Tópicos Selectos en Aprendizaje Maquinal

Guía de Trabajos Prácticos N°2

Clasificación y Regresión con Datos Reales

29 de octubre de 2018

1. Objetivos

- Introducir conceptos básicos de aprendizaje automático.
- Conocer los principales aspectos de las técnicas de validación utilizadas en el reconocimiento de patrones.
- Estudiar y validar el comportamiento de los algoritmos estudiados sobre conjuntos sencillos de datos reales.
- Estimar y comparar el desempeño de los diferentes algoritmos en estos problemas.
- Aprender a utilizar software específico para estas tareas.

2. Ejercicios

Ejercicio 1: *Clasificación de enfermedades (Diabetes)*:

Para este ejemplo se utilizará la base de datos "Diabetes". Esta fue creada en el *National Institute of Diabetes and Digestive and Kidney Diseases* (EE.UU.)[1]. Contiene información sobre mujeres mayores de 21 años descendientes de los indios Pima de Arizona (EE.UU.) Se las ha estudiado especialmente por ser una de las poblaciones con más alta prevalencia de diabetes en el mundo. Hay 768 casos con información en 8 atributos de entrada que se consideran pertinentes para la clasificación. El diagnóstico consiste en una variable binaria que indica si el paciente muestra signos de diabetes mellitus. De estos 768 casos, cerca de la mitad tienen datos faltantes en los atributos presión diastólica, grosor de la piel y concentración de insulina en la sangre. En el Cuadro 1 se muestran los atributos y algunas estadísiticas de la base de datos. Una idea de cuan mezcladas están las clases puede obtenerse a partir de la proyección de los datos en 2 dimensiones (Figura 1).

[1] UCI Repository of Machine Learning Databases and Domain Theories (`ics.uci.edu:pub/machine-learning-databases`).

Utilizando el software `WEKA`[2] entrene con estos datos a los siguientes clasificadores y compare su desempeño (con al menos 2 configuraciones distintas cada uno):

1. Naive Bayes.
2. Perceptrón multicapa (MLP).
3. Red de funciones de base radial (RBF).
4. Árbol de decisión C4.5.
5. Máquina de vectores de soporte (SVM).
6. K-means (comparando los centroides con las etiquetas).
7. Método de selección de características que utilice SVM.
8. Autocodificador basado en MLP seguido de un SVM.
9. MLP con 2 capas ocultas entrenado mediante aprendizaje profundo[3].

Para la validación utilice los métodos de: una sola partición de entrenamiento/prueba (sólo para 7, 8 y 9) y *k-fold cross validation* (para el resto). En cada caso utilice 2 configuraciones diferentes de cada método. Estime los errores de clasificación promedio y la desviación estándar para cada una de las variantes ensayadas.

Ejercicio 2: *Clasificación de plantas (Iris)*:

Iris es el género una planta herbácea con flores que se utilizan en decoración. Dentro de este género existen muy diversas especies entre las que se han estudiado la *Iris setosa*, la *Iris versicolor* y la *Iris virginica* (ver Figura 2).

Estas tres especies pueden distinguirse según las dimensiones de sus pétalos y sépalos. Un grupo de investigadores ha recopilado la información correspondiente a las longitudes y anchos de los pétalos y sépalos de 50 plantas de cada especie. En la base de datos `iris` se encuentran estas mediciones (en cm) junto con un valor numérico que indica la especie reconocida por los investigadores (0=setosa; 1=versicolor; 2=virginica). Para la clasificación de una gran cantidad de estas plantas se desea diseñar un sistema que aprenda de estos 150 patrones para luego realizar la tarea de forma automática:

1. Proceda con estos datos de igual manera que en el punto anterior y seleccione el clasificador con el mejor desempeño.
2. Luego proyecte los datos sobre las dos primeras componentes principales (ver Figura 3).
3. Repita los experimentos con estos datos transformados.
4. Comente brevemente las diferencias encontradas con los experimentos anteriores.

[2] `http://www.cs.waikato.ac.nz/ml/weka/index.html`
[3] `https://deeplearning.cms.waikato.ac.nz/`

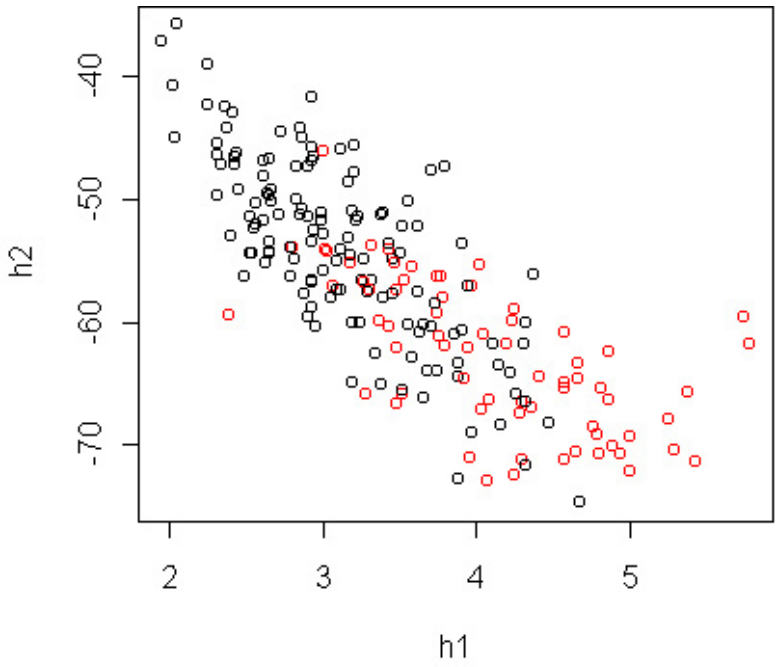

Figura 1: Proyección en $\mathbb{R}^2$ (mediante PCA) de la distribución de clases para la base de datos `diabetes`.

Cuadro 1: Estadísticas de la base de datos `diabetes`.

Atributo	**Mínimo**	**Máximo**	**Valor Medio**	**Desv. Stand.**	**Descripción**
Num_Emb	0	17	3,38	3,32	Embarazos.
Gluc_Plas	68	198	127,62	32,69	Glucosa en plasma.
Pres_Diast	24	110	71,25	13,11	Presión diastólica.
Piel_Triceps	7	63	30,28	10,55	Grosor de la piel.
Insulina	14	846	166,47	121,49	Insulina en sangre.
Masa_Corp	18,2	67,1	33,75	7,23	Masa corporal.
Pedigree	0,085	2,42	0,56	0,37	Pedigree.
Edad	21	81	31,99	10,59	Edad.

Figura 2: Muestra de la especie *Iris virginica*.

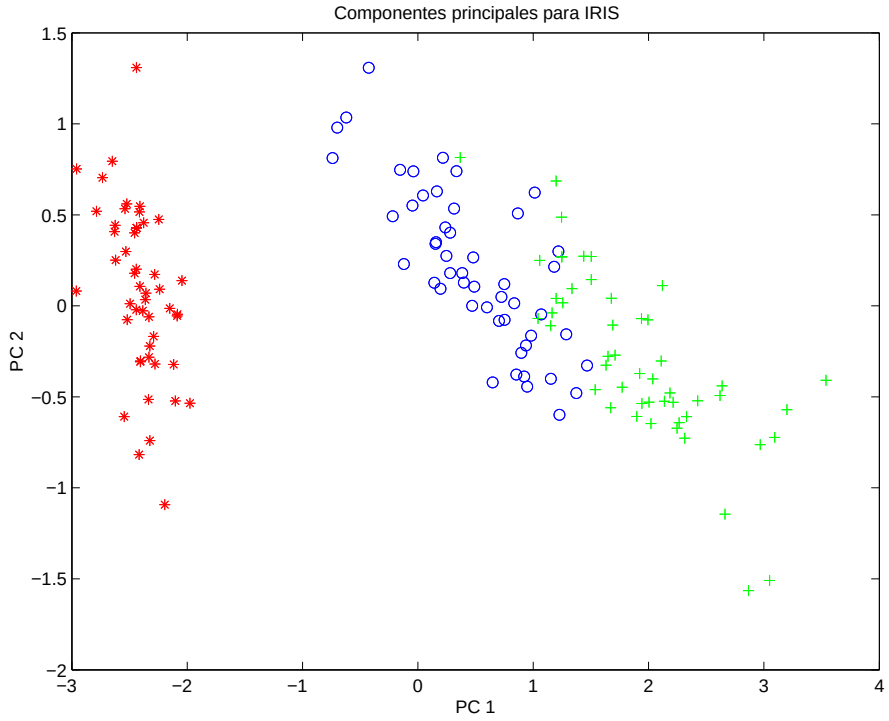

Figura 3: Proyección en $\mathbb{R}^2$ (mediante PCA) de la distribución de clases para la base de datos `iris`.

Referencias

[1] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Amsterdam, 3 edition, 2011.