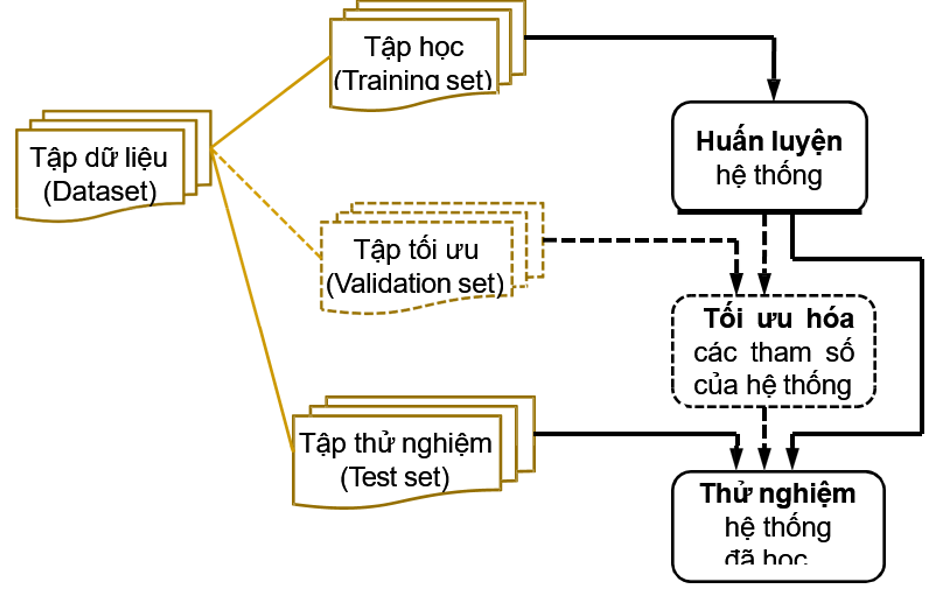
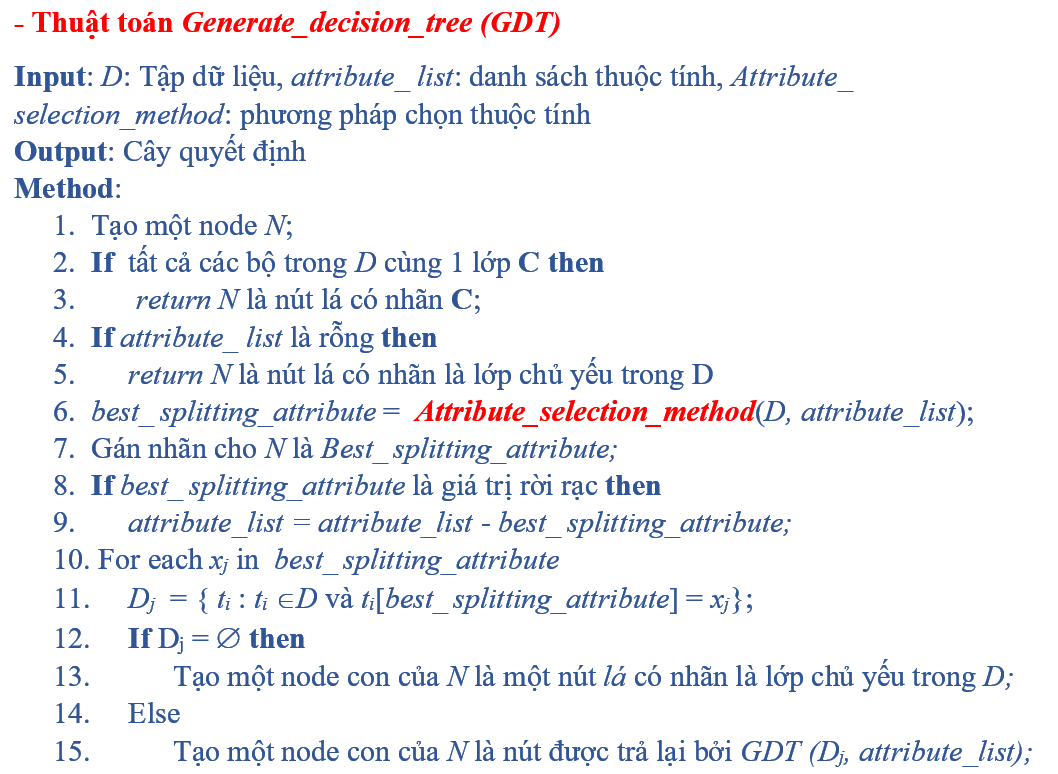
Đề cương ôn tậpChuyên đề khoa học máy tính

1. **Kiến thức chung về học máy**
   1. Trình bày sơ đồ qui trình học máy
   2. Trình bày khái niếm học có giám sát, học nửa giám sát, học tăng cường cho ví dụ

* Học có giám sát (Supervised learning)
  + là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn).
  + Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:
* Classification (Phân loại)
* Regression (Hồi quy)
  + Ví dụ: Thuật toán dò các khuôn mặt trong một bức ảnh. Thời gian đầu, facebook sử dụng thuật toán này để chỉ ra các khuôn mặt trong một bức ảnh và yêu cầu người dùng tag friends - tức gán nhãn cho mỗi khuôn mặt. Số lượng cặp dữ liệu (khuôn mặt, tên người) càng lớn, độ chính xác ở những lần tự động tag tiếp theo sẽ càng lớn.
* Học nửa giám sát (Semi-Supervised Learning)
  + Các bài toán khi chúng ta có một lượng lớn dữ liệu X nhưng chỉ một phần trong chúng được gán nhãn được gọi là học nửa giám sát. Những bài toán thuộc nhóm này nằm giữa hai nhóm học có giám sát và học không giám sát
  + Ví dụ: bài toán phân lớp ảnh có một phần ảnh hoặc văn bản được gán nhãn
* Học tăng cường (Reinforcement learning)
  + là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất. Học tăng cường chủ yếu được áp dụng trong các trò chơi
  + Ví dụ: chương trình chơi cờ vây AlphaGo của google

1. Các phương pháp học máy cho bài toán phân lớp
   1. Trình bày bài toán phân lớp

Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp (model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Quá trình phân lớp là **quá trình gán nhãn** cho đối tượng dữ liệu.

* 1. Cây quyết định
     + Trình bày thuật toán xây dựng cây quyết định
     + Trình bày cây quyết định ID3, C4.5
       - Cây quyết định ID3

*Độ đo lựa chọn thuộc tính Infomation Gain (Gain)*

* + - Dựa trên lý thuyết thông tin của Claude Shannon
    - Một thuộc tính được chọn nếu nó có giá trị Gain lớn nhất

Ta có:

trong đó

* + - Thuộc tính *A* có các giá trị là {*a1, a2, ..., av*}

* + - Độ đo Infomation gain của A

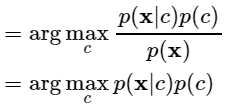
* + - Nhược điểm của độ đo này là có xu hướng lựa chọn thuộc tính có nhiều giá trị => khả năng tạo ra các phân hoạch dữ liệu có tính trong suốt cao => có Gain->min. (Ví dụ: thuộc tính ID)
      * Cây quyết đinh C4.5

*Độ đo lựa chọn thuộc tính Gain ratio*

Khắc phục nhược điểm của độ đo Infomation gain

Một thuộc tính A được chọn nếu nó có giá trị *Gain Ratio* lớn nhất

Tính:

* 1. Thuật toán phân lớp Bayes
* Trình bày thuật toán phân lớp Bayes
* Xét bài toán phân lớp với **C** lớp: {1,2 ,…,**C**}. Giả sử có một điểm dữliệu **x∈Rd**. Hãy tính xác suất để điểm dữ liệu này rơi vào lớp **c**.
* Chính là tính: p(y=c|x)  viết gọn là p(c|x)
* Phương pháp phân lớp dưa trên pp cực đại hàm khả năng (Likelihood)

1. **Phương pháp tối ưu**
   1. Trình bày phương pháp tìm cực trị Gradient

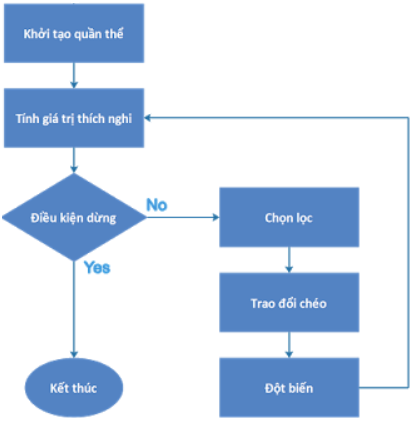
Cho xt ta cần đưa xt về càng gần x\* càng tốt

Nếu đạo hàm của hàm số tại xt : f ’(xt) > 0 thì xt nằm về bên phải so với x\* và ngược lại

Để điểm tiếp theo xt+1 gần với x\* hơn thì ta cần đưa xt về bên trái:

xt+1 = xt+Δ xt+1 = xt − ηf′(xt)

1. **Giải thuật di truyền**
   1. Trình bày sơ đồ khối tổng quát của thuật giải di truyền

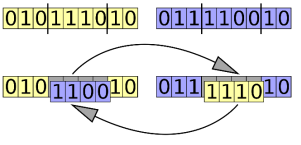


* 1. Trình bày các phương pháp lai ghép của thuật giải di truyền với mã hóa nhị phân
* Lai ghép 1 điểm (one point crossover)

A: 1 0 0 1 0 1 1 | 0 0 0 0 0 => C: 1 0 0 1 0 1 1 | 1 1 0 1 1

B: 0 0 1 1 0 1 0 | 1 1 0 1 1 D: 0 0 1 1 0 1 0 | 0 0 0 0 0

* Lai ghép 2 điểm (two points crossover)

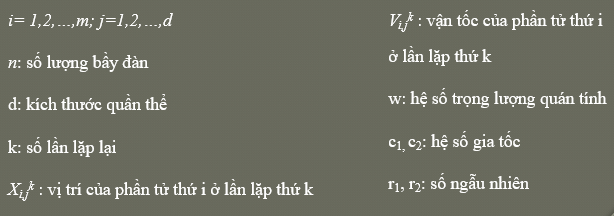


1. **Giải thuật bầy đàn (Particle Swarm Optimization – PSO)**

* PSO là một kỹ thuật tối ưu hóa ngẫu nhiên dựa trên một quần thể và sau đó tìm nghiệm tối ưu bằng cách cập nhật các thế hệ.
* Mỗi cá thể trong quần thể điều chỉnh hướng bay theo vị trí tốt nhất của nó và của cá thể trong quần thể tính tới thời điểm hiện tại
* Quá trình cập nhật các particles:







Pbesti,j: represents personal best *jth* component of *ith* individual

Gbesti,j: represents *jth* component of the best individual of population up to iteration *k*

* Giá trị Pbest và Gbest của mỗi cá thể được cập nhật:

Tại vòng lặp thứ k:

If *f (Xpk+1) < f (Pbestpk)* then *Pbestpk+1 = Xpk+1*

else *Pbestpk+1 = Pbestpk*

If *f (Xpk+1) < f (Gbestk)* then *Gbestk = Xpk+1*

else *Gbestk+1 = Gbestk*

Pbestki: ví trí tốt nhất của cá thể i sau k bước lặp

Gbestk: ví tri tốt nhất của đoàn thể sau k bước lặp

