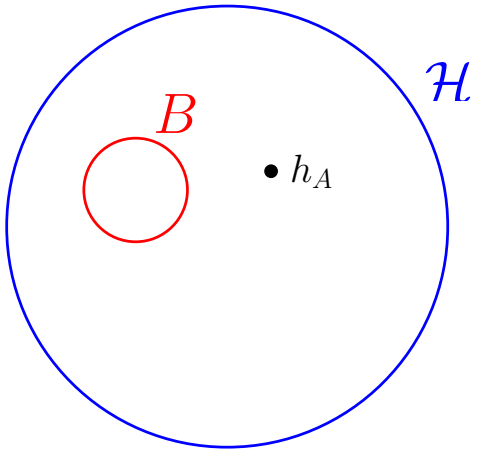
Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \ : \ e(h) > \varepsilon\}$,



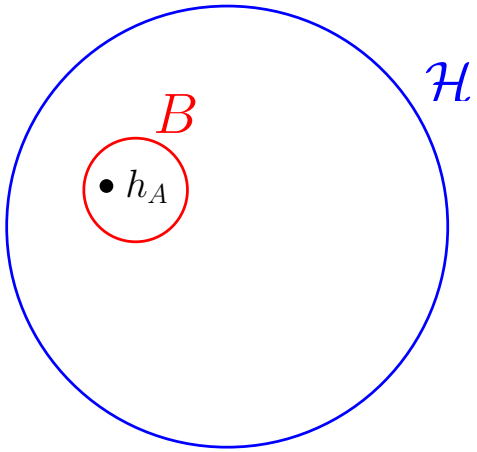
Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \ : \ e(h) > \varepsilon\}$,



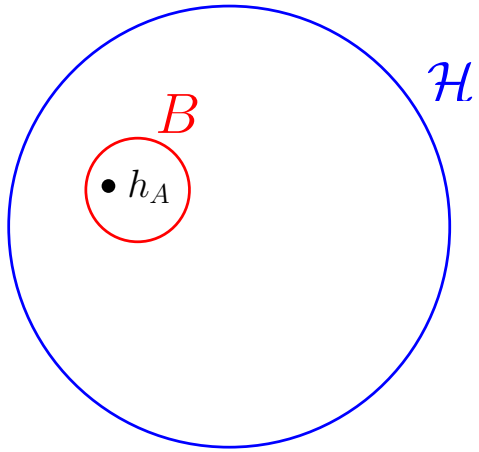
Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \ : \ e(h) > \varepsilon\}$,



Clases de hipótesis finitas

- $|\mathcal{H}| < \infty$
- Algoritmo A: observa m datos y retorna h_A consistente.
- Sea $B = \{h \in \mathcal{H} \ : \ e(h) > \varepsilon\}$,



$$\mathbf{P}_{\mathcal{D}}\left[h_A \in B\right] \leq \mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right]$$

- Para $h \in B$ fija:

$$\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right]$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right]\end{aligned}$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] \leq |B|\,(1-\varepsilon)^m$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m\end{aligned}$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

$$|\mathcal{H}|\,e^{-\varepsilon m} \leq \delta$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

$$|\mathcal{H}|\,e^{-\varepsilon m} \leq \delta \Rightarrow m \geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

- Para $h \in B$ fija:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[h \text{ es consistente}\right] &= \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_1) = c(\mathbf{x}_1) \wedge \cdots \wedge h(\mathbf{x}_m) = c(\mathbf{x}_m)\right] \\ &= \prod_{i=1}^{m} \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}_i) = c(\mathbf{x}_i)\right] \leq (1-\varepsilon)^m\end{aligned}$$

- Sumando sobre todas las posiblidades:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}}\left[\exists h \in B \ : \ h \text{ es consistente con los datos}\right] &\leq |B|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,(1-\varepsilon)^m \\ &\leq |\mathcal{H}|\,e^{-\varepsilon m}\end{aligned}$$

- Para ε, δ dados, podemos calcular:

$$|\mathcal{H}|\,e^{-\varepsilon m} \leq \delta \Rightarrow m \geq \frac{1}{\varepsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

- O para m, δ dados, podemos decir que con probabilidad por lo menos $1-\delta$

$$e(h) \leq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

Clases de clasificadores infinitas

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$
- Medidas de complejidad:

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$
- Medidas de complejidad:
  - Dimensión VC (Vapnik-Chervonenkis).

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$
- Medidas de complejidad:
  - Dimensión VC (Vapnik-Chervonenkis).
  - Números entrópicos.

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$
- Medidas de complejidad:
  - Dimensión VC (Vapnik-Chervonenkis).
  - Números entrópicos.
  - Complejidad de Rademacher (global/local).

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$
- Medidas de complejidad:
  - Dimensión VC (Vapnik-Chervonenkis).
  - Números entrópicos.
  - Complejidad de Rademacher (global/local).
- Complejidad crece con el número de parámetros a ajustar.

Clases de clasificadores infinitas

- En general $|\mathcal{H}| = \infty$
- Medidas de complejidad:
  - Dimensión VC (Vapnik-Chervonenkis).
  - Números entrópicos.
  - Complejidad de Rademacher (global/local).
- Complejidad crece con el número de parámetros a ajustar.
- Machine Learning estadístico/computacional.

Evaluación del clasificador

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

Evaluación del clasificador

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

Evaluación del clasificador

- $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^m$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$.
- Hallamos h a partir de datos de entrenamiento.

Evaluación del clasificador

- $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^{m}$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$.
- Hallamos h a partir de datos de entrenamiento.
- Datos de prueba: $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$ e independientes de los datos de entrenamiento

Evaluación del clasificador

- $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^{m}$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$.
- Hallamos h a partir de datos de entrenamiento.
- Datos de prueba: $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$ e independientes de los datos de entrenamiento
- Criterio de error:

Evaluación del clasificador

- $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^{m}$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$.
- Hallamos h a partir de datos de entrenamiento.
- Datos de prueba: $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$ e independientes de los datos de entrenamiento
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}) \neq y\right]$$

Evaluación del clasificador

- $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^m$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$.
- Hallamos h a partir de datos de entrenamiento.
- Datos de prueba: $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$ e independientes de los datos de entrenamiento
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}) \neq y\right]$$

- Típicamente calculamos el error empírico de h en los datos de prueba:

Evaluación del clasificador

- $(\mathbf{x}, y) \sim \mathcal{D}$
- Datos de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^m$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$.
- Hallamos h a partir de datos de entrenamiento.
- Datos de prueba: $\{\mathbf{x}_i, y_i\}_{i=1}^n$ copias independientes e idénticamente distribuidas de $(\mathbf{x}, y)$ e independientes de los datos de entrenamiento
- Criterio de error:

$$e(h) = \mathbf{P}_{\mathcal{D}}\left[h(\mathbf{x}) \neq y\right]$$

- Típicamente calculamos el error empírico de h en los datos de prueba:

$$\hat{e}(h) = \frac{1}{n}\sum_{i=1}^{n} I_{\{h(\mathbf{x}_i) \neq y_i\}}$$

Cotas de Chernoff (forma aditiva)

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}[X_j = 1] = p$

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}\left[X_j = 1\right] = p$
- $\frac{(X_j - p)}{n} \in$

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}[X_j = 1] = p$
- $\frac{(X_j - p)}{n} \in \left\{ -\frac{p}{n}, \frac{1-p}{n} \right\}$

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}[X_j = 1] = p$
- $\frac{(X_j - p)}{n} \in \left\{-\frac{p}{n}, \frac{1-p}{n}\right\} \Rightarrow b_j - a_j =$

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}[X_j = 1] = p$
- $\frac{(X_j - p)}{n} \in \left\{-\frac{p}{n}, \frac{1-p}{n}\right\} \Rightarrow b_j - a_j = \frac{1}{n}$

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}[X_j = 1] = p$
- $\frac{(X_j - p)}{n} \in \left\{-\frac{p}{n}, \frac{1-p}{n}\right\} \Rightarrow b_j - a_j = \frac{1}{n}$
- $$\mathbf{P}\left[\frac{1}{n}\sum_{j=1}^{n} X_j - p \geq \varepsilon\right] \leq e^{-2\epsilon^2 n}$$

  y

  $$\mathbf{P}\left[\frac{1}{n}\sum_{j=1}^{n} X_j - p \leq -\varepsilon\right] \leq e^{-2\epsilon^2 n}$$

Cotas de Chernoff (forma aditiva)

- $X_j \in \{0,1\}$, $\mathbf{P}[X_j = 1] = p$
- $\frac{(X_j - p)}{n} \in \left\{-\frac{p}{n}, \frac{1-p}{n}\right\} \Rightarrow b_j - a_j = \frac{1}{n}$
- $$\mathbf{P}\left[\frac{1}{n}\sum_{j=1}^{n} X_j - p \geq \varepsilon\right] \leq e^{-2\epsilon^2 n}$$

  y

  $$\mathbf{P}\left[\frac{1}{n}\sum_{j=1}^{n} X_j - p \leq -\varepsilon\right] \leq e^{-2\epsilon^2 n}$$

Ejemplo

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.
- Estimativo $\hat{p} =$ número de caras en n lanzadas.

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.
- Estimativo $\hat{p} =$ número de caras en n lanzadas.
- Cuántas veces tenemos que lanzar la moneda para garantizar con confianza $1 - \delta$ que el estimativo $\hat{p}$ no difiera en más de ϵ de p?

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.
- Estimativo $\hat{p} =$ número de caras en n lanzadas.
- Cuántas veces tenemos que lanzar la moneda para garantizar con confianza $1 - \delta$ que el estimativo $\hat{p}$ no difiera en más de ϵ de p?
- Usando cotas de Chernoff:

$$\mathbf{P}\left[|p - \hat{p}| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.
- Estimativo $\hat{p} =$ número de caras en n lanzadas.
- Cuántas veces tenemos que lanzar la moneda para garantizar con confianza $1 - \delta$ que el estimativo $\hat{p}$ no difiera en más de ϵ de p?
- Usando cotas de Chernoff:

$$\mathbf{P}\left[|p - \hat{p}| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

Queremos $2e^{-2\epsilon^2 n} = \delta$

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.
- Estimativo $\hat{p} =$ número de caras en n lanzadas.
- Cuántas veces tenemos que lanzar la moneda para garantizar con confianza $1 - \delta$ que el estimativo $\hat{p}$ no difiera en más de ϵ de p?
- Usando cotas de Chernoff:

$$\mathbf{P}\left[|p - \hat{p}| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

Queremos $2e^{-2\epsilon^2 n} = \delta$ o despejando $n = \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$.

Ejemplo

- Moneda, estimar probabilidad p de que salga cara.
- Estimativo $\hat{p} =$ número de caras en n lanzadas.
- Cuántas veces tenemos que lanzar la moneda para garantizar con confianza $1 - \delta$ que el estimativo $\hat{p}$ no difiera en más de ϵ de p?
- Usando cotas de Chernoff:

$$\mathbf{P}\left[|p - \hat{p}| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

  Queremos $2e^{-2\epsilon^2 n} = \delta$ o despejando $n = \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$.
- Por ejemplo para confianza del 95 % y precisión 0,05 debemos lanzar la moneda ~ 738 veces.

de vuelta a clasificación...

de vuelta a clasificación...

- Para un dato de prueba $(\mathbf{x}_i, y_i)$, h comete un error con probabilidad $e(h)$.

de vuelta a clasificación...

- Para un dato de prueba $(\mathbf{x}_i, y_i)$, h comete un error con probabilidad $e(h)$.
- $\hat{e}(h)$ es estimativo de $e(h)$.

de vuelta a clasificación...

- Para un dato de prueba $(\mathbf{x}_i, y_i)$, h comete un error con probabilidad $e(h)$.
- $\hat{e}(h)$ es estimativo de $e(h)$.
- Es decir. $\forall \epsilon > 0$,

$$\mathbf{P}\left[|e(h) - \hat{e}(h)| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

de vuelta a clasificación...

- Para un dato de prueba $(\mathbf{x}_i, y_i)$, h comete un error con probabilidad $e(h)$.
- $\hat{e}(h)$ es estimativo de $e(h)$.
- Es decir. $\forall \epsilon > 0$,

$$\mathbf{P}\left[|e(h) - \hat{e}(h)| \geq \epsilon\right] \leq 2e^{-2\epsilon^2 n}$$

- Luego, con $n \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$ datos de prueba, garantizamos con probabilidad por lo menos $1 - \delta$ que $|e(h) - \hat{e}(h)| \leq \epsilon$

"Es mejor tener muchos datos que pocos datos"- Pambelé

"Es mejor tener muchos datos que pocos datos"- Pambelé

- Más datos $\Rightarrow$ mejor aprendizaje.

"Es mejor tener muchos datos que pocos datos"- Pambelé

- Más datos $\Rightarrow$ mejor aprendizaje.
- Modelos más complejos requieren más datos.

"Es mejor tener muchos datos que pocos datos"- Pambelé

- Más datos $\Rightarrow$ mejor aprendizaje.
- Modelos más complejos requieren más datos.
- Datos suficientes para validación/evalución.

Preprocesamiento

Preprocesamiento

- Es necesario obtener una representación apropiada de los datos para buen aprendizaje..

Preprocesamiento

- Es necesario obtener una representación apropiada de los datos para buen aprendizaje..
- Selección de características útiles.

Preprocesamiento

- Es necesario obtener una representación apropiada de los datos para buen aprendizaje..
- Selección de características útiles.
- Normalización, reducción de dimensionalidad.

Preprocesamiento

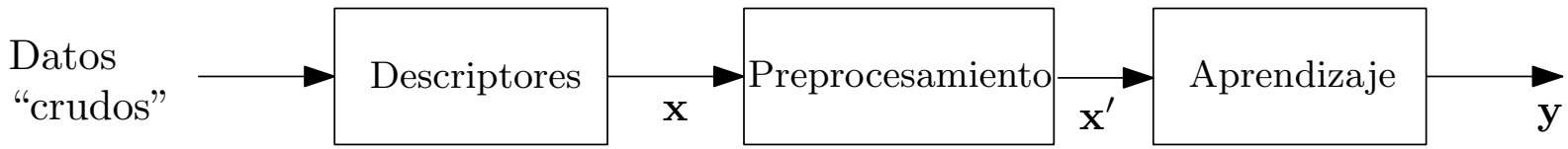
- Es necesario obtener una representación apropiada de los datos para buen aprendizaje..
- Selección de características útiles.
- Normalización, reducción de dimensionalidad.
- Datos faltantes

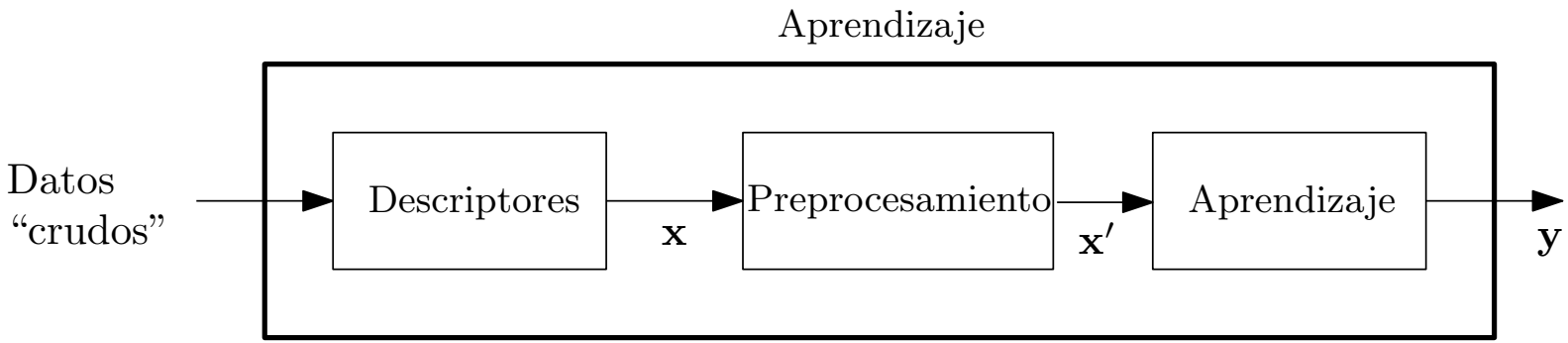
Preprocesamiento

- Es necesario obtener una representación apropiada de los datos para buen aprendizaje..
- Selección de características útiles.
- Normalización, reducción de dimensionalidad.
- Datos faltantes
- Datos ruidosos

Preprocesamiento

- Es necesario obtener una representación apropiada de los datos para buen aprendizaje..
- Selección de características útiles.
- Normalización, reducción de dimensionalidad.
- Datos faltantes
- Datos ruidosos


Datos
“crudos”
Descriptores
x
Preprocesamiento
x′
Aprendizaje
y


Aprendizaje
Datos
“crudos”
Descriptores
x
Preprocesamiento
x′
Aprendizaje
y

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:

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

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:

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

- 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:

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

- Probabilidades a priori de cada clase:

$$\mathbf{P}\left[y = 1\right] = \alpha, \qquad \mathbf{P}\left[y = 0\right] = 1 - \alpha$$

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:

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

- Probabilidades a priori de cada clase:

$$\mathbf{P}\left[y = 1\right] = \alpha, \qquad \mathbf{P}\left[y = 0\right] = 1 - \alpha$$

- 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:

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

- Probabilidades a priori de cada clase:

$$\mathbf{P}\left[y = 1\right] = \alpha, \qquad \mathbf{P}\left[y = 0\right] = 1 - \alpha$$

- Probabilidades marginales

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

$$L(C) \quad = \quad \mathbf{P}\left[y = 1, \mathbf{x} \notin C\right] + \mathbf{P}\left[y = 0, \mathbf{x} \in C\right]$$

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

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

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

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

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

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:

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

Clasificador de Bayes

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

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

Clasificador de Bayes

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

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

Clasificador de Bayes

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

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

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

Clasificador de Bayes

- El clasificador óptimo esta 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 Normales:

Caso Especial

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

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

Caso Especial

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

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

- Tomando logaritmos:

Caso Especial

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

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

- Tomando logaritmos:

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

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

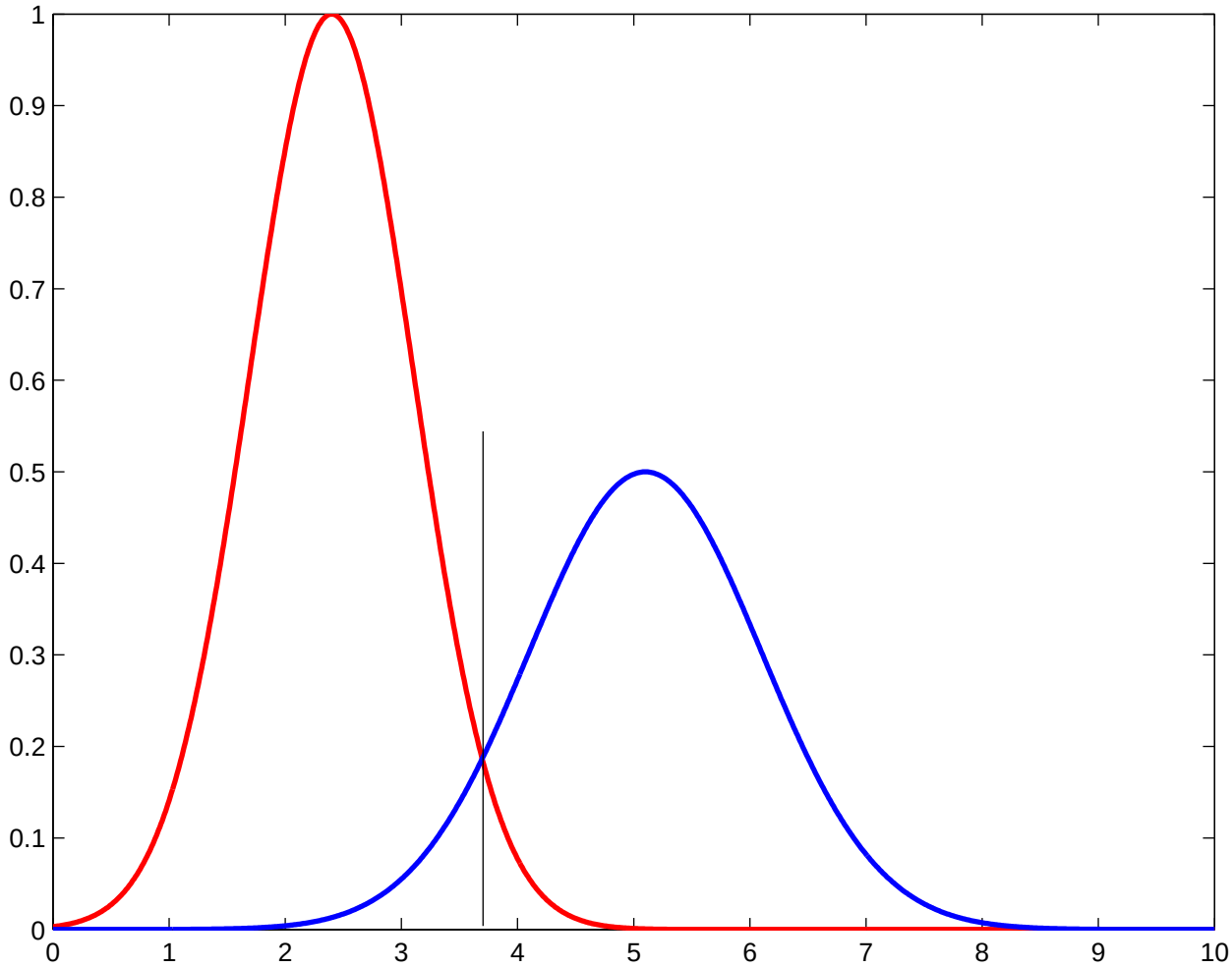
$$\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mathbf{m}_0^T\Sigma^{-1}\mathbf{x} + \mathbf{m}_0^T\Sigma^{-1}\mathbf{m}_0 - \mathbf{x}^T\Sigma^{-1}\mathbf{x} + 2\mathbf{m}_1^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_1^T\Sigma^{-1}\mathbf{m}_1 > 2\ln\left(\frac{1-\alpha}{\alpha}\right)$$

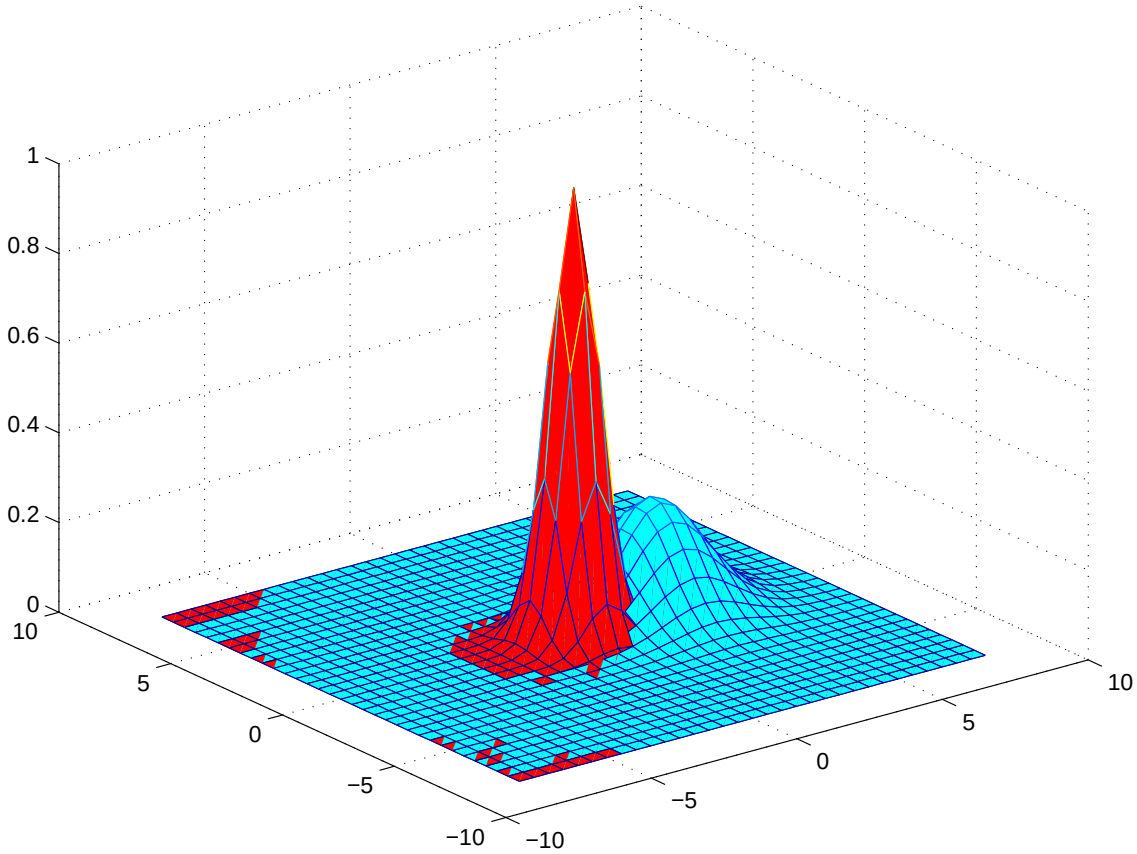
- Si además $\Sigma_0 = \Sigma_1 = \Sigma$:

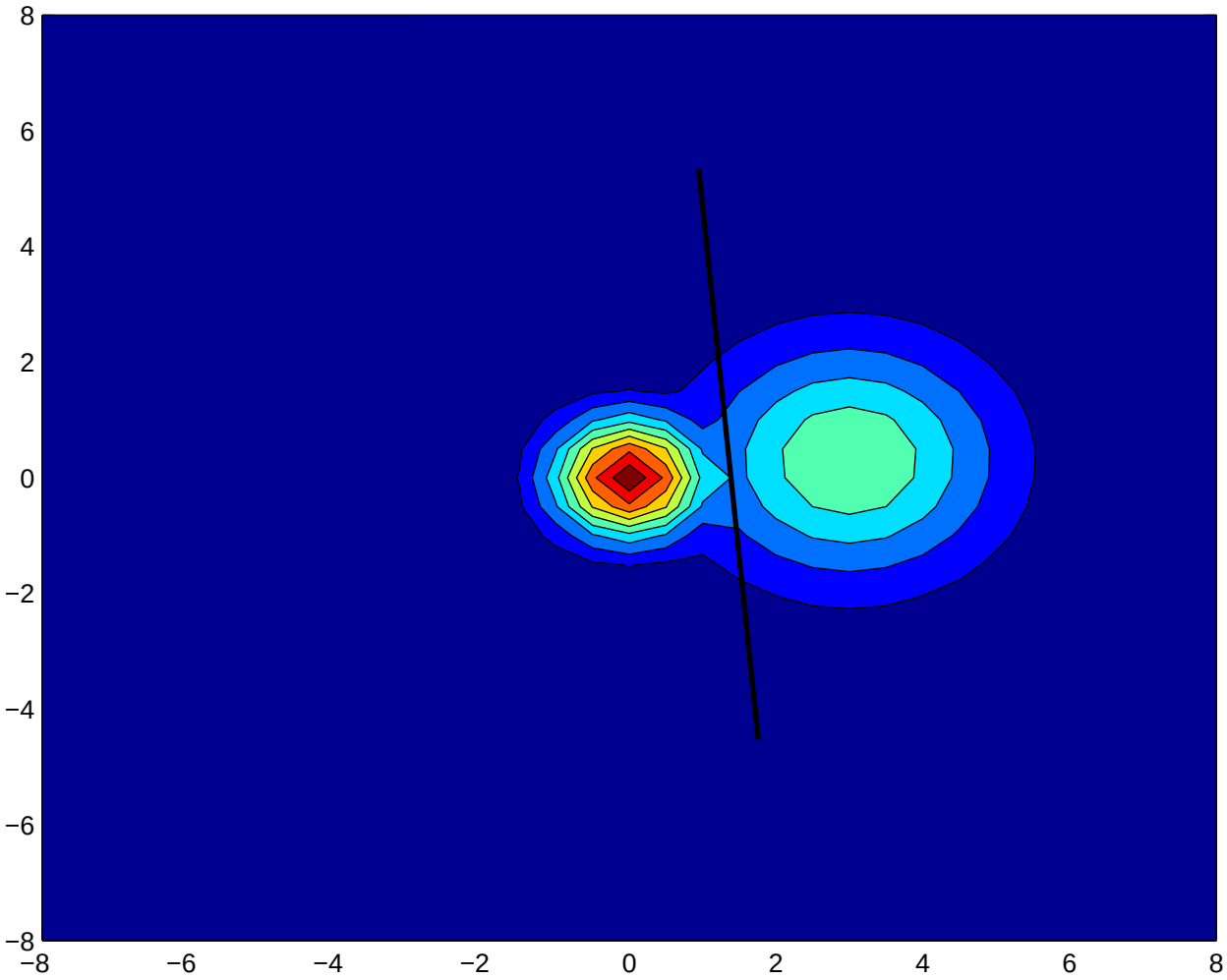
$$\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mathbf{m}_0^T\Sigma^{-1}\mathbf{x} + \mathbf{m}_0^T\Sigma^{-1}\mathbf{m}_0 - \mathbf{x}^T\Sigma^{-1}\mathbf{x} + 2\mathbf{m}_1^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_1^T\Sigma^{-1}\mathbf{m}_1 > 2\ln\left(\frac{1-\alpha}{\alpha}\right)$$

- entonces:

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


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10


1
0.8
0.6
0.4
0.2
0
10
5
0
−5
−10
−10
−5
0
5
10


8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8