Proyecto ML de principio a fin

Alfredo Cuesta Infante

E. T. S. Ingeniería Informática
Universidad Rey Juan Carlos

Master Univ. en Visión Artificial
Reconocimiento de Patrones

Secuencia de un proyecto ML

Preparar los datos

Ingeniería de características

Selección del método de entrenamiento y ejecución

Contenidos

Secuencia de un proyecto ML

Preparar los datos

Ingeniería de características

Selección del método de entrenamiento y ejecución

Contenidos

Secuencia de un proyecto ML

Preparar los datos

Ingeniería de características

Selección del método de entrenamiento y ejecución

Secuencia de un proyecto ML

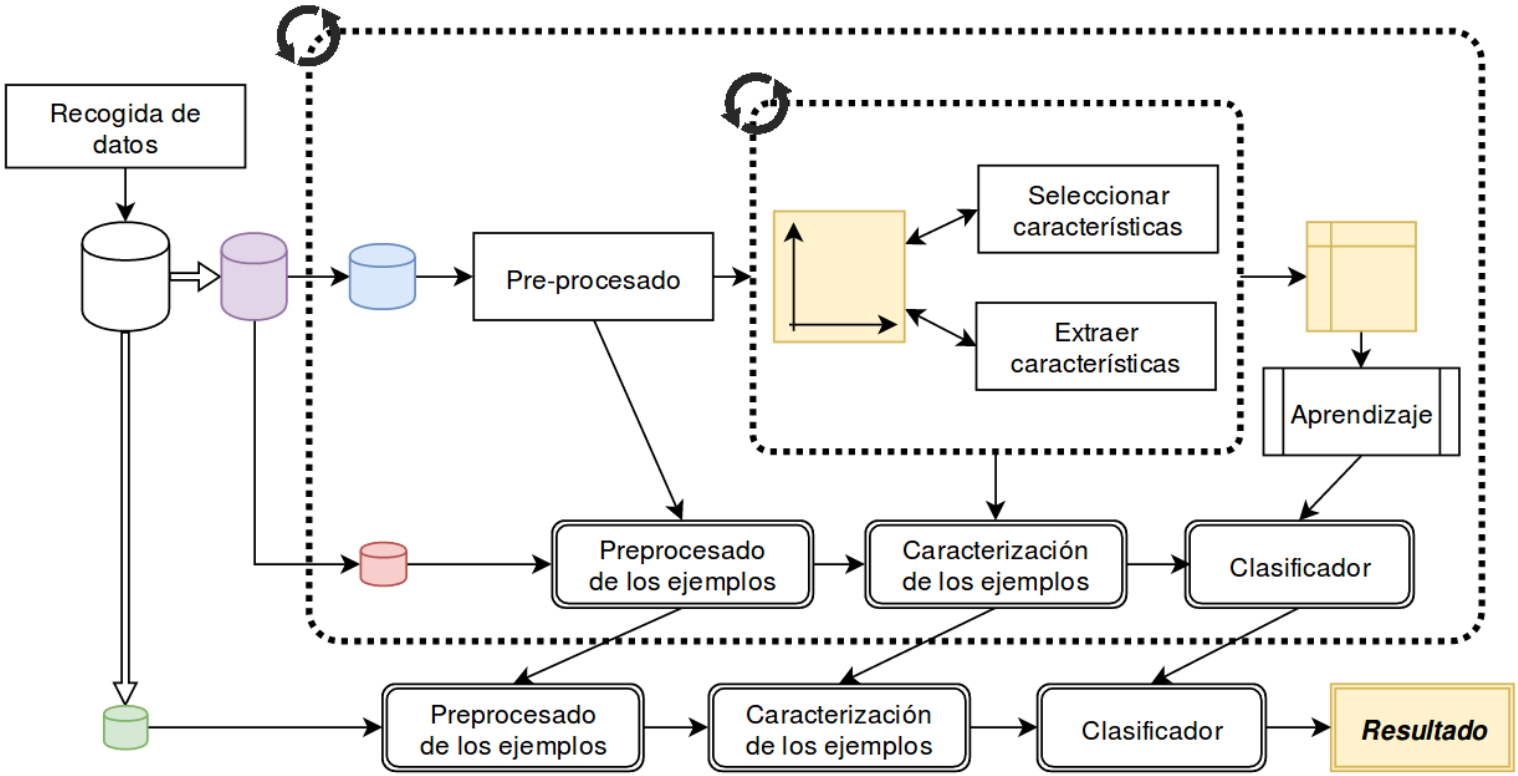


Figura: Diagrama de trabajo para la construcción de un clasificador [Fuente: Original de A. Cuesta]

Secuencia de un proyecto ML

1. Recoger conjunto de ejemplos (en color blanco)
2. Reservar un subconjunto de datos para *Test* (en color verde).
3. Separar el resto de los datos en dos:
   - *Entrenamiento* (en color azul)
   - *Validación* (en color rojo)
4. Casi siempre es necesario preprocesar los datos de entrenamiento.
   - Normalización de los datos (cambio de escala)
   - **Importante**: guardar tanto el método empleado como los parámetros que se utilizaron ya que será necesario para procesar los datos de validación y test del mismo modo.
5. Realizar *ingeniería de características* para lograr una representación de los datos con mayor potencial discriminante.
   - Recomendación: reducir la dimensionalidad a 2, que puede ser representado en un gráfico.
   - Resultado: conjunto de datos sobre el que se puede aplicar el algoritmo de aprendizaje.
   - Formato: tabla donde las filas son ejemplos y las columnas son características
6. Realizar varias iteraciones del entrenamiento con diferentes conjuntos de entrenamiento y validación
7. **Resultado**: clasificador con el que ya podemos probar los datos del conjunto de test, que serán los que den la medida del verdadero rendimiento.

Contenidos

Secuencia de un proyecto ML

Preparar los datos

Ingeniería de características

Selección del método de entrenamiento y ejecución

Explorar los datos

Ejemplo-guía: 0 vs. 1

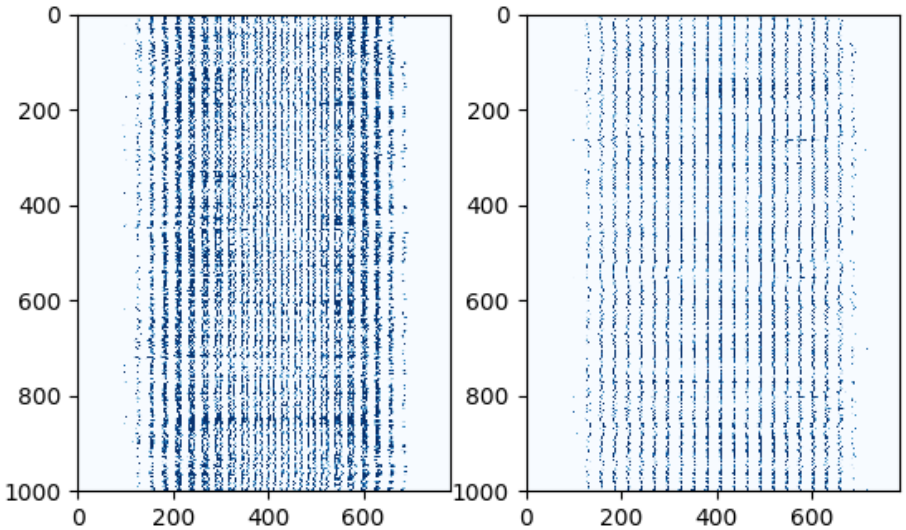


Figura: Conjunto de '0' (izq.) y '1' (der.) proporcionado [Fuente: Original de A. Cuesta]

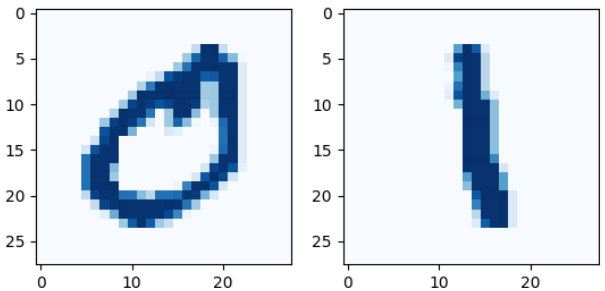


Figura: Fila #1 reordenada como matriz 28×28 del conjunto '0' (izq.) y del conjunto '1' (der.) [Fuente: Original de A. Cuesta]

Explorar los datos

Datos erróneos

- Alguna característica del vector de características en uno o varios ejemplos es incorrecta
- Hay que corregirlo de algún modo
- Ningún método funcionará bien con datos erróneos

Datos perdidos

- Falta alguna característica en el vector de características de uno o varios ejemplos
- Un dato perdido no es necesariamente un error
- Se puede corregir, pero también se puede dejar así

  (hay algoritmos que toleran conjuntos con datos perdidos)

Datos anómalos

- Alguna característica del vector de características en uno o varios ejemplos es anómala
- Esto no significa necesariamente que sea incorrecta
- No se debe corregir ni indicar.

Preparar los datos

Validación cruzada

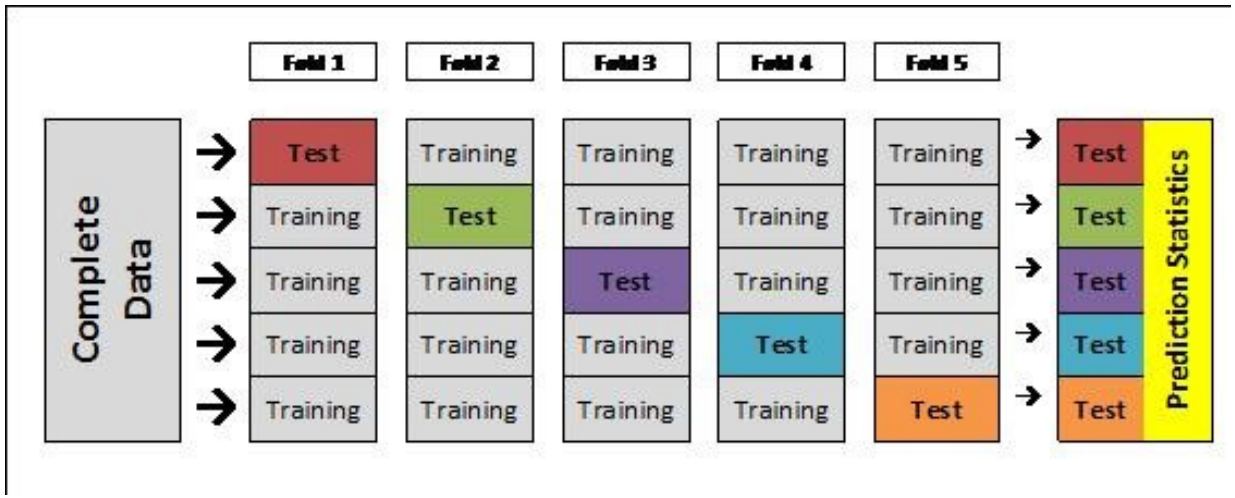


Figura: Ejemplo de 5-*fold* [Fuente: Anónimo @ internet]

Preparar los datos

Escalado lineal

$$Z_i = m \cdot (X_i - \overline{X}_i) + U_i \ , \quad \text{o bien} \quad Z_i = m \cdot (X_i - \underline{X}_i) + L_i,$$

Estandarización

$$Z_i = (X_i - \mu_i)/\sigma_i,$$

Ejemplo

	Datos medidos			Datos escalados en [-1,+1]			Datos estandarizados		
Marca tiempo	Sensor 1	Sensor 2	Sensor 3	Sensor 1	Sensor 2	Sensor 3	Sensor 1	Sensor 2	Sensor 3
0	3.200	17,5	2,25	-0,25	0,13	0,68	-1,45	0,28	1,00
250	4.050	14	2,18	0,57	-0,20	0,59	0,93	-0,74	0,74
500	4.050	11,33	1,83	0,57	-0,45	0,22	0,93	-1,52	-0,47
119500	3.700	19	2,20	0,23	0,27	0,62	-0,05	0,72	
119750	3.875	16,75	2,10	0,40	0,06	0,51	0,44	0,06	0,48
Media =	3.718	16,53	1,96				0,00 ✓	0,00 ✓	0,00 ✓
Desv.=	358	3,43	0,29				1,00 ✓	1,00 ✓	1,00 ✓
Máx=	4.500	26,75	2,55	1,00 ✓	1,00 ✓	1,00 ✓			
Mín=	2.425	5,50	0,70	-1,00 ✓	-1,00 ✓	-1,00 ✓			

Figura: Datos medidos, escalados a [-1,+1] y normalizados. [Fuente: Original de A. Cuesta]

Contenidos

Secuencia de un proyecto ML

Preparar los datos

Ingeniería de características

Selección del método de entrenamiento y ejecución

Ingeniería de características

- Selección de características
- Extracción de características

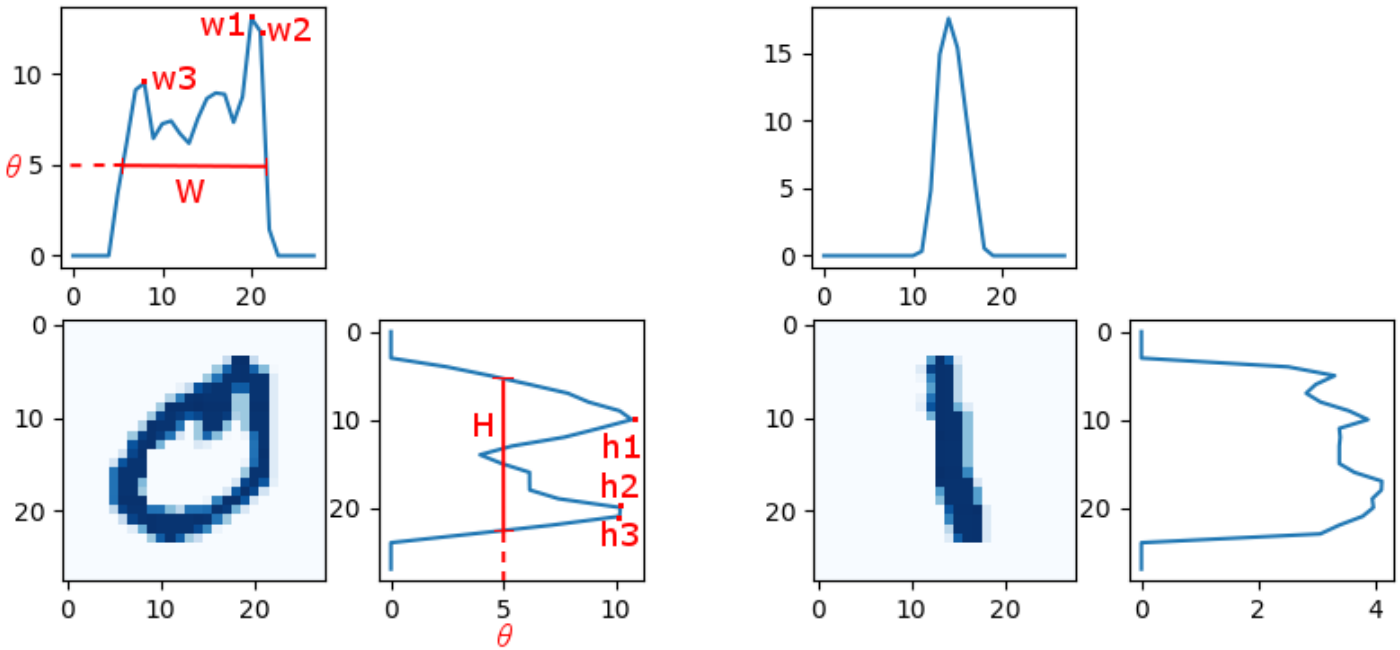
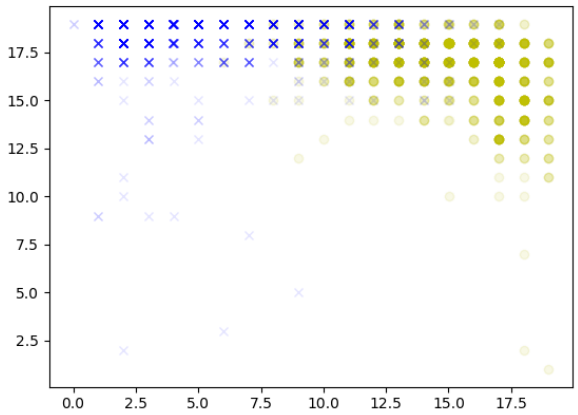


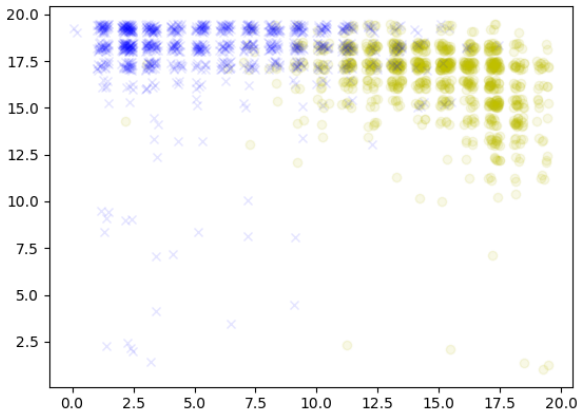
Figura: Proyección horizontal y vertical de un '0' y un '1'. [Fuente: Original de A. Cuesta]

Ingeniería de características

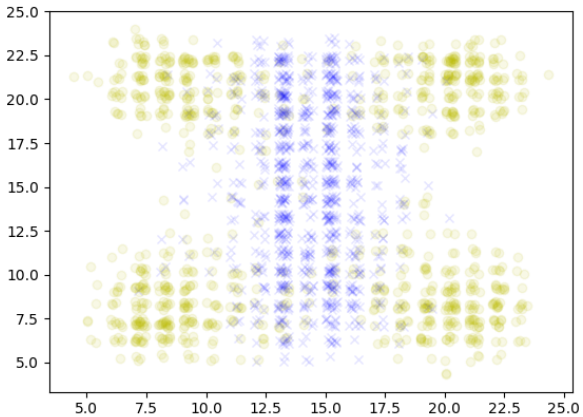
Visualización de características seleccionadas



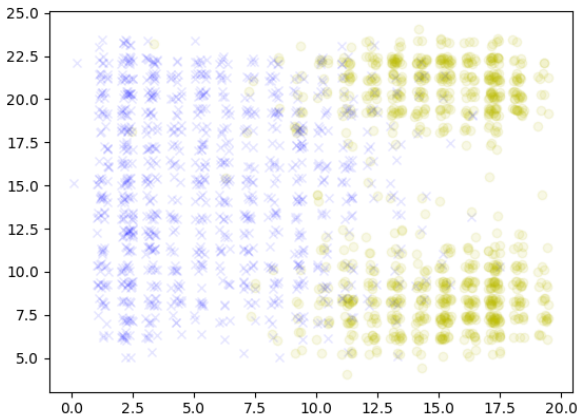
(a) W vs. H sin *jitter*



(b) W vs. H con *jitter*



(c) w_1 vs. h_1 con *jitter*

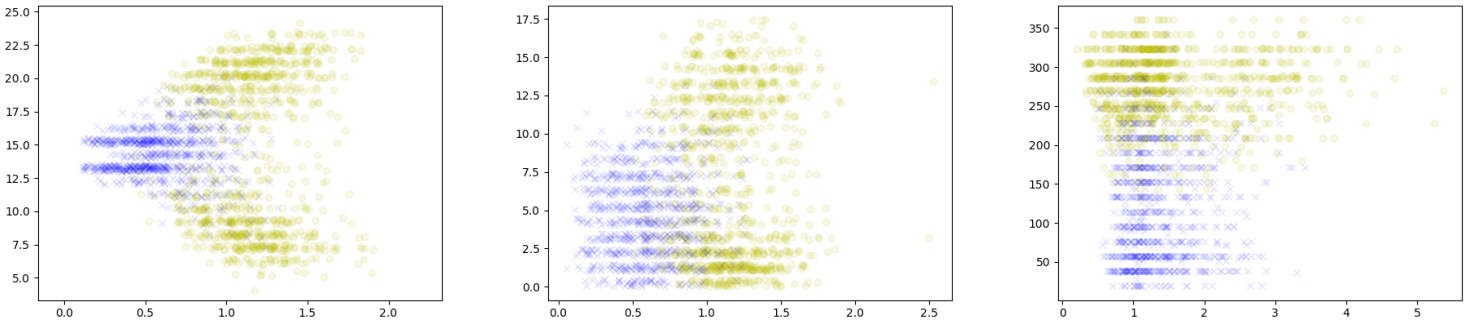


(d) W vs. w_1 con *jitter*

Figura: Conjunto de entrenamiento seleccionando 2 características. [Fuente: Original de A. Cuesta]

Ingeniería de características

Visualización de características extraídas



(a) w_1/W vs. h_1/H (b) w_1/w_2 vs. h_1/h_2 (c) W^2 vs. H^2.

Figura: Conjunto de entrenamiento extrayendo 2 características. [Fuente: Original de A. Cuesta]

Contenidos

Secuencia de un proyecto ML

Preparar los datos

Ingeniería de características

Selección del método de entrenamiento y ejecución

Selección del método de entrenamiento

Expresividad de los clasificadores

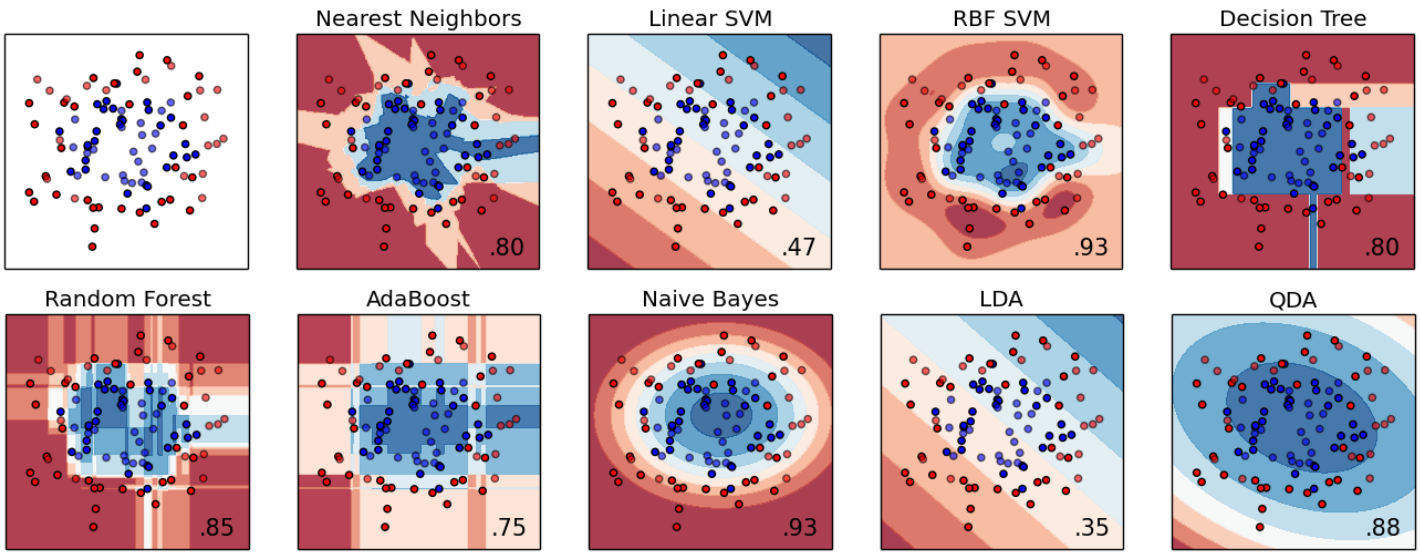


Figura: Expresividad de diferentes métodos de clasificación. [Fuente: SciKit-Learn.org]

Clasificador lineal para ejemplos 2D

Notación

- Cada ejemplo tiene un par de coordenadas (x, y)
- El clasificador se define por el vector de parámetros (w_0, w_1, w_2)

Superficie de decisión

- Es una recta que divide en dos el plano XY
- **Ec. implícita**: $w_0 + w_x x + w_y y = 0$.
- **Ec. explícita**: Despejando y obtenemos,

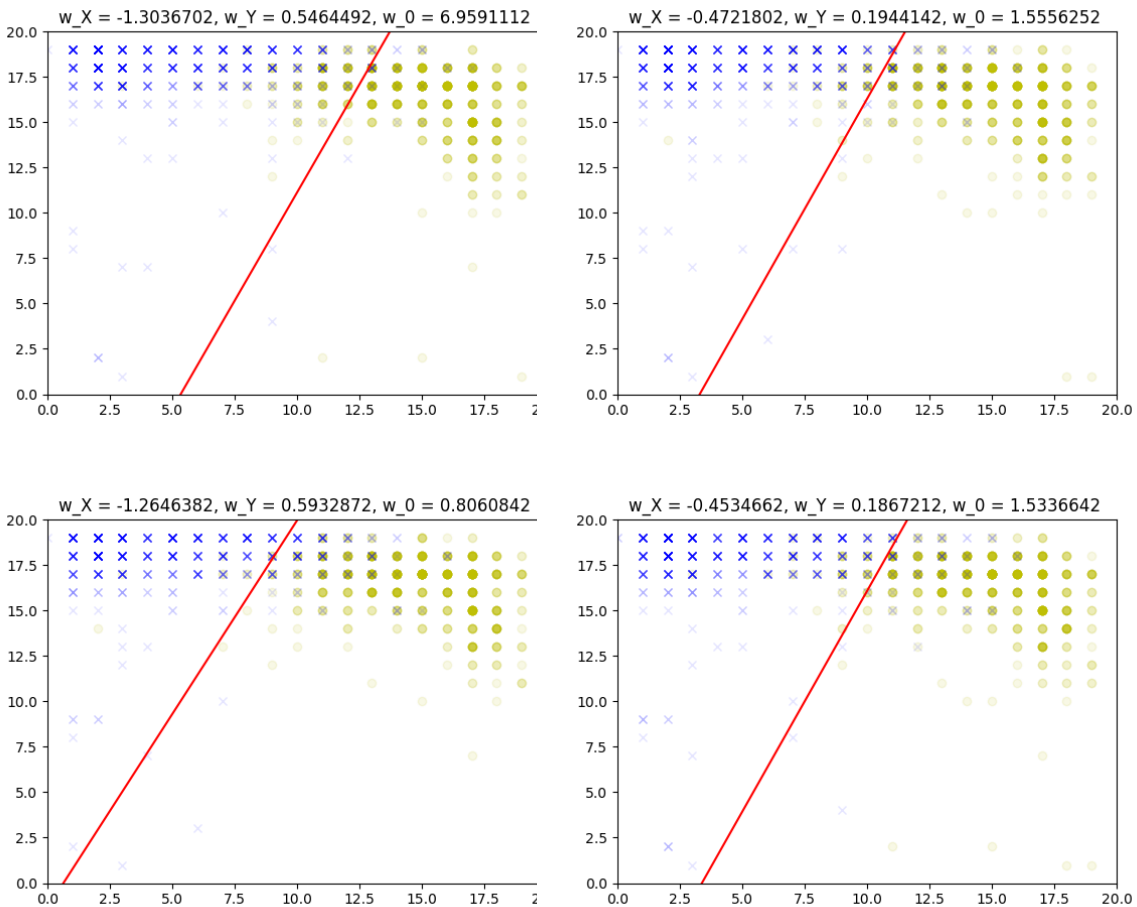
$$y = -\frac{w_x}{w_y} x - \frac{w_0}{w_y}.$$

Método de clasificación

- Descenso del gradiente estocástico (SGD)
- Clasificador basado en Vectores Soporte (SVC)

Ejecución

Usando SGD (sin *jitter*) Usando SVC (sin *jitter*)

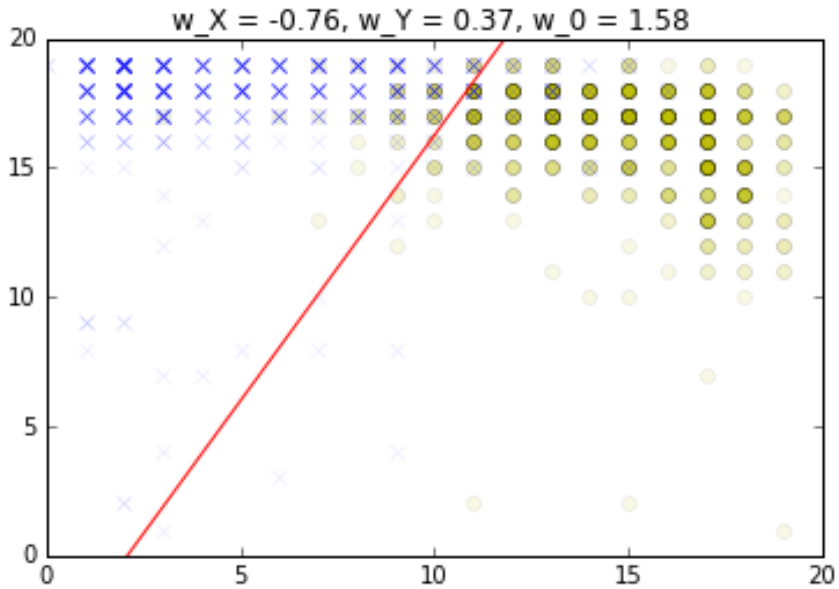


(a) (b)

Figura: Superficie de decisión lineal con dos métodos de clasificación ejecutado dos veces (arriba y abajo). (a) Descenso de gradiente estocástico (SGD) (b) Clasificador basado en vectores soporte (SVC). Se puede apreciar que SGD varía más que SVC. [Fuente: Original de A. Cuesta]

Ejecución con validación cruzada

```
main_2()
```




```
Scores: [ 0.95     0.89375  0.95625  0.8875   0.94375  0.9      0.95625  0.9125
  0.88125  0.93125]
Mean: 0.92125
Standard deviation: 0.028118054698

Test: Hits =  371 ( 92.75%) , Fails = 29 ( 7.25%)
```