

Minería de Datos

IIC2433

Clasificación automática y
evaluación de clasificadores

Vicente Domínguez

Basado en diapositivas del prof. Denis Parra

¿Qué veremos esta clase?

- Clasificación
- Evaluación de clasificadores

Aprendizaje de máquina

(Machine Learning)

Darle a los computadores la habilidad de realizar una actividad, sin programarlos explícitamente.

*La minería de datos y el **aprendizaje de máquina** se traslapan y no tienen límites claros

Aprendizaje de máquina

Tipos de tareas

- Aprendizaje supervisado
 - Clasificación
 - Regresión
- Aprendizaje no supervisado
 - Clustering
 - Aprendizaje por refuerzo
 - etc

Aprendizaje supervisado

Clasificación

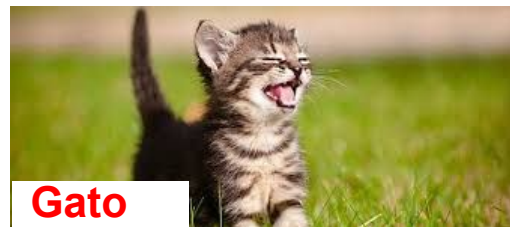
Tarea para el computador:

Decir si en una foto hay un perro o un gato

Aprendizaje supervisado

Clasificación

Conjunto de entrenamiento **etiquetado**



Aprendizaje supervisado

Clasificación

¿Qué es eso?



Perro

Aprendizaje no supervisado

Clustering

Tarea para el computador:

Identificar grupos de elementos similares

Aprendizaje no supervisado

Clustering

Conjunto de datos **no etiquetados**



Aprendizaje de máquina

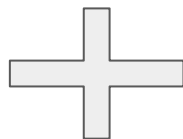
Tipos de tareas

- Aprendizaje supervisado **(necesita etiquetas)**
 - Clasificación
 - Regresión
- Aprendizaje no supervisado **(no necesita etiquetas)**
 - Clustering
 - Aprendizaje por refuerzo
 - etc

Clasificación

1. *Etapas de entrenamiento (Model.fit())*

Conjunto de
entrenamiento



Algoritmo de
clasificación



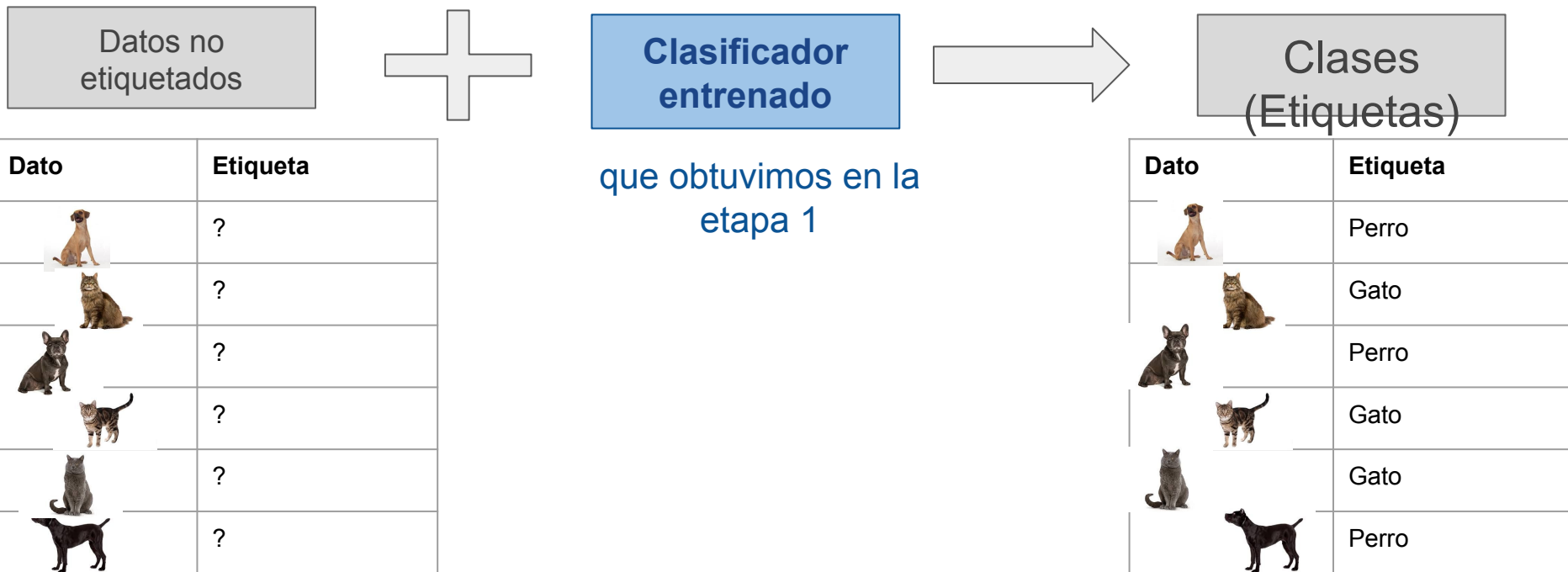
**Clasificador
entrenado**

Dato	Etiqueta
	Gato
	Perro
	Perro
	Gato
	Gato
	Perro

- Árboles de decisión
- Naïve Bayes
- KNN
- SVM
- etc...

Clasificación

2. Etapa de clasificación (*Model.predict()*)



Clasificación

Evaluación




Consideremos este resultado de clasificación:

Dato	Etiqueta
	Perro
	Gato
	Gato

Clasificación

Evaluación

Consideremos este resultado de clasificación:

Dato	Etiqueta
	Perro
	Gato
	Gato Perro

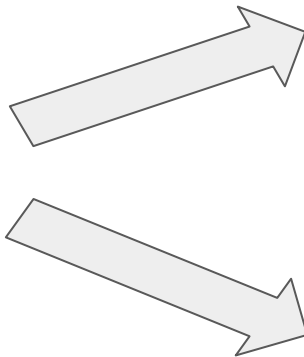
**Lamentablemente, los clasificadores a veces se equivocan...
por esto debemos evaluar su desempeño**

Clasificación

Evaluación: Hold out

1. Dividimos el conjunto etiquetado en dos:

Datos etiquetados	
Dato	Etiqueta
	Gato
	Perro
	Perro
	Gato
	Gato
	Perro



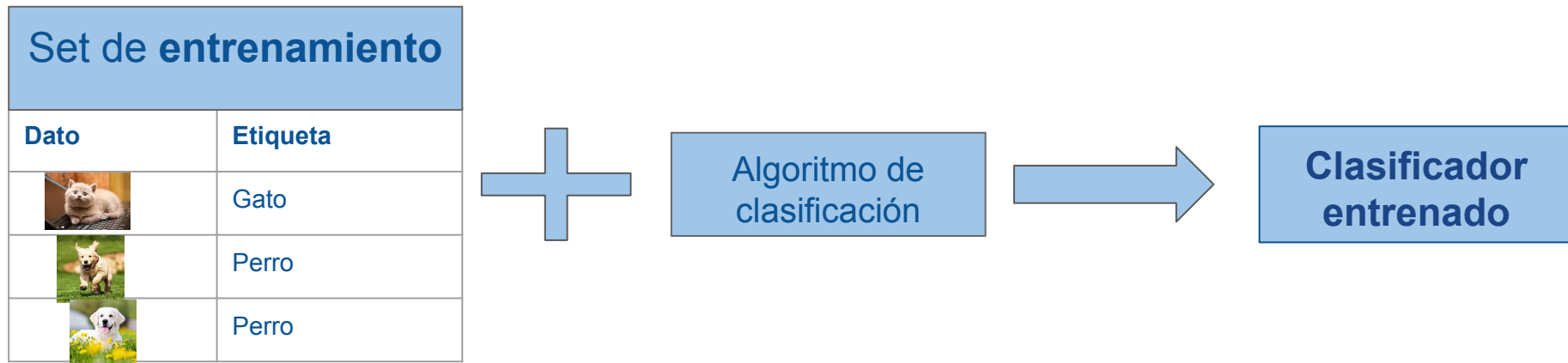
Set de entrenamiento	
Dato	Etiqueta
	Gato
	Perro
	Perro

Set de test	
Dato	Etiqueta
	Gato
	Gato
	Perro

Clasificación

Evaluación: Hold out

2. Entrenamos un clasificador usando el set de entrenamiento



Clasificación

Evaluación: Hold out




3. Usamos el **clasificador entrenado** para clasificar los datos del **set de test**



Clasificación



Evaluación: Hold out

4. Comparamos las etiquetas reales con las etiquetas predichas por el clasificador. A partir de esto calculamos **métricas**.

Dato	Etiqueta real	Etiqueta predicha
	Gato	Gato
	Gato	Perro
	Perro	Perro

Clasificación




Evaluación: Matriz de confusión

Dato	Etiqueta real	Etiqueta predicha
	Gato	Gato
	Gato	Perro
	Perro	Perro

		Predicho	
		Perro	Gato
Real	Perro	1	0
	Gato	1	1

Clasificación

Evaluación: Matriz de confusión

Dato	Etiqueta real	Etiqueta predicha
	Gato	Gato
	Gato	Perro
	Perro	Perro

		Predicho	
		Perro	Gato
Real	Perro	1	0
	Gato	1	1

Clasificación

Evaluación: Matriz de confusión

		Predicho				
		Perro	Gato	Mono	Tortuga	Elefante
Real	Perro	31	5	0	0	0
	Gato	2	42	0	1	0
	Mono	0	0	45	0	0
	Tortuga	0	0	1	23	0
	Elefante	0	0	0	0	15

Clasificación

Evaluación: Matriz de confusión

En algunos casos tenemos una clase que llamamos **positiva** y otra **negativa**

Dato	Etiqueta real	Etiqueta predicha
Examen 1	+	+
Examen 2	+	+
Examen 3	-	+
Examen 4	-	-
Examen 5	-	+

		Predicho	
		Positivo	Negativo
Real	Positivo	2	0
	Negativo	2	1

Clasificación

Evaluación: Matriz de confusión

		Predicho	
		Positivo	Negativo
Real	Positivo	Verdadero positivo	Falso negativo
	Negativo	Falsos positivos	Verdadero negativo

Clasificación

Evaluación: Accuracy (Exactitud)

Porcentaje de elementos datos correctamente clasificados

$$\text{accuracy} = \frac{n^{\circ} \text{ datos bien clasificados}}{n^{\circ} \text{ datos total}}$$

		Predicho				
		Perro	Gato	Mono	Tortuga	Elefante
Real	Perro	31	5	0	0	0
	Gato	2	42	0	1	0
	Mono	0	0	45	0	0
	Tortuga	0	0	1	23	0
	Elefante	0	0	0	0	15

$$\text{accuracy} = \frac{\text{diagonal}}{\text{total}}$$

Clasificación

Evaluación: Accuracy (Exactitud)

Porcentaje de elementos datos correctamente clasificados

$$accuracy = \frac{n^{\circ} \text{ datos bien clasificados}}{n^{\circ} \text{ datos total}}$$

		Predicho	
		Positivo	Negativo
Real	Positivo	Verdadero positivo	Falso negativo
	Negativo	Falsos positivos	Verdadero negativo

$$accuracy = \frac{VP + VN}{total}$$

Clasificación

Evaluación: Accuracy (Exactitud)

¿Qué ocurre con clases no balanceadas?

Clasificación

Evaluación: Precision (Precisión)

Porcentaje de elementos clasificados como una clase que realmente corresponden a la clase

$$\text{Precision (clase } X) = \frac{n^{\circ} \text{ datos } \textbf{bien} \text{ clasificados de la clase } X}{n^{\circ} \text{ total datos clasificados en la clase } X}$$

		Predicho				
		Perro	Gato	Mono	Tortuga	Elefante
Real	Perro	31	5	0	0	0
	Gato	2	42	0	1	0
	Mono	0	0	45	0	0
	Tortuga	0	0	1	23	0
	Elefante	0	0	0	0	15

precision (perro) =

$$\frac{31}{31 + 2 + 0 + 0}$$

Clasificación

Evaluación: Precision (Precisión)

Porcentaje de elementos clasificados como una clase que realmente corresponden a la clase

$$\text{Precision (clase positiva)} = \frac{n^{\circ} \text{ datos bien clasificados como positivos}}{n^{\circ} \text{ total datos clasificados como positivos}}$$

		Predicho	
		Positivo	Negativo
Real	Positivo	Verdadero positivo	Falso negativo
	Negativo	Falsos positivos	Verdadero negativo

$$\text{precision (clase pos)} = \frac{VP}{VP + FP}$$

Clasificación

Evaluación: Recall (Exhaustividad)

Porcentaje de elementos de una clase (real) que fueron bien clasificados

Recall (clase X) = $\frac{n^{\circ} \text{ datos bien clasificados de la clase X}}{n^{\circ} \text{ total datos que pertenecen a la clase X}}$

		Predicho				
		Perro	Gato	Mono	Tortuga	Elefante
Real	Perro	31	5	0	0	0
	Gato	2	42	0	1	0
	Mono	0	0	45	0	0
	Tortuga	0	0	1	23	0
	Elefante	0	0	0	0	15

recall (perro) =

$$\frac{31}{31 + 5 + 0 + 0}$$

Clasificación

Evaluación: Recall (Exhaustividad)

Porcentaje de elementos de una clase (real) que fueron bien clasificados

$$\text{Recall (clase pos)} = \frac{n^{\circ} \text{ datos } \textbf{bien} \text{ clasificados como positivos}}{n^{\circ} \text{ datos realmente positivos}}$$

		Predicho	
		Positivo	Negativo
Real	Positivo	Verdadero positivo	Falso negativo
	Negativo	Falsos positivos	Verdadero negativo

$$\text{recall (clase pos)} =$$

$$\frac{VP}{VP + FN}$$

Clasificación

Otras métricas

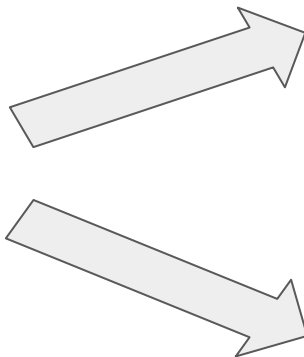
- **F1-score** $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
- **True positive rate** $\frac{vp}{vp + fn}$
- **True negative rate** $\frac{vn}{fp + vn}$
- **False positive rate** $\frac{fp}{fp + vn}$
- **False negative rate** $\frac{fn}{vp + fn}$

Clasificación

Recordemos evaluación “Hold out”

1. Dividimos el conjunto etiquetado en dos:

Datos etiquetados	
Dato	Etiqueta
	Gato
	Perro
	Perro
	Gato
	Gato
	Perro



Set de entrenamiento	
Dato	Etiqueta
	Gato
	Perro
	Perro

Set de test	
Dato	Etiqueta
	Gato
	Gato
	Perro

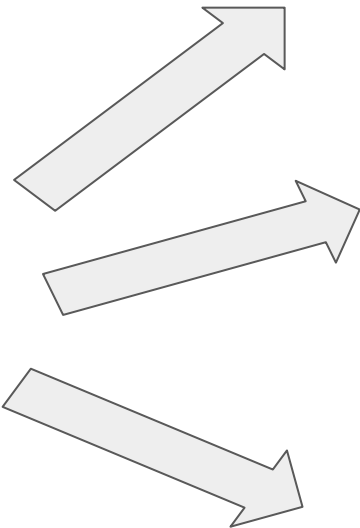
Validación cruzada

(*K-fold Cross Validation*)

Dividimos el conjunto en **K** conjuntos de igual tamaño

1. Supongamos **K = 3**

Datos etiquetados	
Dato	Etiqueta
	Gato
	Perro
	Perro
	Gato
	Gato
	Perro



Set 1	
Dato	Etiqueta
	Gato
	Perro

Set 2	
Dato	Etiqueta
	Perro
	Gato

Set 3	
Dato	Etiqueta
	Gato
	Perro

Validación cruzada





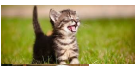

(K-fold Cross Validation)

2. Entrenamos con **set 2** y **set 3**, testeamos con **set 1**

Set 1 (test)	
Dato	Etiqueta
	Gato
	Perro

Set 2 (entrenamiento)	
Dato	Etiqueta
	Perro
	Gato

Set 3 (entrenamiento)	
Dato	Etiqueta
	Gato
	Perro

Dataset		
Dato	Real	Predicha
	Gato	Gato
	Perro	Perro
	Perro	
	Gato	
	Gato	
	Perro	

Validación cruzada

(K-fold Cross Validation)

2. Entrenamos con **set 2** y **set 3**, testeamos con **set 1**



Set 1 (entrenamiento)

Dato	Etiqueta
	Gato
	Perro



Set 2 (test)

Dato	Etiqueta
	Perro
	Gato

Set 3 (entrenamiento)

Dato	Etiqueta
	Gato
	Perro

Dataset

Dato	Real	Predicha
	Gato	Gato
	Perro	Perro
	Perro	Perro
	Gato	Gato
	Gato	
	Perro	

Validación cruzada

(K-fold Cross Validation)

2. Entrenamos con **set 2** y **set 3**, testeamos con **set 1**



Set 1 (entrenamiento)

Dato	Etiqueta
	Gato
	Perro







Set 2 (entrenamiento)

Dato	Etiqueta
	Perro
	Gato

Set 3 (test)

Dato	Etiqueta
	Gato
	Perro

Dataset

Dato	Real	Predicha
	Gato	Gato
	Perro	Perro
	Perro	Perro
	Gato	Gato
	Gato	Gato
	Perro	Perro