Tema 3-1
Clasificación: Conceptos generales

Abdelmalik Moujahid

Grupo de Inteligencia Computacional
Universidad del País Vasco (UPV/EHU)
Curso 2014-2015

Índice

Índice

Introducción

Idea general

- Dada una colección de instancias (**Conjunto de entrenamiento**). Cada instancia se define por un conjunto de atributos; uno de ellos es la clase.
- **Inducir un modelo** para el atributo clase como una función de los valores de los demás atributos.
- **Objetivo**: predecir con la más precisión posible la clase de nuevas instancias (**Conjunto de testeo**).
- La precisión del modelo se estima en base al porcentaje de bien clasificados obtenido en el conjunto de testeo.
- El modelo debe presentar una buena capacidad de generalización.

Introducción

Retos de la clasificación

- ¿Existen realmente éstos grupos en mis Datos?
- ¿Cúales son las similitudes y diferencias entre éstos grupos?
- ¿Las diferencias son suficientemente significativas como para discriminar entre los grupos, y predecir el grupo de una nueva observación?
- ¿Qué variables explican mejor las diferencias entre los grupos?

Introducción

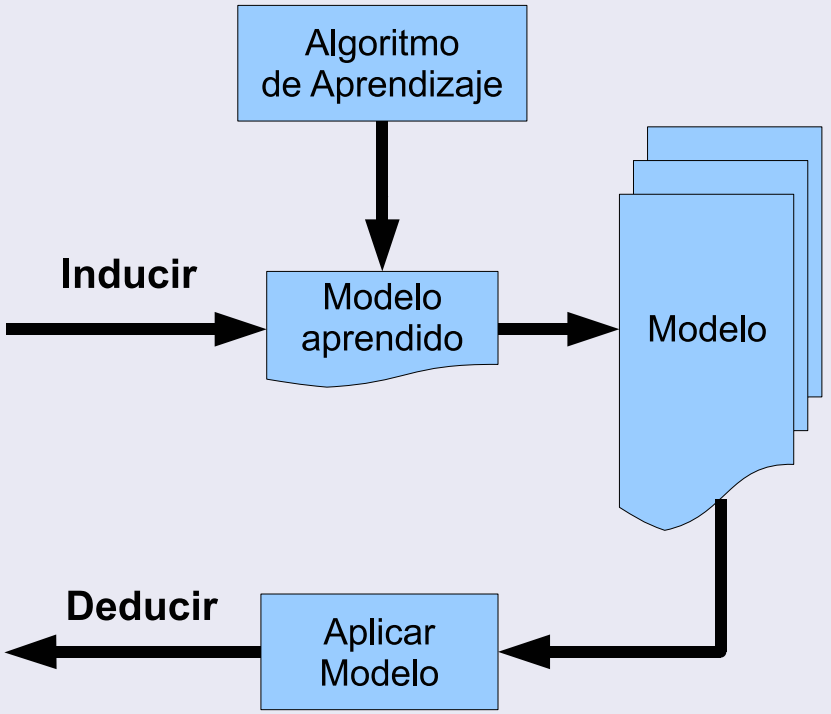
Clasificación: enfoque general

Conjunto de entrenamiento

No	Age	Sex	BP	Cholesterol	Na	K	Drug
1	23.0	F	HIGH	HIGH	0.792535	0.031258	drugY
2	47.0	M	LOW	HIGH	0.739309	0.056468	drugC
3	47.0	M	LOW	HIGH	0.697269	0.068944	drugC
4	28.0	F	NORMAL	HIGH	0.563682	0.072289	drugX
5	61.0	F	LOW	HIGH	0.559294	0.030998	drugY
6	22.0	F	NORMAL	HIGH	0.676901	0.078647	drugX
7	49.0	F	NORMAL	HIGH	0.789637	0.048518	drugY
8	41.0	M	LOW	HIGH	0.766635	0.069461	drugC
9	60.0	M	NORMAL	HIGH	0.777205	0.05123	drugY
10	43.0	M	LOW	NORMAL	0.526102	0.027164	drury

Conjunto de testeo

No	Age	Sex	BP	Cholesterol	Na	K	Drug
11	47.0	F	LOW	HIGH	0.896056	0.076147	?
12	34.0	F	HIGH	NORMAL	0.667775	0.034782	?
13	43.0	M	LOW	HIGH	0.626527	0.040746	?
14	74.0	F	LOW	HIGH	0.792674	0.037851	?
15	50.0	F	NORMAL	HIGH	0.82778	0.065166	?



Esquema general de un modelo de clasificación

Introducción

Classification problem data matrix

	a_1	...	a_j	...	a_m	*Clase*
X_1	x_{11}	...	x_{1j}	...	x_{1m}	c_1
...	...	...	...	...	...	...
X_i	x_{i1}	...	x_{ij}	...	x_{im}	c_i
...	...	...	...	...	...	...
X_n	x_{n1}	...	x_{nj}	...	x_{nm}	c_n
Y	y_1	...	y_2	...	y_m	?

Problema de clasificación supervisada.

Machine learning repository: `http://archive.ics.uci.edu/ml/`

Introducción

Clasificación: pasos a seguir

1. Preparar los datos
2. Elegir un algoritmo de aprendizaje
3. Ajustar el modelo a los datos
4. Elegir un método de validación
5. Examinar el ajuste y adaptar el modelo hasta alcanzar un resultado satisfactorio
6. Utilizar el modelo ajustado para llevar a cabo la predicción

Introducción

Características de los algoritmos de aprendizaje

1. Precisión predictiva
2. Velocidad de ajuste
3. Velocidad de predicción
4. Uso de memoria
5. Facilidad de interpretación

Introducción

Regresión lineal

- Formular el problema en términos de variables predictoras y variable clase
- Definir una función que relaciona la variable clase con las variables predictoras
- Definir el error cuadrático medio (función a optimizar)
- Estimar los parámetros de regresión

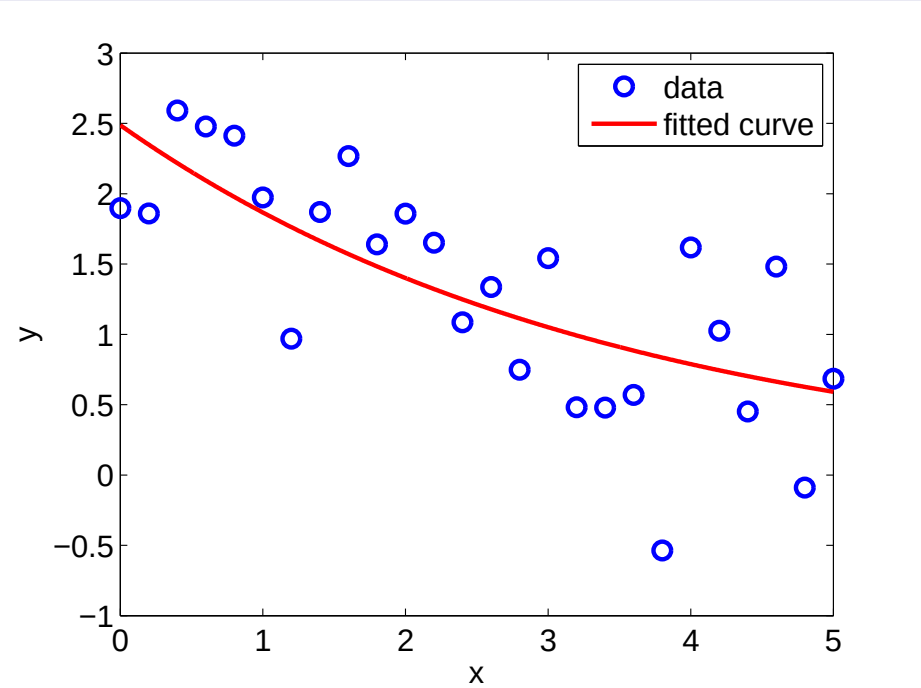
$$f(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

$$f(X) = \beta^T X$$

$$J(\beta) = \frac{1}{2} \sum_{i=1}^{n} (f(X^{(i)}) - C^{(i)})^2$$

Introducción

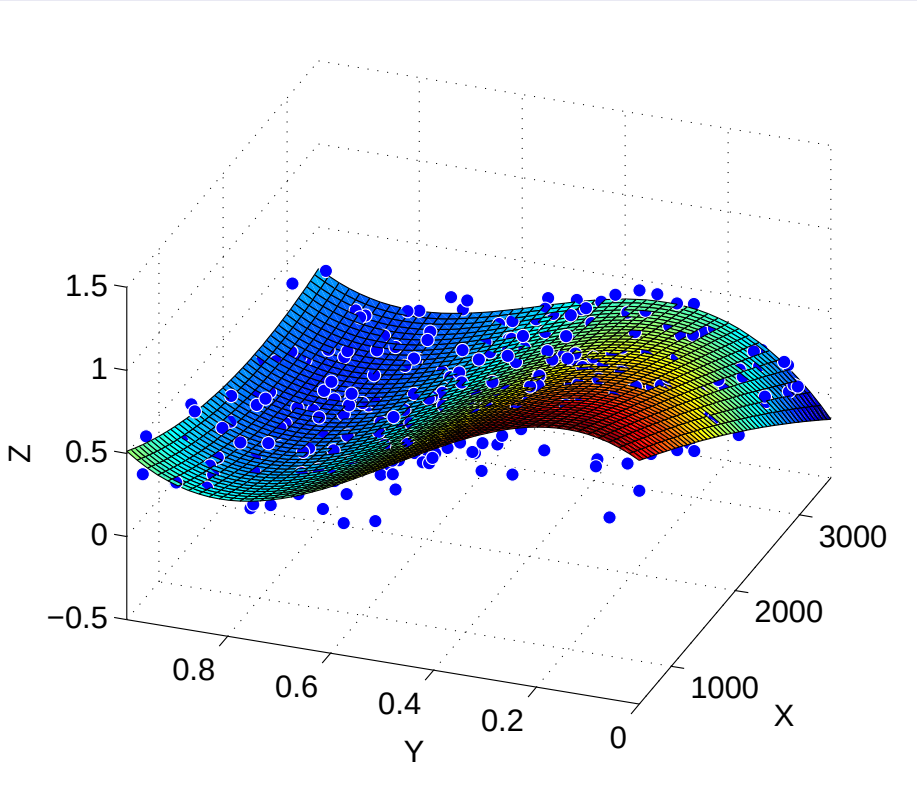
Regresión no lineal: ajuste exponencial



$$f(x) = a * exp(b * x)$$

Introducción

Regresión no lineal: ajuste polinomial



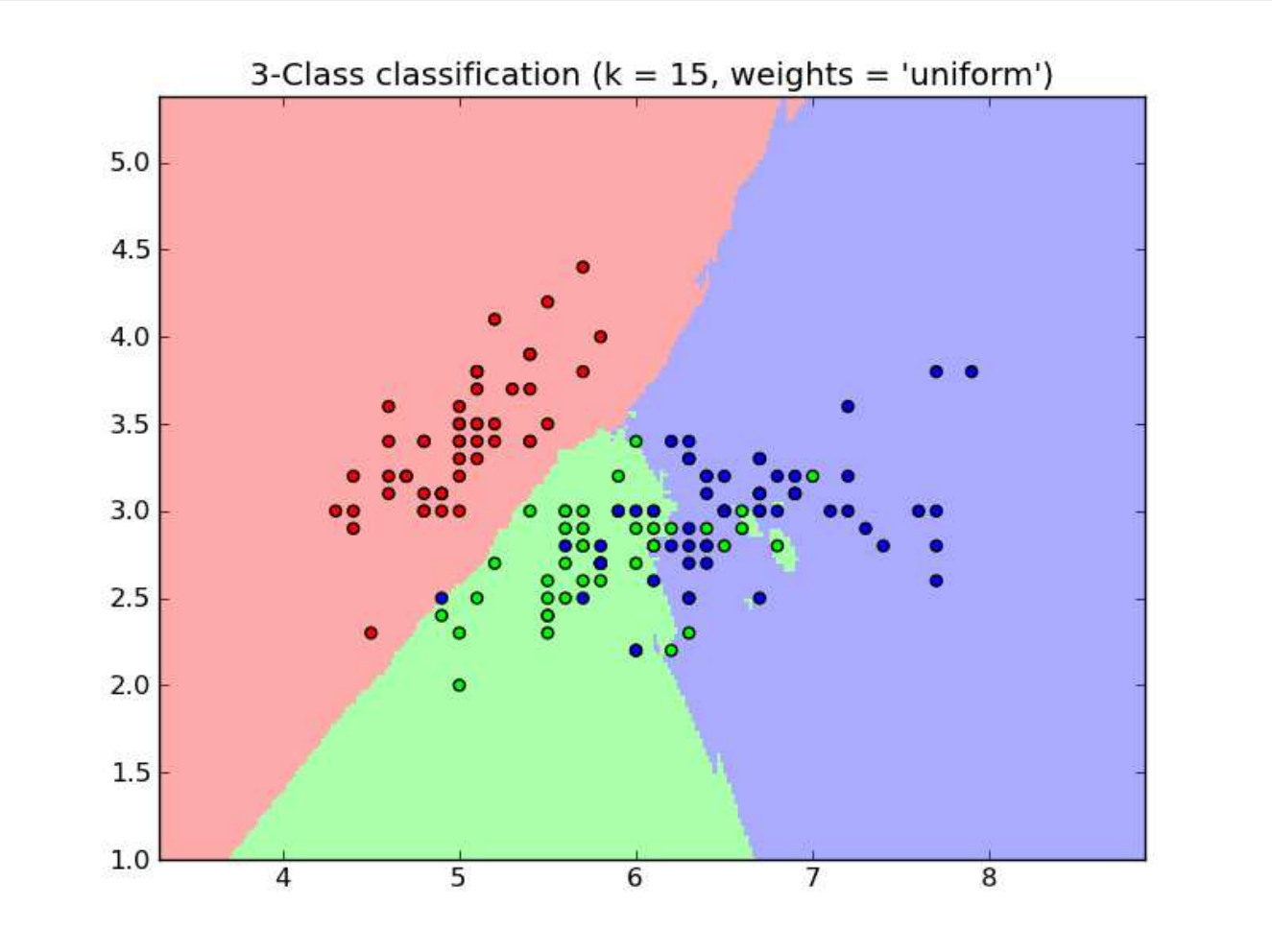
$$sf(x,y) = p_{00} + p_{10}x + p_{01}y + p_{20}x^2 + p_{11}xy + p_{02}y^2 + p_{21}x^2y + p_{12}xy^2 + p_{03}y^3$$

Índice

Algoritmos de clasificación

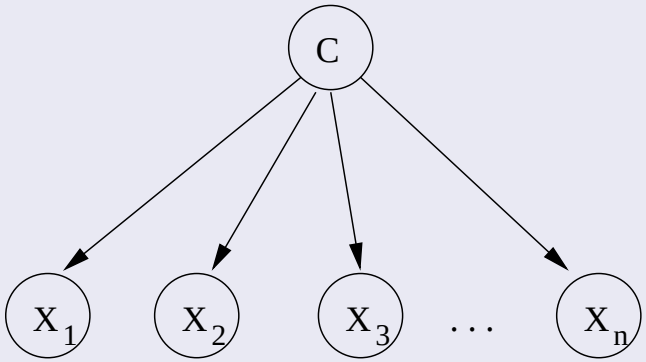
- Métodos basados en vecindad (k-NN)
- Árboles de decición y sistemas de reglas
- Métodos bayesianos
- Redes neuronales
- Métodos paramétricos
- etc

Algoritmos de clasificación

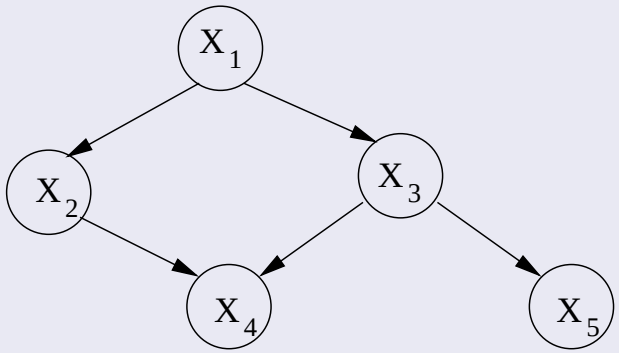


Clasificación basada en vecindad

Algoritmos de clasificación

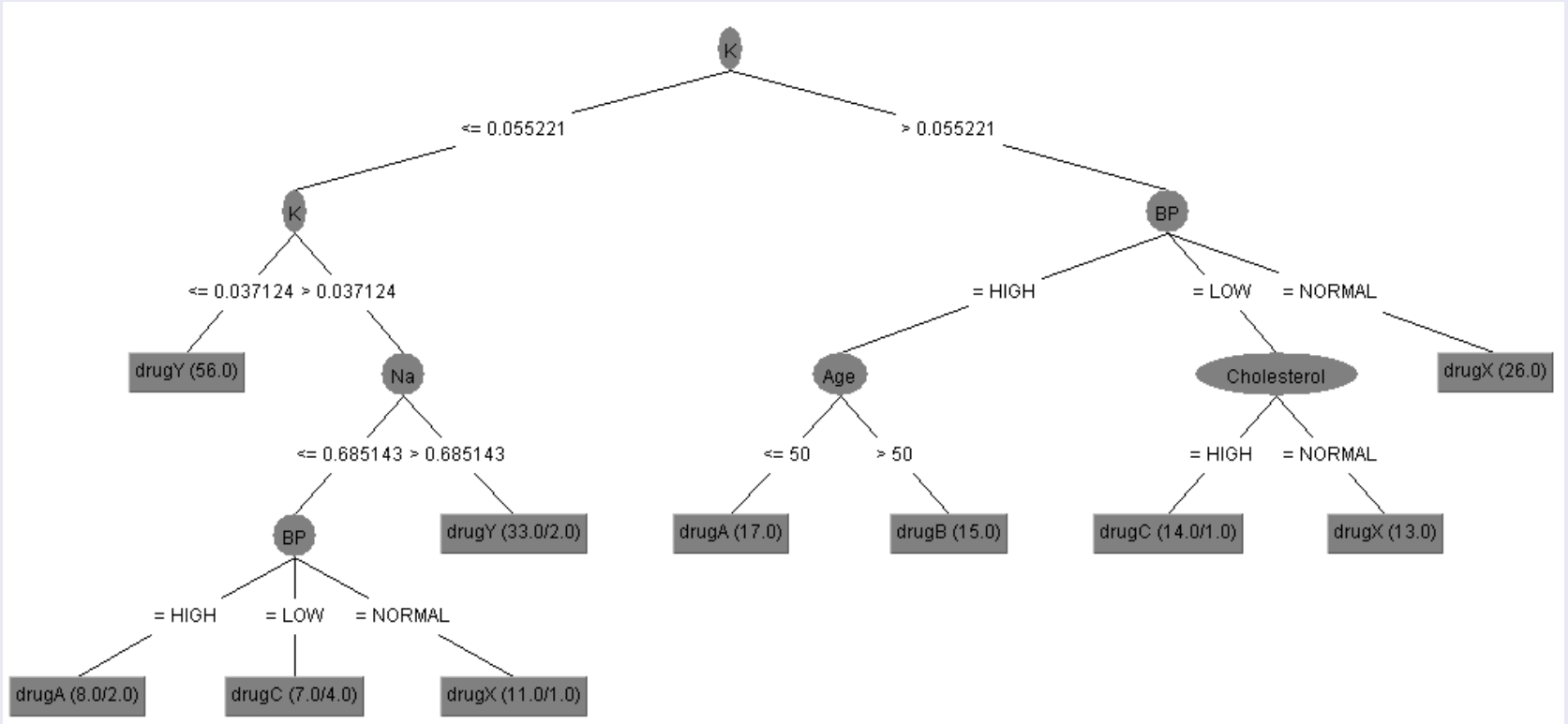


Naive bayes



Red bayesiana simple

Algoritmos de clasificación



Árbol de clasificación

Índice

Técnicas de validación

Validación de un modelo de clasificación

- Métricas para evaluar el rendimiento. ¿Cómo evaluar el rendimiento del modelo?
- Métodos para evaluar el rendimiento. ¿Cómo obtener una estimación fiable?
- Métodos para comparar modelos

Métricas.

Matriz de confusión

		True class: p	True class: n
Predicted class	p	True Positive	False Positive
	n	False Negative	True Negative
		P	N

Métricas de rendimiente más comunes:

- TP rate (sensitivity): $TPR = \frac{TP}{P}$
- FP rate: $FPR = \frac{FP}{N}$
- Precision$= \frac{TP}{TP+FP}$
- Accuracy$= \frac{TP+TN}{P+N}$

Métricas.

Matriz de confusión: Ejemplo 0

a	b	← classified as
180	21	a=no-recurrence-events
58	27	b=recurrence-events

Base de datos:	breast-cancer
Clasificador:	1-NN
Modo de evaluación:	10-fold cross-validation
Accuracy:	72.377
TP rate:	0.896
Precison:	0.756

Métricas.

Matriz de confusión: Ejemplo 1

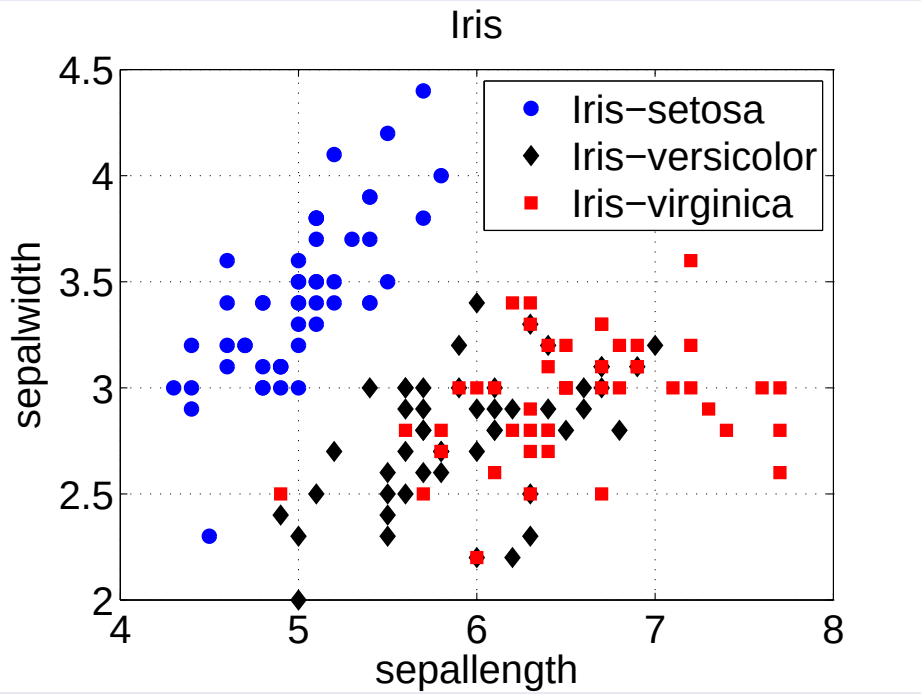
a	b	c	← classified as
50	0	0	a=Iris-setosa
0	49	1	b=Iris-versicolor
0	5	45	c=Iris-virginica

Base de datos: Iris
Clasificador: OneR
Modo de evaluación: sobre el conjunto de entrenamiento
Modelo de clasificación: petalwidth

$< 0{,}08 \rightarrow$ Iris-setosa
$< 1{,}75 \rightarrow$ Iris-versicolor
$>= 1{,}75 \rightarrow$ Iris-virginica

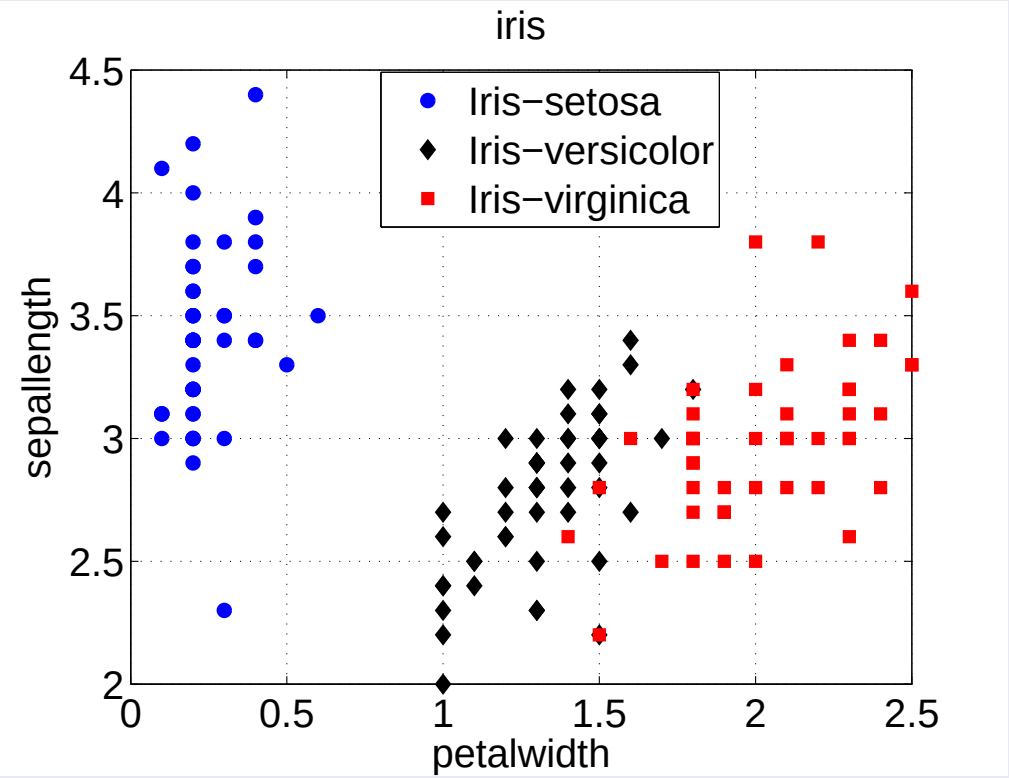
Métricas.

Visualización



Métricas.

Visualización



Métricas.

Matriz de confusión: Ejemplo 2

a	b	c	d	e	← classified as
91	0	0	0	0	a=drugY
16	0	0	0	0	b=drugC
54	0	0	0	0	c=drugX
23	0	0	0	0	d=drugA
16	0	0	0	0	e=drugB

Base de datos: Drugn
Clasificador: ZeroR
Modo de evaluación: sobre el conjunto de entrenamiento
Modelo de clasificación: ZeroR predice la clase mayoritaria, drugY

Métricas.

Limitación de la precisión

a	b	← classified as
0	32	a=die
0	123	b=live

Base de datos: hepatitis
Clasificador: ZeroR
Modo de evaluación: sobre el conjunto de entrenamiento
Modelo de clasificación: ZeroR predice la clase mayoritaria, live
Precisión: 79.3548

- Este modelo tiene una precisión de casi el 80 por ciento, sin embargo, es un modelo que no predice ninguna caso como perteneciente a la clase ”die”

Métricas.

Matriz de coste

COST MATRIX
when there are only two classes

		True class **p**	True class **n**
Predicted class	**p**	C(1,1)	C(1,0)
	n	C(0,1)	C(0,0)

- $C(i/j)$ representa el coste asociado al clasificar un caso de la clase i como perteneciente a la clase j.

Técnicas de validación

- Método Holdout: reserva $2/3$ del conjunto de datos para el entrenamiento y $1/3$ para el testeo
- Submuestreo aleatorio: es el método H repetidas veces
- Validación cruzada (Cross validation): particiona el conjunto de datos en k subconjuntos (folds), reserva (k-1) folds para entrenar y el fold restante para el testeo. El proceso se repite k veces cada vez con un fold de testeo distinto.
- Bootstrapping: muestreo aleatorio con reemplazo

Técnicas de validación

Comparación de modelos.

- Análisis ROC (Receiver Operating Characteristic)
  - Aprender un conjunto de clasificadores y seleccionar el que mejor se comporte para unas cicunstancias o contextos de coste determinados a posteriori.
  - Seleccionar el subconjunto de clasificadores que tienen un comportamiento óptimo en general.
- Test de hípotesis (de significatividad)