Explication de l'algorithme pour la méthode EC

Rivaldi Tristan

2024-03-01

Explication de l'algorithme pour la méthode EC

Le but de cet algorithme est de calculer T_{EC} pour que la confiance moyenne corresponde à la précision moyenne sur l'ensemble de validation. Pour ce faire, on dispose d'un ensemble de validation (x_i, y_i) pour $i = 1, \ldots, n_{\text{val}}$, et d'un classifieur $\hat{f} : X \to \mathbb{R}^K$ où les $x_i \in \mathbb{R}^p$ sont les caractéristiques (les variables) et les $y_i \in \{1, 2, \ldots, K\}$ sont les étiquettes de classes associées à ces caractéristiques. L'algorithme calcule les logits $z_i = f(x_i) \in \mathbb{R}^K$ et produit la sortie $\hat{y}_i = \arg\max_k z_{ik}$.

Le classifieur prend en entrée des données (ici les caractéristiques x_i extraites d'une observation) et y attribue des logits $z_i = (z_{i1}, ..., z_{iK})$ où pour tous k appartenant à $\{1, 2, \ldots, K\}$, chaque $z_{ik} \in \mathbb{R}$. Pour un k fixé z_{ik} correspond au logit (score) associé à la classe k. Le fait de produire la sortie $\hat{y}_i = \arg\max_k z_{ik}$ signifie que la classe correspondant au logit le plus élevé est choisie comme la prédiction. Pour la classe d'appartenance de x_i, l'algorithme choisit de prédire que la classe associée à x_i est $\hat{y}_i$.

Ensuite, l'algorithme calcule la précision moyenne sur l'ensemble de validation :

$$A_{val} = \frac{1}{n_{\text{val}}} \sum_i \delta(y_i = \hat{y}_i)$$

Les logits sont des valeurs brutes, résultant de la dernière couche d'un réseau de neurones avant l'application d'une fonction d'activation. Ces valeurs brutes ne sont pas normalisées et peuvent être n'importe quel nombre réel.

Cependant, avant d'obtenir les probabilités associées à chaque classe, les logits passent généralement par une fonction d'activation softmax. La fonction softmax transforme les logits en probabilités, produisant une distribution de probabilité sur les classes. Les valeurs résultantes après la fonction softmax seront dans l'intervalle [0,1], et leur somme sera égale à 1. La fonction softmax va transformer le vecteur z_i en un vecteur $z_i' = (\sigma(1)(z_{i1}), ..., \sigma(K)(z_{iK}))$ où pour tous k appartenant à $\{1, 2, \ldots, K\}$:

$$\sigma(k)(z_{ik}) = \frac{e^{z_{ik}}}{\sum_{j=1}^{K} e^{z_{ij}}}$$

On a que $\sigma(k)(z_{ik})$ est la probabilité telle qu'estimée par le réseau, que x_i appartienne à la classe k. Pour un $x_i \in \mathbb{R}^p$ la prédiction finale du modèle pour prédire quelle est la classe associée à x_i, est alors donner par $\hat{y}_i = \arg\max_k \sigma(k)(z_{ik})$ pour k appartenant à $\{1, 2, \ldots, K\}$ et la confiance de prédiction associé est définie comme étant: $\max_k \sigma(k)(z_{ik})$. De plus, on a bien :

$$\sum_{k=1}^{K} \sigma(k)(z_{ik}) = \frac{\sum_{k=1}^{K} e^{z_{ik}}}{\sum_{j=1}^{K} e^{z_{ij}}} = 1$$

Pour trouver T_{EC}, on va en fait prendre T_{EC} tel que :

$$\frac{1}{n_{\text{val}}} \sum_i \max_k \sigma(k) \left(\frac{z_{ik}}{T_{\text{EC}}} \right) = A_{\text{val}}$$

De cette manière, T_{EC} permet à ce que la probabilité maximale d'appartenir à une classe après l'application de la fonction softmax soit en accord avec la précision moyenne sur l'ensemble de validation. De cette manière on a bien que la confiance moyenne correspond à la précision moyenne sur l'ensemble de validation.

Dans la pratique

Dans la pratique, le code qui calcule la température T_{EC} est une fonction qu'on applique a un modèle de réseau de neurones et qui prend en entrée :

- les logits: qui sont les sorties brutes du réseau de neurones avant l'application de la fonction softmax.
- les labels: qui sont les étiquettes réelles associées aux données
- T_{min} qui est la température minimale à considérer (0,01)
- T_{max} qui est la température maximale à considérer (10)

La première étape consiste à calculer l'erreur de classification du modèle à l'aide de la fonction get_classification_error(logits, labels). Cela donne la mesure de la performance actuelle du modèle. Par exemple si elle vaut 0.30 cela veux dire que dans 70% des cas le modèle associe correctement les étiquettes et les données du modèle.

Ensuite, une fonction objectif est définie, notée objective(T), qui prend la température T comme paramètre. À l'intérieur de cette fonction, les logits sont divisés par la température T, puis passés à travers la fonction softmax. On extrait ensuite les probabilités maximales pour chaque exemple avec l'aide de la fonction "torch.max". La valeur de retour de la fonction objectif est la moyenne de ces probabilités maximales moins la complémentaire de l'erreur de validation c'est-à-dire la précision moyenne sur l'ensemble de validation. On utilise l'optimiseur "optimize.root_scalar" de la bibliothèque "scipy" pour trouver la racine de cette fonction dans l'intervalle spécifié par $[T_{min}, T_{max}]$.

L'Expected Calibration Error (ECE)

L'Expected Calibration Error (ECE) est une mesure d'évaluation de la calibration d'un modèle de réseau de neurones, particulièrement dans le contexte de la classification probabiliste. La calibration se réfère à la justesse des prédictions de probabilité du modèle, c'est-à-dire à quel point les probabilités prédites correspondent aux fréquences réelles des événements.

On dit qu'un algortithme de classification est calibré si la probabilité prédite $\hat{p}$, correspond à la probabilité réelle que la prédiction soit bonne. Ce qui revient mathématqiquement à:

$$P(\hat{y} = y | \hat{p} = p) = p, \forall p \in [0; 1]$$

Où $\hat{y}$ est la classe prédite et y est la vraie classe. L'erreur de calibration est une valeur qui représente la calibration du modèle sur l'ensemble des prédictions. Il s'agit de l'espérance mathématique de la différence entre la réalité et la confiance du modèle. On a donc:

$$ECE = \mathbb{E}[P(\hat{y} = y | \hat{p} = p) - p]_{\hat{p}}$$

On a donc que une valeur faible de l'ECE indique une bonne calibration, tandis qu'une valeur élevée suggère une mauvaise calibration. En effet, une ECE faible indique que le modèle a une tendance à produire des probabilités proches des véritables probabilités d'appartenance à une classe.