**BÀI TẬP LỚN**

**Môn học: Phân tích và Xử lý Dữ liệu Đa chiều**  
**Đề tài: Phân tích Đa chiều Đặc tính Hóa lý và Quang phổ của Các mẫu Ngũ cốc từ Nhiều Tỉnh Thành**

**1. Mục tiêu**

* Hiểu và áp dụng các phương pháp phân tích đa chiều (PCA, MCA, MFA) để khám phá dữ liệu thực tế.
* Sử dụng hồi quy đa biến và hồi quy logistic để dự đoán và phân loại các đặc tính của ngũ cốc.
* Rèn luyện kỹ năng tiền xử lý, trực quan hóa và diễn giải kết quả từ dữ liệu đa chiều trong lĩnh vực khoa học nông nghiệp.

**2. Dữ liệu**

Dữ liệu được cung cấp bao gồm:

* **Thông số hóa lý:** Độ trắng (Whiteness), chiều cao (Height), chiều dài (Length), độ ẩm (Moisture), tro (Ash), tinh bột (Starch), amylose (Amylose), lipid (Lipid), protein (Protein).
* **Dữ liệu quang phổ:** Độ hấp thụ của nước (Water) và ethanol (Etanol) trong khoảng bước sóng từ 190 nm đến 1100 nm (mỗi bước sóng là một cột, ví dụ: Water\_190, Water\_191,..., Ethanol\_190, Ethanol\_191,...).
* **Biến phân loại:** Giống ngũ cốc (Variety), tỉnh thành (Province).
* **Số lượng mẫu:** Dữ liệu chứa nhiều mẫu, đại diện cho các giống và tỉnh thành khác nhau.

**3. Nhiệm vụ**

**3.1. Tiền xử lý dữ liệu**

* **Kiểm tra dữ liệu:** Xác định và xử lý các giá trị thiếu (nếu có).
* **Chuẩn hóa:** Chuẩn hóa các biến số (hóa lý và quang phổ) về cùng thang đo (ví dụ: z-score) để đảm bảo tính đồng nhất.
* **Mã hóa biến phân loại:** Chuyển đổi các biến Variety và Province thành dạng số để sử dụng trong các phân tích sau.
* **Xử lý dữ liệu quang phổ:** Áp dụng các bước tiền xử lý đặc thù cho dữ liệu phổ (xem chi tiết trong mục 6 - Gợi ý).

**3.2. Phân tích khám phá**

* **Phân tích thành phần chính (PCA):** 
  + Áp dụng PCA lên dữ liệu quang phổ đã tiền xử lý (Water và Ethanol, từ 190 nm đến 1100 nm) để giảm chiều và xác định các thành phần chính (PCs).
  + Xác định số lượng PCs cần thiết để giải thích một phần lớn phương sai (gợi ý: mục tiêu khoảng 70-80%, nhưng có thể điều chỉnh tùy theo dữ liệu). Vẽ biểu đồ scree plot để hỗ trợ quyết định này.
  + Vẽ biplot để trực quan hóa sự phân bố của các mẫu và đóng góp của các bước sóng.
  + ***Câu hỏi định hướng:*** 
    1. *Các bước sóng nào (hoặc vùng bước sóng nào) có đóng góp lớn nhất vào PC1 và PC2? Liệu chúng có liên quan đến các thành phần hóa học cụ thể của ngũ cốc (như nước, protein, hoặc tinh bột) không?*
    2. *Có sự phân nhóm rõ ràng nào giữa các mẫu dựa trên dữ liệu quang phổ không? Nếu có, điều này có liên quan đến Variety hay Province không?*
    3. *Đồ thị quang phổ trung bình của các nhóm Variety hoặc Province có khác biệt đáng kể ở vùng bước sóng nào không?*
    4. *Số lượng PCs bạn chọn giải thích được bao nhiêu phần trăm phương sai? Liệu ngưỡng này có đủ để đại diện cho dữ liệu quang phổ không, hay cần thêm PCs?*
* Phân tích tương quan đa biến (MCA):
  + Sử dụng MCA để phân tích mối quan hệ giữa Variety, Province và các biến hóa lý (nếu cần phân loại chúng thành nhóm, ví dụ: cao/thấp).
  + Vẽ biểu đồ phân tán MCA để nhận diện các nhóm giống hoặc tỉnh thành có đặc điểm tương đồng.
  + ***Câu hỏi định hướng:*** 
    1. *Các giống ngũ cốc (Variety) nào có xu hướng tương đồng về đặc tính hóa lý? Điều này có thể phản ánh đặc điểm di truyền hay điều kiện môi trường không?*
    2. *Các tỉnh thành (Province) có phân bố gần nhau trên biểu đồ MCA không? Nếu có, điều này có thể liên quan đến yếu tố địa lý hoặc khí hậu không?*
* Phân tích nhân tố đa biến (MFA):
  + Tích hợp hai nhóm dữ liệu: (1) thông số hóa lý và (2) dữ liệu quang phổ (có thể dùng PCs từ PCA).
  + Xác định đóng góp của từng nhóm biến vào cấu trúc tổng thể.
  + Vẽ biplot MFA để trực quan hóa mối quan hệ giữa các mẫu, biến hóa lý và quang phổ.
  + ***Câu hỏi định hướng:*** 
    1. *Nhóm biến hóa lý hay quang phổ đóng góp nhiều hơn vào sự khác biệt giữa các mẫu? Điều này có ý nghĩa gì về việc sử dụng quang phổ để phân tích ngũ cốc?*
    2. *Có mối quan hệ nào giữa các biến hóa lý (như Starch, Protein) và các vùng bước sóng cụ thể trong dữ liệu quang phổ không?*
    3. *Các mẫu từ cùng một Variety hoặc Province có xu hướng tập trung gần nhau trên biplot MFA không? Nếu không, điều gì có thể giải thích sự phân tán này?*

**3.3. Mô hình hóa**

* **Hồi quy đa biến:**
  + Xây dựng mô hình hồi quy đa biến để dự đoán hàm lượng tinh bột (Starch) dựa trên các biến đầu vào:
    - Các thành phần chính từ PCA của dữ liệu quang phổ (Water hoặc Ethanol).
    - Một số biến hóa lý khác (ví dụ: Moisture, Protein).
  + Đánh giá mô hình bằng R² và RMSE.
  + ***Câu hỏi định hướng:***
    - *Các thành phần chính từ dữ liệu quang phổ có cải thiện khả năng dự đoán hàm lượng tinh bột so với chỉ sử dụng các biến hóa lý không?*
    - *Biến nào (hóa lý hay quang phổ) có hệ số hồi quy lớn nhất? Điều này có phù hợp với kỳ vọng về mối quan hệ vật lý giữa các biến không?*
    - *Nếu RMSE cao, đâu có thể là nguyên nhân (nhiễu trong dữ liệu, biến đầu vào không đủ mạnh, hay mô hình chưa tối ưu)?*
* **Hồi quy logistic:**
  + Xây dựng mô hình hồi quy logistic để phân loại giống ngũ cốc (Variety) thành hai nhóm (ví dụ: chọn hai giống phổ biến nhất từ dữ liệu, như Variety 8 và Variety 30).
  + Sử dụng các biến đầu vào: PCs từ PCA và một số biến hóa lý (ví dụ: Whiteness, Protein).
  + Đánh giá mô hình bằng độ chính xác (accuracy) và ma trận nhầm lẫn (confusion matrix).
  + ***Câu hỏi định hướng:***
    - *Các biến đầu vào nào (PCs hay hóa lý) có ảnh hưởng lớn nhất đến khả năng phân loại giống ngũ cốc? Điều này có ý nghĩa gì về đặc trưng của các giống?*
    - *Nếu mô hình phân loại sai nhiều mẫu, đâu là đặc điểm chung của các mẫu bị nhầm lẫn (ví dụ: giá trị hóa lý tương tự, quang phổ gần giống)?*
    - *Liệu việc thêm nhiều biến đầu vào (như các PCs khác) có cải thiện độ chính xác không, hay gây overfitting?*
* **Phân loại cơ bản:**
  + Dựa trên ngưỡng của PC1 từ PCA (ví dụ: trên/dưới trung vị), phân loại các mẫu thành hai nhóm và so sánh với Variety hoặc Province thực tế.
  + Tính tỷ lệ chính xác của phương pháp phân loại này và so sánh với hồi quy logistic.
  + ***Câu hỏi định hướng:***
    - *Phương pháp phân loại đơn giản dựa trên PC1 có hiệu quả đến đâu so với hồi quy logistic? Điều này cho thấy PC1 có nắm bắt được đặc trưng chính của dữ liệu không?*
    - *Các mẫu bị phân loại sai có xu hướng nằm ở vùng biên của ngưỡng không? Nếu có, điều này gợi ý gì về sự phân bố dữ liệu?*

**3.4. Trực quan hóa và báo cáo**

* **Trực quan hóa:**
  + Vẽ biplot cho PCA và MFA.
  + Vẽ biểu đồ phân tán cho MCA.
  + Vẽ đồ thị quang phổ trung bình của các nhóm Variety hoặc Province sau khi tiền xử lý.
  + Vẽ biểu đồ dự đoán thực tế so với dự đoán từ hồi quy đa biến và hồi quy logistic.
* **Báo cáo:**
  + Mô tả ngắn gọn các bước thực hiện (tiền xử lý, phân tích, mô hình hóa).
  + Trình bày kết quả chính từ PCA, MCA, MFA, hồi quy đa biến và hồi quy logistic, kèm theo câu trả lời cho các câu hỏi định hướng.
  + Kết luận:
    - Đặc điểm nổi bật của các giống ngũ cốc hoặc tỉnh thành dựa trên dữ liệu.
    - Hiệu quả của các phương pháp phân tích và mô hình hóa đã sử dụng.

**4. Yêu cầu nộp bài**

* **Báo cáo:** File PDF (tối đa 15 trang), bao gồm nội dung theo các mục trên, kèm hình ảnh minh họa và câu trả lời cho các câu hỏi định hướng.
* **Mã nguồn:** File mã (R, Python, hoặc phần mềm khác) được sử dụng để phân tích và vẽ biểu đồ.
* **Thời hạn:** 4 tuần kể từ ngày giao bài.

**5. Tiêu chí đánh giá**

* **Tiền xử lý dữ liệu (20%):** Độ chính xác và hợp lý trong xử lý dữ liệu, đặc biệt là dữ liệu quang phổ.
* **Phân tích khám phá (30%):** Ứng dụng đúng PCA, MCA, MFA và diễn giải kết quả, bao gồm trả lời các câu hỏi định hướng.
* **Mô hình hóa (30%):** Xây dựng và đánh giá mô hình hồi quy đa biến, hồi quy logistic, phân loại cơ bản, kèm insight từ câu hỏi định hướng.
* **Trình bày và báo cáo (20%):** Tính rõ ràng, logic và trực quan của báo cáo, mức độ sâu sắc của các câu trả lời.

**6. Gợi ý**

**6.1. Làm việc với dữ liệu quang phổ**

Dữ liệu quang phổ (Water và Ethanol) là tập hợp các giá trị độ hấp thụ tại hàng trăm bước sóng (190 nm đến 1100 nm), tạo thành một ma trận lớn với số cột rất nhiều. Để làm việc hiệu quả với dữ liệu này, cần thực hiện các bước sau:

* **Bước 1: Kiểm tra dữ liệu phổ:**
  + Vẽ đồ thị quang phổ trung bình (mean spectrum) của tất cả các mẫu để quan sát xu hướng chung (trục x: bước sóng, trục y: độ hấp thụ).
  + Kiểm tra các giá trị bất thường (outliers) bằng cách xem xét các mẫu có độ hấp thụ quá cao hoặc quá thấp so với trung bình.
* **Bước 2: Tiền xử lý dữ liệu phổ:**
  + **Loại bỏ nhiễu (Smoothing):** Sử dụng phương pháp trung bình trượt (moving average) hoặc Savitzky-Golay filter để làm mịn dữ liệu phổ, giảm nhiễu do đo lường. Ví dụ: Trong R, dùng gói prospectr với hàm savitzkyGolay; trong Python, dùng scipy.signal.savgol\_filter.
  + A mathematical equation with black text

    AI-generated content may be incorrect.**Chuẩn hóa (Normalization):** Áp dụng chuẩn hóa z-score (trừ trung bình và chia cho độ lệch chuẩn) cho từng bước sóng để loại bỏ sự khác biệt về thang đo giữa các mẫu.

Công thức:

Trong đó x là giá trị độ hấp thụ, μ là trung bình, σ là độ lệch chuẩn của cột bước sóng.

* + **Loại bỏ nền (Baseline Correction):** Trừ đi đường nền (baseline) bằng cách áp dụng phương pháp đường thẳng (linear baseline) hoặc đa thức (polynomial baseline). Ví dụ: Trong Python, dùng scipy.signal.detrend hoặc gói pybaselines.
  + **Chọn vùng phổ quan trọng (nếu cần):** Nếu dữ liệu quá lớn, có thể chọn một số vùng bước sóng đặc trưng (ví dụ: 400-700 nm hoặc 900-1100 nm) dựa trên đồ thị phổ trung bình hoặc kiến thức chuyên môn.
* **Bước 3: Kiểm tra sau tiền xử lý:**
  + Vẽ lại đồ thị quang phổ sau khi tiền xử lý để so sánh với dữ liệu gốc, đảm bảo nhiễu đã được giảm và dữ liệu vẫn giữ được đặc trưng.

**6.2. Công cụ và thư viện**

* **R:**
  + Gói FactoMineR (PCA, MCA, MFA), ca (MCA), prospectr (xử lý phổ).
  + Gói stats (hồi quy đa biến, hồi quy logistic), ggplot2 (trực quan hóa).
* **Python:**
  + Gói sklearn (PCA, hồi quy), mca hoặc prince (MCA, MFA).
  + Gói scipy và pybaselines (xử lý phổ), matplotlib hoặc seaborn (trực quan hóa).

**6.3. Lưu ý chung**

* Nếu dữ liệu phổ quá lớn (hàng trăm cột), hãy thử áp dụng PCA trước để giảm chiều trước khi tích hợp với dữ liệu hóa lý trong MFA.
* Tập trung vào diễn giải kết quả và trả lời các câu hỏi định hướng để tìm insight, thay vì chỉ trình bày số liệu thô.
* Nếu gặp khó khăn, có thể chọn một tập hợp con (ví dụ: 50 mẫu đầu tiên) để thử nghiệm trước khi phân tích toàn bộ dữ liệu.

====HẾT====