Óptimización II

Hoja de ejercicios Parte I: Teoría básica

Francisco Javier Mercader Martínez

1) Regresión Lineal. Posiblemente, el modelo de predicción más sencillo en Machine Learning (ML) es de la regresión lineal. Partimos de un conjunto de datos (dataset) de la forma

$$S = \left\{ \left(x_i^j, y^j \right), 1 \leq i \leq d, 1 \leq j \leq N \right\}.$$

Usando la terminología habitual en ML, la matriz $X = \left(x_i^j \right)$ se llama matriz de las características (features, en inglés) y el vector $y = (y^j)$ es el llamado vector de etiquetas (label). Nótese que estamos usando el índice i para denotar cada característica, y el j para la muestra (instance, en la jerga del ML).
Empezamos dividiendo nuestro dataset en dos partes: la primera de ellas contiene aproximadamente el 80% de las filas de X y el mismo número de componentes en el vector de y. Este conjunto constituye el llamado datos de entrenamiento (training dataset), es decir,

$$S^{\text{train}} = \left\{ \left(x_i^j, y^j \right), 1 \leq i \leq d, 1 \leq j \leq N^{\text{train}} \right\}.$$

El resto de filas de X y sus etiquetas asociadas conforman los llamados datos de test (test dataset). Estos datos se utilizan para validar el modelo de predicción que se definirá a continuación.
Sin entrar demasiado en detalles, la industria del ML consiste en proponer un modelo de predicción, el cual incluye un conjunto de parámetros que se fijan usando los datos de entrenamiento, de modo que dicho modelo sea capaz de predecir los datos del test dataset. En el caso de la regresión lineal, el modelo que se propone toma la forma

$$\hat{y} = \omega^\intercal x$$

donde $\omega \in \mathbb{R}^n$ son los **parámetros** del modelo, $x \in \mathbb{R}^n$ es el vector de entradas (input) y el escalar $\hat{y}$ es la salida (output) que el modelo predice a partir del input x. El objetivo es calcular ω de modo que el llamado error cuadrático medio en el training dataset sea mínimo. Este proceso se conoce con el nombre de entrenamiento (training process) y se concreta en la resolución del sisguiente problema de optimización no lineal:

$$\text{Minimizar en } \omega \in \mathbb{R}^d: \quad \text{MSE}_{\text{train}}(\omega) = \frac{1}{N^{\text{train}}} \sum_{j=1}^{N^{\text{train}}} \left(\hat{y}^j(imega) - y^j \right)^2, \tag{1}$$

donde

$$\hat{y}^j(\omega) = \sum_{i=1}^{d} x_i^j \omega_i.$$

Escribe la condición necesaria de optimalidad para el problema (1).
Al valor de la función objetivo en el óptimo $\text{MSE}_{\text{train}}(\omega^*)$ se le llama **error de entrenamiento** (training error) y se dice que el modelo está bien entrenado si este valor es pequeño.

2) Regresión linea: continuación. Bien podría suceder que para un modelo que está bien entrenado, cuando se usa sobre los datos del test dataset el error que se produce (llamado **error de generalización**) sea grande. De un modo natural, hemos de proponer algún otro modelo que tenga un mejor comportamiento. Continuacndo con la regresión lineal, la primera modificación que se suele hacer es incorporar al modelo un escalar que permita corregir el hecho de que los datos estén traslados (que tengan sesgo, dicen los profesionales de la Estadística). Este escalar se llama sesgo

(bias, en inglés) y consituye un nuevo parámetro que se añade a los ω. El modelo de predicción resultante es

$$\hat{y} = \omega^\intercal x + b$$

Escribe las condiciones necesarias de optimalidad en este caso.

3) Regresión lineal: regularización. Si las cosas siguen yendo mal (error de generalización grande) con este nuevo modelo que incorpora el bias, bien podría ser que nuestros datos no estén distribuidos de manera lineal, con lo que modelos de predicción lineales como los propuestos anteriormente difícilmente van a ser útiles. O incluso si el modelo generaliza bien, podría suceder que nuestro modelo de predicción fuese lento porque incorporarse un número de parámetros muy grande. Una propuesta para intentar solventar ambas dificultades consiste en añadir a la función objetivo un término (llamado de regularización) de la forma

$$\alpha\|(\omega, b)\|_p^p, \quad \alpha > 0, \quad 1 \leq p < \infty,$$

de modo que la función objetivo resultante es

$$\mathrm{MSE}_{\mathrm{train}}(\omega, b) = \frac{1}{N^{\mathrm{train}}} \sum_{j=1}^{N^{\mathrm{train}}} \left(\hat{y}^j(\omega) - y^j\right)^2 + \alpha\|(\omega, b)\|_p^p.$$

Explica qué efecto puede provocar usar la norma $p = 1$.
Analiza la naturaleza matemática (convexidad, coercividad, etc.) del problema anterior.

4) Sea A una matriz de tamaño $n \times n$, simétrica y definida positiva (todos los autovalores son estrictamente positivos), y B otra matriz de talla $m \times n$ con rango igual a m. Resuelve el problema

$$\begin{cases} \text{Minimizar} & J(x) = \dfrac{1}{2} Ax \cdot x - b \cdot x \\ \text{Sujeto a} & Bx = 0 \end{cases}$$

donde $x = (x_1, \ldots, x_n)$ y $b = (b_1, \ldots, b_n)$.

5) Dado el problema de programación no lineal

$$(\mathrm{PPNL}) \begin{cases} \text{Minimizar} & f(x_1, \ldots, x_n) \\ \text{Sujeto a} & h_1(x_1, \ldots, x_n) = 0 \\ & \cdots\cdots\cdots\cdots\cdots \\ & h_d(x_1, \ldots, x_n) = 0 \end{cases}$$

se define el Lagrangiano asociado a dicho problema como

$$\mathcal{L}(x, \lambda) = f(x) + \lambda \cdot h(x)$$

donde $h = (h_1, \ldots, h_d)$ y $\lambda = (\lambda_1, \ldots, \lambda_d)$. Comprueba que si x^0 es un mínimo de (PPNL), entonces x^0 es un punto estacionario para $\mathcal{L}$, es decir,

$$\frac{\partial \mathcal{L}}{\partial x_i}(x^0, \lambda) = 0 \quad \text{y} \quad \frac{\partial \mathcal{L}}{\partial \lambda_j}(x^0, \lambda), \quad 1 \leq i \leq n,\ 1 \leq j \leq d.$$

6) Demuestra las siguientes afirmaciones:

a) Toda aplicación lineal es convexa, pero no estrictamente convexa.

b) Toda combinación lineal finita de funciones convexas con coeficientes no negativos define una función convexa.

c) Si $T : \mathbb{R}^n \to \mathbb{R}^m$ es lineal, y $g : \mathbb{R}^m \to \mathbb{R}$ es convexa, entonces la composición $f(x) = g(Tx)$ es convexa.

d) Si $g : K \subset \mathbb{R}^n \to \mathbb{R}$ es convexa y $h : \mathbb{R} \to \mathbb{R}$ es convexa y no descreciente, entonces la composición $f(x) = h(g(x))$ es también convexa.

e) Las funciones $f(x) = \|x\|$ y $g(x) = \|x\|^p$, $p \geq 1$ son convexas.

f) Consideremos el problema de programación no lineal siendo f una función lineal no nula y $\mathcal{K} \subset \mathbb{R}^n$ un conjunto convexo y compacto. Demuestra que si x^0 es solución de este problema, entonces x^0 es un punto de la frontera del conjunto $\mathcal{K}$.

7) Sea $c \in \mathbb{R}^n, c \neq 0$. Utiliza las condiciones de KKT para encontrar la solución óptima del problema

$$\begin{cases} \text{Maximizar} & f(x) = c^\intercal x \\ \text{Sujeto a} & \|x\|^2 \leq 1 \end{cases}$$

Indicación: recuerda el apartado (f) del ejercicio anterior.

8) Consideremos el problema

$$\begin{cases} \text{Minimizar} & f(x_1, x_2) = x_1 \\ \text{Sujeto a} & x_2 - x_1^3 \leq 0 \\ & -x_2 \leq 0 \end{cases}$$

Comprueba que el punto $(0, 0)$ es un mínimo global y que, sin embargo dicho punto no cumple las condiciones necesarias de optimalidad de Karush-Kuhn-Tucker. Explica qué está sucediendo.

9) Resolver el siguiente PPNL y comprobar que se satisfacen las condiciones suficientes de optimalidad:

$$\begin{cases} \text{Maximizar} & \|\mathbf{x}\|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \\ \text{Sujeto a} & \mathbf{a} \cdot \mathbf{x} = c \end{cases}$$

donde $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ es un vector no nulo de n componentes, c es una constante y $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

10) Consideremos el problema

$$(\text{PPNL}) = \begin{cases} \text{Minimizar} & f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 \\ \text{Sujeto a} & x_1^2 + x_2^2 + 3x_3 \leq -\dfrac{5}{2} \\ & x_1 + x_2 + x_3 = -2 \end{cases}$$

Se pide:

a) Estudiar las propiedades de convexidad de la función coste f y del conjunto de admisibilidad (o factible)

$$\Omega = \left\{ (x_1, x_2, x_3) : x_1^2 + x_2^2 + 3x_3 \leq -\frac{5}{2},\ x_1 + x_2 + x_3 = -2 \right\}$$

b) Comprueba que la solución de (PPNL) es $x_1 = x_2 = -\dfrac{1}{2}, x_3 = -1$.

11) Consideremos el problema de propamación no lineal

$$(\text{PPNL}) = \begin{cases} \text{Minimizar} & f(x_1, x_2) = x_1^2 + x_2^2 \\ \text{Sujeto a} & x_1 + x_2 \geq 1 \\ & 0.25 \leq x_1, x_2 \leq 4 \end{cases}$$

Se pide:

a) Interpreta geométricamente este problema, es decir, explica el significado físico de la función coste f y dibuja el conjunto de admisibilidad (o factible).

b) Escribe las condiciones necesarias de optimalidad de Kuhn-Tucker y resuélvelas.

c) Estudia las propiedades de convexidad del problema.

d) Demuestra que dicho problema tiene solución.