

Iris Dataset

Iris Dataset

- Base de datos de la flor de iris
- Consta de un número de registros para los cuales se midió largo y ancho de pétalos y sépalos, y se caracterizó (manualmente) la especie
- Tres especies posibles:
 - Setosa, Virginica, Versicolor
- Es un dataset clásico (simple) en machine learning



Clasificación

- La clasificación es una de la clase de problemas más importantes en machine learning:
- Dadas una serie de características K de individuos o items, se quiere predecir el valor de una característica adicional k'
 - k' debe ser *discreta*



Métricas de evaluación

- Consideremos un problema de clasificación P , que intenta predecir la clase c_i para ítems arbitrarios de una población, y sea Prog un clasificador.
- Los clasificadores suelen no ser perfectos, es decir, predicen clases incorrectas para algunos individuos de la población
 - $\text{Prog}(x)$ no coincide con la clase correcta que corresponde a x
- La precisión (accuracy) de un clasificador se suele medir, para cada clase c_k , en relación a la tasa de
 - True positives: casos en los cuales la clase c_k predicha coincide con la real
 - False positives: casos en los cuales la clase predicha c_k no coincide con la clase del individuo correspondiente
 - True negatives: casos en los cuales la clase predicha no es c_k , y efectivamente los individuos correspondientes no son de clase c_k
 - False negatives: casos en los cuales la clase predicha por el clasificador no es c_k , pero los individuos correspondientes corresponden a la clase c_k

		Estimate		
		$c_0 \dots c_{k-1}$	c_k	$c_{k+1} \dots c_n$
annotated ground truth	$c_{k+1} \dots c_n$	TN	FP	TN
	c_k	FN	TP	FN
	$c_0 \dots c_{k-1}$	TN	FP	TN

TN

 true negative

TP

 true positive

FN

 false negative

FP

 false positive

Entrenamiento vs Evaluación

- Cuando los clasificadores son basados en aprendizaje automático, éstos se construyen a partir de datos.
- Los datos provienen de un dataset (anotado)
- Para evaluar la precisión (accuracy) de un modelo, se suele separar los datos en conjuntos disjuntos de datos de entrenamiento, y datos de evaluación
- Una tasa usual es usar 70% de los datos para entrenamiento, y 30% para evaluación

