Класификација применом логистичке регресије и методе к најближих суседа на проблем одређивања кулинарског рецепта на основу састојака

Студент: Јован Гашпар, IN60/2018 , e-mail:jbutea@gmail.com

I. Увод

Мотивација за израду овог задатка је обучавање класификатора који ће на основу задатих састојака моћи да процени која је кухиња у питању. Ово истраживање је првенствено намењено љубитељима разних кухиња и куварима који би уносом сопствених рецепата могли да упореде своје укусе са укусима других народа света. Читаоци такође могу наћи инспирацију за своје будуће кулинарске подвиге.

II. База података

База представља скуп рецепата који припадају различитим националним кухињама. У бази се налази укупно 10565 узорака. Сваки од узорака саджи обележја која се односе на то да ли рецепт садржи одређени састојак или не. Таква овележја садрже вредности „0“ уколико рецепт не садржи одређени састојак и „1“ уколико рецепт садржи одређени састојак. Таквих обележја укупно има 150. База садржи још једно обележје*(назива:„country“)* које носи назив имена одређене националне кухиње. То је категоричко обележје чију ће вредност обучени модели да предвиђају.

У задатој бази је било 9 различитих назива кухиња:„јужноамеричка“;
„британска“;„таиландска“;„мексичка“;„јапанска“;
„кинеска“;
„италијанска“;„грчка“ и „француска“.

III. Подела података

Улазни подаци су након прегледа базе подељени троструком поделом на три подскупа: подскуп над којим ће бити обучен модел, валидациони подскуп и тест подскуп. Ова три подскупа су дисјунктивна и омогућавају проверавање перформанси модела на валидационом скупу пре обучавања финалног модела.

Након поделе података извршена је фаза обуке међумодела за обе класификационе методе.

IV. Обучавање модела логистичке регресије

Логистичка регресија за решавање класификационог проблема своди се на предвиђање апостериорне вероватноће класа за неки узорак.

У случају решавања вишекласног проблема као што је овај потребно је поделити излазни опсег вредности на више подопсега при чему се подразумева неки поредак међу њима.

У овом случају је можда могао да постоји географски поредак али није узет у обзир приликом обуке модела.

Функција која се користи да испита да ли одређени узорак припада некој класи је логистичка функција дефинисана изразом:

$$g(x)=1/(1+e^{**}(-x))$$

Овај проблем се може решити следећим методом опадања градијента. Израз за одређивање функције цене се своди на решавање следећег проблема:

$$J(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)}\log\left(h_{\theta}(\mathbf{x}^{(n)})\right) + (1-y^{(n)})\log(1-h_{\theta}(\mathbf{x}^{(n)}))\right)$$

Код методе *логистичке регресије* кориштени су следећи параметри:

НАЗИВ ПАРАМЕТРА	СКУП ВРЕДНОСТИ
***Num**:број итерација*	**{100,200,500,1000}**
***Solver:**оптимизациони алгоритам*	**{’newton-cg’, ’lbfgs’,’sag’,’saga’}**
Way: *принцип решавања проблема логистичке регресије за вишекласни проблем*	**{‘ovr’,’multinomial’}**

Solver решава проблем опадања градијента на различите начине.

Way примењује принципе „један протиц свих“ и мултиномијалну логистичку регресију.

Num представља број итерација за које ће ***Solver*** конвергирати.

С обзиром да је у питњању вишекласни проблем примењује се више модела логиситчке регресије.

За сваки од горенаведених параметара дефинисан је класификациони модел који садржи по један елемент из скупова (***NUM,SOlVER,WAY***)

Односно класификациони модел логистичке регресије представља тројку M=(N,S,W) где важи:

(N ∈ ***NUM)***, (S ∈***SOLVER),(***W ∈***WAY).***

V. Одабир финалног модела применом методе логистичке регресије

Након обуке свих модела извршен је одабир модела

Са највећом Ф мером која представља хармонијску средину прецизности и осетљивости.

(осетљивост представља удео исправно класификованих из класе позитива)

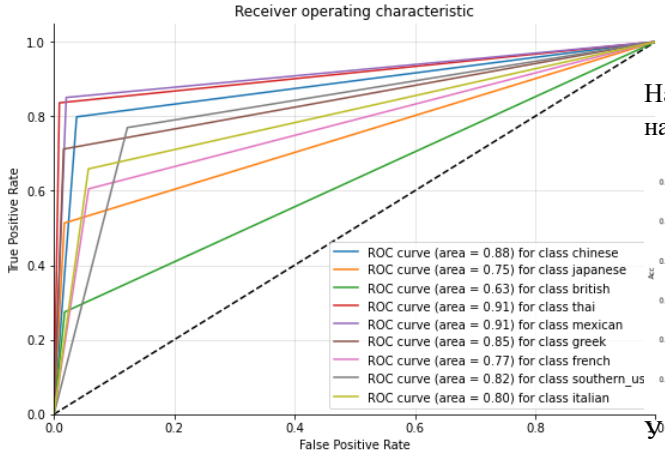
(прецизност= удео исправно класификованих узорака међу узорцима предвиђеним као позитивним)

Треба напоменути да је скоро половина изгенерисаних модела конвергирала до истих вредности. Ипак као финални модел је одабран један од њих то јест следећи модел: **M=(1000,'lbfgs','multinomial')**

На слици испод имамо приказ оперативне карактеристике(*енгл. Receiver operating characteristic*)

Која приказује зависност између исправно класификованих позитива у свим позитивима

(*енгл. True positive rate*) од удела лажних позитива у свим негативима. (*енгл. False positive rate)*



Слика 1. Оперативне карактеристике: површина испод сваке криве треба да буде што већа (*енгл. area*)
то јест да тежи јединици. Како је урађен приказ за све класе паралелно јасно се види да се мексичка и таиландска кухиња најлакше препознају

Мере успешности финалног модела логистичке регресије:

Проценат погођених узорака: 0.6991
Прецизност микро: 0.6991
Прецизност макро: 0.6967
Осетљивост микро: 0.6991
Осетљивост макро: 0.6686
Ф мера микро: 0.6991
Ф мера макро: 0.6781

VI. Обучавање модела К најближих суседа

Класификација на основу К најближих суседа се своди на то да се проверава класа к најближих узорака и на основу тога се додељује класа посматраном узорку.

Код методе к најближих суседеа кориштени су следећи параметри:

НАЗИВ ПАРАМЕТРА СКУП ВРЕДНОСТИ

Број суседа (NN)	**{1,2,3,4,5,6,7,8,9,10}**
Алгоритам (SOLVER)	**{'ball_tree','brute'}**
Метрика (METRIC)	**{'hamming','euclidean'}**
Тежине (WEIGHTS)	**{'uniform','distance'}**

Алгоритам представља начин на који ће се примењивати метрика при одабиру најближег суседа.

Метрика дефинише одабрану функцију растојања. Тежинама се подеришу суседни узорци. Овај параметар служи за одређивњае значаја за сваког од к суседа у одлучивању.

За сваки од горенаведених параметара дефинисан је класификациони модел који садржи по један елемент из скупова (NN,SOlVER,METRIC,WEIGHTS)

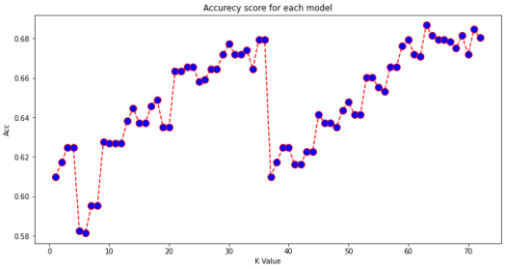
Односно класификациони модел логистичке регресије представља четворку M=(K,S,W,Me) где важи:

(K ∈ NN), (S ∈SOLVER),(W ∈WEIGHTS),

(Me ∈ METRICS).

VII. Одабир финалног модела применом методе К најближих суседа

Након обуке свих модела извршен је одабир модела са највећом микро прецизношћу.



У овом случају постоји модел који се показао боље у одсносу на друге моделе.

У питању је следећи модел:

M1=(K=7,S='brute',W='distance',Me='hamming')

Мере успешности финалног модела к најближих суседа:

Проценат погођених узорака: 0.6866
Прецизност микро: 0.6866
Прецизност макро: 0.7025
Осетљивост микро: 0.6866
Осетљивост макро: 0.6369

Ф мера микро: 0.6866
Ф мера макро: 0.6571

VIII. Закључак

Након завршених обука модела за обе класификационе методе поставља се једно здраворазумско питање: ,,Који је модел бољи и због чега?“ Зависи од скупа података на ком се модел примењује. Ако посматрамо случај логистичке регресије, модел М, приликом проширења тренинг скупа М неће дати боље резултате од ових који су дати изнад. За разлику од модела М, модел М1 за проширен тенинг скуп даје боље резултате. Међутим ако се не мењају тренинг скупови остаје бољи модел М.

У овом конкрентном случају с обзиром да се могу појавити и нови рецепти који би могли бити укључени у базу или чак када се би се додале и нове националне кухиње класификатор к најближих суседа би се показао боље.