eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 11

EJERCICIOS

Hay dos conjuntos de datos a utilizar. Tienes que llevar a cabo un análisis discriminante y la construcción de un árbol binario. Puedes hacer las dos cosas sobre el mismo conjunto de datos, o una sola cosa sobre cada uno, a tu elección.

1. El fichero `landsat.dge` (lee con `dget`) contiene datos ASCII procedentes de una escena LANDSAT.

   Información general. Los satélites de la serie LANDSAT giran sobre la tierra en órbita aproximadamente polar y a relativa baja altura. Disponen de sensores tanto para la luz con longitudes de onda en el espectro visible como para el infrarrojo cercano. En la época en que se recopilaron los datos que se te ofrecen, eran cuatro bandas de frecuencia las recogidas[1]: dos en el visible (correspondiendo aproximadamente a los colores verde y rojo) y dos en el infrarrojo cercano.

   Los satélites transmiten la información que recogen "fotografiando" la superficie de la tierra. Esta información se procesa, corrige y alinea para producir "escenas", especie de mapas en que para cada punto se tiene la intensidad de la luz reflejada (o radiación infrarroja) en las bandas de frecuencia aludidas.

   Esta información tiene mucha aplicación. Unos pocos ejemplos son: cartografiar usos del suelo, diagnosticar daños tras un incendio forestal, evaluar la innivación sobre grandes áreas (y la previsible disponibilidad de recursos hidráulicos la primavera siguiente), examinar las condiciones del océano (incluida temperatura superficial) sobre grandes áreas, y otros muchos similares.

   Cada punto de la imagen o *pixel* (*pixel* = *pic*ture *el*ement) corresponde a un cuadrado de unos 80 metros de lado: la resolución es muy buena, sin llegar a la que ofrecen los satélites de observación militar[2]. El valor de la intensidad de la luz (o radiación infrarroja) para cada pixel se discretiza y codifica en un byte: 0 es el mínimo y 255 el máximo.

   Cada imagen o escena consiste en 2340 x 3380 pixels, para cada uno de los cuales se registran cuatro valores entre 0 y 255.

   Descripción de los datos. La muestra consiste en una pequeña region (de 82 x 100 pixels) de una escena como las descritas. Cada línea en cualquiera de los dos ficheros contiene 36 valores correspondientes a la radiación registrada en los 9 pixels de un cuadradito de 3 x 3 pixels. Como último campo tienes un código de clasificación de la naturaleza del terreno verificada en el pixel central, según la clave en la Tabla 1.

[1] Posteriormente fueron cinco.
[2] Y posteriormente ha mejorado.

Cuadro 1: Claves y tipos de suelo en el fichero `landsat.dge`.

Código	Tipo de terreno
1	tierra roja
2	sembrado de algodón
3	tierra gris
4	tierra gris húmeda
5	rastrojos
6	vegetación variada (retirada de la muestra)
7	tierra gris pantanosa

Trabajo a realizar. Tu trabajo consiste en elaborar un procedimiento de clasificación. Entre los comentarios al final tienes alguna sugerencia. No olvides proporcionar información sobre cómo funciona el procedimiento que utilices (por ejemplo, en forma de una tabla de confusión).

2. Utiliza los datos de CIRES que ya empleaste en una tarea anterior para ajustar un árbol binario a un subconjunto de variables de tu interés.

Lectura recomendada. Cualquiera de los manuales utilizados en el curso trata, con mayor o menor desarrollo, el tema de análisis discriminante. El capítulo 13 de [9], por ejemplo, puede serte de utilidad. Monografías especializadas son, entre otras, [6] (antiguo, pero todavía de útil y agradable lectura), [3], [4] y [8]. [1] tiene también en su Capítulo 3 una presentación interesante del análisis discriminante, desde el punto de vista de las redes neuronales. Otro libro con la misma orientación —clasificación, reconocimiento de pautas y redes neuronales— es [10], cuyo capítulo 7 trata además específicamente de árboles binarios.

La referencia esencial sobre árboles binarios de regresión y clasificación continúa siendo [2]. Hay buena documentación *on line* en R (paquete `rpart`).

AYUDAS, SUGERENCIAS, COMENTARIOS

1. Sobre estos datos obtendrás resultados buenos tanto con los métodos tradicionales (análisis discriminante, en cualquiera de sus variedades) como con el uso de árboles binarios. Merece la pena que emplees los dos procedimientos y compares.

2. Una alternativa a medio camino entre las dos anteriores es la propuesta por [7]. Si tienes interés, puedes leer el artículo (y la nada condescendiente réplica de los proponentes del método CART) y sacar tus conclusiones.

3. El libro [2] ha sido muy influyente. Puso en circulación ideas que han resultado muy productivas —entre ellas, la de "podar" en lugar de buscar criterios de parada temprana y la de divisiones delegadas para tratar valores perdidos—. Pero previamente había existido bastante trabajo sobre árboles, parte de él realizado extramuros de la Estadística. Se propusieron diversos métodos (AID, y otros varios relacionados identificables por el sufijo 'AID') que han sido utilizados y se utilizan. Un exponente más moderno de esta línea de trabajo es FIRM (véase [5]).

4. La percepción remota *(remote sensing)* ha adquirido un enorme desarrollo, y emplea técnicas puramente estadísticas o combinadas con otras propias de la Inteligencia Artificial.

   Observa que si empleas los métodos que se te sugieren estás infrautilizando la información: los datos no proceden de una urna en orden arbitrario, sino que tienen un orden espacial[3]. Este orden puede explotarse: por

[3] En la muestra que se te proporciona, los registros han sido permutados y el orden perdido.

ejemplo, si un pixel está en la vecindad de otros cultivados de algodón, cabe cierta presunción de que también está cultivado de algodón: en general los usos del suelo tienden a agruparse en áreas contiguas.

5. Las escenas LANDSAT son caras. Los datos que utilizas han sido ofrecidos gratuitamente por Ashwin Srinivasan de la Universidad de Strathclyde y proceden en última instancia de una escena comprada por The Centre for Remote Sensing, University of New South Wales, Australia. Personal de esta última institución realizó la identificación sobre el terreno de los usos del suelo.

   Lo ideal sería tener una escena más amplia, e identificación *(ground truth data)* para cada *pixel*, pero ello no ha sido posible.

6. Como ya se comentó en clase, si empleas datos de CIRES es fácil que tropieces con ejemplos inabordables: hay muchas variables y algunas contienen muchas categorías. Puedes ampliar el campo de los ejercicios factibles si variables que tienen orden natural las defines como factores ordenados: una variable nominal con K categorías, obliga a probar 2^K divisiones, en tanto una ordenada sólo requiere $K - 1$.

Referencias

[1] C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1996.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, California, 1984.

[3] D.J. Hand. *Discrimination and Classification*. Wiley, 1981.

[4] D.J. Hand. *Construction and Assessment of Classification Rules.* Wiley, 1997.

[5] D.M. Hawkins. Firm: Formal inference-based recursive modeling. Technical Report 546, University of Minnesota, School of Statistics, 1997.

[6] P.A. Lachenbruch. *Discriminant Analysis*. Hafner Press, New York, 1975.

[7] Wei-Yin Loh and Nunta Vanichsetakul. Tree-structured clasification via generalized discriminant analysis. 83:715–728, 1988.

[8] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 1992.

[9] D. Peña. *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.

[10] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. 519.237.8 RIP.