Práctica 1

Nombre del curso: Teoría Moderna de la Detección y Estimación

Autores: Jesús Cid Sueiro, Jerónimo Arenas García

Universidad
Carlos III de Madrid

Practica 1: Estimación máquina

Teoría Moderna de la Detección y la Estimación

September 23, 2013

1 Introducción

En esta práctica el alumno implementará diferentes métodos de regresión máquina para predecir los precios diarios establecidos por una compañía aérea en sus billetes a partir de los ofrecios por otras nueve compañías aéreas en los mismos días.

En el archivo `data_P1.mat` se proporcionan los datos necesarios para la práctica. Concretamente, tras cargar dicho fichero verá que tiene disponible en su espacio de trabajo Matlab las seis variables siguientes:

- `X_train`: Matriz con las observaciones para los datos de entrenamiento. Su dimensión es 375×9, correspondiente a 375 puntos o patrones de dimensión 9; es decir, cada vector $\mathbf{X}$ consta de 9 observaciones diferentes.
- `s_train`: Vector de longitud 375 con los valores de la variable a estimar S para cada uno de los patrones de entrenamiento, siguiendo el mismo orden que en `X_train`. Es decir, el elemento k-ésimo de `s_train` contiene el valor de S para la fila k-ésima de `X_train`.
- `X_val`: Matriz con las observaciones para los datos de validación. Su dimensión es 200×9, correspondiente a 200 puntos o patrones de dimensión 9.
- `s_val`: Vector de longitud 200 con los valores de la variable a estimar S para cada uno de los patrones contenidos en `X_val`.
- `X_test`: Matriz con las observaciones para los datos de test. Su dimensión es 375×9, correspondiente a 375 puntos o patrones de dimensión 9.
- `s_test`: Vector de longitud 375 con los valores de la variable a estimar S para cada uno de los patrones contenidos en `X_test`.

Cada fila de las variables `X_train`, `X_val` contiene los precios establecidos en un mismo día por 9 compañías para una determinada ruta aérea, siendo el objetivo de los estimadores a diseñar predecir el valor para una décima ruta y compañía. Para el diseño de los estimadores máquina que se considerarán en esta práctica se dispone de los valores objetivos reales correspondientes a cada una de las filas de las matrices anteriores en las variables `s_train`, `s_val`, respectivamente.

A lo largo de la práctica usaremos `X_train` y `s_train` para el ajuste de los regresores, mientras que `X_val` y `s_val` se usarán para validar el modelo con el fin de asegurar una correcta generalización, es decir, que el estimador funcione correctamente cuando se utilice con datos diferentes de los vistos durante el entrenamiento.

Una vez diseñado el estimador, éste puede aplicarse para estimar valores desconocidos de S. En esta práctica, las variables `X_test` y `s_test` se proporcionan para que los alumnos puedan simular cuál sería el error de generalización de los estimadores diseñados.

Los objetivos de esta práctica son:

- Diseñar estimadores lineales y no lineales. Se considerará regresión lineal, semilineal polinómica, y basada en k-vecinos más cercanos.
- Ilustrar el concepto de sobreajuste. Podrá comprobar durante la práctica que el diseño de un estimador con excesivos grados de libertad puede conducir a prestaciones deficientes cuando se mida el error sobre los datos de test.
- Estudiar cómo el uso de un conjunto de validación permite seleccionar un modelo que proporcione una adecuada generalización.

2 Visualización y normalización de los datos

1. Cargue el archivo de datos y considere, por el momento, solamente los datos de entrenamiento. Con objeto de tener una primera visión (aunque parcial) del problema, represente gráficamente la variable a estimar, S, en función de cada observación, X_i (es decir, los pares $\{x_i(k), s(k)\}$ del conjunto de entrenamiento). A lo largo de esta práctica nos referiremos a estas gráficas como diagramas de dispersión de los datos.
2. Con el fin de asegurar la estabilidad numérica en la resolución de secciones posteriores, es conveniente llevar a cabo una normalización de los datos. Las siguientes líneas de código realizan una transformación lineal que hace que los datos de entrenamiento tengan media nula y varianza unidad. Exactamente la misma transformación ha de aplicarse a los datos de validación y de test:

   ```
   >> mx = mean(X_train); stdx = std(X_train);
   X_train = (X_train - ones(n_train,1)*mx)./(ones(n_train,1)*stdx);
   X_val = (X_val - ones(n_val,1)*mx)./(ones(n_val,1)*stdx);
   X_test = (X_test - ones(n_test,1)*mx)./(ones(n_test,1)*stdx);
   ```

   donde las variables `n_train`, `n_val` y `n_test` contienen, respectivamente, el número de muestras de entrenamiento, validación y test.

3 Regresión lineal

Determine los coeficientes de regresión lineal de mínimo Error Cuadrático Promedio (ECP), basado en los datos de entrenamiento, y determine el ECP obtenido tanto sobre el conjunto de datos de entrenamiento como en el de test. Compare estos resultados con el ECP que se obtendría con un método "base" que no utilizase ninguna observación, i.e., $\hat{S} = w_0$.

4 Regresión polinómica con una única variable

1. Analice en detalle el caso de regresión polinómica con una sola variable. Para ello, determine, para cada variable, los coeficientes de regresión de los modelos polinómicos de mínimo ECP de grado 0 (sin variables), 1 (lineal), 2, . . . , 10:

$$\hat{S} = w_0 + \sum_{g=1}^{G} w_g X_i^g \tag{1}$$

   donde G es el grado del polinomio y X_i la variable seleccionada ($i = 1, \ldots, 9$). Represente gráficamente, también para cada variable, las curva de ECP de entrenamiento y validación en función del grado del polinomio. Seleccione para el resto de la sección la variable que muestre un menor error de validación.

2. Repita el apartado anterior para la variable seleccionada, considerando ahora un grado máximo del modelo desde 0 a 25. Represente en una misma figura la evolución del ECP de entrenamiento y validación en función del grado del polinomio. A la vista de la curva de validación, seleccione un valor de G que asegure una buena generalización del regresor. Calcule para dicho valor G el ECP de test del regresor diseñado.

3. Represente, la función de regresión del modelo construido en la sección anterior conjuntamente con el diagrama de dispersión de dicha variable. Represente sobre la misma gráfica las funciones de regresión de un modelo polinómico basado en la misma variable que sufra subajuste y de otro que sufra sobreajuste.

5 Regresión no paramétrica de k vecinos más cercanos

En su versión más sencilla, el estimador de k vecinos más cercanos calcula la estimación asociada a un vector de observaciones $\mathbf{x}_t$ mediante el siguiente procedimiento:

- Búsqueda en el conjunto de entrenamiento de los k vectores de observaciones más cercanos a $\mathbf{x}_t$. En esta práctica consideraremos únicamente la distancia euclídea.
- Se estima el valor de $\hat{s}_t$ como promedio de los valores de s para los vecinos de $\mathbf{x}_t$.

Las prestaciones de este estimador dependen fundamentalmente del valor de k, que suele escogerse utilizando un conjunto de validación.

1. En primer lugar trabajaremos con una única observación, la misma que fue seleccionada en el apartado anterior, a fin de poder representar las curvas de regresión sobre dicha variable. Utilizando el método de los k vecinos más cercanos, calcule el ECP de validación para los valores de k en el rango [1,50]. Seleccione el valor de k que minimiza el ECP de validación, y calcule el ECP de test del estimador correspondiente.

2. Represente, la función de regresión del modelo para el valor de k seleccionado por validación, conjuntamente con el diagrama de dispersión de la variable utilizada. Represente sobre la misma gráfica las funciones de regresión asociadas a valores de k para los que se observe subajuste y sobreajuste.

3. Trabaje por último de forma conjunta con todas las observaciones disponibles. Calcule el ECP de validación en función del valor de k. Seleccione el valor de k que minimiza el ECP de validación, y calcule el ECP de test del estimador correspondiente.