Reconocimiento de patrones: Práctica 1

Descripción de los datos

Para la realización de la tarea de clasificación se ha elegido el wine dataset (https://archive.ics.uci.edu/ml/datasets/wine). Este dataset cuenta con 178 muestras y 3 categorías. A la categoría 1 pertenecen 59, a la categoría 2 pertenecen 71 y a la categoría 3 pertenecen 48 muestras. Cada muestra del dataset contiene 13 categorías.

- Alcohol
- Ácido málico
- Ceniza
- Alcalinidad de ceniza
- Magnesio
- Fenoles totales
- Flavonoides
- Fenoles no flavonoides
- Glucósidos de una proantocianidina
- Intensidad del color
- Matiz
- OD280 / OD315 de vinos diluidos
- Prolina

Estos atributos suponen una buena representación de las propiedades y características de un vino. Dado que no tenemos conocimientos de enología hemos decido utilizar todos los atributos como características.

En la partición del dataset para elaboración de los grupos de entrenamiento, validación y test hemos decidido asignar un 20% de los datos para test. Dado que vamos a aplicar validación cruzada, para los datos de entrenamiento y validación hemos decidido partir los datos restantes en 5 bloques, 4 para entrenamiento y uno para validación.

Dado que los datos tienen rangos de valores muy diferentes, hemos decidido normalizarlos todos a la unidad. Esto es altamente recomendable al usar SVM ya que la clasificación depende de la magnitud de los datos

Descripción de los métodos de clasificación

Como métodos de clasificación hemos decidido utilizar SVM y arboles de decisión, que a su vez los entrenaremos uno contra todos y uno contra uno para ver qué método de entrenamiento nos genera una mayor precisión.

Selección de los hiperparámetros

El hiperparámetro del clasificador de SVM es el parámetro C para regular como de estricto es el margen permitiendo que haya muestras dentro de él. Además del hiperparámetro C, también es necesario definir el kernel que se va a emplear.

- En el SVM lineal no es requerido el ajuste de otro hiperparámetro
- En el SVM polinomial se encuentran gamma, r y d
- En el SVM sigmoide gamma y r
- En el SVM RBF gaussiano solo gamma

Gamma es el coeficiente del kernel, r es el término independiente por lo que normalmente se considera cero y d, que indica el orden del polinomio.

Los hiperparámetros de los árboles de decisión, son el Coeficiente de Gini (G), que tiene como objetivo definir qué porcentaje de muestras se permite que queden mal clasificadas en un nudo del árbol y la profundidad del árbol.

Rango de los hiperparámetros:

- C: [1000000, 10000, 1000, 100, 10, 1, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001]
- d: [8, 7, 6, 5, 4, 3, 2]
- gamma: [1/13, 1/12, 1/11, 1/10, 1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2]

- Gini: [0, 0.05555556, 0.11111111, 0.16666667, 0.22222222, 0.27777778, 0.33333333, 0.38888889, 0.44444444, 0.5]
- Profundidad: [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]

Iterando con estos hiperparámetros y con la tarea de clasificación de uno contra uno y uno contra todos se han obtenido los siguientes resultados:

Modelo	**Descripción**	**Precisión**
Modelo 1	SVM lineal One vs One, el mejor hiperparámetro C es 1	0.959666666667
Modelo 2	SVM lineal One vs All, el mejor hiperparámetro C es 1000000	0.967666666667
Modelo 3	SVM con kernel RBF gaussiano One vs One, el mejor hiperparámetro C es 10, el mejor hiperparámetro gamma es 0.125	0.983666666667
Modelo 4	SVM con kernel RBF gaussiano One vs All, el mejor hiperparámetro C es 1000000, el mejor hiperparámetro gamma es 0.09090909090909091	0.976
Modelo 5	SVM con kernel sigmoide One vs One, el mejor hiperparámetro C es 100, el mejor hiperparámetro gamma es 0.125	0.975666666667
Modelo 6	SVM con kernel Sigmoide One vs All, el mejor hiperparámetro C es 1000000, el mejor hiperparámetro gamma es 0.1	0.984
Modelo 7	SVM con kernel Polinomial One vs One, el mejor hiperparámetro C es 1000, el mejor hiperparámetro gamma es 0.125, el grado 8	0.992
Modelo 8	SVM con kernel Polinomial One vs All, el mejor hiperparámetro C es 1, el mejor hiperparámetro gamma es 0.3333333333333333, el grado 6	0.984
Modelo 9	Árbol de decisión con One vs One, con una profundidad de 10.0, un Gini de 0.0555555555556	0.927333333333
Modelo 10	Árbol de decisión con One vs All, con una profundidad de 10.0, un Gini de 0.0	0.903666666667

Los clasificadores seleccionados arriba, al realizarse mediante validación cruzada han sido seleccionados por ser el clasificador más cercano a la media de los cinco clasificadores generados.

Validación cruzada

Para entrenar el modelo hemos utilizado el método de K-fold, es decir para cada iteración *k* hemos entrenado el modelo. Después de entrenar *k* clasificadores hemos elegido el clasificador que más se acerca a la media ya que si elegimos el clasificador que posea más precisión podríamos estar sobreajustándolo.

Selección a priori de un candidato

Tras evaluar los clasificadores anteriormente descritos, decidimos quedarnos con el que mayor precisión ha obtenido asumiendo el riesgo de que puede que el modelo conseguido se encuentre sobreajsutado a los datos de entrenamiento. Por lo tanto, seleccionamos el modelo 7.

Resultados de test y conclusiones

Para comprobar si hemos realizado la elección correcta del modelo a implementar, calculamos la precisión con los datos de test para todos los modelos obteniendo los siguientes resultados:

Modelo	**Precisión validación**	**Precisión test**
Modelo 1	0.959666666667	1
Modelo 2	0.967666666667	1
Modelo 3	0.983666666667	1
Modelo 4	0.976	0.981481481481
Modelo 5	0.975666666667	0.981481481481
Modelo 6	0.984	0.981481481481
Modelo 7	**0.992**	**1**
Modelo 8	0.984	1
Modelo 9	0.927333333333	0.925925925926
Modelo 10	0.903666666667	0.888888888889

Como muestra la tabla, nuestra selección del modelo ha sido correcta. Esto puede deberse a los pocos datos de test que hay, 35. Tras comprobar que tanto en train como en validación, todo parece indicar que los datos son muy linealmente separables por lo que, con modelos basados en SVM se consiguen muy buenos resultados.