Introduction to Big Data Techniques

October 26, 2023

[ ]:
```
from IPython.display import Image
```

1 How is Fintech used in Quantitative Investment Analysis?

1.1 Fintech

Theo nghĩa rộng nhất, **Fintech** đề cập đến sự đổi mới công nghệ diễn ra trong lĩnh vực tài chính nói chung. Cụ thể hơn, Fintech đề cập đến sự đổi mới công nghệ trong thiết kế và cung cấp các sản phẩm và dịch vụ tài chính

Các lĩnh vực phát triển liên quan trực tiếp đến phân tích định lượng trong đầu tư bao gồm:

- Phân tích dữ liệu lớn: tích hợp phân tích dữ liệu từ các nguồn truyền thống như giá chứng khoán, dữ liệu từ báo cáo tài chính, chỉ số kinh tế; cũng như nguồn phi truyền thống như mạng xã hội, cảm biến... vào quá trình ra quyết định đầu tư
- Công cụ phân tích: Áp dụng kỹ thuật AI trong xác định các mối quan hệ phi tuyến phức tạp so với các phương pháp phân tích định lượng truyền thống

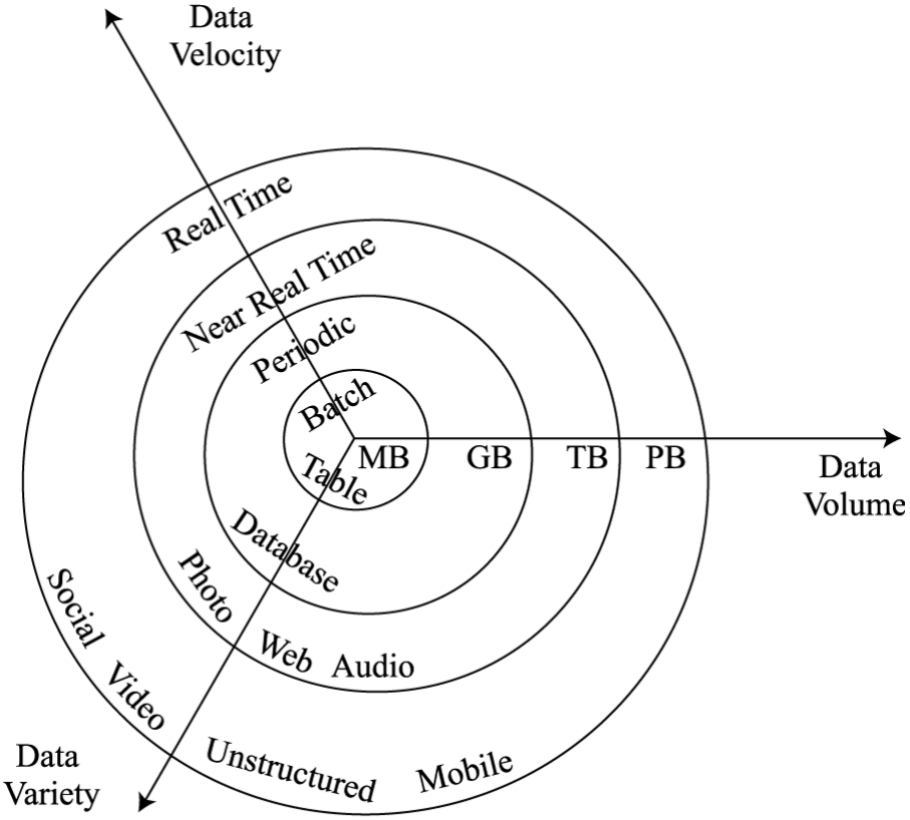
1.2 Big Data

Dữ liệu lớn bao gồm các dữ liệu được tạo ra từ nguồn truyền thống cũng như phi truyền thống, với các đặc trưng:

- Khối lượng (`Volume`): Lượng dữ liệu trong các tệp (files), bản ghi (records), bảng (tables) cực lớn, có thể lên đến hàng tỷ điểm dữ liệu
- Vận tốc (`Velocity`): Tốc độ và tần suất ghi và truyền dữ liệu ngày càng gia tăng. Dữ liệu thời gian thực hoặc tiệm cận thời gian thực trở thành tiêu chuẩn trong nhiều lĩnh vực
- Sự đa dạng (`Variety`): Dữ liệu được thu thập từ nhiều nguồn khác nhau với nhiều định dạng khác nhau (dữ liệu cấu trúc, phi cấu trúc hoặc bán cấu trúc)

[ ]:
```
# Đặc trưng của dữ liệu lớn
Image(filename = "Pictures/01.png")
```

[ ]:



1.3 Nguồn dữ liệu lớn

Dữ liệu lớn bao gồm các dữ liệu đến từ các nguồn sau:

- Thị trường tài chính (vốn cổ phần, thu nhập cố định, chứng khoán phái sinh…)
- Doanh nghiệp (tài chính doanh nghiệp, giao dịch thương mại, tín dụng…)
- Chính phủ (dữ liệu thương mại, kinh tế, việc làm…)
- Cá nhân (tín dụng cá nhân, đánh giá sản phẩm, nhật ký duyệt web, bài đăng trên MXH…)
- Internet of Things (IoT) (dữ liệu được đo lường từ các cảm biến, ví dụ như trong tòa nhà thông minh, vệ tinh, vườn công nghệ cao…)

2 Advanced Analytical Tools: Artificial Intelligence and Machine Learning

2.1 Artificial Intelligence (AI)

Trí tuệ nhân tạo (AI) có khả năng thực hiện các nhiệm vụ mà bình thường đòi hỏi trí thông minh con người. Sự phát triển của AI cho phép tạo ra các hệ thống máy tính có khả năng nhận thức và

ra quyết định tương đương hoặc vượt trội so với con người

2.2 Machine Learning (ML)

Học máy (ML) đề cập đến các kỹ thuật dwuaj trên máy tính nhằm tìm cách trích xuất lượng kiến thức từ lượng lớn dữ liệu mà không đưa ra bất kỳ giả định nào về phân phối cơ sở của dữ liệu. Mục tiêu của học máy là tự động hóa các quy trình ra quyết định bằng việc khái quát hóa, hoặc "học hỏi" từ các ví dụ đã biết để xác định cấu trúc cơ bản trong dữ liệu. Một cách hiểu đơn giản, thuật toán học máy được sử dụng với mục đích tìm mẫu và áp dụng mẫu

Học máy yêu cầu lượng dữ liệu khổng lồ để đào tạo. Do đó, mặc dù một số kỹ thuật học máy đã tồn tại trong nhiều năm, nhưng việc thiếu dữ liệu đã hạn chế sự ứng dụng rộng rãi của học máy. Tuy nhiên, sự phát triển của dữ liệu lớn cung cấp cho các thuật toán học máy, bao gồm cả mạng thần kinh (`neutral network`) đủ dữ liệu để cải thiện độ chính xác của mô hình và dự đoán, đồng thời có thể áp dụng nhiều kỹ thuật học máy hơn

Dưới đây là một số đặc trưng của học máy:

- Thuật toán học máy được cung cấp đầu vào và có thể được cung cấp đầu ra. Thuật toán cố gắng học từ dữ liệu các tốt nhất để mô tả đầu ra dựa vào đầu vào, hoặc xác định/mô tả cấu trúc dữ liệu cơ bản nếu không có đầu ra mẫu
- Học máy liên quan đến việc chia tập dữ liệu thành ba tập con: tập huấn luyện (xác định quan hệ), tập xác thực (tinh chỉnh mô hình) và tập thử nghiệm (kiểm tra dự đoán)
- Học máy yêu cầu sự phán đoán và hiểu biết của con người về dữ liệu và kỹ thuật để lựa chọn kỹ thuật học máy phù hợp trong phân tích dữ liệu
- Mô hình học máy có thể không hoạt động tốt nếu thiếu dữ liệu đào tạo
- Phân tích viên cần nhận biết được các lỗi có thể phát sinh từ việc dữ liệu khớp quá mức (`overfitting`). Vấn đề này diễn ra nếu dữ liệu đầu vào và đầu ra quá khớp với nhau, dẫn đến mô hình được đào tạo quá mức (`overtrained`), từ đó đưa ra các kết quả dự đoán sai hoặc không có căn cứ.
- Các kỹ thuật học máy có vẻ là các kỹ thuật tiếp cận mù, hoặc "hộp đen", có thể đưa ra các kết quả không thể hiểu và/hoặc giải thích hoàn toàn

Học máy có thể giúp xác định mối quan hệ giữa các biến, phát hiện các mẫu xu hướng và tạo cấu trúc từ dữ liệu, bao gồm phân loại dữ liệu. Học máy có thể được chia làm ba loại kỹ thuật: Học có giám sát, Học không giám sát và Học sâu

3 Tackling Big Data with Data Science

Khoa học dữ liệu là lĩnh vực liên ngành khai thác những tiến bộ của khoa học máy tính (AI, ML, NLP…), thống kê và các ngành khác nhằm khám phá thông tin từ dữ liệu

Để xác định kỹ thuật quản lý dữ liệu tốt nhất cho phân tích dữ liệu lớn, phân tích viên áp dụng nhiều **phương pháp xử lý dữ liệu** ở các bước khác nhau:

- Thu thập (`Capture`): thu thập dữ liệu và chuyển đổi chúng sang định dạng có thể phân tích
- Quản lý (`Curation`): quá trình đảm bảo chất lượng và độ chính xác của dữ liệu thông qua làm sạch dữ liệu

- Lưu trữ (`Storage`): đề cập đến cách dữ liệu được ghi lại, lưu trữ và truy cập, cũng như thiết kế cơ sở dữ liệu cơ bản
- Tìm kiếm (`Search`): cách truy vấn dữ liệu
- Truyền (`Transfer`): cách thức dữ liệu di chuyển từ nguồn dữ liệu và/hoặc nơi lưu trữ đến công cụ phân tích

Trực quan hóa dữ liệu là một công cụ quan trọng để hiểu dữ liệu lớn. Trực quan hóa đề cập đến cách dữ liệu sẽ được định dạng, hiển thị và tóm tắt dưới dạng đồ họa

Dữ liệu truyền thống có thể được hiển thị bằng bảng, biểu đồ… trong khi dữ liệu phi cấu trúc yêu cầu các kỹ thuật mô tả mới (đồ họa 3D, tương tác, tag cloud…). Hơn nữa, có nhiều giải pháp khác nhau để phản ánh cấu trúc của dữ liệu thông qua trực quan hóa với đồ họa tương tác (bản đồ nhiệt, sơ đồ cây, biểu đồ mạng…)

Fintech phát triển và được sử dụng trong nhiều lĩnh vực quản lý đầu tư, ví dụ như phân tích văn bản, xử lý ngôn ngữ tự nhiên (NLP), phân tích rủi ro, giao dịch theo thuật toán…