Báo cáo: Phân tích Phân cụm Dữ liệu Chứng khoán Sử dụng Thuật toán Expectation-Maximization (EM) với Gaussian Mixture Models (GMM)

Slide 1: Giới thiệu và Mục tiêu Dự án

- **Tiêu đề:** Phân tích Đặc điểm Thị trường Chứng khoán bằng Kỹ thuật Học Máy Không Giám sát.
- **Nội dung:**
  - Thị trường chứng khoán biến động phức tạp, việc xác định các "trạng thái" hoặc "chế độ" thị trường tiềm ẩn có thể mang lại thông tin giá trị cho nhà đầu tư.
  - Dự án này sử dụng thuật toán Expectation-Maximization (EM) kết hợp với Gaussian Mixture Models (GMM) để phân cụm dữ liệu giao dịch chứng khoán.
  - **Mục tiêu:**
    - Khám phá các cụm (clusters) tiềm ẩn trong dữ liệu dựa trên giá đóng cửa điều chỉnh ('Adj Close') và khối lượng giao dịch ('Volume').
    - Hiểu rõ đặc điểm của từng cụm được phát hiện.

Slide 2: Mô tả Bộ Dữ liệu (data.csv)

- **Tiêu đề:** Tổng quan về Dữ liệu Đầu vào
- **Nội dung:**
  - **Nguồn dữ liệu:** Tệp `data.csv`.
  - **Số lượng bản ghi:** 10409 (trong ví dụ mã là 1000 mẫu).
  - **Số lượng thuộc tính:** 7 cột.
  - **Các cột dữ liệu chính:**
    - `Date`: Ngày giao dịch (kiểu object).
    - `Open`: Giá mở cửa (float64).
    - `High`: Giá cao nhất (float64).
    - `Low`: Giá thấp nhất (float64).
    - `Close`: Giá đóng cửa (float64).
    - `Adj Close`: Giá đóng cửa điều chỉnh (float64). **Đặc trưng được chọn**

Slide 3: Phương pháp Luận - Expectation-Maximization (EM)

- **Tiêu đề:** Giới thiệu Thuật toán Expectation-Maximization (EM)
- **Nội dung:**
    - EM là một thuật toán lặp để tìm ước lượng hợp lý cực đại (Maximum Likelihood Estimation - MLE) hoặc ước lượng hậu nghiệm cực đại (Maximum A Posteriori - MAP) cho các tham số trong mô hình thống kê có chứa các biến ẩn (latent variables).
    - **Quy trình lặp của EM gồm 2 bước:**
        a. **Bước Expectation (E-step):** Tính toán giá trị kỳ vọng của log-likelihood của các biến ẩn, dựa trên các quan sát hiện tại và ước lượng tham số hiện tại. Trong GMM, bước này tính toán "trách nhiệm" (xác suất) mỗi điểm dữ liệu thuộc về mỗi cụm.
        b. **Bước Maximization (M-step):** Cập nhật lại các tham số của mô hình bằng cách tối đa hóa giá trị kỳ vọng của log-likelihood (đã tính ở E-step). Trong

Slide 4: Phương pháp Luận - Gaussian Mixture Models (GMM)

- **Tiêu đề:** Áp dụng GMM cho Phân cụm
- **Nội dung:**
  - GMM giả định rằng tất cả các điểm dữ liệu được tạo ra từ một hỗn hợp của một số hữu hạn các phân phối Gaussian với các tham số chưa biết.
  - Mỗi phân phối Gaussian tương ứng với một cụm.
  - **Ưu điểm của GMM:**
    - Phân cụm "mềm" (soft clustering): Mỗi điểm dữ liệu có một xác suất thuộc về mỗi cụm, thay vì gán cứng vào một cụm duy nhất.
    - Linh hoạt về hình dạng cụm: Có thể mô hình hóa các cụm có hình dạng elip (do sử dụng ma trận hiệp phương sai).
  - Trong dự án này, GMM được sử dụng để mô hình hóa các trạng thái thị trường khác nhau, với mỗi trạng thái là một thành phần Gaussian.

Slide 5: Tiền xử lý Dữ liệu

- **Tiêu đề:** Các Bước Chuẩn bị Dữ liệu
- **Nội dung:**
    i. **Tải dữ liệu:** Đọc tệp `data.csv` vào DataFrame của Pandas.
    ii. **Lựa chọn đặc trưng:** Chọn các cột 'Adj Close' và 'Volume' làm đầu vào cho mô hình GMM.
        - `features = ['Adj Close', 'Volume']`
    iii. **Xử lý giá trị thiếu (Missing Values):** Kiểm tra và điền giá trị thiếu (nếu có) bằng giá trị trung bình của cột tương ứng (trong code mẫu, nếu có).
        - `X = X.fillna(X.mean())`
    iv. **Chuẩn hóa dữ liệu (Data Scaling):**
        - Sử dụng `StandardScaler` từ `sklearn.preprocessing`.
        - Mục đích: Đưa các đặc trưng về cùng một thang đo (trung bình 0, phương

Slide 6: Huấn luyện Mô hình GMM & Kết quả Bước Expectation (Cuối cùng)

- **Tiêu đề:** Áp dụng GMM và Phân tích Bước Expectation
- **Nội dung:**
  - **Huấn luyện mô hình:**
    - Khởi tạo `GaussianMixture` với `n_components=3`, `random_state=42`, `covariance_type='full'`.
    - Huấn luyện mô hình trên dữ liệu đã chuẩn hóa: `gmm.fit(X_scaled)`.
  - **Kết quả Bước Expectation (E-step) - Sau khi hội tụ:**
    - E-step tính toán xác suất mỗi điểm dữ liệu thuộc về từng cụm (component). Đây là các "trách nhiệm" (responsibilities).
    - Trong `sklearn`, sau khi `fit()`, chúng ta có thể lấy các xác suất này bằng `gmm.predict_proba(X_scaled)`.

Slide 7: Kết quả Bước Maximization (Cuối cùng)

- **Tiêu đề:** Phân tích Bước Maximization - Tham số Mô hình
- **Nội dung:**
    - **Bước Maximization (M-step) - Sau khi hội tụ:**
        - M-step cập nhật các tham số của mô hình (trọng số, trung bình, hiệp phương sai của mỗi cụm) dựa trên các trách nhiệm đã tính ở E-step, nhằm tối đa hóa likelihood của dữ liệu.
        - Các tham số này được lưu trong mô hình `gmm` sau khi huấn luyện:
            - **Trọng số cụm (`gmm.weights_`):** Tỷ lệ của mỗi cụm trong hỗn hợp.
                - Ví dụ: Cụm 0: 0.33, Cụm 1: 0.35, Cụm 2: 0.32
            - **Trung bình cụm (`gmm.means_`):** Vector trung tâm của mỗi cụm (cho dữ liệu đã chuẩn hóa).
                - Ví dụ (đã chuẩn hóa):
                    - Cụm 0: `[-0.9, -1.2]`

Slide 8: Trực quan hóa Kết quả - Biểu đồ Phân cụm

- **Tiêu đề:** Trực quan hóa các Cụm Dữ liệu
- **Nội dung:**
  - Biểu đồ scatter plot của hai đặc trưng ('Adj Close' và 'Volume' - đã chuẩn hóa) với các điểm được tô màu theo cụm được gán bởi GMM.
  - Các hình ellipse được vẽ chồng lên biểu đồ, đại diện cho đường đồng mức (contour) của các phân phối Gaussian cho mỗi cụm.
    - Ellipse thể hiện trung tâm (mean) và hình dạng/độ phân tán (covariance) của mỗi cụm.
  - **Mục đích:** Giúp hình dung vị trí, kích thước và hình dạng tương đối của các cụm trong không gian đặc trưng.
- **Hình ảnh:** Biểu đồ scatter plot với các cụm và ellipse GMM từ output của script.
  - `[Image of Biểu đồ phân cụm GMM với Adj Close và Volume]`

Slide 9: Trực quan hóa Kết quả - Phân phối Đặc trưng theo Cụm

- **Tiêu đề:** Phân tích Đặc điểm của Từng Cụm
- **Nội dung:**
  - **Biểu đồ Histogram/KDE:**
    - Vẽ biểu đồ phân phối (histogram có đường ước lượng mật độ Kernel - KDE) cho từng đặc trưng ('Adj Close', 'Volume' - đã chuẩn hóa), được tách riêng cho mỗi cụm.
    - Giúp so sánh sự khác biệt trong phân phối giá trị của các đặc trưng giữa các cụm.
  - **Biểu đồ Pairplot:**
    - Hiển thị mối quan hệ giữa tất cả các cặp đặc trưng, được tô màu theo cụm (sử dụng dữ liệu gốc để dễ diễn giải).
    - Cung cấp cái nhìn tổng quan về cách các cụm phân tách trong không

Slide 10: Kết luận và Hướng Phát triển

- **Tiêu đề:** Tóm tắt Kết quả và Định hướng Tương lai
- **Nội dung:**
  - **Tóm tắt kết quả:**
    - Thuật toán EM-GMM đã được áp dụng thành công để phân cụm dữ liệu giao dịch chứng khoán thành 3 cụm (ví dụ) dựa trên 'Adj Close' và 'Volume'.
    - Đã xác định được các tham số (trọng số, trung bình, hiệp phương sai) cho từng cụm, mô tả đặc điểm của chúng.
    - Các biểu đồ trực quan hóa cho thấy sự phân tách tương đối rõ ràng giữa các cụm.
  - **Ý nghĩa:** Các cụm có thể đại diện cho các "trạng thái" hoặc "chế độ" thị trường khác nhau (ví dụ: thị trường tăng trưởng ổn định với khối lượng thấp, thị

Lưu ý:

- Các giá trị cụ thể (ví dụ: trọng số, trung bình) trong slide chỉ mang tính minh họa và sẽ phụ thuộc vào kết quả chạy thực tế của script với dữ liệu `data.csv` của bạn.
- Bạn có thể tùy chỉnh số lượng slide, thêm/bớt nội dung chi tiết cho phù hợp với yêu cầu báo cáo.