11 OPTIMITZACIÓ

(Resum teòric)

Índex

11.1. Mètode del descens del gradient per a optimització de funcions escalars de diverses variables.

El mètode del descens del gradient és un dels algoritmes d'optimització que s'utilitzen sovint en aprenentatge automàtic, particularment en les xarxes neuronals. És un mètode iteratiu que serveix per trobar mínims locals d'una funció real de diverses variables diferenciable. La idea és fer passos repetits en la direcció oposada del vector gradient de la funció en cada punt, perquè aquesta és la direcció de baixada més pronunciada. Per contra, si es fan passos repetits en la direcció del gradient conduiran a un màxim local d'aquesta funció; el procediment es coneix llavors com a mètode de l'ascens del gradient.

Considerem un conjunt U de $\mathbb{R}^n$, una funció real $f\colon U \to \mathbb{R}$ de classe C^1 en U i un punt $\mathbf{x}^0 \in U$. Recordem que la recta de $\mathbb{R}^n$ que passa per $\mathbf{x}^0$ amb vector director unitari $\mathbf{u} \in \mathbb{R}^n$, $||\mathbf{u}|| = 1$, és el lloc geomètric $r_{\mathbf{x}^0,\mathbf{u}}$ definit per

$$r_{\mathbf{x}^0,\mathbf{u}} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{x}^0 + t\mathbf{u}, t \in \mathbb{R}\},$$

Quan $t = 0$ estem al punt $\mathbf{x}^0$, quan $|t| > 0$ ens allunyem de $\mathbf{x}^0$ sobre la recta i $d(\mathbf{x}^0 + t\mathbf{u}, \mathbf{x}^0) = ||\mathbf{x}^0 + t\mathbf{u} - \mathbf{x}^0|| = ||t\mathbf{u}|| = |t|$. Aquest valor $|t|$ s'anomena *pas* de l'allunyament de $\mathbf{x}^0$ sobre la recta $r_{\mathbf{x}^0,\mathbf{u}}$. Hem estudiat que:

- La *direcció de màxim decreixement* de f al punt $\mathbf{x}^0$ és la direcció del vector oposat al vector gradient de f al punt $\mathbf{x}^0$, és a dir la del vector unitari $-\dfrac{\nabla f(\mathbf{x}^0)}{||\nabla f(\mathbf{x}^0)||}$.
- La *direcció de màxim creixement* de f al punt $\mathbf{x}^0$ és la direcció del vector gradient de f al punt $\mathbf{x}^0$, és a dir la del vector unitari $\dfrac{\nabla f(\mathbf{x}^0)}{||\nabla f(\mathbf{x}^0)||}$.

En això es basen el mètode de minimització del descens del gradient i el mètode de maximització de l'ascens del gradient.

Algoritme de descens del gradient

- Sigui $\mathbf{x}^0 \in U$ el punt inicial i sigui $\mathbf{x}^*$ un punt on f assoleix un mínim.
- Ens allunyem de $\mathbf{x}^0$ sobre la recta que té direcció oposada al vector gradient de f al punt $\mathbf{x}^0$ amb un pas de longitud $\tilde{t}^0$:

$$\mathbf{x}^1 = \mathbf{x}^0 - \tilde{t}^0 \frac{\nabla f(\mathbf{x}^0)}{||\nabla f(\mathbf{x}^0)||} \equiv \mathbf{x}^0 - t^0 \nabla f(\mathbf{x}^0),$$

  on $t^0 = \dfrac{\tilde{t}^0}{||\nabla f(\mathbf{x}^0)||}$. Observem que t^0 ha de complir que $f(\mathbf{x}^1) < f(\mathbf{x}^0)$; si no ho compleix, es repeteix el càlcul amb un t^0 més petit (per exemple, la meitat).
- A continuació, ens allunyem de $\mathbf{x}^1$ sobre la recta que té direcció oposada al vector gradient de f al punt $\mathbf{x}^1$ amb un pas de longitud $\tilde{t}^1$:

$$\mathbf{x^2} = \mathbf{x}^1 - \tilde{t}^1 \frac{\nabla f(\mathbf{x}^1)}{||\nabla f(\mathbf{x}^1)||} \equiv \mathbf{x}^1 - t^1 \nabla f(\mathbf{x}^1),$$

  on $t^1 = \dfrac{\tilde{t}^1}{||\nabla f(\mathbf{x}^1)||}$ i s'ha de complir que $f(\mathbf{x^2}) < f(\mathbf{x}^1)$; si no ho compleix, es repeteix el càlcul amb un t^1 més petit.
- Iterant aquest procediment, la iteració k és

$$\boxed{\mathbf{x}^k = \mathbf{x}^{k-1} - t^{k-1} \nabla f(\mathbf{x}^{k-1}),} \quad \text{descens del gradient, iteració } k$$

  amb t^{k-1} tal que $f(\mathbf{x}^k) < f(\mathbf{x}^{k-1})$, $\forall k \geq 1$.

La fórmula de l'ascens del gradient és anàloga:

$$\boxed{\mathbf{x}^k = \mathbf{x}^{k-1} + t^{k-1} \nabla f(\mathbf{x}^{k-1}),} \quad \text{Ascens del gradient, iteració } k$$

amb t^{k-1} tal que $f(\mathbf{x}^k) > f(\mathbf{x}^{k-1})$, $\forall k \geq 1$.

Quan el valor de t és constant ($t^k = t, \forall k \geq 1$)es diu que el mètode del descens del gradient té *pas únic*. Quan t varia en cada pas, el mètode s'anomena *multi-pas*.

La condició de parada de l'algoritme és una de les següents:

- L'error relatiu entre dues iteracions consecutives està per sota d'una tolerància fixada $T > 0$:

$$\frac{||\mathbf{x}^k - \mathbf{x}^{k-1}||}{||\mathbf{x}^k||} < T$$

- k ha arribat a un nombre màxim d'iteracions k_{max}.

Quan es tria la segona condició, de vegades es pot no arribar a l'extrem de f amb prou precisió, per aquest motiu, aquest criteri de parada s'usa quan es prefereix rapidesa a precisió. Normalment es fixa un valor de k_{max} el suficientment alt per assegurar-nos d'arribar a una solució amb bona precisió en cas de convergència de l'algoritme i evitar que l'algoritme iteri de forma infinita en cas de la no-convergència d'aquest.

Observem que si la funció té dos o més mínims locals, depenent de l'elecció de $\mathbf{x}_0$ l'algoritme pot trobar un mínim local qualsevol i no necessàriament el mínim global (en cas d'existir).