TRƯỜNG ĐẠI HỌC CẦN THƠ

KHOA CNTT & TRUYỀN THÔNG

BỘ MÔN KHOA HỌC MÁY TÍNH

KHAI KHOÁNG DỮ LIỆU

Chủ đề:

PHÂN LOẠI NẤM

Giảng viên hướng dẫn:

TS. Lưu Tiến Đạo

Sinh viên thực hiện:

1. Huỳnh Thanh Toàn (B1404453)
2. Phạm Công Tâm (B1609841)

HỌC KỲ 2, NH 2019-2020

NỘI DUNG

CHƯƠNG I: GIỚI THIỆU

1. Đặt vấn đề
2. Mô tả dữ liệu
3. Mục tiêu
4. Phương pháp đánh giá

CHƯƠNG II: MÔ TẢ GIẢI THUẬT

1. Mô tả giải thuật
2. Vấn đề liên quan đến bài toán
3. Giải pháp cho bài toán

CHƯƠNG III: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận
2. Hướng phát triển

CHƯƠNG IV: TÀI LIỆU THAM KHẢO

1. Tài liệu tham khảo

CHƯƠNG I: GIỚI THIỆU

1. Đặt vấn đề

a. Lý do chọn đề tài

- Nấm là một loại thực phẩm rất phổ biến, tuy nhiên, ngoài các loại nấm được sử dụng cho các bữa ăn còn có các loại nấm gây độc. Việc phân biệt giữa nấm độc và không độc rất khó, thậm chí là không thể phân biệt được, nhất là các loại nấm mọc hoang ở vườn, ruộng, nấm hái trong rừng...

- Rất nhiều người thường có những lầm tưởng về nấm độc như nấm độc là loại nấm có màu sặc sỡ; nấm bị sâu bọ ăn là nấm không độc; thử nấm bằng thìa, đũa, dây chuyền có thể phát hiện nấm độc; thử cho động vật (chó, mèo) ăn sau 1 – 2 giờ, nếu không có vấn đề gì thì đó là nấm không độc… Đây là nhận định hoàn toàn sai lầm.

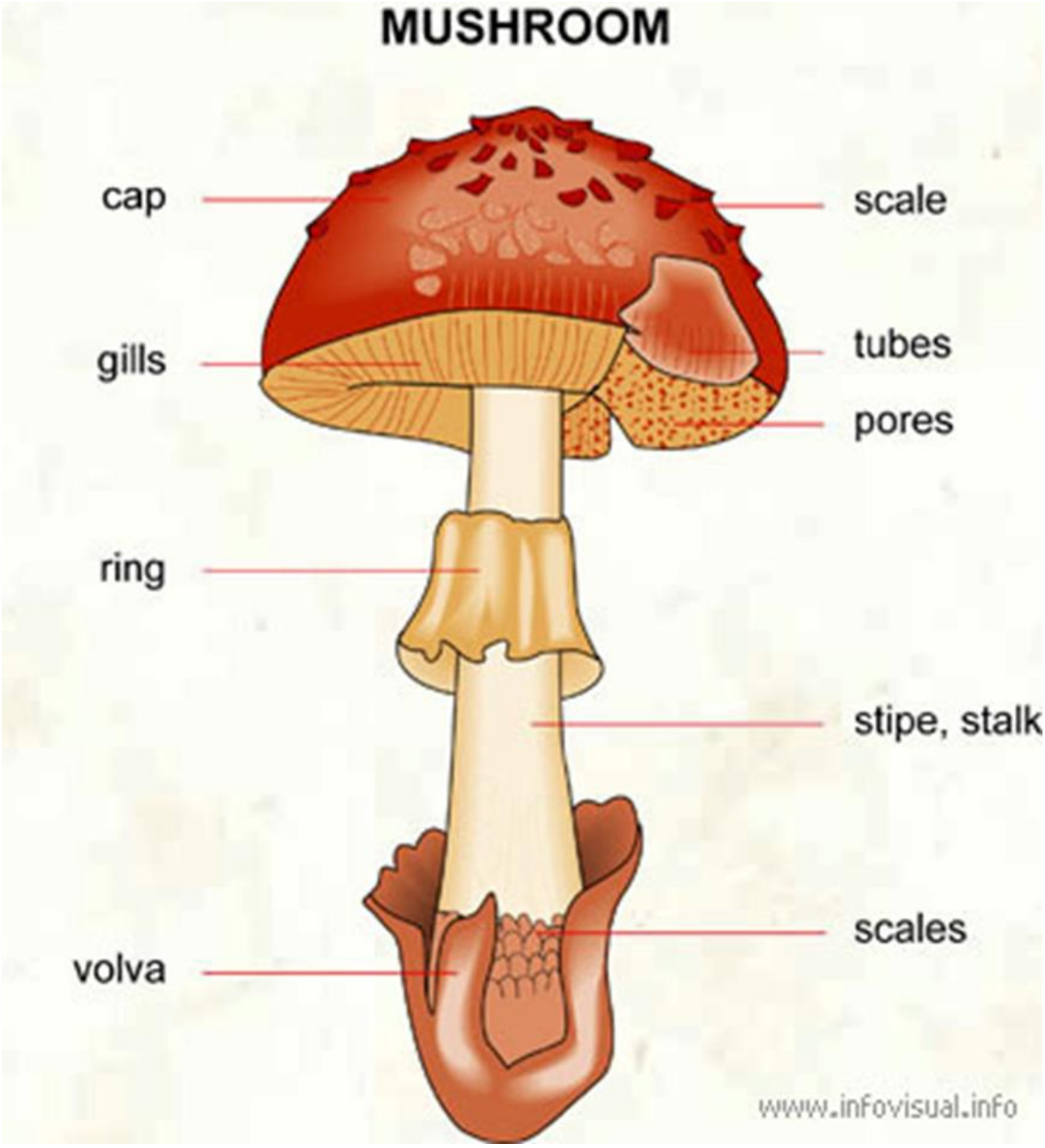
- Để hiểu rõ và xác định cái gì làm cho nấm ăn được hoặc không ăn được, nhóm đã quyết định chọn đề tài phân tích, khai thác dữ liệu về các đặc tính của nấm.

2. Mô tả dữ liệu:

a. Giới thiệu chung

- Đây là bộ dữ liệu mô tả về các mẫu giả định tương ứng với 23 loài nấm được nghiên cứu bởi The Audubon Society Field Guide về các loại nấm Bắc Mĩ (1981). Bộ dữ liệu này nghiên cứu và cho ra kết luận về khả năng ăn được hoặc độc hại của từng mẫu thử nấm.

- Bộ dữ liệu gồm 8124 dòng dữ liệu và 22 cột thuộc tính.



b. Kiểu dữ liệu

STT	Tên thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	classes	Phân loại	String
2	cap-shape	Hình dạng mũ nấm	String
3	cap-surface	Bề mặt mũ nấm	String
4	cap-color	Màu mũ nấm	String

5	bruises	Vết thâm	String
6	odor	Mùi hương	String
7	gill-attachment	Lá tia đính kèm	String
8	gill-spacing	Mật độ lá tia	String
9	gill-size	Kích cỡ lá tia	String
10	gill-color	Màu lá tia	String
11	stalk-shape	Hình dạng cuống	String
12	stalk-root	Hình dạng cuống rễ	String
13	stalk-surface-above-ring	Bề mặt cuống trên vòng	String
14	stalk-surface-below-ring	Bề mặt cuống dưới vòng	String
15	stalk-color-above-ring	Màu cuống trên vòng	String
16	stalk-color-below-ring	Màu cuống dưới vòng	String
17	veil-type	Loại mạng	String
18	veil-color	Màu mạng	String
19	ring-number	Số vòng	String
20	spore-print-color	Màu bào tử	String
21	population	Mật độ	String
22	Habitat	Môi trường sống	String

c. Tiền xử lý dữ liệu

Thuộc tính	Các giá trị		
	Dữ liệu gốc	Dữ liệu sau khi tiền xử lý	Ý nghĩa
class	e	edible	ăn được
	p	poisonous	không ăn được, có độc
cap-shape	b	bell	hình chuông
	c	conical	hình nón

	x	convex	hình lồi
	f	flat	bằng phẳng
	k	knobbed	hình dáng bị đâm thủng
	s	sunken	trủng xuống
cap-surface	f	fibrous	có thớ sợi
	g	grooves	rãnh đường khuyết
	y	scaly	vảy
	s	smooth	trơn
cap-color	n	brown	nâu
	b	buff	vàng sẫm
	c	cinnamon	nâu vàng
	g	gray	xám
	r	green	xanh lá
	p	pink	hồng
	u	purple	tím
	e	red	đỏ
	w	white	trắng
	y	yellow	vàng
bruises	t	bruises	có vết thâm
	f	no	không có
odor	a	almond	mùi hạnh nhân
	l	anise	hương cây hồi
	c	creosote	creosote
	y	fishy	mùi tanh
	f	foul	mùi hôi
	m	musty	mốc
	n	none	không mùi
	p	pungent	vị cay
	s	spicy	có gia vị
gill-attachment	a	attached	đính kèm
	d	descending	hướng xuống
	f	free	tự do
	n	notched	vết nứt
gill-spacing	c	close	dày đặc
	w	crowded	chật nít
	d	distant	rải rác
gill-size	b	broad	rộng
	n	narrow	hẹp
gill-color	k	black	đen
	n	brown	nâu
	b	buff	vàng sẫm
	h	chocolate	màu socola
	g	gray	xám
	r	green	xanh lá
	o	orange	cam

	p	pink	hồng
	u	purple	tím
	e	red	đỏ
	w	white	trắng
	y	yellow	vàng
stalk-shape	e	enlarging	to lớn
	t	tapering	thon dài
stalk-root	b	bulbous	hình củ
	c	club	chum
	u	cup	hình chén
	e	equal	hình dạng giống nhau
	z	rhizomorphs	hình dạng sợi nấm
	r	rooted	ăn sâu vào đất
	?	missing	không thấy
stalk-surface-above-ring	f	fibrous	có sợi thớ
	y	scaly	vảy
	k	silky	mềm mịn
	s	smooth	trơn
stalk-surface-below-ring	f	fibrous	có sợi thớ
	y	scaly	vảy
	k	silky	mềm mịn
	s	smooth	trơn
stalk-color-above-ring	n	brown	nâu
	b	buff	vàng sẫm
	c	cinnamon	nâu vàng
	g	gray	xám
	r	green	xanh lá
	p	pink	hồng
	u	purple	tím
	e	red	đỏ
	w	white	trắng
	y	yellow	vàng
stalk-color-below-ring	n	brown	nâu
	b	buff	vàng sẫm
	c	cinnamon	nâu vàng
	g	gray	xám
	r	green	xanh lá
	p	pink	hồng
	u	purple	tím
	e	red	đỏ
	w	white	trắng
	y	yellow	vàng
veil-type	p	partial	Một phần
	u	universal	Toàn phần
veil-color	n	brown	nâu

	o	orange	cam
	w	white	trắng
	y	yellow	vàng
ring-number	n	none	không có
	o	one	một
	t	two	hai
ring-type	c	cobwebby	mạng nhện
	e	evanescent	không hiện rõ
	f	flaring	hiện rõ
	l	large	rộng lớn
	n	none	không có
	p	pendant	hình mặt dây chuyền, tua
	s	sheathing	bao bên ngoài
	z	zone	phân theo từng vùng
spore-print-color	k	black	đen
	n	brown	nâu
	b	buff	vàng sẫm
	h	chocolate	màu socola
	r	green	xanh lá
	o	orange	cam
	p	pink	hồng
	u	purple	tím
	w	white	trắng
	y	yellow	vàng
population	a	abundant	phong phú
	c	clustered	từng đám
	n	numerous	nhiều
	s	scattered	rải rác
	v	several	một vài
	y	solitary	cô độc, riêng lẻ
habitat	g	grasses	có cỏ
	l	leaves	có lá
	m	meadows	có đồng cỏ
	p	paths	gần đường mòn
	u	urban	thành thị
	w	waste	ô nhiễm
	d	woods	trong rừng

d. Dữ liệu gốc

class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ring
p	x	s	n	t	p	f	c	n	k	e	e	s
e	x	s	y	t	a	f	c	b	k	e	c	s
e	b	s	w	t	l	f	c	b	n	e	c	s
p	x	y	w	t	p	f	c	n	n	e	e	s
e	x	s	g	f	n	f	w	b	k	t	e	s
e	x	y	y	t	a	f	c	b	n	e	c	s
e	b	s	w	t	a	f	c	b	g	e	c	s
e	b	y	w	t	l	f	c	b	n	e	c	s
p	x	y	w	t	p	f	c	n	p	e	e	s
e	b	s	y	t	a	f	c	b	g	e	c	s
e	x	y	y	t	l	f	c	b	g	e	c	s
e	x	y	y	t	a	f	c	b	n	e	c	s
e	b	s	y	t	a	f	c	b	w	e	c	s
p	x	y	w	t	p	f	c	n	k	e	e	s
e	x	f	n	f	n	f	w	b	n	t	e	s
e	s	f	g	f	n	f	c	n	k	e	e	s
e	f	f	w	f	n	f	w	b	k	t	e	s
p	x	s	n	t	p	f	c	n	n	e	e	s
p	x	y	w	t	p	f	c	n	n	e	e	s

e. Dữ liệu sau khi tiền xử lý

class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ring
poisonous	convex	smooth	brown	bruises	pungent	free	close	narrow	black	enlarging	equal	smooth
edible	convex	smooth	yellow	bruises	almond	free	close	broad	black	enlarging	club	smooth
edible	bell	smooth	white	bruises	anise	free	close	broad	brown	enlarging	club	smooth
poisonous	convex	scaly	white	bruises	pungent	free	close	narrow	brown	enlarging	equal	smooth
edible	convex	smooth	gray	no	none	free	crowded	broad	black	tapering	equal	smooth
edible	convex	scaly	yellow	bruises	almond	free	close	broad	brown	enlarging	club	smooth
edible	bell	smooth	white	bruises	almond	free	close	broad	gray	enlarging	club	smooth
edible	bell	scaly	white	bruises	anise	free	close	broad	brown	enlarging	club	smooth
poisonous	convex	scaly	white	bruises	pungent	free	close	narrow	pink	enlarging	equal	smooth
edible	bell	smooth	yellow	bruises	almond	free	close	broad	gray	enlarging	club	smooth
edible	convex	scaly	yellow	bruises	anise	free	close	broad	gray	enlarging	club	smooth
edible	convex	scaly	yellow	bruises	almond	free	close	broad	brown	enlarging	club	smooth
edible	bell	smooth	yellow	bruises	almond	free	close	broad	white	enlarging	club	smooth
poisonous	convex	scaly	white	bruises	pungent	free	close	narrow	black	enlarging	equal	smooth
edible	convex	fibrous	brown	no	none	free	crowded	broad	brown	tapering	equal	smooth
edible	sunken	fibrous	gray	no	none	free	close	narrow	black	enlarging	equal	smooth
edible	flat	fibrous	white	no	none	free	crowded	broad	black	tapering	equal	smooth
poisonous	convex	smooth	brown	bruises	pungent	free	close	narrow	brown	enlarging	equal	smooth
poisonous	convex	scaly	white	bruises	pungent	free	close	narrow	brown	enlarging	equal	smooth
poisonous	convex	smooth	brown	bruises	pungent	free	close	narrow	black	enlarging	equal	smooth
edible	bell	smooth	yellow	bruises	almond	free	close	broad	black	enlarging	club	smooth

3. Mục tiêu:

- Huấn luyện mô hình phân loại nấm có độc và không độc
- So sánh độ chính xác của các giải thuật

- Đánh giá mô hình thông qua biểu đồ

4. Phương pháp thực hiện

Bước 1: Chuẩn bị tập dữ liệu huấn luyện và rút trích đặc trưng. Công đoạn này được xem là công đoạn quan trọng trong các bài toán về ML. vì đây là input cho việc học đẻ tìm ra mô hình của bài toán. Chúng ta phải biết cần chọn ra những đặc trưng tốt của dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu. Ước lượng số chiều của dữ liệu bao nhiêu là tốt hay nói cách khác là chọn bao nhiêu feature. Nếu số nhiều quá lớn gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn giữ được độ chính xác của dữ liệu. Ở bước này chúng ta cũng chuẩn bị bộ dữ liệu để test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập dataset thành hai phàn, một phần phục vụ cho training và phần còn lại phục vụ cho mục đích testing trên mô hình. Có hai cách thường sử dụng trong cross-validation là splitting và k-fold.

Bước 2: Xây dựng mô hình phân lớp Mục đích của mô hình huấn luyện là tìm ra hàm F(x) và thông qua hàm f tìm được để chúng ta gán nhãn cho dữ liệu. Bước này thường được gọi là học hay training.

$$F(x)= y$$

Trong đó: x là các feature hay input đầu vào của dữ liệu Y là nhãn dán lớp hay output đầu ra Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát như KNN, NN, SVM, Decision tree, navies bayers

Bước 3: Kiểm tra dữ liệu với mô hình

Bước 4: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất Bước cuối cùng chúng ta sẽ đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu testing và dữ liệu training thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn của chúng ta thì phải thay đổi các tham số của thuật toán học để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. và cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán của chúng ta

CHƯƠNG II: MÔ TẢ GIẢI THUẬT

1. Mô tả giải thuật

*** Cây quyết định**

a. Định nghĩa

- Cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng.

- Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó.

- Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó.

- Các kiểu cây quyết định

+ Cây hồi quy (Regression tree)

+ Cây phân loại (Classification tree): phân loại nấm áp dụng cây phân loại (có độc và không độc)

b. Công thức

- Gini impurity:

+ Dùng trong thuật toán CART (Classification and Regression Trees). Nó dựa vào việc bình phương các xác suất thành viên cho mỗi thể loại đích trong nút. Giá trị của nó tiến đến cực tiểu (bằng 0) khi mọi trường hợp trong nút rơi vào một thể loại đích duy nhất.

+ Giả sử y nhận các giá trị trong {1, 2, ..., m} và gọi f (i, j) là tần suất của giá trị j trong nút i. Nghĩa là f (i, j) là tỉ lệ các bản ghi với y=j được xếp vào nhóm i.

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2$$

- Entropy:

+ Dùng trong các thuật toán sinh cây ID3, C4.5 và C5. Số đo này dựa trên khái niệm entropy trong lý thuyết thông tin (information theory)

$$I_E(i) = -\sum_{j=1}^{m} f(i,j) \log_2 f(i, approxeq. dashv. sdot)$$

c. Ưu và nhược điểm

- Ưu điểm:

+ Cây quyết định dễ hiểu.

+ Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết.

+ Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại.

+ Có thể thẩm định một mô hình bằng các kiểm tra thống kê.

+ Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn.

- Khuyết điểm:

+ Khó giải quyết được những vấn đề có dữ liệu phụ thuộc thời gian liên tục.

+ Dễ xảy ra lỗi khi có quá nhiều lớp chi phí tính toán để xây dựng mô hình cây quyết định cao.

*** K láng giềng (KNN)**

a. Định nghĩa

- K-nearest neighbor (KNN) là một trong những thuật toán học có giám sát đơn giản nhất trong Machine Learning. Ý tưởng của KNN là tìm ra output của dữ kiệu dựa trên thông tin của những dữ liệu training gần nó nhất.

b. Quy trình làm việc thuật toán k láng giềng

- Bước 1: xác định tham số K= số láng giềng gần nhất.

- Bước 2: tính khoảng cách đối tượng cần phân lớp với tất cả các đối tượng trong training data.

- Bước 3: sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với đối tượng cần phân lớp

- Bước 4: lấy tất cả các lớp của K láng giềng gần nhất.

- Bước 5: dựa vào phần lớn lớp của K để xác định lớp cho đối tượng cần phân lớp.

c. Ưu và nhược điểm

- Ưu điểm:

+ Thuật toán đơn giản, dễ dàng triển khai.

+ Độ phức tạp tính toán nhỏ.

+ Xử lý tốt với tập dữ liệu nhiễu

- Nhược điểm:

+ Với K nhỏ dễ gặp nhiễu dẫn tới kết quả đưa ra không chính xác.

+ Cần nhiều thời gian để thực hiện do phải tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu.

+ Cần chuyển đổi kiểu dữ liệu thành các yếu tố định tính.

***Giải thuật máy học hỗ trợ vecto (SVM)**

a. Định nghĩa

- Máy vectơ hỗ trợ (SVM) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy.

- SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào hai lớp khác nhau. Do đó SVM là một thuật toán phân loại nhị phân.

b. Công thức

- Vectơ xác định một siêu phẳng sử dụng trong SVM là một tổ hợp tuyến tính của các vectơ dữ liệu luyện tập trong không gian mới với các hệ số αi.

- Với siêu phẳng lựa chọn như trên, các điểm x trong không gian đặc trưng được ánh xạ vào một siêu mặt phẳng là các điểm thỏa mãn:

$$\Sigma_i \alpha_i K(x_i,x) = \text{hằng số}.$$

- Nếu K (x, y) nhận giá trị ngày càng nhỏ khi y xa dần khỏi x thì mỗi số hạng của tổng trên được dùng để đo độ tương tự giữa x với điểm xi tương ứng trong dữ liệu luyện tập.

c. Ưu và nhược điểm

- Ưu điểm:

+ Xử lý trên không gian số chiều cao.

+ Tiết kiệm bộ nhớ.

+ Tính linh hoạt.

- Nhược điểm:

+ Bài toán số chiều cao.

+ Chưa thể hiện rõ tính xác suất.

2. Vấn đề liên quan đến bài toán.

a. Accuracy

- Cách đơn giản và hay được sử dụng nhất là accuracy (độ chính xác).

- Cách đánh giá này đơn giản tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

b. Confusion matrix

- Cách tính sử dụng accuracy như ở trên chỉ cho chúng ta biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác.

- Để có thể đánh giá được các giá trị này, chúng ta sử dụng một ma trận được gọi là **confusion matrix.**

c. True/False Positive/Negative

- Cách đánh giá này thường được áp dụng cho các bài toán phân lớp có hai lớp dữ liệu. Cụ thể hơn, trong hai lớp dữ liệu này có một lớp nghiêm trọng hơn lớp kia và cần được dự đoán chính xác.

- Ví dụ, trong bài toán xác định có bệnh ung thư hay không thì việc không bị sót (miss) quan trọng hơn là việc chẩn đoán nhầm âm tính thành dương tính. Trong bài toán xác định có mìn dưới lòng đất hay không thì việc bỏ sót nghiêm trọng hơn việc báo động nhầm rất nhiều. Hay trong bài toán lọc email rác

thì việc cho nhầm email quan trọng vào thùng rác nghiêm trọng hơn việc xác định một email rác là email thường.

- Trong những bài toán phân loại nấm, ta định nghĩa lớp dữ liệu quan trọng hơn cần được xác định đúng là lớp Positive (P-có độc), lớp còn lại được gọi là Negative (N-không độc). Ta định nghĩa True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN)

d. Precision và Recall

- Với bài toán phân loại mà tập dữ liệu của các lớp là chênh lệch nhau rất nhiều, có một phép đó hiệu quả thường được sử dụng là Precision-Recall.

- Trước hết xét bài toán phân loại nhị phân. Ta cũng coi một trong hai lớp là positive, lớp còn lại là negative.

- Với một cách xác định một lớp là positive, Precision được định nghĩa là tỉ lệ số điểm true positivetrong số những điểm được phân loại là positive (TP + FP).

- Recall được định nghĩa là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN).

- Theo cách toán học, Precison và Recall là hai phân số có tử số bằng nhau nhưng mẫu số khác nhau :

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

- Ta có thể nhận thấy rằng TPR và Recall là hai đại lượng bằng nhau. Ngoài ra, cả Precision và Recall đều là các số không âm nhỏ hơn hoặc bằng một.

- Precision cao đồng nghĩa với việc độ chính xác của các điểm tìm được là cao. Recall cao đồng nghĩa với việc True Positive Rate cao, tức tỉ lệ bỏ sót các điểm thực sự positive là thấp.

e. F1-score

- F_1 score, hay F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không):

$$\frac{2}{F_1} = \frac{1}{precision} + \frac{1}{recall} \; hay \; F_1 = 2\frac{1}{\frac{1}{precision} + \frac{1}{recall}} = 2\frac{precision\,.recall}{precision + recall}$$

VD:

precision	recall	F1
1	1	1

f. K-fold

- Để tránh việc trùng lặp giữa các tập kiểm thử (một số ví dụ cùng xuất hiện trong các tập kiểm thử khác nhau)

- K-fold cross-validation

- Tập toàn bộ các ví dụ A được chia ngẫu nhiên thành k tập con không giao nhau (gọi là "fold") có kích thước xấp xỉ nhau

- Mỗi lần (trong số k lần) lặp, một tập con được sử dụng làm tập kiểm thử, và (k-1) tập con còn lại được dùng làm tập huấn luyện

- K giá trị lỗi (mỗi giá trị tương ứng với một fold) được tính trung bình cộng để thu được giá trị lỗi tổng thể

- Các lựa chọn thông thường của k: 10, hoặc 5

- Thông thường, mỗi tập con (fold) được lấy mẫu phân tầng (xấp xỉ phân bố lớp) trước khi áp dụng quá trình đánh giá Cross-validation

3. Giải pháp cho bài toán

- Để chạy chương trình cần thực hiện bước sau:

+ Gán nhãn dữ liệu trong cột class

+ Xóa bỏ thuộc tính rỗng và các dữ liệu

+ Biến đổi giá trị các dữ liệu tất cả các cột về kiểu số thực

+ Huấn luyện dữ liệu với 3 giải thuật: cây quyết định, K láng giềng và SVM.

CHƯƠNG III: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

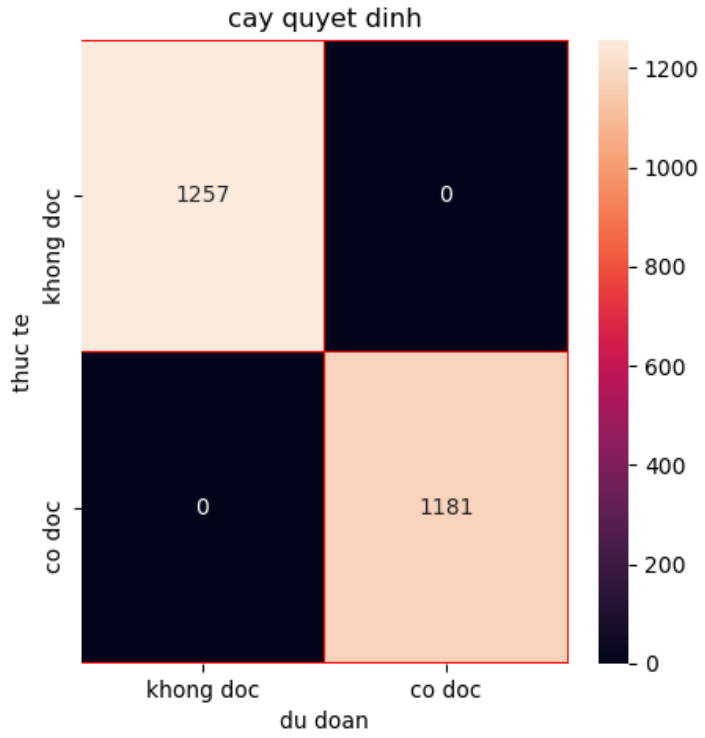
* Bảng so sánh các thuật toán

Thuật toán	Sai lầm loại 1	Sai lầm loại 2	F1	K-fold
Cây quyết định	1	1	1	0,966
KNN	1	1	1	0,967
SVM	1	1	1	0,956

* Kết quả đạt được

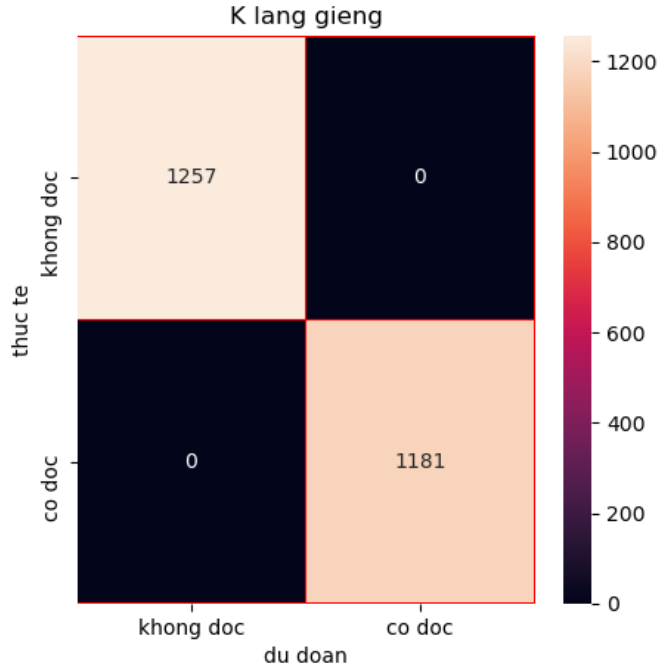
- Qua quá trình phân loại từ tập dữ liệu nấm rơm với 8124 phần tử, sử dụng phương pháp hold-out chia tập dữ liệu thành 2 phần: dữ liệu huấn luyện và dữ liệu kiểm tra theo tỉ lệ 7/3. Kết quả dự đoán nhãn thu được tại tập dữ liệu kiểm tra:

- Giải thuật cây quyết định



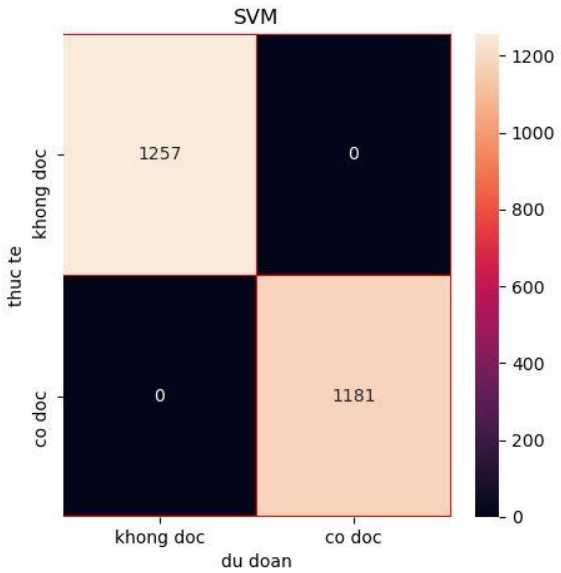
- Giải thuật cây quyết định với nghi thức kiểm tra hold-out cho kết quả dự đoán nhãn phân loại nấm là 100% độ chính xác. Sử dụng ma trận hỗn loạn thể hiện chi tiết kết quả dự đoán: số lượng nấm không độc của tập dữ liệu kiểm tra là 1257 và kết quả dự đoán cho số lượng nấm không độc là 1257, số lượng nấm có độc của tập dữ liệu kiểm tra là 1181 và kết quả dự đoán số lượng nấm có độc cho tập kiểm tra là 1181.

- Giải thuật K láng giềng



- Giải thuật K láng giềng với nghi thức kiểm tra hold-out cho kết quả dự đoán nhãn phân loại nấm là 100% độ chính xác. Sử dụng ma trận hỗn loạn thể hiện chi tiết kết quả dự đoán: số lượng nấm không độc của tập dữ liệu kiểm tra là 1257 và kết quả dự đoán cho số lượng nấm không độc là 1257, số lượng nấm có độc của tập dữ liệu kiểm tra là 1181 và kết quả dự đoán số lượng nấm có độc cho tập kiểm tra là 1181.

- Giải thuật máy học hỗ trợ vectơ



Giải thuật máy học hỗ trợ vectơ với nghi thức kiểm tra hold-out cho kết quả dự đoán nhãn phân loại nấm là 100% độ chính xác. Sử dụng ma trận hỗn loạn thể

hiện chi tiết kết quả dự đoán: số lượng nấm không độc của tập dữ liệu kiểm tra là 1257 và kết quả dự đoán cho số lượng nấm không độc là 1257, số lượng nấm có độc của tập dữ liệu kiểm tra là 1181 và kết quả dự đoán số lượng nấm có độc cho tập kiểm tra là 1181.

***Kết luận**

- Qua các kết quả đạt được, giải thuật K láng giềng cho kết quả dự đoán chính xác nhất về việc phân loại nấm có độc và không có độc. Góp phần hỗ trợ việc phân loại nấm với số lượng lớn và nhanh chóng.

2. Hướng phát triển

- Từ dữ liệu phân loại nấm không độc đề xuất phân loại nấm ăn được và không ăn được.

CHƯƠNG IV: TÀI LIỆU THAM KHẢO

1. Tài liệu tham khảo.

- https://vi.wikipedia.org/
- https://viblo.asia/
- https://github.com/topics/mushroom