Laura Martinez Londoño
Sebastián Montoya J.

4.6

Taller de Aprendizaje de Máquina: Estimación no Paramétrica

Julián D. Arias Londoño, Juan Felipe Pérez
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Medellín, Colombia
jdarias@udea.edu.co

September 16, 2015

1 Marco teórico

Como se ha visto hasta ahora en clase, los dos problemas básicos de aprendizaje de máquina: el problema de clasificación y el problema de regresión, requieren ambos de la estimación de funciones de densidad de probabilidad. Las aproximaciones básicas de regresión polinomial y de estimación de valores medios y desviaciones estándar, hacen parte del conjunto de los modelos de aprendizaje paramétricos en los cuales se hacen suposiciones sobre la funciones de densidad de probabilidad (*fdp*) que representan el conjunto de datos de entrenamiento. En el otro costado se encuentran los modelos de aprendizaje no paramétricos, en los cuales no se hacen suposiciones sobre la forma de la *fdp* que representa los datos, aunque a diferencia de lo que puede pensarse por el nombre, dichos modelos requieren incluso de un número de parámetros mayor que el que se requiere en varios tipo de modelos paramétricos.

A continuación se van a presentar 3 métodos de estimación de *fdp* no parámetricos. Todos los métodos presentados permiten llevar a cabo tareas tanto de regresión como de clasificación. Es importante tener en cuenta que siempre partimos de un conjunto de entrenamiento $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ (estamos abordando tareas de aprendizaje supervisado), y dependiendo de la tarea necesitaremos llevar a cabo uno de los siguientes cálculos:

- **Regresión** Debemos encontrar un modelo $f(\mathbf{x})$ a partir de cual podamos calcular:

$$y = f(\mathbf{x}) = E[p(y|\mathbf{x})]$$

  En el caso discreto el valor esperado de la ecuación anterior se puede expresar como

Nota. Recuerde que es necesario realizar una previa normalización de las variables, por cuanto ambos métodos requieren el cálculo de medidas de distancia que se ven afectadas debido al sesgo que puede introducir las diferencias en los rangos dinámicos de las variables utilizadas.

2 Ejercicios

1. Adjunto a este taller encontrará los archivos: `Main.m`, `normalizar.m`, `gaussianKernel.m`, `vecinosCercanos.m`, `ventanaParzenClass.m` y `ventanaParzenRegress.m`. El archivo `Main.m` es el script principal, desde el cual se ejecutan todas las instrucciones para realizar el taller. Una vez corra el Script principal se le solicitará ingresar el numeral del punto que desea resolver (es decir 3,4,5 o 6). Analice con cuidado el script y comprenda como esta construido.

2. Describa la base de datos utilizada en el problema de regresión. Cuantas son las muestras de entrenamiento y validación y cuantas son las características.

   R/: El número total de muestras es 1030, donde se usan 721 para entrenar (70%) y 309 para validar (30%). El número de características en las muestras son 8 pero solo se seleccionan las 6 primeras. ✓

   Para poder resolver el problema de regresión con vecinos cercanos debe completar el archivo `vecinosCercanos.m`, el cual tiene indicado las lineas donde se debe implementar el cálculo para determinar los K vecinos más cercanos de una muestra.

   Una vez haya completado el código, ejecute varias veces el proceso de entrenamiento y evaluación cambiando el parámetro k, el cual es el número de vecinos, y complete la siguiente tabla con los valores del error cuadrático medio (ECM) obtenidos:

Número de vecinos	Error Cuadrático Medio (ECM)
1	269.111
2	199.5921
3	172.5643
4	179.147
5	164.0136
6	159.5951
7	157.4445
100	187.7005

✓

Responda las siguientes preguntas:

- ¿Cuál es el porcentaje de la base de datos usado en el conjunto de prueba?
  R/: Se utiliza un 0.3 ó 30% de todas las muestras para la validación.

- ¿Por qué cree que se obtiene este resultado con 100 vecinos?
  R/: Porque a mayor número de vecinos se aumenta la posibilidad de incluir muestras que no deberían ser de dicho grupo o cluster de modo que las salidas de dichas muestras erroneas alteran o cambian la salida

3. **Ahora resuelva el problema de regresión con el método de ventana de parzen. Para hacerlo debe completar el archivo `ventanaParzenRegress.m`, el cual tiene indicado las lineas donde se debe implementar la función de predicción de Nadaraya-Watson.**

   Una vez haya completado el código, ejecute varias veces el proceso de entrenamiento y evaluación cambiando el parámetro h, donde h es el ancho de la ventana de suavizado, y complete la siguiente tabla con los valores del error cuadrático medio (ECM) obtenidos:

Ventana de suavizado	Error Cuadrático Medio (ECM)
0.05	165.4973
0.1	162.9778
1	173.3412
10	283.2141

Responda las siguientes preguntas:

- ¿Cuál es la formula usada para calcular el ECM?
  R/:

$$ECM = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2$$

Donde N no son todas las muestras, sino tan solo las muestras reservadas para la validación.

- **¿Por qué cree que se obtiene este resultado con h igual a 0.01?**
  R/: h = 0.01 → NaN

Este tamaño de ventana hace que la predicción sea muy ruidosa y con discontinuidades, lo que hace que se incremente en gran medida el error. Pruebas de las discontinuidades son algunos valores NaN en el vector y estimado. Estos valores aparecen por la división de 0/0 y lo cual significa que el h le da un peso muy pequeño a los muestras más alejados.

Una ventana pequeña reducirá la parcialidad de la función de densidad de probabilidad.

Cuando h es pequeño la ventana se hace muy angosta y es posible que le asigne un peso de 0 a todas las muestras, por lo que la división toma valores 0/0.

Laura Martinez Londoño
Sebastián Montoya Jiménez

4.5

4. Describa la base de datos utilizada en los problemas de clasificación. Responda las siguientes preguntas: ¿Cuántas son las muestras de entrenamiento y validación?, ¿Cuántas son las características? ¿Cuantas son las clases del problema? y ¿Cuántas son las muestras por cada clase?

R/: Son en total 150 muestras de las cuales se destinan 105 para entrenamiento (70%) y 45 para la validación (30%). Se tienen 4 características pero tan solo se cargan 3; El número de clases son 3 donde cada una tiene 50 muestras.

Para poder resolver el problema de clasificación con vecinos cercanos debe completar el archivo *vecinosCercanos.m*, el cual tiene indicado las lineas donde se debe implementar el cálculo para determinar los K vecinos cercanos de una muestra.

Una vez haya completado el código, ejecute varias veces el proceso de entrenamiento y evaluación cambiando el parámetro k, correspondiente al número de vecinos, y complete la siguiente tabla con los valores de eficiencia y error de clasificación obtenidos:

Número de vecinos	Eficiencia	Error de clasificación
1	0.93333	0.066667
2	0.91111	0.088889
3	0.93333	0.066667
4	0.88889	0.11111
5	0.86667	0.13333
6	0.88889	0.11111
7	0.88889	0.11111

Responda las siguientes preguntas:

- ¿Por que se usa la moda en el caso de clasificación con el método de los K vecinos?

  R/: Se usa la moda porque la clase que más se repite entre los K-vecinos de la muestra que estoy evaluando tiene más probabilidad de que sea mi clase porque se supone que en los modelos no parametricos los datos tienen un comportamiento suave.

  Porque la variable clase es de tipo categórica

- ¿Por qué cree que se deben armar los conjuntos de entrenamiento y prueba de forma aleatoria?

  R/: Porque al seleccionar muestras de manera aleatoria se evita la probabilidad del sesgamiento e imbalance de clases a la hora de entrenar el sistema, de modo que se evita patrones y los predicciones se aproximan mejor a la realidad, o sea, se entrena para generalizar y no memorizar.

Este no se evita.

5. Ahora resuelva el problema de clasificación con el método de ventana de Parzen. Debe completar el archivo `ventanaParzenClass.m`, el cual tiene indicadas las lineas donde se debe implementar el cálculo de la función de probabilidad.

   Una vez haya completado el código, ejecute varias veces el proceso de entrenamiento y evaluación cambiando el parámetro h y complete la siguiente tabla con los valores de eficiencia y error de clasificación obtenidos:

Ventana de suavizado	Eficiencia	Error de clasificación
0.05	0.93333	0.066667
0.1	0.93333	0.066667
1	0.84444	0.15556
10	0.84444	0.15556

   Responda las siguientes preguntas:

   - ¿Por qué el modelo de ventana de Parzen es un modelo no paramétrico?
     R/: Porque no se asume ninguna forma para la función ya que a partir de los datos es que se crea esa forma, a diferencia de los modelos paramétricos que parten de una forma para sus modelos

6. Utilice la función classify de MatLab para llevar a cabo la clasificación utilizando un modelo basado en funciones discriminantes Gaussianas. Modifique el archivo Main.m para incluir el experimento con dicho modelo y compare con los resultados obtenidos por los dos modelos anteriores.

	Eficiencia	Error de clasificación
Linear	0.95556	0.044444
Diaglinear	0.91111	0.088889
Quadratic	0.91111	0.088889
Diagquadratic	0.91111	0.088889
Mahalanobis	0.93333	0.066667

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. New Jersey, NY, USA: Wiley-Interscience, 2nd ed., 2000.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

Comparando los resultados de esta tabla con las anteriores, se puede observar que el mejor resultado es el obtenido con la función discriminante lineal.