Aprendizaje Supervisado

Fernando Lozano

Universidad de los Andes

23 de agosto de 2017

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.

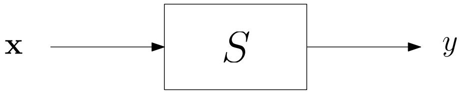
Aprendizaje Supervisado

- 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

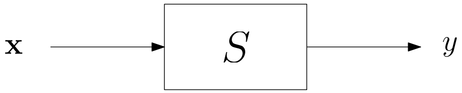
- Entrada $\mathbf{x}$, salida y.
- Queremos un sistema que prediga el valor de y a partir de $\mathbf{x}$.
- Existe un supervisor o maestro que conoce la respuesta correcta para patrones de entrada.
- Conjunto de entrenamiento: $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$

Visión Conceptual



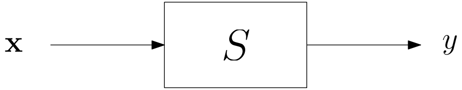
- Queremos modelar S.

Visión Conceptual



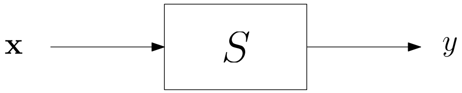
- Queremos modelar S.
- No es fácil obtener un modelo analítico.

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



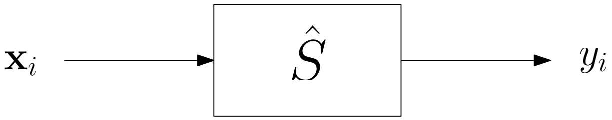
- Queremos modelar S.
- No es fácil obtener un modelo analítico.
- Usar modelo para predecir valores de la salida para nuevas entradas.

Elementos

$$\mathbf{x}_i \longrightarrow \boxed{\hat{S}} \longrightarrow y_i$$

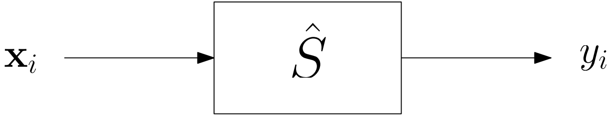
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.

Elementos



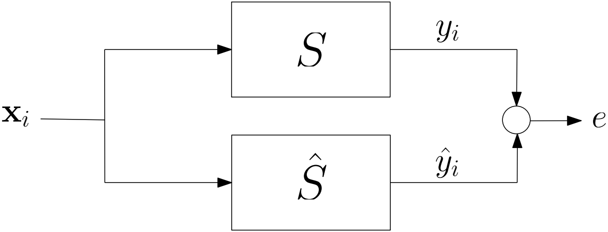
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
- Conjunto de modelos a utilizar.

Elementos



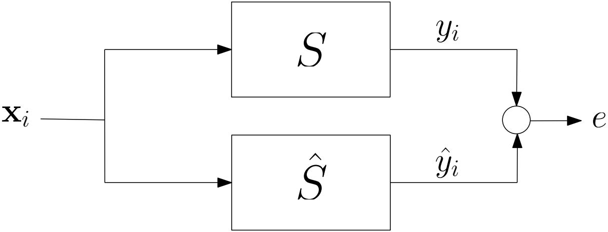
- Conjunto de datos de entrenamiento $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$.
- Conjunto de modelos a utilizar.
- Conjunto de datos de prueba $\{\mathbf{x}_i, y_i\}_{i=1}^{q}$.

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

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

- 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 $\mathbf{P}\left[\eta = 1\right] = p$.
  - $\mathbf{x} \sim \mathcal{D}$ y $\mathbf{P}\left[y = 1 \mid \mathbf{x}\right] = \alpha(\mathbf{x})$.
- 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 $y = c(\mathbf{x})$ 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 $y = c(\mathbf{x})$ para alguna función determinística f desconocida.
  - $\mathbf{x} \sim \mathcal{D}$ y $y = c(\mathbf{x}) + \eta$ donde $\eta \sim \mathcal{D}_\eta$

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

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

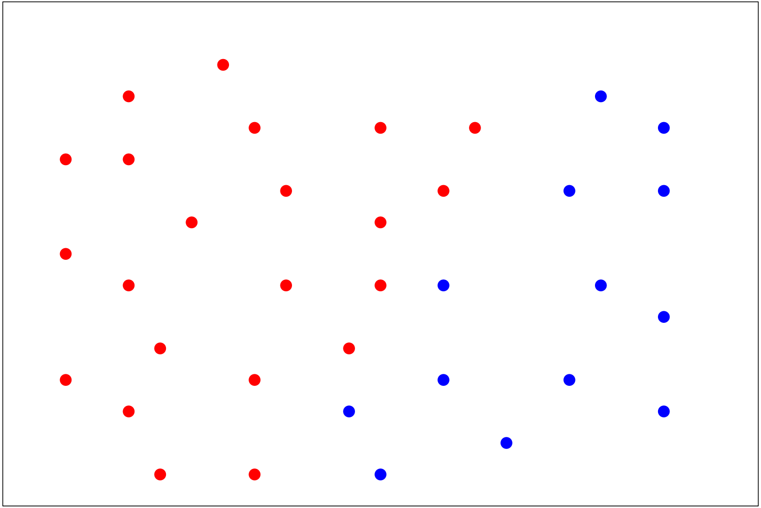
Existencia: Es posible solucionar en principio el problema usando cualquier modelo? (es decir, está el problema bien definido?).

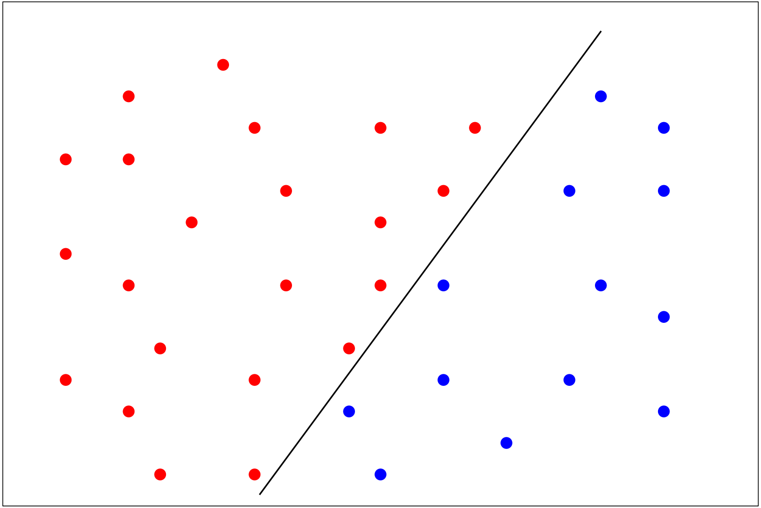
Capacidad de Representación: Es posible solucionar el problema usando una clase de modelos dada?

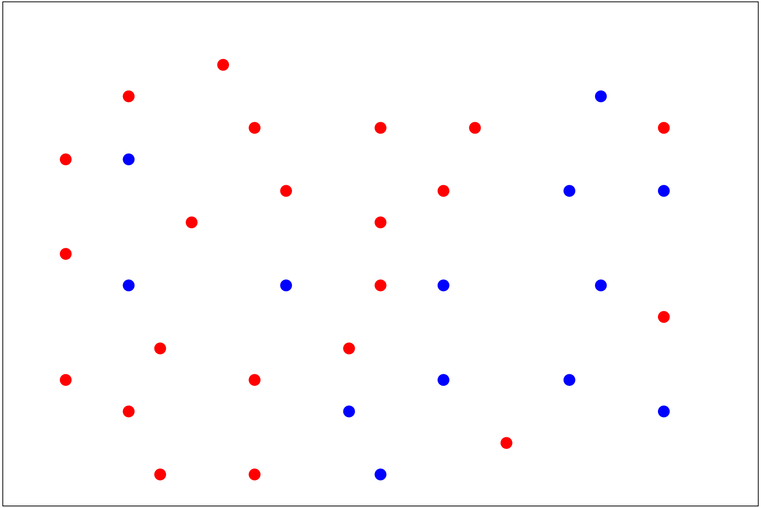
Estimación: Es posible determinar el modelo a partir de un conjunto de datos?

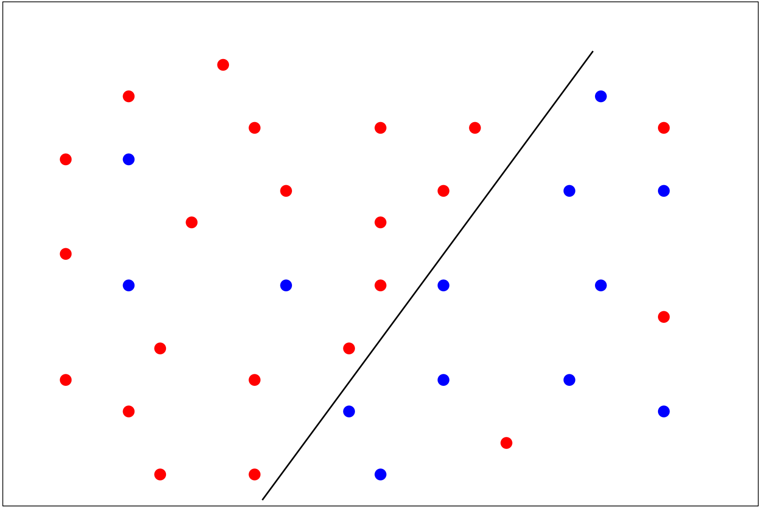
Computación: Es posible determinar el modelo eficientemente?

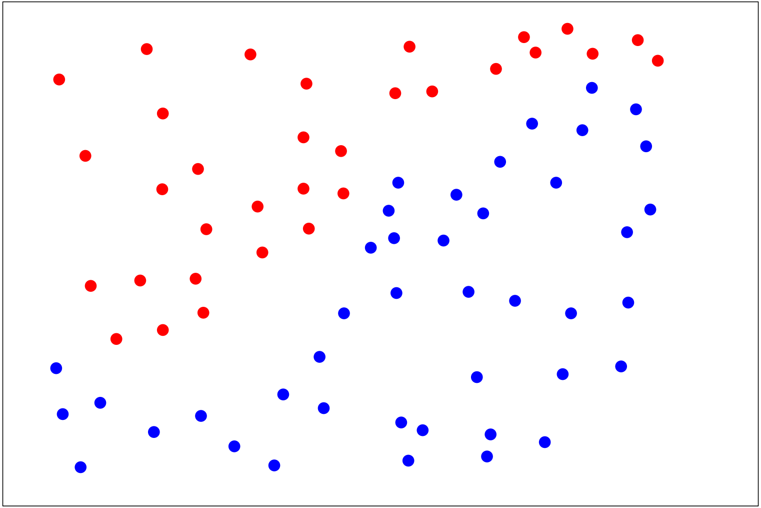
Implementación: Es posible diseñar e implementar el modelo usando precisión finita?

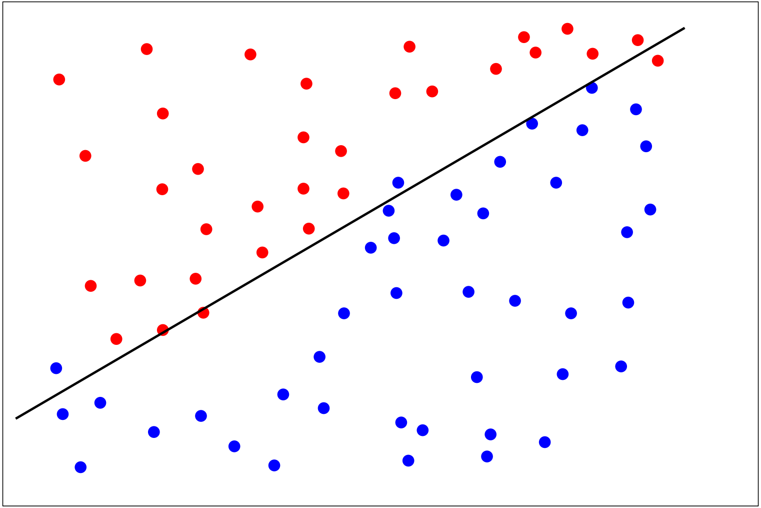


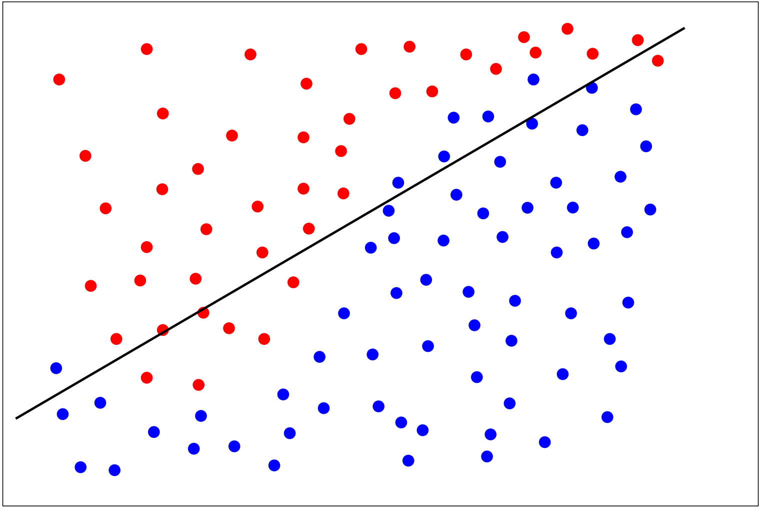


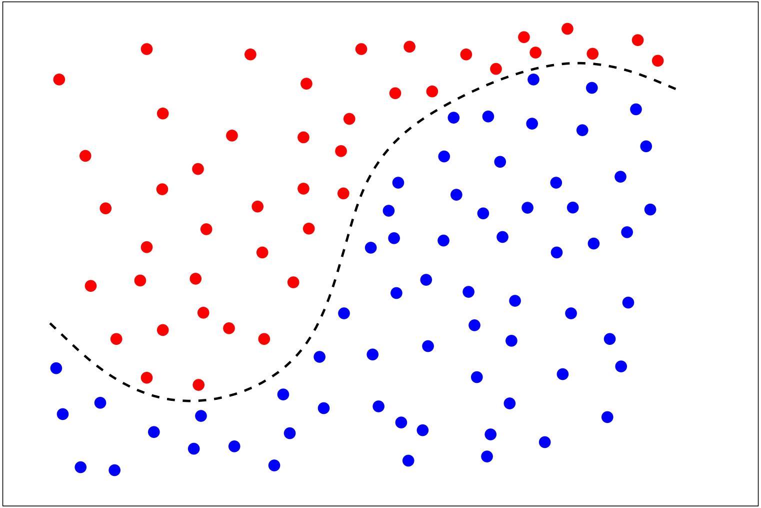


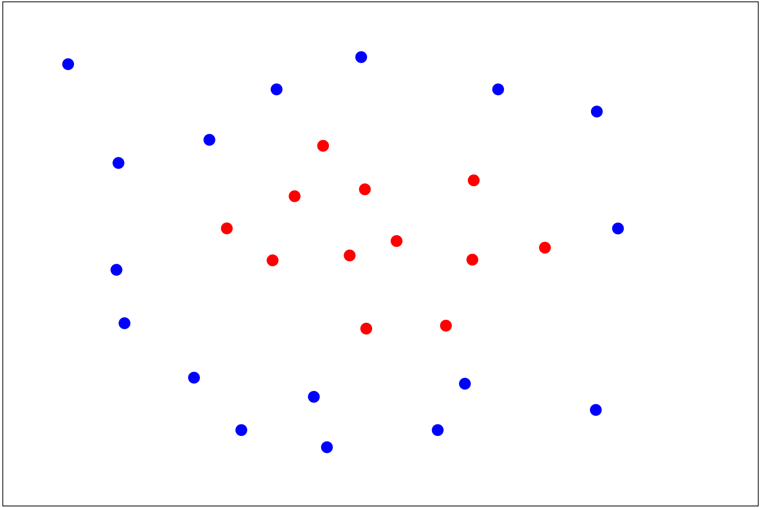


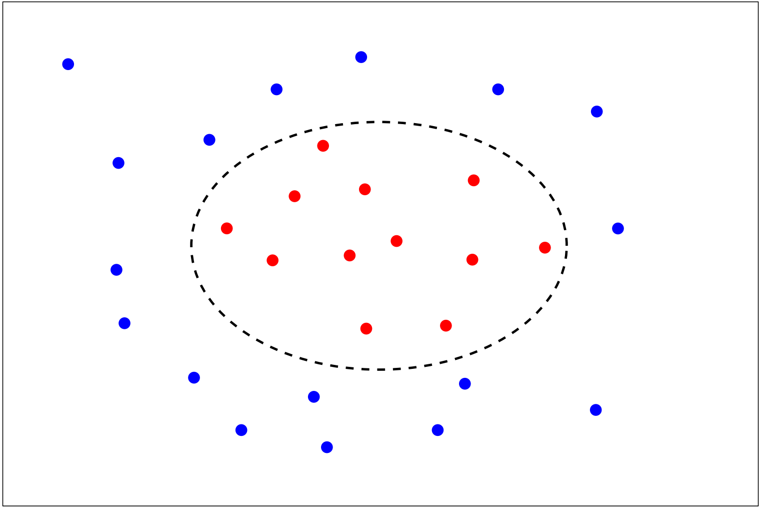


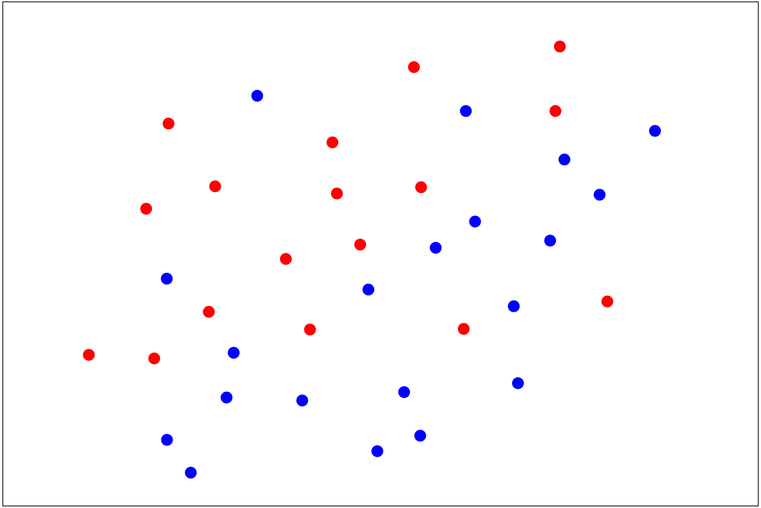


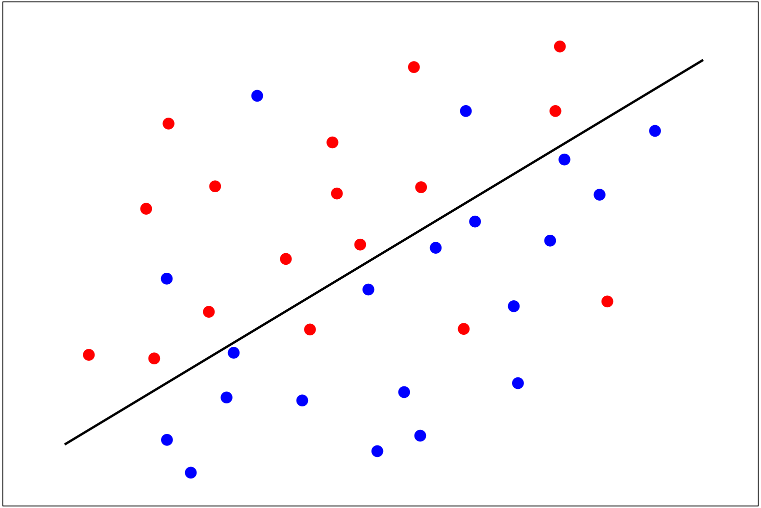


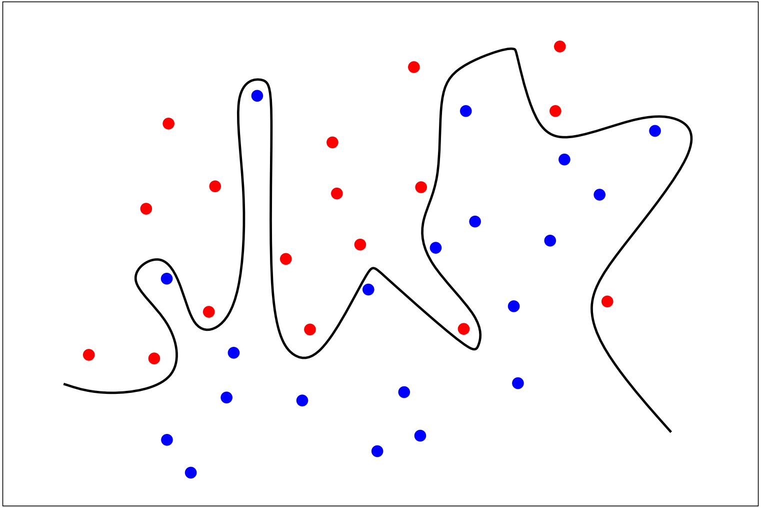


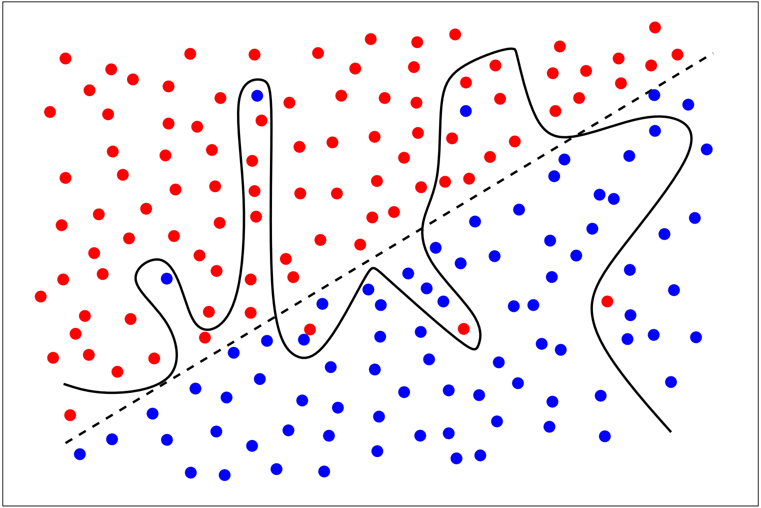


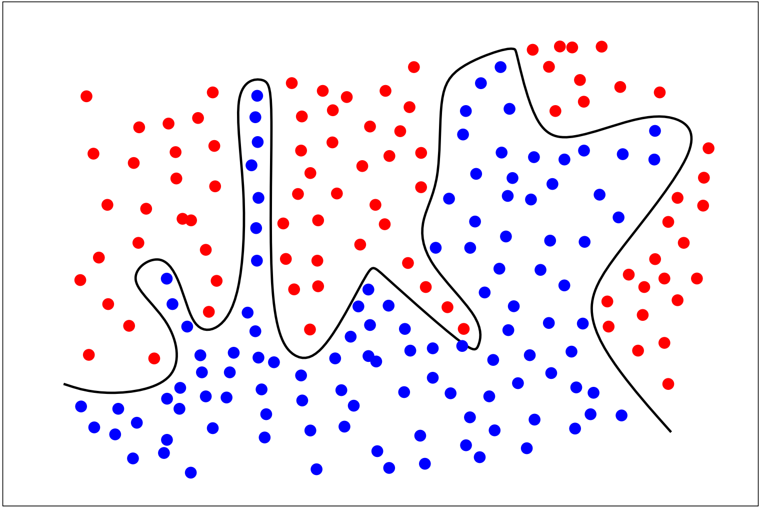


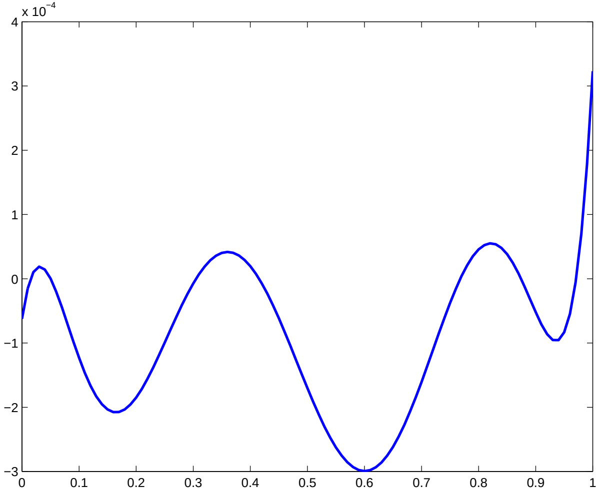


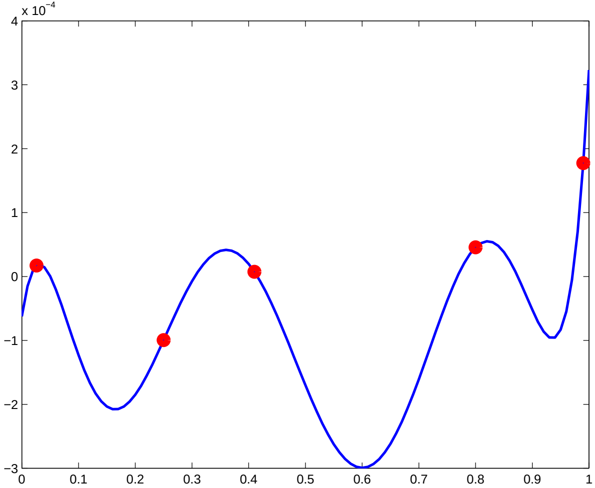


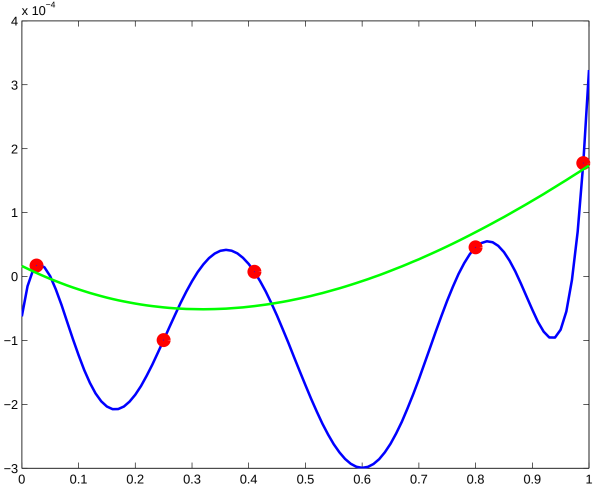


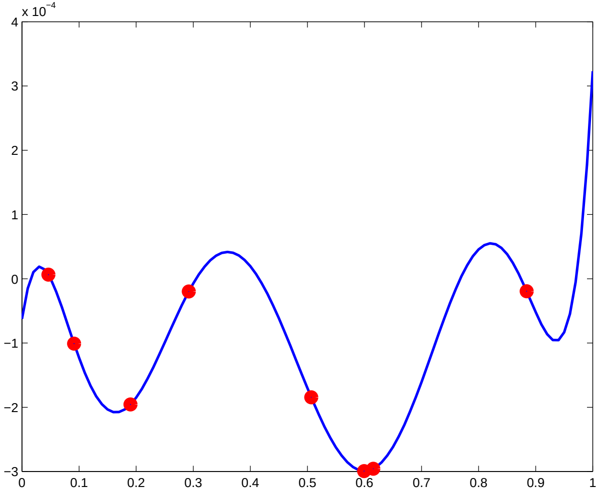


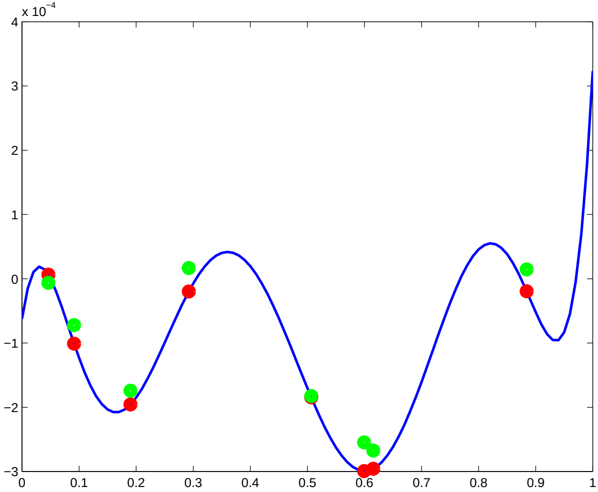


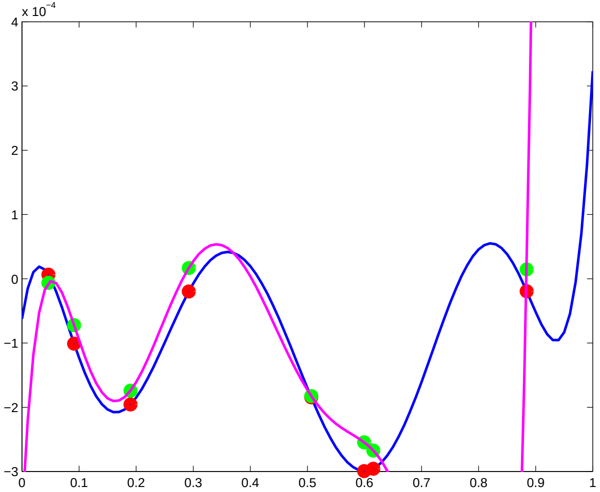

x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

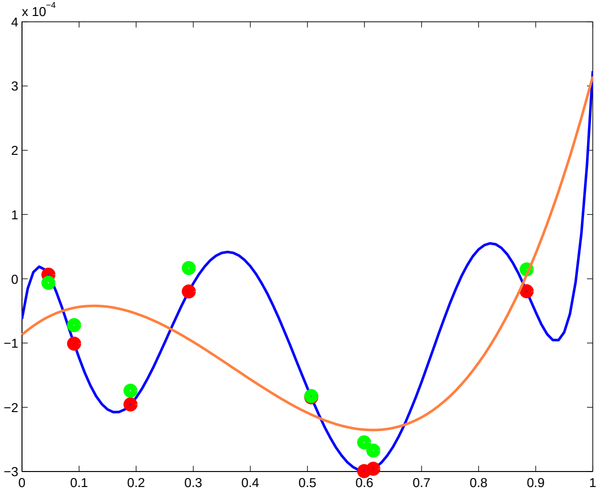

x 10^−4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


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

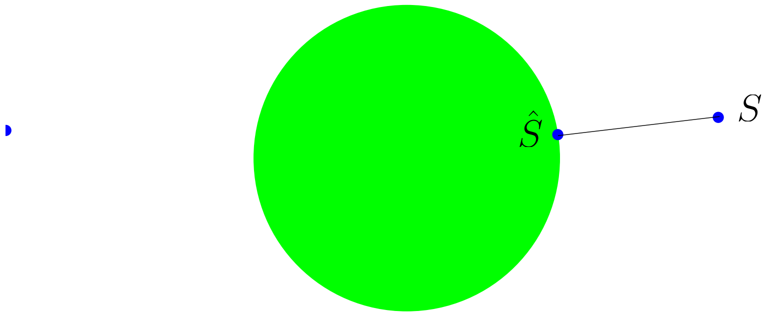

x 10^{-4}
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


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

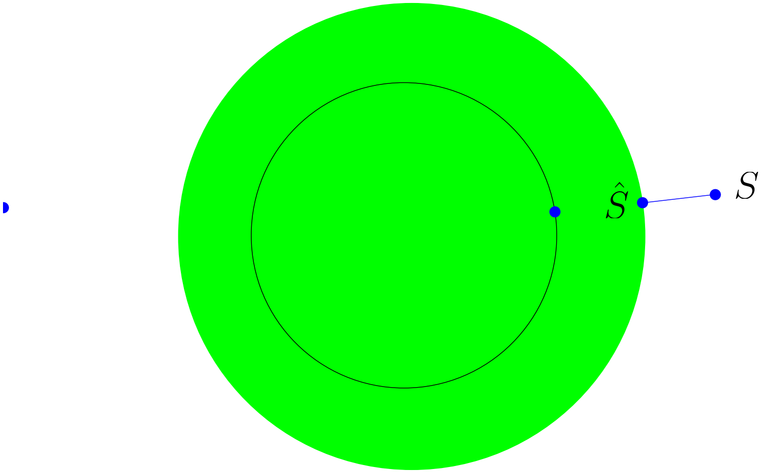

x 10^-4
4
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


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

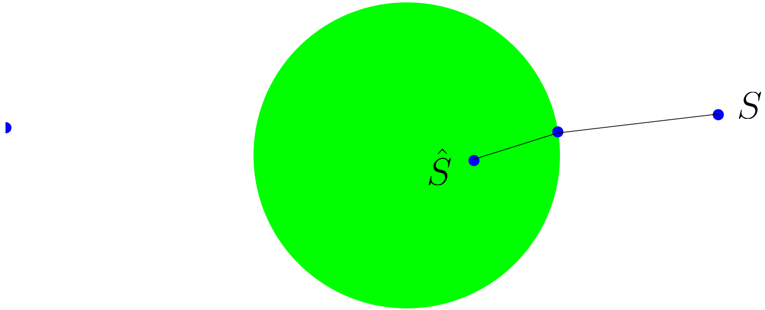
Error de Aproximación



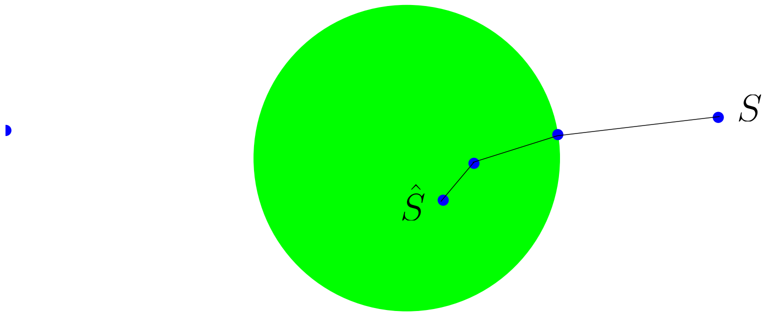
Incrementar complejidad



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 de optimización (con/sin restricciones)

Dificultades Adicionales

- Modelos altamente no lineales:
  - Aprendizaje es lento (comparado por ejemplo con modelos lineales).
  - Algoritmos de optimización (con/sin restricciones)
  - Difícil interpretación.

Dificultades Adicionales

- Modelos altamente no lineales:
  - Aprendizaje es lento (comparado por ejemplo con modelos lineales).
  - Algoritmos de optimización (con/sin restricciones)
  - 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 de optimización (con/sin restricciones)
  - Difícil interpretación.
- Dimensionalidad alta de la entrada ("maldición" de la dimensionalidad).
- 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?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^{n}$ son i.i.d.

Con respecto a los datos

- Cuáles?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^{n}$ son i.i.d.
  - Datos de prueba provienen de la misma distribución.

Con respecto a los datos

- Cuáles?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^n$ 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?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^{n}$ 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?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^{n}$ 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?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^n$ 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?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^n$ 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?
  - $\{\mathrm{x}_i, y_i\}_{i=1}^{n}$ 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:

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

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:

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

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

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

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

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

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

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

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

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

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:

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

- El clasificador óptimo recibe el nombre de clasificador de Bayes.
- $l(\mathbf{x})$ 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 $\Sigma_0 = \Sigma_1 = \Sigma$:

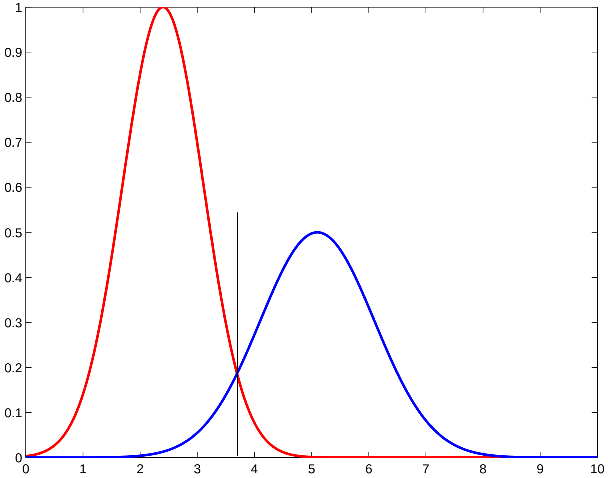
$$
\begin{aligned}
\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mathbf{m}_0^T\Sigma^{-1}\mathbf{x} + \mathbf{m}_0^T\Sigma^{-1}\mathbf{m}_0 - \mathbf{x}^T\Sigma^{-1}\mathbf{x} \\
+ 2\mathbf{m}_1^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_1^T\Sigma^{-1}\mathbf{m}_1 > 2\ln\left(\frac{1-\alpha}{\alpha}\right)
\end{aligned}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

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

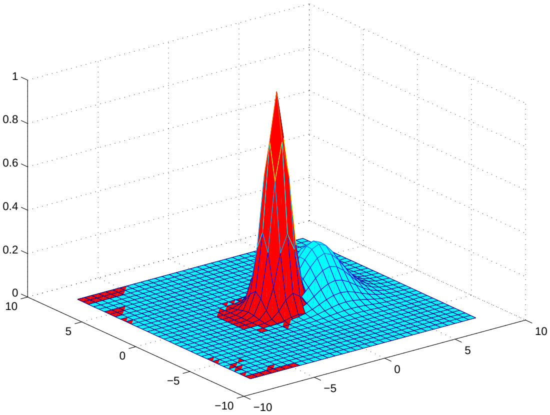
$$\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mathbf{m}_0^T\Sigma^{-1}\mathbf{x} + \mathbf{m}_0^T\Sigma^{-1}\mathbf{m}_0 - \mathbf{x}^T\Sigma^{-1}\mathbf{x} + 2\mathbf{m}_1^T\Sigma^{-1}\mathbf{x} - \mathbf{m}_1^T\Sigma^{-1}\mathbf{m}_1 > 2\ln\left(\frac{1-\alpha}{\alpha}\right)$$

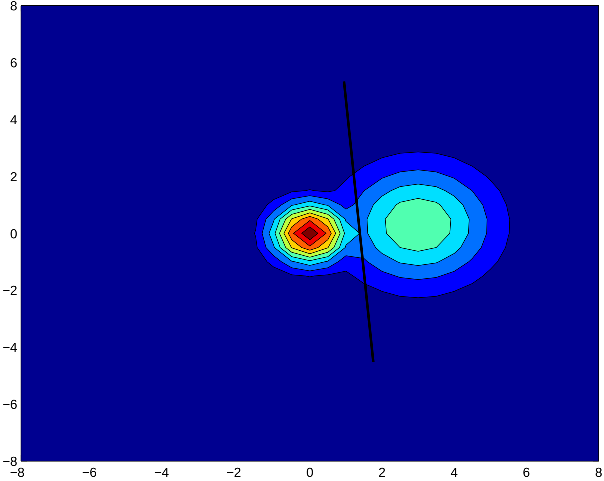
- entonces:

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



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

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:

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

Desigualdad de Markov

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$.

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$. Entonces:

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}$$

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$. Entonces:

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}$$

Demostración.

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$. Entonces:

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}$$

Demostración.

$$\mathbf{P}[X \geq a]$$

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$. Entonces:

$$\mathbf{P}\left[X \geq a\right] \leq \frac{\mathbf{E}\left[X\right]}{a}$$

Demostración.

$$\mathbf{P}\left[X \geq a\right] = \mathbf{E}\left[I_{\{X \geq a\}}\right]$$

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$. Entonces:

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}$$

Demostración.

$$\mathbf{P}[X \geq a] = \mathbf{E}\left[I_{\{X \geq a\}}\right] \leq \mathbf{E}\left[\frac{X}{a}\right]$$

Desigualdad de Markov

Teorema

Sea X una variable aleatoria con $X \geq 0$ casi seguramente, y $a > 0$. Entonces:

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}$$

Demostración.

$$\mathbf{P}[X \geq a] = \mathbf{E}\left[I_{\{X \geq a\}}\right] \leq \mathbf{E}\left[\frac{X}{a}\right] = \frac{\mathbf{E}[X]}{a}$$

$\square$

Desigualdad de Chebyshev

Teorema

Sea X una variable aleatoria con $\mathbf{E}[X] = \mu$, $\mathbf{E}\left[(X-\mu)^2\right] = \sigma^2$.

Desigualdad de Chebyshev

Teorema

Sea X *una variable aleatoria con* $\mathbf{E}[X] = \mu$, $\mathbf{E}\left[(X-\mu)^2\right] = \sigma^2$. *Entonces* $\forall a > 0$

$$\mathbf{P}[|X-\mu| \geq a] \leq \frac{\sigma^2}{a^2}$$

Desigualdad de Chebyshev

Teorema

Sea X una variable aleatoria con $\mathbf{E}\left[X\right] = \mu$, $\mathbf{E}\left[(X-\mu)^2\right] = \sigma^2$. Entonces $\forall a > 0$

$$\mathbf{P}\left[|X-\mu| \geq a\right] \leq \frac{\sigma^2}{a^2}$$

Demostración.

$$\mathbf{P}\left[|X-\mu| \geq a\right]$$

Desigualdad de Chebyshev

Teorema

Sea X *una variable aleatoria con* $\mathbf{E}[X] = \mu$, $\mathbf{E}\left[(X-\mu)^2\right] = \sigma^2$.
Entonces $\forall a > 0$

$$\mathbf{P}\left[|X-\mu| \geq a\right] \leq \frac{\sigma^2}{a^2}$$

Demostración.

$$\mathbf{P}\left[|X-\mu| \geq a\right] = \mathbf{P}\left[|X-\mu|^2 \geq a^2\right]$$

Desigualdad de Chebyshev

Teorema

Sea X *una variable aleatoria con* $\mathbf{E}[X] = \mu$, $\mathbf{E}\left[(X-\mu)^2\right] = \sigma^2$.
Entonces $\forall a > 0$

$$\mathbf{P}\left[|X-\mu| \geq a\right] \leq \frac{\sigma^2}{a^2}$$

Demostración.

$$\mathbf{P}\left[|X-\mu| \geq a\right] = \mathbf{P}\left[|X-\mu|^2 \geq a^2\right] \underset{\substack{\downarrow \\ \text{Markov}}}{\leq} \frac{\sigma^2}{a^2}$$

$\square$

Desigualdad de Hoeffding

Teorema

Sean $X_1, X_2, \ldots, X_n$ *variables aleatorias independientes, con*

Desigualdad de Hoeffding

Teorema

Sean $X_1, X_2, \ldots, X_n$ *variables aleatorias independientes, con*

- $\mathbf{E}[X_j] = 0$ *para* $j = 1, \ldots, n$.

Desigualdad de Hoeffding

Teorema

Sean $X_1, X_2, \ldots, X_n$ *variables aleatorias independientes, con*

- $\mathbf{E}[X_j] = 0$ *para* $j = 1, \ldots, n$.
- $a_j \leq X_j \leq b_j$, *con* $a_j, b_j \in \mathbb{R}$ *para* $j = 1, \ldots, n$.

Desigualdad de Hoeffding

Teorema

Sean $X_1, X_2, \ldots, X_n$ *variables aleatorias independientes, con*

- $\mathbf{E}[X_j] = 0$ *para* $j = 1, \ldots, n$.
- $a_j \leq X_j \leq b_j$, *con* $a_j, b_j \in \mathbb{R}$ *para* $j = 1, \ldots, n$.

entonces:

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \leq -\epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

Desigualdad de Hoeffding

Teorema

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes, con

- $\mathbf{E}[X_j] = 0$ *para* $j = 1, \ldots, n$.
- $a_j \leq X_j \leq b_j$, *con* $a_j, b_j \in \mathbb{R}$ *para* $j = 1, \ldots, n$.

entonces:

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \leq -\epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

o combinando:

$$\mathbf{P}\left[\left|\sum_{j=1}^{n} X_j\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

Caso particular: Cotas de Chernoff

Caso particular: Cotas de Chernoff

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

Caso particular: Cotas de Chernoff

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

Caso particular: Cotas de Chernoff

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

Caso particular: Cotas de Chernoff

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

Caso particular: Cotas de Chernoff

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

Caso particular: Cotas de Chernoff

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

Caso particular: Cotas de Chernoff

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

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

y

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

Caso particular: Cotas de Chernoff

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

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

y

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

- Estas son las cotas de Chernoff en forma aditiva.

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 $|e(h) - \hat{e}(h)| \leq \epsilon$

Demostración de Hoeffding

Lema

Sea X una variable aleatoria con media cero tal que $a \leq X \leq b$, entonces $\forall s > 0$

$$\mathbf{E}\left[e^{sX}\right] \leq e^{s^2 \frac{(b-a)^2}{8}}$$

Demostración de Hoeffding

Lema

Sea X una variable aleatoria con media cero tal que $a \leq X \leq b$, entonces $\forall s > 0$

$$\mathbf{E}\left[e^{sX}\right] \leq e^{s^2 \frac{(b-a)^2}{8}}$$

Demostración.

- Por convexidad de la función exponencial:

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

Demostración de Hoeffding

Lema

Sea X una variable aleatoria con media cero tal que $a \leq X \leq b$, entonces $\forall s > 0$

$$\mathbf{E}\left[e^{sX}\right] \leq e^{s^2 \frac{(b-a)^2}{8}}$$

Demostración.

- Por convexidad de la función exponencial:

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

- Tomando valor esperado:

Demostración de Hoeffding

Lema

Sea X una variable aleatoria con media cero tal que $a \leq X \leq b$, entonces $\forall s > 0$

$$\mathbf{E}\left[e^{sX}\right] \leq e^{s^2 \frac{(b-a)^2}{8}}$$

Demostración.

- Por convexidad de la función exponencial:

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

- Tomando valor esperado:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}$$

Demostración de Hoeffding

Lema

Sea X una variable aleatoria con media cero tal que $a \leq X \leq b$, entonces $\forall s > 0$

$$\mathbf{E}\left[e^{sX}\right] \leq e^{s^2 \frac{(b-a)^2}{8}}$$

Demostración.

- Por convexidad de la función exponencial:

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$

- Tomando valor esperado:

$$\mathbf{E} e^{sX} \leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}$$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u}$$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u})$$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^u) = e^{\phi(u)}$$

con $\phi(u) = -pu + \ln(1-p+pe^u)$.

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1-p+pe^{u})$.
- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a} e^{\frac{a}{b-a}u} - \frac{a}{b-a} e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1 - p + pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1 - p + pe^{u})$.
- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a} e^{\frac{a}{b-a}u} - \frac{a}{b-a} e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1-p+pe^{u})$.

- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a} e^{\frac{a}{b-a}u} - \frac{a}{b-a} e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1-p+pe^{u})$.
- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1-p+pe^{u})$.
- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor
  - $\phi(0) =$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1-p+pe^{u})$.
- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor
  - $\phi(0) = 0$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

con $\phi(u) = -pu + \ln(1-p+pe^{u})$.

- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor
  - $\phi(0) = 0$
  - $\phi'(p) = -p + \frac{p}{p+(1-p)e^{-u}}$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

con $\phi(u) = -pu + \ln(1-p+pe^{u})$.

- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor
  - $\phi(0) = 0$
  - $\phi'(p) = -p + \frac{p}{p+(1-p)e^{-u}} \Rightarrow \phi'(0) = 0$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

con $\phi(u) = -pu + \ln(1-p+pe^{u})$.

- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor
    - $\phi(0) = 0$
    - $\phi'(p) = -p + \frac{p}{p+(1-p)e^{-u}} \Rightarrow \phi'(0) = 0$
    - $\phi''(p) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2}$

- cambio de variable $u = s(b-a)$:

$$\mathbf{E}e^{sX} \leq \frac{b}{b-a}e^{\frac{a}{b-a}u} - \frac{a}{b-a}e^{\frac{b}{b-a}u}$$

- denotando $p = -\frac{a}{b-a}$:

$$\mathbf{E}e^{sX} \leq (1-p)e^{-pu} + pe^{(1-p)u} = e^{-pu}(1-p+pe^{u}) = e^{\phi(u)}$$

  con $\phi(u) = -pu + \ln(1-p+pe^{u})$.
- Queremos probar $e^{\phi(u)} \leq e^{\frac{u^2}{8}}$, es decir $\phi(u) \leq \frac{u^2}{8}$
- Expandir $\phi(u)$ en serie de Taylor
  - $\phi(0) = 0$
  - $\phi'(p) = -p + \frac{p}{p+(1-p)e^{-u}} \Rightarrow \phi'(0) = 0$
  - $\phi''(p) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq \frac{1}{4}$ (porque $\frac{xy}{(x+y)^2} \leq \frac{1}{4}\ \forall x, y$)

- Por serie de Taylor con residuo, para algún $\theta \in [0, u]$,

$$\phi(u) =$$

- Por serie de Taylor con residuo, para algún $\theta \in [0, u]$,

$$\phi(u) = \phi(0)$$

- Por serie de Taylor con residuo, para algún $\theta \in [0, u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta)$$

- Por serie de Taylor con residuo, para algún $\theta \in [0, u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8}$$

$\square$

... de vuelta a Hoeffding

... de vuelta a Hoeffding

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] = \mathbf{P}\left[e^{s\sum_{j=1}^{n} X_j} \geq e^{s\epsilon}\right]$$

... de vuelta a Hoeffding

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] = \mathbf{P}\left[e^{s\sum_{j=1}^{n} X_j} \geq e^{s\epsilon}\right]$$

$$\underset{\substack{\downarrow \\ \text{Markov}}}{\leq} \frac{\mathbf{E}\left[e^{s\sum_{j=1}^{n} X_j}\right]}{e^{s\epsilon}}$$

... de vuelta a Hoeffding

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] = \mathbf{P}\left[e^{s\sum_{j=1}^{n} X_j} \geq e^{s\epsilon}\right]$$

$$\underset{\text{Markov}}{\underset{\downarrow}{\leq}} \frac{\mathbf{E}\left[e^{s\sum_{j=1}^{n} X_j}\right]}{e^{s\epsilon}}$$

$$= e^{-s\epsilon}\mathbf{E}\left[\prod_{j=1}^{n} e^{sX_j}\right]$$

... de vuelta a Hoeffding

$$
\begin{aligned}
\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] &= \mathbf{P}\left[e^{s\sum_{j=1}^{n} X_j} \geq e^{s\epsilon}\right] \\
&\underset{\substack{\downarrow \\ \text{Markov}}}{\leq} \frac{\mathbf{E}\left[e^{s\sum_{j=1}^{n} X_j}\right]}{e^{s\epsilon}} \\
&= e^{-s\epsilon}\mathbf{E}\left[\prod_{j=1}^{n} e^{sX_j}\right] \underset{\substack{\downarrow \\ X_j\,\text{ind.}}}{=} e^{-s\epsilon}\prod_{j=1}^{n}\mathbf{E}\left[e^{sX_j}\right]
\end{aligned}
$$

... de vuelta a Hoeffding

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] = \mathbf{P}\left[e^{s\sum_{j=1}^{n} X_j} \geq e^{s\epsilon}\right]$$

$$\underset{\text{Markov}}{\underset{\downarrow}{\leq}} \frac{\mathbf{E}\left[e^{s\sum_{j=1}^{n} X_j}\right]}{e^{s\epsilon}}$$

$$= e^{-s\epsilon}\mathbf{E}\left[\prod_{j=1}^{n} e^{sX_j}\right] \underset{X_j\,\text{ind.}}{\underset{\downarrow}{=}} e^{-s\epsilon}\prod_{j=1}^{n}\mathbf{E}\left[e^{sX_j}\right]$$

$$\underset{\text{lema}}{\underset{\downarrow}{\leq}} e^{-s\epsilon}\prod_{j=1}^{n} e^{s^2\frac{(b_j-a_j)^2}{8}}$$

... de vuelta a Hoeffding

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] = \mathbf{P}\left[e^{s\sum_{j=1}^{n} X_j} \geq e^{s\epsilon}\right]$$

$$\underset{\text{Markov}}{\underset{\downarrow}{\leq}} \frac{\mathbf{E}\left[e^{s\sum_{j=1}^{n} X_j}\right]}{e^{s\epsilon}}$$

$$= e^{-s\epsilon}\mathbf{E}\left[\prod_{j=1}^{n} e^{sX_j}\right] \underset{X_j\text{ind.}}{\underset{\downarrow}{=}} e^{-s\epsilon}\prod_{j=1}^{n}\mathbf{E}\left[e^{sX_j}\right]$$

$$\underset{\text{lema}}{\underset{\downarrow}{\leq}} e^{-s\epsilon}\prod_{j=1}^{n} e^{s^2\frac{(b_j-a_j)^2}{8}} = \exp\left(-s\epsilon + \frac{s^2}{8}\sum_{j=1}^{n}(b_j - a_j)^2\right)$$

- Tenemos que

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-s\epsilon + \frac{s^2}{8}\sum_{j=1}^{n}(b_j - a_j)^2\right)$$

para cualquier $s > 0$.

- Tenemos que

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-s\epsilon + \frac{s^2}{8}\sum_{j=1}^{n}(b_j - a_j)^2\right)$$

para cualquier $s > 0$.

- Para obtener la mejor cota posible, minimizamos el lado derecho con respecto a s.

- Tenemos que

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-s\epsilon + \frac{s^2}{8}\sum_{j=1}^{n}(b_j - a_j)^2\right)$$

  para cualquier $s > 0$.
- Para obtener la mejor cota posible, minimizamos el lado derecho con respecto a s.
- Derivando e igualando a cero, tenemos $s = \frac{4\epsilon}{\sum_{j=1}^{n}(b_j - a_j)^2}$.

- Tenemos que

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-s\epsilon + \frac{s^2}{8}\sum_{j=1}^{n}(b_j - a_j)^2\right)$$

  para cualquier $s > 0$.
- Para obtener la mejor cota posible, minimizamos el lado derecho con respecto a s.
- Derivando e igualando a cero, tenemos $s = \frac{4\epsilon}{\sum_{j=1}^{n}(b_j - a_j)^2}$. Reemplazando tenemos:

$$\mathbf{P}\left[\sum_{j=1}^{n} X_j \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right)$$

Teorema (McDiarmid(1989))

Sean $X_1, X_2, \ldots, X_n$ *variables aleatorias independientes que toman valores en un conjunto* A, *y asuma que* $f : A^n \to \mathbb{R}$ *satisface:*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_i' \in A}} \left| f(x_1, \ldots x_n) - f(x_1, \ldots x_{i-1}, x_i', x_{i+1}, \ldots x_n) \right| \leq c_i$$

Entonces, para todo $\epsilon > 0$,

$$\mathbf{P}\left[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}}$$

y

$$\mathbf{P}\left[\mathbf{E}f(X_1, \ldots, X_n) - f(X_1, \ldots, X_n) \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}}$$

Teorema (McDiarmid(1989))

Sean $X_1, X_2, \ldots, X_n$ *variables aleatorias independientes que toman valores en un conjunto* A*, y asuma que* $f : A^n \to \mathbb{R}$ *satisface:*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_i' \in A}} \left| f(x_1, \ldots x_n) - f(x_1, \ldots x_{i-1}, x_i', x_{i+1}, \ldots x_n) \right| \leq c_i$$

Entonces, para todo $\epsilon > 0$,

$$\mathbf{P}\left[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}}$$

y

$$\mathbf{P}\left[\mathbf{E}f(X_1, \ldots, X_n) - f(X_1, \ldots, X_n) \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}}$$

- Si la variación de f con respecto al cambio de una variable es pequeña, entonces la variable aleatoria $f(X_1, \ldots X_n)$ es concentrada.