Fondement Théorique de l'informatique

Exposée sur les applications des mathématiques dans le domaine informatique

THEME : CONCEPT MATHEMATIQUES DERRIERE LES ALGORITHMES DE CLASSIFICATION DANS LE MACHINE LEARNING ET DANS L'ANTISPAMING

Membre du groupe

AKADE Sem Parfait
KOTY Bertrand
SEWAVI Gaël

TRAVAIL DIRIGE PAR Mr EDARH-BOSSOU

Contents

Préambule :

Les méthodes de classification sont souvent utilisées dans le machine learning afin de pouvoir réaliser des prévisions en se basant sur un certains nombres de paramètre. Le but d'un système de classification est d'effectuer la tâche de classification et de le faire avec un degré raisonnable d'exactitude. Ses méthodes découlent essentiellement des principes mathématiques. Dans cet exposée nous allons principalement voir le classificateur Naïf de Bayes.

Introduction :

Le classificateur Naïf de Bayes est un modèle d'apprentissage automatique et simple généralement utilisé dans les problèmes de classification. Le calcul derrière cela est assez facile à comprendre et les principes sous-jacents sont assez intuitifs. Pourtant, ce modèle fonctionne étonnamment bien dans de nombreux cas et ce modèle et ses variations sont utilisés dans de nombreux problèmes. Dans cet exposée, nous allons donc expliquer les mathématiques et la logique du modèle et implémenter également d'une part un classificateur Naïf Bayes en Python avec Scikit-Learn (bibliothèques basée sur le machine learning python et codée essentiellement sur le principe des algorithmes de Bayes) et montrer aussi son rôle dans le filtrage de spam d'autre part

Résumé du contenu général

Les tâches de classification dans le Machine Learning sont chargées de mapper une série d'entrées X = [x1, x2,…, xn] à une série de probabilités Y = [y1, y2,…, ym]. Cela signifie que pour un ensemble particulier d'observations X = (x1, x2,…, xn), nous devons découvrir quelle est la manière dont Y soit yi et pour obtenir une classification, il suffit de choisir le yi le plus élevé.

Nous pouvons prendre un exemple illustratif et très basique en vue d'expliquer ce qui est dit dans le dernier paragraphe.

Ayons ce tableau fictif que nous pouvons utiliser pour prédire si une ville connaîtra un embouteillage.

Temps	Temps dans la semaine	Temps dans le jour	Embouteillage
Claire	Jour ouvrable	Matin	Oui
Claire	Jour ouvrable	Déjeuner	Non
Claire	Jour ouvrable	Soir	Oui
Claire	Week-end	Matin	Non
Claire	Week-end	Déjeuner	Non
Claire	Week-end	Soir	Non
Pluvieux	Jour ouvrable	Matin	Oui
Pluvieux	Jour ouvrable	Déjeuner	Oui
Pluvieux	Jour ouvrable	Soir	Oui
Pluvieux	Week-end	Matin	Non
Pluvieux	Week-end	Déjeuner	Non
Pluvieux	Week-end	Soir	Non
Neigeux	Jour ouvrable	Matin	Oui
Neigeux	Jour ouvrable	Déjeuner	Oui
Neigeux	Jour ouvrable	Soir	Oui
Neigeux	Week-end	Matin	Oui
Neigeux	Week-end	Déjeuner	Non
Neigeux	Week-end	Soir	Oui

Donc, dans une tâche de classification, notre objectif serait de former un modèle de classificateur qui peut prendre des informations de la gauche (la météo à l'extérieur, quel genre de jour c'est et l'heure de la journée) et peut prédire si la ville connaîtra un trafic avec embouteillage.

Remarque: ce tableau semble simple, car nous n'avons que quelques points de données. Mais dans des situations du monde réel, nous aurions plus d'informations et chaque point de données aurait beaucoup plus de valeurs. J'utilise ce tableau simple pour expliquer uniquement par souci de simplicité.

Revenant à l'explication avant, nous aurons

X = [Claire, Jour ouvrable, matin]

À notre modèle et notre modèle nous reviendrait

Y = [y1, y2]

Où *y1* est la probabilité qu'il n'y ait pas d'embouteillage et *y2* est la probabilité qu'il y ait un embouteillage. Il nous suffit de choisir la probabilité la plus élevée et nous avons terminé, nous avons obtenu notre prédiction.

Une manière plus formelle d'exprimer cela est que nous devons calculer une probabilité conditionnelle, la probabilité que Y soit y étant donné que X = (x1, x2,…, xm).

P (Y = y | X = (x1, x2,…, xm))

Classificateur Naïve Bayes

Le théorème de Bayes nous dit comment nous pouvons calculer cette probabilité conditionnelle. Voyons l'équation pour cela.

$$P(A \vee B) = \frac{P(B \vee A)P(A)}{P(B)}$$

Classificateur Naïve Bayes- Le théorème de Bayes. Source: Wikipédia

- P (A | B) est une probabilité conditionnelle qui nous donne la probabilité que l'événement A se produise étant donné que l'événement B s'est produit
- P (B | A) est une autre probabilité conditionnelle et il est clair maintenant qu'elle nous donne la probabilité que l'événement B se produise étant donné que l'événement A s'est produit.
- P (A) et P (B) sont les probabilités que les événements A et B se produisent.

Ce que nous savons de la théorie des probabilités, c'est que si X1 et X2 sont des valeurs indépendantes (ce qui signifie que, par exemple, le fait que le temps soit pluvieux et qu'aujourd'hui soit un jour de week-end sont totalement indépendants, il n'y a pas de relation conditionnelle entre eux), alors nous pouvons utiliser cette équation.

$$P(Y) = P(Y) * P(X_2 \vee Y)$$

Indépendamment des probabilités
Maintenant, dans notre exemple, cette hypothèse est vraie. Il n'y a absolument aucun moyen que le fait qu'aujourd'hui soit pluvieux soit influencé par le fait qu'aujourd'hui est samedi. Mais d'une manière générale, cette hypothèse n'est pas vraie dans la plupart des cas. Si nous observons un grand nombre de variables pour une tâche de classification, il est probable qu'au moins certaines de ces variables soient dépendantes (par exemple, le niveau d'éducation et le revenu mensuel).

Mais le classificateur Naïve Bayes est appelé naïf simplement parce qu'il fonctionne sur la base de cette hypothèse. Nous considérons toutes les variables observées comme indépendantes, car l'utilisation de l'équation ci-dessus nous aide à simplifier les étapes suivantes.

Revenons donc à notre table pour voir ce qui se passe. Essayons de voir quelle est la probabilité qu'il y ait un embouteillage étant donné que le temps est clair, aujourd'hui est une journée de travail et c'est l'heure du matin (première ligne de notre tableau).

$$P(Y = Oui \,|X = (Claire, Jour\ ouvrable, matin) = \frac{P(Y = Oui) * P(X = Claire, Jour\ ouvrable, matin\,|Y = Oui}{P(X = (Claire, jour\ ouvrable, Matin)}$$

Probabilité qu'il y ait un embouteillage
Non, nous devons seulement développer cela pour pouvoir transformer cette équation en une équation contenant uniquement des probabilités de base.

$$P(Y = Oui|X = (Claire, jour\ ouvrable, Matin) = \frac{P(Y = Oui) * P(Y = Oui) * P(Y = Yes) * P(X = Matin|Y = Oui)}{P(X = (Claire, Jour\ ouvrable, matin)}$$

Équation élargie
À partir de là, nous pouvons déjà calculer chaque probabilité, comme par exemple:

$$P(Y = Yes) = \frac{\sum(Oui)}{\sum(Oui) + \sum(Non)} = \frac{10}{18} = 0.55$$

Classificateur Naïve Bayes - Probabilité qu'il y ait un embouteillage
Vous pouvez voir que cela devient déjà un processus douloureux. Vous pourriez avoir des doutes parce que l'intuition derrière ce modèle semble très simple (bien que calculer autant de probabilités puisse vous donner un mal de tête) mais cela fonctionne simplement très bien et il est utilisé dans de nombreux cas d'utilisation. Voyons certains d'entre eux.

Nous utiliserons une implémentation Scikit-Learn en Python et jouerons avec les données du tableau afin d'illustrer de façon concrète la méthode d'apprentissage de Bayes.

Dans une seconde partie nous montrerons le principe d'extraction et de classification des textes basée sur la méthode de Bayes sur

Définition

Il n'y a aucune définition exacte de spam. La plupart du spam peut être nommé comme le courrier électronique indésirable, mais pas tous les courriers électroniques indésirables sont des spam. Un autre terme serait le courrier électronique commercial non sollicité, mais spam malheureusement non seulement fait de la publicité. Le Spam peut être aussi défini comme le courrier indésirable mais il implique la question : quel est un courrier indésirable ? Bien que la plupart des utilisateurs de courrier électronique connaissent quel est le spam, mais il n'est pas évident comment définir le spam et le publipostage excessif. Wikipédia, la plus grande encyclopédie sur Internet donne les définitions suivantes : Spam : "spam de Courrier électronique (...) Implique l'envoi de messages presque identiques aux milliers (ou des millions) de destinataires."

Le spam est généralement considéré comme l'envoi massif répété de non-sollicités messages commerciaux par un expéditeur qui masque ou falsifie son identité. " [...] Le spam est défini comme la messagerie électronique non sollicité, indépendamment de son contenu. Cette définition prend en compte les caractéristiques de l'e-mail en vrac [...] ».

Le classificateur naïf bayésien

Soient ϕ = (X1,……Xj) l'ensemble des descripteurs, Y la variable à prédire (l'attribut

classe comportant K modalités). En apprentissage supervisé, pour un individu ω à classer, la règle bayésienne d'affectation optimale revient à maximiser la probabilité a posteriori d'appartenance aux classes c.-à-d. :

$$y(w) = yk \Leftrightarrow yk = \arg\max \quad P[\ Y = yk/(w)\ ]$$

La décision repose donc sur une estimation viable de la probabilité conditionnelle P(Y/X). Cette dernière peut s'écrire d'une manière différente

$$P[\ Y = yk/(w)\ ] = \frac{P[\ Y = yk/x(w))\ ]\ \ X\ \ P[\ x(w)/\ \ y=yk\ ]}{P(x(w)}$$

Comme l'objectif est de détecter le maximum de cette quantité selon yk, et que le dénominateur n'en dépend pas, nous pouvons ré écrire la règle d'affectation ci-dessus

$$y(w) = yk \Leftrightarrow yk = \arg\max \quad P[\ Y = yk\ x\ P[x(w)\ /\ y = yk\ ]\ ]$$

Hypothèse : l'indépendance conditionnelle des descripteurs La quantité P (Y = y k) est facile à estimer à partir d'un échantillon d'observations. Il suffit de calculer les proportions de chaque modalité de la variable cible. En pratique, on utilise souvent la « m probabilité estime » pour « lister » les estimations sur les petits effectifs. Si nk est le nombre d'individu de la modalité yk dans un échantillon de n observations, nous utilisons n

$$P[\ Y = yk\ ] pk = \frac{nk+m}{n+mXK}$$

Lorsque nous fixons m = 1, nous obtenons l'estimateur laplacien des probabilités.

La véritable difficulté réside finalement dans la production d'une estimation viable de la quantité $\mathbf{P[\ k(\omega)\ /\ Y = y\ k]}$. Nous sommes souvent obligés d'introduire des hypothèses pour rendre le calcul réalisable. L'analyse discriminante paramétrique stipule que la distribution est gaussienne, la régression logistique binaire Y ∈ {+, -} par sur l'idée que le rapport P appartient à une famille de lois particulière.

Dans le cadre du classificateur bayésien naïf, on considère que les descripteurs sont deux à deux indépendant conditionnellement aux valeurs de la variable cible. Par conséquent,

$$P[\ k(\omega)\ /\ Y\ =\ y\ k]\ = \prod_{j=1}^{j} \quad P[X\ j(\omega)/\ Y = yk]$$

Le nombre de paramètre à estimer est réduit de manière drastique. Pour une variable quelconque X, comportant L valeurs, nous utiliserons l'estimation suivante.

$$P[X = l/Y = yk] = p_{l/k} = \frac{n_{kl}}{n_k + m \times l}$$

Usuellement, nous fixons m = 1. On peut toujours vouloir produire une valeur « Optimale » de la constante m, mais c'est assez illusoire. Il faut surtout qu'elle soit supérieur à 0 pour éviter les probabilités estimées nulles de P(X/Y) qui auraient pour conséquence de rendre caduc le calcul de la probabilité conditionnelle $\mathbf{P[\ k(\omega)\ /\ Y = y\ k]}$

Pourquoi le classificateur bayésien naïf est-il performant ?

Le modèle bayésien naïf est un classificateur linéaire. Beaucoup s'étonnent qu'une méthode reposant sur une hypothèse (apparemment) aussi farfelue présente d'excellentes performances en prédiction, comparables aux autres techniques dont l'efficacité est reconnue. Mieux même, elle est très utilisée dans la communauté des chercheurs. Parce que d'une part elle est très facile à programmer, sa mise en œuvre est aisée ; d'autre part, parce que l'estimation de ses paramètres, la construction du modèle, est très rapide sur de très grandes bases de données, que ce soit en nombre de variables ou en nombre d'observations. Il en va autrement auprès des praticiens. Il y a beaucoup de méfiance vis-à-vis du classificateur bayésien naïf parce qu'il est mal compris. De plus, la règle d'affectation n'est pas (semble-t-il) explicite, le modèle n'est pas interprétable. On ne voit pas très bien de quelle manière chaque variable pèse sur la décision. L'interprétation des résultats n'est pas aisée. Cette opinion, largement répandue, est pourtant erronée. En se penchant attentivement sur les formules, on se rend compte que le modèle bayésien naïf est un classificateur linéaire. Il propose un biais de représentation similaire à celui de l'analyse discriminante, de la régression logistique ou des SVM (support vector machine) linéaires !!! Ce qui explique en grande partie sa bonne tenue en prédiction, comparable souvent à ces techniques

Classificateur Naïve Bayes - Applications et cas d'utilisation

- Classification en temps réel - parce que le classificateur Naïve Bayes fonctionne très très rapidement (incroyablement rapide par rapport à d'autres modèles de classification), il est utilisé dans des applications qui nécessitent des réponses de classification très rapides sur des ensembles de données de petite à moyenne taille.
- Filtrage du spam - c'est le cas d'utilisation que vous entendrez le plus souvent en ce qui concerne ce classificateur. Il est largement utilisé pour identifier si un e-mail est un spam.
- Classification de texte - le classificateur Naïve Bayes fonctionne très bien dans les méthodes de classification de texte.
- Le classificateur Naïve Bayes fonctionne généralement très bien avec la classification multi-classes et même s'il utilise cette hypothèse très naïve, il surpasse toujours les autres méthodes.

Algorithme de Bayes

Apprendre-NB (TRN)

Pour chaque classe

ci; Estimer *p(ci)*

sur TRN Pour

chaque attribut aj

Pour chaque valeur vjk de aj , estimer p (aj=vjk/ci) sur TRN

Classes—NB (TST)

Pour chaque example de TST

Pour chaque c1asser ci , estimer p (ci)$\prod$ $jp\left(aj = \frac{vjk}{ci}\right)$

Retourner ci ayant ppi maximal

Conclusion :

En sommes, Le but d'un système de classification est d'effectuer la tâche de classification et de le faire avec un degré raisonnable d'exactitude. Ses méthodes découlent essentiellement des principes mathématiques. L'un de ses principes est le classificateur naïve de Bayes qui est utilisé dans des applications qui nécessitent des réponses de classification très rapides sur des ensembles de données de petite à moyenne taille notamment dans le filtrage de spam et dans les tâches de classification dans le machine learning .

Bibliographie :

www.CoursGratuit.com
www.ICHI.PRO