**Chapter 1: Introduction**

**Outline**

[1.1 Question: 2](#_qw6i9d49txt7)

[1.1 Answer: 2](#_ckbyn0oba4i4)

[[Đạt] 2](#_n5xueotpywi8)

[[Trâm] 3](#_3rdonq9qhwq2)

[1.2 Question 3](#_jemdbhup0bjz)

[1.2 Answer 3](#_wdfi7vvs62ms)

[[Đạt] 3](#_n3r6razfqs9n)

[1.3 Question 4](#_ybno0r2swcfi)

[1.3 Answer 5](#_ji4kzc7ljb6m)

[[Đạt] 5](#_puj13gkx8tpx)

[[Trâm] 5](#_1q5o2j4jv544)

#### **1.1 Question:**

Specify whether the following use statistical or syntactic pattern recognition

(a) The patterns are vectors formed by the features. The nearest neighbor based classifier is used.

(b) The patterns are complex and are composed of simple sub-patterns which are themselves built from simpler sub-patterns.

(c) Support vector machines are used to classify the patterns.

(d) The patterns can be viewed as sentences belonging to a language.

*Chỉ định xem phần sau sử dụng nhận dạng mẫu thống kê hay cú pháp*

*(a) Các mẫu là các vectơ được hình thành bởi các tính năng. Trình phân loại dựa trên hàng xóm gần nhất được sử dụng.*

*(b) Các mẫu phức tạp và bao gồm các mẫu con đơn giản mà bản thân chúng được xây dựng từ các mẫu con đơn giản hơn.*

*(c) Các máy vectơ hỗ trợ được sử dụng để phân loại các mẫu.*

*(d) Các mẫu có thể được xem như các câu thuộc về một ngôn ngữ.*

#### **1.1 Answer:**

##### **[Đạt]**

* Syntactic pattern recognition (nhận dạng mẫu ngữ pháp) xử lý các đối tượng được mô tả bằng các đặc điểm định tính mô tả các mối quan hệ cấu trúc hoặc cú pháp vốn có trong đối tượng.
* Statistical pattern recognition (nhận dạng mẫu thống kê) sử dụng không gian vectơ (không gian con/phép chiếu và sự giống nhau giữa các điểm theo các phép đo khoảng cách) để biểu diễn các mẫu và lớp kết hợp với các phép trừu tượng để xử lý mật độ/phân phối xác suất của các điểm trong không gian nhiều chiều.
* Từ 2 khái niệm trên:

1. The patterns are vectors formed by the features. The nearest neighbor based classifier is used → sử dụng Statistical pattern recognition
2. The patterns are complex and are composed of simple sub-patterns which are themselves built from simpler sub-patterns → Sử dụng Statistical pattern recognition hoặc một mô hình hybrid kết hợp 2 phương pháp.
3. Support vector machines are used to classify the patterns → sử dụng Statistical pattern recognition
4. The patterns can be viewed as sentences belonging to a language → sử dụng Syntactic pattern recognition (phương pháp cổ điển) hoặc Statistical pattern recognition (tiếp cận hiện đại - deep learning).

##### **[Trâm]**

(a**) Chỉ sử dụng nhận dạng mẫu thống kê**, vì nó đề cập đến việc sử dụng tính năng để hình thành các vectơ và sử dụng phân loại dựa trên hàng xóm gần nhất.

(b) **Không phải là nhận dạng mẫu thống kê**, mà có thể được coi là một hình thức của khai thác dữ liệu hoặc học máy.

(c) **Chỉ sử dụng nhận dạng mẫu thống kê,** vì nó đề cập đến việc sử dụng máy vectơ hỗ trợ để phân loại các mẫu.

(d) **Không phải là nhận dạng mẫu thống kê**, mà có thể được coi là một hình thức của xử lý ngôn ngữ tự nhiên hoặc trí tuệ nhân tạo.

#### **1.2 Question**

Explain the following terms: pattern, class, classifier, feature, decision rule, and decision boundary.

*Giải thích các thuật ngữ sau: mẫu, lớp, phân loại, tính năng, quy tắc quyết định và ranh giới quyết định.*

#### **1.2 Answer**

##### **[Đạt]**

+ Pattern (mẫu): Trong học máy, cụ thể là trong nhận dạng mẫu (pattern recognition), mẫu là tất cả mọi thứ xung quanh trong thế giới kỹ thuật số này. Một mẫu có thể được nhìn thấy trên phương diện vật lý hoặc có thể được quan sát dưới dạng toán học bằng cách áp dụng các thuật toán. Ví dụ: màu sắc trên quần áo, mẫu giọng nói, các cá thể chó mèo,…

+ Class (lớp): Trong học máy, lớp liên kết một nhóm các cá thể (instance) có cùng điểm chung. Ví dụ đối với bài toán phân loại thư rác, ta có hai lớp là spam và không spam.

+ Classifier: là một loại thuật toán học máy (machine learning algorithm) được sử dụng để gán nhãn (class label) cho các điểm dữ liệu (data point). Ví dụ: nhận dạng hình ảnh để gắn nhãn cho xe ô tô.

+ Feature (đặc trưng): Các đặc trưng trong học máy là các biến độc lập hoạt động như một đầu vào trong hệ thống của bạn. Trên thực tế, trong khi đưa ra dự đoán, các mô hình sử dụng các đặc trưng đó để đưa ra dự đoán. Nói một cách đơn giản thì đặc trưng là một cột dữ liệu trong bộ dữ liệu đầu vào. Ví dụ: Mô hình đang cố gắng dự đoán loại vật nuôi mà ai đó sẽ chọn, các đặc trưng đầu vào có thể bao gồm tuổi, khu vực quê hương, thu nhập gia đình.

+ Decision rule (Quy tắc ra quyết định): Các quy tắc ra quyết định là một khái niệm được sử dụng khi xây dựng trong cây quyết định (decision tree) trong đó kết quả là nội dung của các nút lá (leaf node). Quy tắc quyết định có thể được tạo ra bằng cách xây dựng các luật kết hợp (association rule) với biến mục tiêu (target variable) nhằm biểu thị các mối quan hệ nhân quả hoặc thời gian.

+ Decision boundary: Trong học máy, khi có hai đặc trưng trở lên, decision boundary là một đường (trong trường hợp có hai đặc trưng (feature)), trong đó hầu hết các sample của một lớp nằm ở một phía của đường đó và tất cả các mẫu của lớp khác nằm ở phía đối diện. Đường này giúp phân cách một lớp với lớp kia. Trong trường hợp có nhiều hơn hai đặc trưng, decision boundary không phải là một đường thẳng, mà là một siêu phẳng (hyperplane) trong kích thước không gian đặc trưng (feature space).

#### **1.3 Question**

What is a training set? How it is chosen? Does the size of the training set influence to the performance of the classification process?

*Tập huấn luyện (training set) là gì? Nó được chọn như thế nào? Kích thước của tập huấn luyện có ảnh hưởng đến hiệu suất của quá trình phân lớp không?*

#### **1.3 Answer**

##### **[Đạt]**

* Training set: là một tập dữ liệu có kích thước lớn, được dùng để huấn luyện mô hình trong quá trình máy học, đây cũng chính là tập dữ liệu mà máy dùng để học và rút trích được những những điểm quan trọng để ghi nhớ lại.
* Về cách lựa chọn, tập training set sẽ có 2 phần:
  + Input: sẽ là những dữ liệu đầu vào, sẽ được chọn lựa phù hợp với bài toán đặt ra,. Ví dụ: bài toán nhận dạng hình ảnh những con mèo thì input là những bức hình mèo.
  + Output: sẽ là những kết quả tương ứng với input. Ví dụ, gán nhãn dán cho những tấm hình, nếu input là hình ảnh chứa con chó thì output sẽ cho ra kết quả label là “con chó”.
* Thông thường, khi tăng độ lớn của tập huấn luyện sẽ cho kết quả tốt hơn cho quá trình phân loại, giúp mô hình giảm overfitting và tăng độ chính xác của mô hình. Tuy nhiên cần lưu ý là không nhất thiết tăng kích thước của tập huấn luyện để cải thiện hiệu suất của mô hình vì ở ngoài thực tế việc này khá là tốn tài nguyên, và cũng có thể có nhiều phương pháp khác giúp cải thiện mô hình tốt hơn.

##### **[Trâm]**

*Tập huấn luyện (training set) là gì?*

* Training set là một tập dữ liệu được sử dụng để huấn luyện mô hình máy học.
* Training set bao gồm các mẫu dữ liệu đầu vào và đầu ra tương ứng (nếu có) được sử dụng để huấn luyện mô hình máy học, với mục đích học cách suy luận hoặc dự đoán đầu ra từ đầu vào.
* Quá trình huấn luyện mô hình thường bao gồm việc tối ưu hóa các tham số của mô hình để giảm thiểu sai số giữa các dự đoán đầu ra của mô hình và các giá trị đầu ra thực tế tương ứng trong tập dữ liệu huấn luyện.
* Sau khi được huấn luyện trên training set, mô hình có thể được áp dụng để dự đoán đầu ra cho các dữ liệu mới không có sẵn trong tập huấn luyện.

*Nó được chọn như thế nào?*

* Việc lựa chọn tập huấn luyện phụ thuộc vào mục đích và đặc điểm của bài toán. Tập huấn luyện cần được chọn sao cho nó đủ lớn và đại diện cho các dữ liệu mà mô hình sẽ được sử dụng để dự đoán trong tương lai.
* Đối với các bài toán phân loại và dự đoán, tập huấn luyện cần có đầy đủ các lớp hoặc phân bố đầu vào để mô hình có thể học được các đặc trưng đa dạng. Ngoài ra, tập huấn luyện cần có các kết quả đầu ra đúng được gán nhãn chính xác để mô hình có thể học từ chúng.
* Đối với các bài toán mà mô hình cần học từ thời gian, tập huấn luyện cần chứa các dữ liệu từ quá khứ đến hiện tại để mô hình có thể học được xu hướng và mối liên hệ giữa các biến trong quá khứ và hiện tại.
* Các thuật toán máy học cũng đòi hỏi các tập huấn luyện khác nhau. Ví dụ, các mô hình deep learning thường cần tập huấn luyện lớn hơn và đòi hỏi nhiều thời gian để đào tạo hơn so với các thuật toán khác như SVM hay Decision Tree.
* Tuy nhiên, một điểm quan trọng cần lưu ý là tập huấn luyện không nên quá lớn hoặc quá nhỏ. Tập huấn luyện quá lớn có thể dẫn đến tình trạng overfitting, trong đó mô hình học quá nhiều từ tập huấn luyện và không thể dự đoán tốt cho dữ liệu mới. Ngược lại, tập huấn luyện quá nhỏ có thể dẫn đến tình trạng underfitting, trong đó mô hình không thể học đầy đủ các đặc trưng của dữ liệu.

*Kích thước của tập huấn luyện có ảnh hưởng đến hiệu suất của quá trình phân lớp không?*

* Kích thước của tập huấn luyện có ảnh hưởng đến hiệu suất của quá trình phân lớp.
* Nếu kích thước tập huấn luyện quá nhỏ, mô hình có thể không học đủ các đặc trưng của dữ liệu và sẽ có xu hướng underfitting, tức là không thể phân loại chính xác cho các dữ liệu mới.
* Ngược lại, nếu kích thước tập huấn luyện quá lớn, mô hình có thể bị overfitting, tức là học quá nhiều từ tập huấn luyện và không thể tổng quát hóa tốt cho các dữ liệu mới. Trong trường hợp này, mô hình có thể trở nên quá phức tạp và nhạy cảm với nhiễu trong dữ liệu huấn luyện.
* Vì vậy, để đạt được hiệu suất tốt nhất, kích thước tập huấn luyện cần được lựa chọn sao cho đủ lớn để mô hình có thể học được đầy đủ các đặc trưng của dữ liệu, nhưng không quá lớn để tránh tình trạng overfitting. Điều này có thể đòi hỏi nhiều thử nghiệm và điều chỉnh với các bộ tham số khác nhau để đạt được hiệu suất tốt nhất cho mô hình.