LES 12 TRAVAUX D'ASTERIX - EXAM MATHS

Allakere Hormo Maxime

02/02/2021

OPTIMISATION DES HYPERPARAMETRES

Auteur Kabirou Lien du github

Introduction

Le travail s'oriente sur le domaine de la santé dans lequel on utilise des hyperparametres dans la modelisation et l'optimisation des modeles dans la resolution des pathologies, la retinopathie diabetique (RD), qui est une cause importante deficience visuelle chez les personnes agées 25 à 74 ans. POur la detection de cette pathologie on utilise defois les modeles de machines learning avec des hyperparametres, le modele en question est le HPTI-v4 appliqué pour extraire les fonctionnalités requises de l'image segmentée et il subit par la suite une classification par l'utilisation d'un perceptron multicouche (MLP).

Dans son étude, Kabirou a tenu a brievement presenté le modèle en soit et la partie optimisation des hyperparamètres et montrer son utilité pour améliorer des modèles Deep Learning en les couplant à d'autres techniques.

Égalisation adaptative de l'histogramme à contraste limité (CLAHE)

L'égalisation d'histogramme adaptatif (AHE) est une technique de traitement d'image informatique utilisée pour améliorer le contraste des images. L'AHE ordinaire a tendance à suramplifier le contraste dans les régions quasi constantes de l'image, puisque l'histogramme dans ces régions est très concentré. En conséquence, AHE peut provoquer une amplification du bruit dans des régions quasi constantes. Contrast Limited AHE (CLAHE) est une variante de l'égalisation adaptative d'histogramme dans l'amplification du contraste est limitée, de manière à réduire ce problème d'amplification du bruit.

MLP (Perceptron multicouche)

Le perceptron multicouche (multicouche perceptron MLP) est un type de réseau neuronal artificiel organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement; il s'agit donc d'un réseau à propagation directe (feedforward). Chaque couche est constituée d'un nombre variable de neurones, les neurones de la dernière couche (dite «de sortie») étant les sorties du système global.

Hyperparamètres

Dans l'apprentissage automatique, un hyperparamètre est un paramètre dont la valeur est utilisée pour contrôler le processus d'apprentissage. En revanche, les valeurs des autres paramètres (généralement la pondération de nœuds) sont obtenues par apprentissage.

Les hyperparamètres peuvent être classés comme étant des hyperparamètres de modèle, qui ne peuvent pas être déduits en ajustant la machine à l'ensemble d'entraînement parce qu'ils s'appliquent à la tâche de la sélection du modèle, ou des hyperparamètres d ' algorithmes, qui en principe n'ont aucune influence sur la performance du modèle mais affectent la rapidité et la qualité du processus d'apprentissage.

Des exemples d'hyperparamètres dans le cas du Deep Learning sont:

- **DropOut** : il s'agit d'une technique de régularisation pour éviter le surentrainement
- **Taux d'apprentissage** : il définit l'aptitude d'un réseau à mettre à jour ses paramètres
- **Nombre d'époques** : il s'agit du nombre de fois que les données d'entrainement sont présentées au réseau construit pour le processus d'entrainement
- **Batch Size** : c'est le nombre de sous-échantillons donnés au réseau après lequel la mise à jour des paramètres se produit.

Processus

Le processus de traitement se divise en deux phases, la premiere celle d'entrainement du modele et la deuxieme celle du test.

Formulation mathematiques

$$h(g_k) = \ fracn_kN, k = 0, 1, ..., L - 1).$$

Les parametres :

- $ {n_ {k}} $: nombre de pixels
- $ {g_ {k}} $: couleur g à la position k
- $ N $: nombre total de pixels

Cette equation montre le processus de segmentation basé sur un histogramme, on essaie de segmenter l'image par pixels, position etc..

nous n'allons pas decrire l'equation suivante mais il s'agit d'une equation de base d'un MLP (Multicouche Perceptron) :

$$y = f(x) = (b + \ mathbf{f}B \ times(a + \ mathbf{f}A \ timesx))$$

Evaluation et Conclusion

Tres Bon travail avec un sujet tres interessant, quelques notions seront à completer mais l'explication est fluide et la formulation mathematique devrait etre developper un peu plus meme si le sujet n'est pas evident. Je recommanderai vivement ce travail pour comprendre les hyperparametres.

LA CLASSIFICATION DE NAIVES BAYES

Auteur : Jordy Hounsinou Lien du Github

Introduction

La classification Naive Bayes est un ensemble d'algorithmes couramment utilisé dans l'apprentissage automatique. II s'agit d'une collection d'algorithmes de classification basée sur le théorème de Bayes. Son objectif est donc de pouvoir résoudre les problèmes de classification dont on fait face dans la vie courante en se basant sur des variables totalement indépendantes entre elles, d'où son appellation "Naif". Une de ses applications les plus connues est le filtre anti-spam. Pour mieux comprendre son fonctionnement, dans la suite de cet article, nous ferons dans un premier temps un zoom sur la loi de Bayes, puis nous expliquerons son fonctionnement avec des exemples succincts.

Formulations Mathematiques

La loi de Bayes

$$P(A\ midB) = \ fracP(B\ midA) * P(A)P(B)$$

avec $ A $ et $ B $ des événements, $ P (A) $ la conversion de $ A $ et $ P (A \ m id B) $ la conversation conditionnelle de A sachant B.

Dans cette illustration, j'ai preferé maintenir les mots utilisés par l'auteur pour pas modifier le sens de son exemple et garder l'idée principale.

Illustration : Pour mieux expliquer ce phénomène, nous allons utiliser les circonstances sanitaires actuelles pour les tests de dépistage Covid-19 sont devenus ordinaires. Nous allons supposer les faits suivants:

- 1 personne sur 1000 attrape le covid-19
- La précision du test génique PCR est de 99 $ % $

Une personne devant voyager décide de faire le test recommandé et malheureusement se retrouve avec un test positif. Quelle est la conversation qu'elle soit vraiment porteuse du covid-19? On aurait tout de suite tenté de dire qu'elle est très forte $ (\ m athrm {P}> 50 %) $ au vu de la précision du test. Pourtant, la conversation à priori fausse totalement cette pensée.

Si on est d'avance certain de ne pas avoir le virus, le fait d'avoir un résultat positif fait d'avantage penser qu'on est dans les $ 1 % $ de marge d'erreur et pas le contraire. C'est donc pour cela qu'il est nécessaire de prendre en compte la conversation à priori qui est dans notre cas, 1 personne sur 1000 contracte le virus. La bonne démarche est la suivante:

Soit A l'évènement avoir le covid-19. B l'évènement résultat de test positif $ P (A) = 0,001 $ $ P (B \ m id A) = 0,99 $ Avec $ \ o verline { \ t ext {A}} $ complémentaire de $ \ m athrm {A}, $ sur un

$$begintableaurP(B) = P(B\ midA)*P(A)+P(B\ mid\ barA)*P(\ barA)P(B) = 0,99^*0,001+0,01^*0,999P(B)\ equiv0,01098P(A$$

On se rend compte que la publication d'avoir le virus sachant que le test est positif est de $ 9 % , $ et qu'il est très faible par rapport à la pensée eue de prime abord.

L'application de ce probleme avec cette illustration sur le covid revient donc a formaliser tout cela sous la loi de Bayes avec la formule ci-dessous :

$$\ mathrmP(\ mathrmC\ midF1,\ ldots, Fn) = \ frac\ mathrmP(\ mathrmC) * \ \ mathrmP\ left(\ left.\ MathrmC\ right|_{F1,\ ldots,$$

où C est une variable de classe dépendante dont les instances ou classes sont peu nombreuses, conditionnée par plusieurs variables caractéristiques $ F_ {1}, \ l points, F_ {n} $

Evaluation et Conclusion

Un travail tres bien expliqué accompagné d'une illustration sur un sujet d'actualité qui nous fait saisir l'utilité de la classification de Bayes. Rien à ajouter ou à retirer à ce travail, parfait.

REGRESSION LINEAIRE SIMPLE ET MULTIPLE

Auteur : Nina ZOUMANIGUI Lien du Github

Introduction

Il s'agit d'une notion tres importante en mathematiques, elle permet de voir la variation des variables quantitaives ou qualitatives et d'en deduire la cause et consequence defois. cette methode sert à etudier des variables qu'on suppose liées, le travail effectué par Nina est de mettre en reflief la regression dite lineaire simple et lineaire multiples.

Regression lineaire simple

le modele de regression lineaire simple est sans doute l'une des methodes de regression les plus connues ou l'on etudie la variation de deux variables de façon simple et classique.

L'on emet dans le cadre des etudes de test en statistiques, des hypotheses, et dans le cadre d'une regression l'hypothe initiale dite nulle est celle selon laquelle il y aurait pas de relation entre les variables contre l'hypothese alternative selon laquelle les variables seraient liées ou dependantes.

Pour illustrer un graphique de regression, l'auteur a laissé un jeu de données denommé "Examen" sur son github qu'il utilise pour tracer un graphique de regression avec ce bloc de code ci-dessous :

```
# g=read.csv("C:\\Users\\ninaz\\OneDrive\\Documents\\Exament\\Examen.csv",sep=";", header=T)
# attach(g)
#
# #definir les variables
#
# y=g$Age
# x=g$Elevel
# plot(y,x)
#
# #Autre application
#
# data=read.csv("C:\\Users\\ninaz\\OneDrive\\Bureau\\R\\mtcars.csv",sep = ";", header= T)
# attach(data)
# head(data)
#
# plot(mpg~wt,pch=20)
# fit= lm(mpg~wt,data=data)
# fit
# abline(fit,col="red",lwd=2)
```

Regression multiple lineaire

Ce modele de regression contrairement à la premiere n'est pas delimité par un nombre fini de variable à prendre en compte, soit : $yi = b0 + b1xi1 + b2xi2 + ůůů + bpxip + i, i = 1, ůůů, n.$

Evaluation et Conclusion

Travail tres interessant sur des modeles tres utilisés en maths en revanche l'accent a plus été mis sur l'explication et la comparaisons des deux sans inclure leur formalisation mathematiques et l'utilité réelle de ces modeles.

CRYPTOGRAPHIE ET THEORIE DES NOMBRES

Auteurs : Arsic Marko et RObache William Lien du Github

Introduction

La crypto est une science qui a su trouvé son utilité au fur et à mesure chez le grand public autrefois reservé au domaine militaire, aujourd'hui ce sont les banques, les reseaux sociaux, les chaines de télé etc. qui utilisent cette technologie pour crypter les données de leurs clients ou crypter certains services pour une question de securité et confidentialité.

les auteurs ont fait un travail decrivant les types de cryptage, entre autre :

Cryptage à clé symetrique

il s'agit d'un systeme de cryptage ou la seule solution de la decrypter est de reutiliser la clé de cryptage initiale.

Codage cesar Il s'agit du plus ancien des systemes qui tire son nom de Jules Cesar, un systeme de cryptage basé sur la codification de l'alphabet.

Codage Vigenere Le codage vigenere est une complication du codage de Cesar, meme principe qui differe du fait que les lettres sont decalés avec un modulo 3 c'est a dire un rassemble de 3 lettres par trois.

Principe de chiffrement

Le récepteur du message annonce publiquement sa fonction de cryptage $ f_C $ et conserve dans un lieu secret sa fonction de décryptage $ f_D $.
Tout émetteur peut alors envoyer un message $ m $ en procédant ainsi:
Au lieu d'envoyer le message $ m $ par la poste en prenant le risque que quelqu'un n'intercepte le message, il envoie $ M = f_C (m) $.
Ainsi, toute personne interceptant ce message ne peut le lire, le seul récepteur peut décrypter ce message en utilisant sa fonction de décryptage privée $ f_D $.
Effectivement,

$$f_D(M) = f_D[f_C(m)] = m$$

Nous avons dans le travail de l'auteur une cuite de principe et systeme de chiffrement.

Evaluation et conclusion

Une branche des maths qui fait enorment parler d'elle et donc son utilité est visible et important aujourd'hui (chiffrement de nos discussions sur whatsapp etc.), le travail a été bien expliqué de façon large en et super fluide avec des exemple en revanche manque de formulation mathemqatiques pour comprendre l'importance des formules de chiffrements ou codages.

REGRESSION SUR UNE VARIABLE FONCTIONNELLE

Auteur : Allakere Hormo Maxime Lien du Github

Introduction

Dans la nature, nous avons des données dit de nature numeriques, alphanumeriques etc. qu'on arrive à etudier facilement lorsque ceux-ci sont finis mais nous avons des types de données dit fonctionnelles ce sont par exemple de données de type meteorologique, medicales, imagerie satellite etc. et cesc données font l'objet des plusieurs etudes de nos jours, nous allons donc plus nous attarder sur l'aspect regression et test de structure de ces données. pour mieux comprendre ces notions mathematiques de regression, test de structure etc. j'ai dans mon travail essayé de vulgarisé au maximum ces termes scientifiques.

Les methodes d'analyses de données fonctionnelles

Un peu dit precedemment dans l'introduction, il existe plsuieurs types d'analyses de données fonctionnelles, des analyses factorielles, des analyses exploratoires (etude de moyenne, mediane, ecart-type etc.), la regression celui sur lequel j'ai basé mon travail.

Modele de regression sur variable fonctionnelle

Un modèle de régression s'écrit donc de cette façon: $ Y = r (X) + epsilon $

la variable Y est la valeur réelle tandis que la variable explicative X est à valeur dans un espace semi-métrique (E, d) de dimension infinie. $ epsilon $ représente le résidu.

Modele Parametriques et non parametriques de regression sur variable fonctionnelles

Nous allons présenter des modèles de régressions sur la variable fonctionnelle au travers des estimateurs paramétriques et non paramétriques.

Considérons les trois équations suivantes:

(1) `$ Y = aX + b + \epsilon , X \sim N( \mu , \gamma^2 ) $ et $ N ( 0 , \sigma^2 ) $`

(2) `$ Y = aX + b + \epsilon , E ( \epsilon | X) = 0 $`

(3) `$ Y = r (X) + \epsilon , r \ dans C ( \RR ) $ et $ E ( \epsilon | X) = 0 $`

L'équation (1) est une équation paramétrique, l'équation (3) est une équation non-paramétrique tandis que la (2) est une équation paramétrique et non paramétrique selon les deux points de vue suivant:

- Non-Paramétrique si la loi du couple (X, Y) est supposé appartenir à un espace indexé par un nombre fini de paramètres réels.
- Paramétrique si le traitement de régression est supposé appartenir à un espace indexé par un nombre fini de paramétriques réels.

Test de structure pour variable fonctionnelle

Nous avons souligné le fait qu'il existe très peu de résultats concernant les tests de structures en régression sur la variable fonctionnelle, on ne trouve que très peu d'articles scientifiques qui développent le sujet.

Il existe pas une approche théorique générale permettant de tester si un modèle possède une structure particulière: linéaire, à indice simple etc. Nous allons donc nous pencher sur le cas le plus simple ou l'on teste une hypothèse nulle soit H0 et voir si cette hypothèse est acceptée ou rejetée, en cas de rejet de l'hypothèse alternative H1 est acceptée.

Tester si $ r = r_ {0} $ - le cas ou l'on souhaite tester l'hypothèse nulle.

$ H_ {0}: (PP (r (X) = r_ {0} (X)) = 1)

Ou $ r_ {0} $ est un opérateur connu contre l'alternative locale H 1 :

$ H_ {1}: (|| r-r_ {0} || _ {L ^ 2 (wdP_ {x})} geq eta _ {n})

Cela revient à faire un test de non effet, la façon dont $ eta _ {n} $ tend vers 0 la capacité de notre test à détecter les différences de plus en plus petites entre r et $ r_ {0 } $ lorsque n croit.

Evaluation et conclusion

J'ai essayé le plus possible de vulgariser ce concept relativement compliqué de variable fonctionnelle et etre plus litteraire sur le contenu que la formalisation mathematiques et en toute objectivité, le travail aurait été complet si j'avais inserer des applications de test sur ces variables fonctionnelles.

NB: quelques soucis à compiler les packages sur Latex pour afficher certaines formules vu que j'ai realisé mon travail sur rmarkdown et par consequent se referer à mon travail sur le github concernant les variables fonctionnelles.