Classification supervisée

Rémy Degenne

31 janvier 2023

Ces notes de cours contiennent uniquement un bref résumé de ce qui a été dit en cours. Reportez vous à vos notes pour plus de détails.

1 Cours 2: le neurone formel et le perceptron

1.1 Le neurone formel

Un neurone biologique possède des dendrites, un noyau et un axone. Les dendrites reçoivent des excitations électriques. Ces nombreuses sources d'excitation sont agrégées, et un potentiel électrique est créé dans l'axone, ou pas. L'axone est connecté aux dendrites d'autres neurones et transmet ainsi une excitation électrique à son tour.

Attention: ceci est une présentation très (trop) simplifiée du fonctionnement d'un neurone biologique.

Le neurone formel est une approximation très grossière du neurone biologique.

- Il possède $P+1$ entrées $(e_i)_{i\in\{0,\dots,P\}}$. L'entrée e_0 est appelée le biais et vaut toujours 1.
- Il possède une sortie $s \in \mathbb{R}$.
- Il est paramétré par $P+1$ poids $(w_i)_{i\in\{1,\dots,P\}} \in \mathbb{R}^{P+1}$.
- Une fonction, dite d'*activation* et notée ϕ, détermine la valeur de s en fonction des poids et entrées:

$$s = \phi\Big(\sum_{i=1}^{P+1} w_i e_i\Big)\,. \tag{1}$$

Note : dans les librairies de réseaux de neurones, la fonction d'activation est parfois séparée du neurone, qu'on appelle alors neurone linéaire.

Fonctions possibles pour ϕ:

- identité (neurone linéaire): $\phi(x) = x$.
- fonction de Heavyside: $\phi(x) = 1$ si $x \geq 0$, $\phi(x) = 0$ (ou -1) sinon.
- logistique: $\phi(x) = \frac{1}{1+e^{-ax}}$.
- tangente hyperbolique: $\phi(x) = \tanh(x) \in [-1, 1]$.

- ReLU: $\phi(x) = \max(0, x)$.

Les fonctions autres que Heavyside sont continues et dérivable, ce qui sera important pour l'entrainement des classeurs utilisant des neurones.

Un classeur: le perceptron Le perceptron est un classeur constitué d'un seul neurone. Sa sortie s est utilisée pour prédire l'étiquette. Ses P entrées autres que le biais prennent la valeur des attributs de la donnée.

Pour que s puisse être une classe, il faut que $s \in \{0, 1\}$ ou $s \in \{-1, 1\}$. On peut choisir n'importe quel fonction d'activation ϕ, mais il faudra la composer avec un seuil si elle a des valeurs dans $\mathbb{R}$.

- Construction: à l'aide des exemples, trouver des poids qui permettent de bien classer ces exemples. Processus itératif d'optimisation.
- Evaluation: calcul de s à partir d'une nouvelle donnée. $s = \phi(\sum_{i=0}^{P} w_i e_i)$.

Interprétation géométrique du neurone linéaire (où ϕ est l'identité): l'équation $s = \sum_{i=0}^{P} w_i e_i$ définit un hyperplan dans l'espace $\mathbb{R}^{P+2}$. On classe les données en fonction du côté de l'hyperplan où elles se trouvent.

S'il est possible de trouver un hyperplan qui classe les exemples correctement, on dit qu'elles sont linéairement séparables. Souvent, ce n'est pas le cas.

1.2 Apprendre les poids

Construction du perceptron = trouver de bons poids.

Initialisation: poids aléatoires, par exemple Gaussiens centrés en zéro. Question: quelle variance? Peut avoir une grande influence sur le processus d'apprentissage.

Pour trouver les poids on utilise une méthode itérative, avec plusieurs *épisodes*.

Optimisation Optimisation = résoudre le problème

$$\min_x f(x)$$

Pour le perceptron avec un seul neurone, suivi d'un seuil θ, le perceptron classe une donnée $x = (x_0, \ldots, x_P)$ dans la classe 1 si $\phi(\sum_{i=0}^{P} w_i x_i) \geq \theta$ et dans la classe 0 sinon. C'est à dire que le perceptron définit une classe $\hat{y}_w(x) = \mathbb{I}\{\phi(\sum_{i=0}^{P} w_i x_i) \geq \theta\}$. On a un ensemble d'exemples $\mathcal{X}$ et on peut vouloir minimiser l'erreur du perceptron sur ces données d'entrainement (parce qu'on se dit qu'elles sont représentatives des autres données).

Concept: minimisation du risque empirique. On veut optimiser

$$f(w) = \sum_{(x,y)\in\mathcal{X}} \mathbb{I}\{\hat{y}_w(x) \neq y\} .$$

Comment optimiser une fonction ? On se place dans un premier temps en dimension 1, avec $f : \mathbb{R} \to \mathbb{R}$.

- Ordre 0 : on peut évaluer la fonction. (Au tableau : dessin et début d'algorithme)
- Ordre 1 : on peut évaluer la fonction et sa dérivée (si elle est dérivable!). Idée: on peut suivre la pente pour trouver un point où la fonction est plus petite.

Descente de gradient:

- Entrée: $x_0 \in \mathbb{R}$, $\alpha > 0$.
- A chaque temps $t \in \mathbb{N}$, tant qu'on ne s'est pas arrêté :
  - Evaluer $f'(x_t)$
  - Calculer $x_{t+1} = x_t - \alpha f'(x_t)$
  - Vérifier si on s'arrête (comment ?)
- A la fin: retourner x_T, où T est le temps final.

Comment s'arrêter ? Autour d'un minimum, la dérivée est faible. On choisit $\varepsilon > 0$ et on s'arrête si $|f(x_t)| \leq \varepsilon$.

Application à un réseau de neurone avec un seul poids (qui ne correspond pas à ce qu'on a vu au dessus):

$$f(w) = \sum_{(x,y)\in\mathcal{X}} \mathbb{I}\{\mathbb{I}\{\phi(wx) \geq \theta\} \neq y\}$$

Problème : cette fonction n'est pas dérivable. On pourrait utiliser une méthode d'optimisation à l'ordre 0, mais c'est beaucoup plus lent que l'ordre 1.

Solution: on remplace $f(w)$ par une autre fonction, qui est aussi minimisée quand les données sont bien classées.

$$f(w) = \sum_{(x,y)\in\mathcal{X}} \ell(\phi(wx), y)$$

avec $\ell(a, b)$ dérivable par rapport à a. La dérivée de f est $f'(w) = \sum_{(x,y)\in\mathcal{X}} (\frac{\partial}{\partial a}\ell(\phi(wx), y))\phi'(wx)x$.

Exemple pour ℓ : $\ell(a, b) = (a - b)^2$.

En dimension plus grande que 1:

Pour $w \in \mathbb{R}^{P+1}$, le gradient $\nabla f(w) = (\frac{\partial f}{\partial w_i}(w))_{i\in P+1}$ est l'analogue de la dérivée f'. C'est le vecteur tel que

$$f(w + h) = f(w) + \sum_{i=0}^{P} (\nabla f(w))_i h_i + o(\|h\|) .$$

On peut appliquer la descente de gradient: on calcule $w_{t+1} = w_t - \alpha\nabla f(w_t)$ et on s'arrête quand $\|\nabla f(w_t)\| \leq \varepsilon$.

Première méthode pour les réseaux de neurones: descente de gradient
On choisit une fonction de perte $\ell(s, y) \geq 0$ et la perte totale de l'algorithme sur les exemples $\mathcal{X}$ est

$$L(w, \mathcal{X}) = \sum_{i=1}^{N} \ell(\hat{y}_i, y_i)$$

où $\hat{y}_i$ est la valeur obtenue en évaluant le neurone quand l'entrée est fixée à x_i.

La fonction de perte $w \mapsto L(w, \mathcal{X})$ est une fonction réelle de $P+1$ variables. On peut l'optimiser, c'est à dire chercher un minimum. Les paramètres w qui réalisent le minimum sont ceux qui permettent de mieux classer les exemples dans $\mathcal{X}$.

La méthode d'optimisation que l'on va utiliser est la descente de gradient.

- Entrée: w aléatoires; $\alpha \geq 0$; $\mathcal{X}$, ensemble de N données x_i avec P attributs $x_{i,j}$ et une classe y_i. Pour chaque donnée, $x_{i,0}$ est pris égal à 1.
- Entrée: un seuil $\varepsilon > 0$.
- Initialiser $E = +\infty$
- Tant que $E > \varepsilon$,
  - calculer $\nabla_w L(w, \mathcal{X})$ où $L(w, \mathcal{X}) = \sum_{i=1}^{N} \ell(\hat{y}_i, y_i)$
  - calculer $E = \|\nabla_w L(w, \mathcal{X})\|$
  - mettre à jour $w \leftarrow w - \alpha \nabla_w L(w, \mathcal{X})$

Le gradient est $\nabla_x \ell(s_i, y_i) = (\frac{\partial \ell}{\partial w_j})_{j \in \{0,\dots,P\}}$

Exemple de perte: $\ell(s_i, y_i) = \frac{1}{2}(s_i - y_i)^2 = \frac{1}{2}(y_i - \phi(\sum_{j=0}^{P} w_j x_{i,j}))$.

Gradient pour l'exemple: $\frac{\partial \ell}{\partial w_j}(s_i, y_i) = (\phi(\sum_{j=0}^{P} w_j x_{i,j}) - y_i)\phi'(\sum_{j=0}^{P} w_j x_{i,j})(-x_{i,j})$.

En pratique: calculer le gradient est une tâche qui peut être réalisée par un algorithme.

Inconvénient de cette méthode d'optimisation: il faut calculer la perte de tous les exemples pour pouvoir faire un pas de gradient (une modification de w). C'est coûteux. Si on a un très grand nombre d'exemples, on peut imaginer qu'un sous-ensemble plus petit des exemples pourrait suffire à avoir une bonne estimation du gradient.

Gradient stochastique On n'utilise pas tous les exemples pour calculer le gradient, mais juste un. On obtient beaucoup plus de mises à jour des poids à moindre coût, mais chaque mise à jour est une moins bonne estimation du gradient.

- Entrées: w, aléatoires; $\alpha \geq 0$; $\mathcal{X}$, ensemble de N données x_i avec P attributs $x_{i,j}$ et une classe y_i. Pour chaque donnée, $x_{i,0}$ est pris égal à 1.
- Entrée: un seuil $\varepsilon > 0$.
- Initialiser $E = +\infty$
- Tant que $E > \varepsilon$,
  - mélanger les données

- $E \leftarrow 0$
- Pour chaque donnée x_i:
  - calculer $\nabla_w \ell(s_i, y_i)$
  - mettre à jour $w \leftarrow w - \alpha \nabla_w \ell(s_i, y_i)$
  - $E \leftarrow E + \|\nabla \ell(s_i, y_i)\|/N$

Propriété : cette règle d'apprentissage converge asymptotiquement. C'est à dire que si on utilise assez d'épisodes les poids se stabilisent à une valeur.

Remarque : ajouter ou enlever un exemple dans $\mathcal{X}$ peut changer beaucoup la valeur limite de w.

1.3 Perceptron multi-couche: le premier réseau de neurones

Plusieurs neurones sont organisés par couches. Exemple pour la classification binaire:

- Premier niveau: $P + 1$ entrées correspondant aux attributs. N_1 neurones avec autant de sorties.
- Deuxième niveau: $N_1 + 1$ entrées (il y a toujours un biais), N_2 sorties.
- Troisième niveau: un neurone avec $N_2 + 1$ entrées et une sortie.

On n'est pas limité à trois niveaux mais on peut en empiler autant qu'on veut.

Optimisation: gradient stochastique comme pour le perceptron mais il faut pouvoir calculer le gradient de la fonction de perte par rapport à tous les poids du réseau. Ce calcul peut être fait automatiquement, par un algorithme appelé "backpropagation".

Grace aux fonctions d'activation, le réseau de neurones n'est pas un classeur linéaire.