Tópicos Selectos en Aprendizaje Maquinal

Guía de Trabajos Prácticos N°2

Problemas de Clasificación con Datos Reales

17 de octubre de 2020

1. Objetivos

- Introducir conceptos básicos de aprendizaje automático.
- Conocer los principales aspectos de las técnicas de evaluación de desempeño y 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.
- Apreciar las ventajas distintos métodos de selección de características.
- Aprender a utilizar herramientas y librerías para implementar soluciones a problemas de aprendizaje maquinal.

2. Ejercicios

Ejercicio 1: *Diagnóstico de 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 estadísticas de la base de datos. Una idea de cuan mezcladas están

[1] Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database).

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.

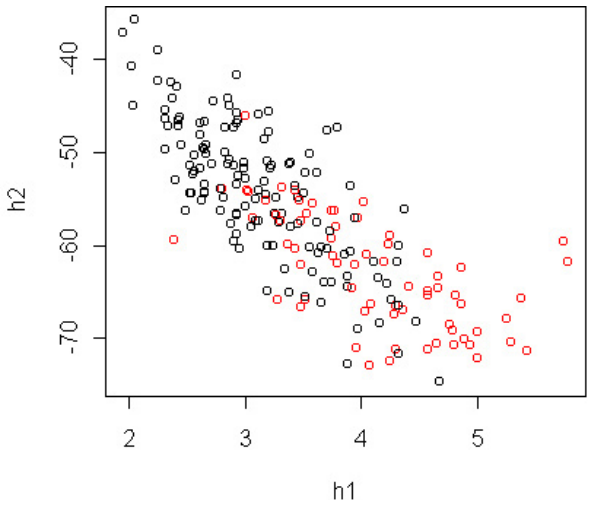

Figura 1: Proyección en $\mathbb{R}^2$ (mediante PCA) de la distribución de clases para la base de datos `diabetes`.

las clases puede obtenerse a partir de la proyección de los datos en 2 dimensiones (Figura 1).

a) Entrene con estos datos a los siguientes clasificadores y compare su desempeño (con al menos 2 configuraciones distintas cada uno)[2]:

1. Naive Bayes.
2. Perceptrón multicapa (MLP).
3. Red de funciones de base radial (RBF).
4. Árbol de decisión.
5. Random forest.
6. Máquina de vectores de soporte (SVM).
7. Un método de selección de características y SVM.
8. Autocodificador basado en MLP seguido de un SVM.

[2]Para esto se recomienda utilizar la librería `scikit-learn` en Python (`https://scikit-learn.org/`).

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

9. K-means (asignando etiquetas a los clústers).
10. Autocodificadores apilados y un perceptrón.

b) Describa detalladamente la estructura, los parámetros y la configuración empleada en cada caso. Luego analice y compare los resultados obtenidos con las diferentes configuraciones de cada clasificador.

c) Para la validación utilice los métodos de: particionamiento simple y *k-fold cross validation* (con 2 diferentes configuraciones cada uno). En el segundo caso estime la media y la desviación estándar de las medidas de *accuracy*, *unweighted accuracy* (también conocida como *balanced accuracy* o *average recall*) y *recall*. Realice un BoxPlot para cada una de las métricas y analice los resultados obtenidos, comparando los métodos de validación.

d) Grafique la curva ROC y calcule el área bajo la curva (AUC) para el MLP y el SVM, variando umbrales sobre la estimación de probabilidades (salida lineal) y la función de decisión, respectivamente. Para esto utilice el método *k-fold* y construya la curva ROC promedio.[3] Compare ambos clasificadores en función de estas métricas.

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* (Figura 2).

[3]Para esto se recomienda emplear las funciones de `scikit-learn` (https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html).

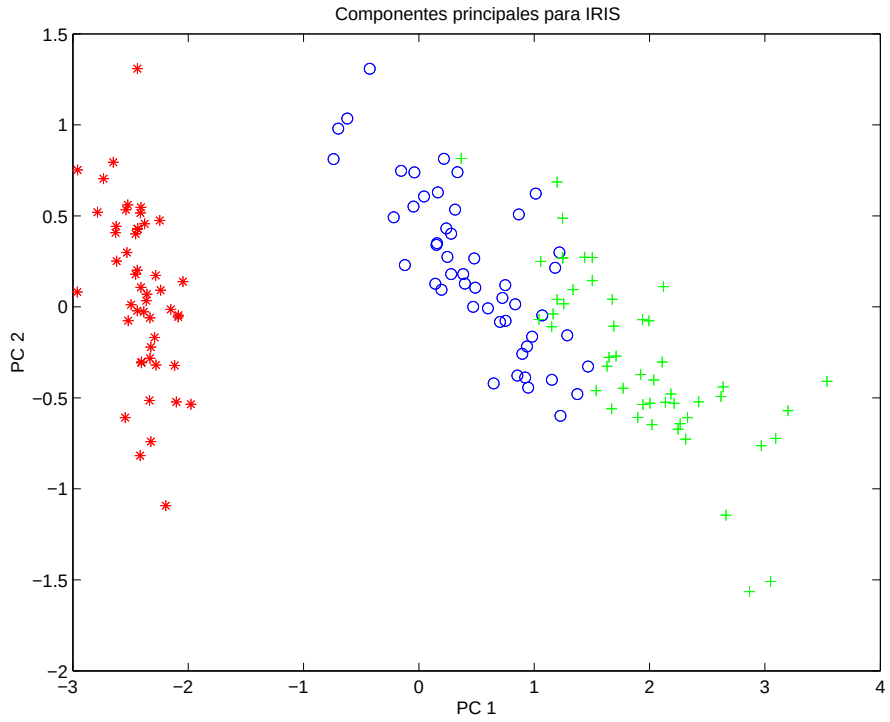

Figura 3: Proyección en $\mathbb{R}^2$ (mediante PCA) de la distribución de clases para la base de datos `iris`.

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 ejercicio anterior, comparando el desempeño de los clasificadores con el método *k-fold*.
2. Estime la significancia estadística de la diferencia de desempeño entre pares de clasificadores y determine si en algún caso la mejora es significativa.
3. Seleccione el clasificador con mejor desempeño y repita el proceso de validación, empleando la técnica *k-fold* estratificada y el método de particionamiento simple repetido k veces[4]. Compare con los resultados anteriores, teniendo en cuenta la media y la varianza del *accuracy*.
4. Proyecte los datos sobre las dos primeras componentes principales y compare el desempeño de los clasificadores con estos datos transformados.

[4]Para esto se recomienda emplear las funciones de `scikit-learn` (https://scikit-learn.org/stable/modules/classes.html?highlight=model_selection#module-sklearn.model_selection).

5. Represente los datos transformados gráficamente, como en la Figura 3, y compare esta proyección con la obtenida mediante el autocodificador.

Ejercicio 3: *Clasificación de tipos de leucemia*:

En un trabajo[5] se realizó un análisis de los datos de expresión génica obtenidos a partir de micromatrices de ADN para la clasificación de tipos de cáncer. Se construyó un conjunto de datos con 7129 mediciones de expresión génica en las clases ALL (leucemia linfocítica aguda) y AML (leucemia mielógena aguda). El problema es distinguir entre estas dos variantes de leucemia (ALL y AML).

Los datos se dividen originalmente en dos subconjuntos: un conjunto de entrenamiento y un conjunto de prueba independiente. El conjunto de entrenamiento consta de 38 muestras (27 ALL y 11 AML) y el conjunto de prueba tiene 34 muestras (20 ALL y 14 AML). Las muestras fueron preparadas en diferentes condiciones experimentales e incluyen 24 muestras de médula ósea y 10 muestras de sangre. Todas las muestras consisten en un total de 7129 características, correspondientes a valores de expresión génica extraídos de la imagen de microarreglo.

Considerando los conjuntos de entrenamiento y prueba de este dataset[6]:

1. Evalúe el desempeño de los clasificadores considerando el conjunto de características completo.
2. Seleccione subconjuntos de características mediante:
   a) Un método de ranking.
   b) El método de eliminación recursiva de características (RFE).
   c) El método Relief.
3. Vuelva a evaluar el desempeño de los clasificadores considerando los subconjuntos de características obtenidos en cada caso. Compare el desempeño con los resultados obtenidos en el punto 1. En la comparación tenga en cuenta las medidas de *accuracy* y *unweighted accuracy*, además del tiempo de procesamiento de los clasificadores y de los métodos de selección de características.

[5] Golub, T.R. et al, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, American Association for the Advancement of Science, Vol. 286, Num. 5439, 1999.

[6] http://tsam-fich.wikidot.com/local--files/apuntes/leukemia_train.csv
http://tsam-fich.wikidot.com/local--files/apuntes/leukemia_test.csv