Regresión logística regularizada 1

Introducción

La finalidad del proyecto es ilustrar el funcionamiento de una técnica de regresión logística regularizada con la norma ℓ_2. El ejemplo de prueba consiste en un conjunto de 596 casos de tumores de seno etiquetados como benignos/malignos. Cada caso está caracterizado por 30 atributos. Los datos se representan mediante una matriz X de $n = 596$ renglones con $m = 30$ columnas. Las etiquetas se representan mediante un vector y de ceros y unos; cero indica *tumor benigno*, mientras que uno indica *tumor maligno*.

El proyecto está dividido en dos partes. La primera está dedicada a estudiar el concepto de las técnicas de regularización. La segunda parte consiste en combinar técnicas de optimización con técnicas de regularización.

El problema por resolver se puede formular como

$$\text{minimizar} \qquad \phi(w) = f(w) + \frac{\gamma}{2}||w||_2^2, \tag{1}$$

en donde a f se le conoce como la función de pérdida; γ es una constante real positiva; w es el vector de parámetros a optimizar. El término $\frac{\gamma}{2}||w||_2^2$ tiene como efecto controlar el posible crecimiento de las componentes de w; otro efecto deseable es reducir el número de componentes de w *diferentes de cero*; en estos casos es preferible utilizar $||w||_1$.

El método propuesto para resolver (1) es el método de Newton globalizado mediante una búsqueda lineal con condiciones fuertes de Wolfe. Los sistemas de ecuaciones lineales se resuelven en forma aproximada con el algoritmo de gradiente conjugado protegido contra posibles singularidades en la Hessiana de ϕ debidas al uso de aritmética de punto flotante.

Regularización

Considerar los datos del problema *Filip* descrito en

`http://www.itl.nist.gov/div898/strd/lls/data/Filip.shtml`

El problema consiste en ajustar un polinomio de grado 10 a un conjunto de $n = 82$ observaciones $\{(x_i, y_i), \quad i = 1, \ldots, n\}$. Todos los paquetes de software comerciales explorados (SPSS, Stata, R, S-plus, SAS) tienen dificultades numéricas para resolver este problema. En esencia, la matriz de diseño es muy mal condicionada y los procedimientos numéricos en los paquetes de software comercial son utilizadas en forma inapropiada. El problema se puede resolver fácilmente en MATLAB con el comando `polyfit`, el cual hace uso de la factorización QR-Householder en forma correcta. Los coeficientes del polinomio calculados con `polyfit` coinciden en al menos 6 cifras decimales con los reportados por NIST. Es importante notar que NIST obtiene los coeficientes reportados utilizando aritmética de punto flotante simulada de 500 dígitos decimales. Los procesadores convencionales utilizan aritmética de punto flotante de 16 dígitos.

Procedimiento

1. Ajustar polinomios de grado creciente $k = 2, \ldots, 10$ mediante `polyfit`. Graficar $||\beta||_2$, la norma euclidiana del vector de coeficientes, y el error de ajuste como funciones del grado del polinomio

$$p_k(x) = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$
$$e_k = \sum_{j=1}^{n} (p_k(x_i) - y_i)^2.$$

2. Ajustar un *spline* cúbico natural al conjunto de observaciones. Considerar el *spline* como la función verdadera que describe las observaciones.

3. Evaluar los polinomios en $\ell = 100$ puntos aleatorios en el intervalo (x_1, x_n) y comparar contra el valor *verdadero* obtenido con el *spline*. Acumular y graficar el error de predicción como función del grado del polinomio.

$$\bar{e}_k = \sum_{j=1}^{\ell} (s(x_j) - p_k(x_j))^2,$$

   en donde s es el *spline* natural.

4. Ahora construir los polinomios $k = 2, \ldots, 10$ utilizando la formulación

$$\text{minimizar} \qquad f(\beta) + \frac{\gamma}{2}||\beta||_2^2$$

   en donde f_k, la función de pérdida, es la muy conocida

$$f_k(\beta) = \sum_{i=1}^{n} (y_i - p_k(x_i))^2$$

   Utilizar el método de Newton globalizado descrito en la primera parte.