Tecnicas de Optimizacion para Aprendizaje profundo y Analisis de datos

Abraham Zamudio

Grupo de Modelamiento Matematico y Simulacion Numerica - CTIC UNI

10 de agosto del 2018

Nociones sobre funciones convexas

Condiciones de optimalidad para problemas no lineales irrestrictos

Perceptron Multicapa

Función Convexa

Recordemos que una función $f : \mathbb{R} \to \mathbb{R}^n$ es **convexa** en un conjunto S si

$$f(\lambda x_1+(1-\lambda)x_2) \leq \lambda f(x_1)+(1-\lambda)f(x_2), \qquad \forall x_1, x_2 \in S, \quad \lambda \in [0,1]$$

S la desigualdad anterior es estricta para todo par de puntos distintos entonces la función f se dirá estrictamente convexa.

Ademas un conjunto S se dirá convexo si para todo $x_1, x_2 \in S$ y para todo escalar $\lambda \in (0,1)$ se tiene que $\lambda x_1 + (1-\lambda)x_2 \in S$.

Principales teoremas sobre funciones convexas

Teorema

Sean $S \subset \mathbb{R}^n$ convexo y no vacio y $f : \mathbb{R}^n \to \mathbb{R}$ una funcion convexa en S. Entonces todo minimo local f es S es un minimo global de f en S.

Teorema

Sean $S \subset \mathbb{R}^n$ convexo y no vacio y $f : \mathbb{R}^n \to \mathbb{R}$ una funcion estrictamente convexa en S. Entonces f tiene a lo más un minimo (global)en S.

Principales teoremas sobre funciones convexas

Lema

Sean $S \subset \mathbb{R}^n$ convexo y no vacio y $f : \mathbb{R}^n \to \mathbb{R}$ una función convexa en S.Pruebe que el conjunto $S_\alpha(f) = \{x \in S : f(x) \leq \alpha\}$ es convexo .

Lema

Sean $S \subset \mathbb{R}^n$ convexo y no vacio y $f : \mathbb{R}^n \to \mathbb{R}$ una función estrictamente convexa en S.Sean $a \in S$ y $d \in \mathbb{R}^n/\{0\}$ para los cuales existe $\delta > 0$ tal que para cada $\lambda \in (0, \delta)$, $a + \lambda d \in S$. Prueba que $\dfrac{f(a + \lambda d) - f(a)}{\lambda}$ es monótono creciente (estricta) en λ. En consecuencia se tiene

$$\lim_{\lambda \to 0^+} \frac{f(a + \lambda d) - f(a)}{\lambda} = \lim_{\lambda \in \mathbb{R}_+} \frac{f(a + \lambda d) - f(a)}{\lambda}$$

y por lo tanto el limite existe o es igual a $-\infty$.

Principales teoremas sobre funciones convexas

Sea S convexo abierto no vacio en $\mathbb{R}^n$ y $f : \mathbb{R}^n \to \mathbb{R}$, diferenciable en S. Entonces f es convexa en S si y solo si

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^t (x - \bar{x}), \quad \forall x, \bar{x} \in S.$$

Además, f es estrictamente convexa si y solo si la desigualdad anterior es estricta.

Nociones sobre funciones convexas

Condiciones de optimalidad para problemas no lineales irrestrictos

Perceptron Multicapa

Condicion necesaria de primer orden

Regla de Fermat

Sea $f : \mathbb{R}^n \to \mathbb{R}$ diferenciable en $x \in int(Dom(f))$. Si x es un mínimo local de f entonces $\nabla f(x) = 0$. Si además f es convexa, entonces esta condición es también suficiente y el mínimo es global

Condicion necesaria de segundo orden

Sea $f : \mathbb{R}^n \to \mathbb{R}$ dos veces continuamente diferenciable en $x \in int(Dom(f))$. Si x es minimo local de f entonces $\nabla f(x) = 0$ y $\nabla^2 f(x)$ es semidefinida positiva.

Condicion suficiente de segundo orden

Sea $f : \mathbb{R}^n \to \mathbb{R}$ dos veces continuamente diferenciable en $x \in int(Dom(f))$. Si $\nabla f(x) = 0$ y $\nabla^2 f(x)$ es definida positiva, entonces x es mínimo local aislado de f.

Nociones sobre funciones convexas

Condiciones de optimalidad para problemas no lineales irrestrictos

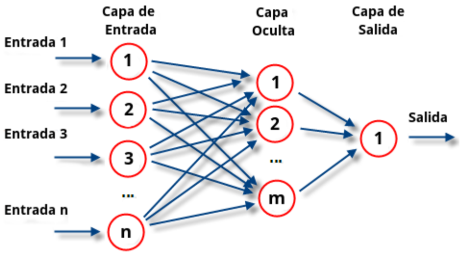
Perceptron Multicapa

Introduccion

La modificación de los pesos y bias con un algoritmo de entrenamiento de un perceptrón multicapa es un problema clásico de programación no lineal irrestricto. El Perceptron multicapa es una generalización del Perceptron simple y surgió como consecuencia de las limitaciones de dicha arquitectura en lo referente al problema de la separabilidad no lineal. Minsky y Papert mostraron en 1969 que la combinación de varios Perceptrones simples -inclusión de neuronas ocultas- podía resultar una solución adecuada para tratar ciertos problemas no lineales. Diferentes autores han demostrado independientemente que el Perceptron multicapa es un aproximador universal, en el sentido de que cualquier función contínua en un espacio $\mathbb{R}^n$ puede aproximarse con un Perceptron multicapa, con al menos una capa oculta de neuronas.

Arquitectura

La arquitectura de Perceptron multicapa se caracteriza porque tiene sus neuronas agrupadas en capas de diferentes niveles. Cada una de las capas está formada por un conjunto de neuronas y se distinguen tres tipos de capas diferentes: la capa de entrada, las capas ocultas y la capa de salida.



Modelo

Propagacion de los patrones de entrada

El Perceptron multicapa define una relación entre las variables de entrada y las variables de salida de la red. Esta relación se obtiene propagando hacia adelante los valores de las variables de entrada. Para ello, cada neurona de la red procesa la información recibida por sus entradas y produce una respuesta o activación que se propaga, a través de las conexiones correspondientes, hacia las neuronas de la siguiente capa. A continuación, se muestran las expresiones para calcular las activaciones de las neuronas de la red.

Modelo

Propagacion de los patrones de entrada

Sea un perceptron multicapa con C capas (C-2 capas ocultas) y n_c reuronas en la capa c, para $c = 1, 2, \ldots, C$. Sea $W^c = (w_{ij}^c)$ la matriz de pesos donde w_{ij}^c representa el peso de la conexion de la neurona i de la capa c para $c = 2, \ldots, C$. Denotaremos a_i^c a la activacion de la neurona i de la capa c . Estas activaciones se calculan con el modelo siguiente :

Modelo

Propagacion de los patrones de entrada

- **Activación de las neuronas de la capa de entrada** (a_i^1) . . Las neuronas de la capa de entrada se encargan de transmitir hacia la red las señales recibidas desde el exterior. Por tanto:

$$a_i^1 = x_i \text{ para } i = 1, 2, \ldots, n_1 \tag{1}$$

donde $X = (x_1, x_2, \ldots, x_{n_1})$ representa el vector o patrón de entrada a la red.

- **Activación de las neuronas de la capa oculta c** (a_i^c). Las neuronas ocultas de la red procesan la información recibida aplicando la función de activación f a la suma de los productos de las activaciones que recibe por sus correspondientes pesos, es decir:

$$a_i^c = f\left(\sum_{j=1}^{n_{c-1}} w_{ij}^{c-1}\, a_j^{c-1} + u_i^c\right) \text{ para } i = 1, 2, \ldots, n_c \qquad (2)$$

  y $c = 2, 3, \ldots, C-1$. Donde a_j^{c-1} son las activaciones de las neuronas de la capa $c-1$.

- **Activación de las neuronas de la capa de salida** (a_i^C). . Al igual que en el caso anterior, la activación de estas neuronas viene dada por la función de activación f aplicada a la suma de los productos de las entradas que recibe por sus correspondientes pesos:

$$y_i = a_i^C = f\left(\sum_{j=1}^{n_{C-1}} w_{ij}^{C-1}\, a_j^{C-1} + u_i^C\right) \text{ para } i = 1, 2, \ldots, n_c \tag{3}$$

Donde $Y = (y_1, y_2, \ldots, y_{n_C})$ es el vector de salida de la red.

Modelo

Propagacion de los patrones de entrada

La funcion f es llamada funcion de activacion. Para el perceptron multicapa , las funciones de activacion mas utilizadas son la funcion *sigmoidal* y la funcion *tangente hipérbolica* . Dichas funciones poseen como imagen un intervalo continuo de valores dentro de [0,1] y [-1,1] , respectivamente, y vienen dadas por las siguientes reglas de correspondencias :

- Funcion sigmoidal :$f_{sigm}(x) = \dfrac{1}{1 + e^{x}}$
- Funcion tangente hiperbolica : $f_{thip}(x) = \dfrac{1 + e^{-x}}{1 + e^{-x}}$