**Chapter 1**

**The Machine Learning Landscape**

1. **How would you define Machine Learning?** *Định nghĩa về ML?*

* *ML có nghĩa là xây dựng 1 chương trình có thể học từ dữ liệu. Học ở đây có nghĩa là thực hiện 1 nhiệm vụ tốt hơn dựa trên thang đo hiệu suất.*

1. **Can you name four types of problems where it shines?**

* *4 vấn đề mà ML có thể giải quyết?*
* *Không có thuật toán để giải.*
* *Thay thế 1 danh sách dài bằng tay.*
* *Xây dựng hệ thống mô thỏng theo dao động của môi trường.*
* *Giúp con người trong việc học, khai phá dữ liệu.*

1. **What is a labeled training set?** *Dữ liệu được gán nhãn là gì?*

* *Là dữ liệu chứa kết quả mong muốn cho từng mẫu.*

1. **What are the two most common supervised tasks?** *2 nhiệm vụ chính của học có giám sát là gì?*

* *Regression and classification*

1. **Can you name four common unsupervised tasks?** *4 nhiệm vụ của học không giám sát?*

* *Clustering, visualization, dimensionaltity reduction, association rule learning.*

1. **What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?**

* ***Học tăng cường (Reinforcement learning - RL)****là một nhánh của****học máy (Machine learning - ML)****, nghiên cứu cách thức một****tác nhân (Agent)****trong một****môi trường (Environment)****đang ở một****trạng thái (State)****thực hiện một****hành động (Action)****để tối ưu hóa một****phần thưởng (Reward)****chung.*

1. **What type of algorithm would you use to segment your customers into multiple groups?** *Thuật toán được sử dụng để phân vùng khách hàng của bạn với 1 nhóm?*

* *Nếu là 1 nhóm biết trước thì ta sẽ sử dụng* ***classification****.*
* *Nếu là 1 nhóm không xác định thì ta có thể sử dụng* ***clustering****.*

1. **Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?** *Phát hiện spam là supervised hay unsupervised?*

* *Tất nhiên là supervise vì nó là phân loại spam hoặc không spam.*

1. **What is an online learning system?** *Học trực tuyến là gì?*

* *Một hệ thống học trực tuyến có thể học tăng dần, trái ngược với hệ thống học hàng loạt. Điều này làm cho nó có khả năng thích ứng nhanh chóng với cả dữ liệu thay đổi và hệ thống tự trị, cũng như đào tạo trên lượng dữ liệu rất lớn.*

1. **What is out-of-core learning?**

* *Các thuật toán ngoài lõi có thể xử lý lượng dữ liệu khổng lồ không thể chứa vừa trong bộ nhớ chính của máy tính. Một thuật toán học tập bên ngoài sẽ cắt dữ liệu thành các đợt nhỏ và sử dụng các kỹ thuật học trực tuyến để học từ các đợt nhỏ này.*

1. **What type of learning algorithm relies on a similarity measure to make predictions?**

* *Một hệ thống học tập dựa trên phiên bản học thuộc lòng dữ liệu huấn luyện; sau đó, khi được cung cấp một phiên bản mới, nó sử dụng thước đo độ tương tự để tìm các phiên bản đã học tương tự nhất và sử dụng chúng để đưa ra dự đoán.*

1. **What is the difference between a model parameter and a learning algorithm’s hyperparameter?**

* *Một mô hình có một hoặc nhiều tham số mô hình xác định những gì nó sẽ dự đoán trong một trường hợp mới (ví dụ: độ dốc của mô hình tuyến tính). Thuật toán học cố gắng tìm các giá trị tối ưu cho các tham số này để mô hình tổng quát hóa tốt cho các trường hợp mới. Siêu tham số là một tham số của chính thuật toán học tập, không phải của mô hình (ví dụ: mức độ chính quy hóa cần áp dụng).*

1. **What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?**

* *Các thuật toán học dựa trên mô hình tìm kiếm một giá trị tối ưu cho các tham số mô hình sao cho mô hình sẽ khái quát hóa tốt cho các trường hợp mới.*
* *Chúng tôi thường đào tạo các hệ thống như vậy bằng cách giảm thiểu hàm* ***loss*** *đo lường mức độ kém của hệ thống khi đưa ra dự đoán trên dữ liệu đào tạo, cộng với hình phạt cho độ phức tạp của mô hình nếu mô hình được chuẩn hóa.*
* *Để đưa ra dự đoán, chúng tôi cung cấp các tính năng của phiên bản mới vào chức năng dự đoán của mô hình, sử dụng các giá trị tham số mà thuật toán học tìm được.*

1. **Can you name four of the main challenges in Machine Learning?**

* *Một số thách thức chính trong ML là :*
* *Thiếu dữ liệu, chất lượng dữ liệu kém.*
* *Dữ liệu không mang tính đại diện,*
* *Các tính năng không mang tính thông tin,*
* *Các mô hình quá đơn giản hoặc quá phức tạp không phù hợp với dữ liệu huấn luyện.*

1. **If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?**

* *Trong trường hợp này thì mô hình đã bị* ***overfitting****.*
* *Để giải guyết vấn đề overfitting có 3 cách:*
* *Thu thập thêm nhiều dữ liệu hơn tăng tính đa dạng cho dữ liệu.*
* *Đơn giản hóa mô hình, giảm tham số, chuẩn hóa*
* *Giảm nhiễu trong quá trình huấn luyện*

1. **What is a test set and why would you want to use it?**

* *Bộ kiểm thử được sử dụng để ước tính lỗi tổng quát hóa mà mô hình sẽ mắc phải trên tập dữ liệu mới trước khi đưa mô hình vào quá trình sản xuất.*

1. **What is the purpose of a validation set?**

* *Bộ thẩm định được dùng để so sánh các mô hình để chọn ra mô hình tốt nhất và điều chỉnh các siêu tham số.*

1. **What can go wrong if you tune hyperparameters using the test set?**

* *Nếu thay đổi siêu tham số trong quá trình kiểm thử thì sẽ làm cho mô hình quá khớp với dữ liệu kiểm thử và sẽ không có tính tổng quát.*

1. **What is cross-validation and why would you prefer it to a validation set?**

* *Xác thực chéo là một kỹ thuật cho phép so sánh các mô hình (để lựa chọn mô hình và điều chỉnh siêu tham số) mà không cần bộ xác thực riêng. Điều này tiết kiệm dữ liệu đào tạo quý giá.*