Machine Learning for Policy Analysis

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

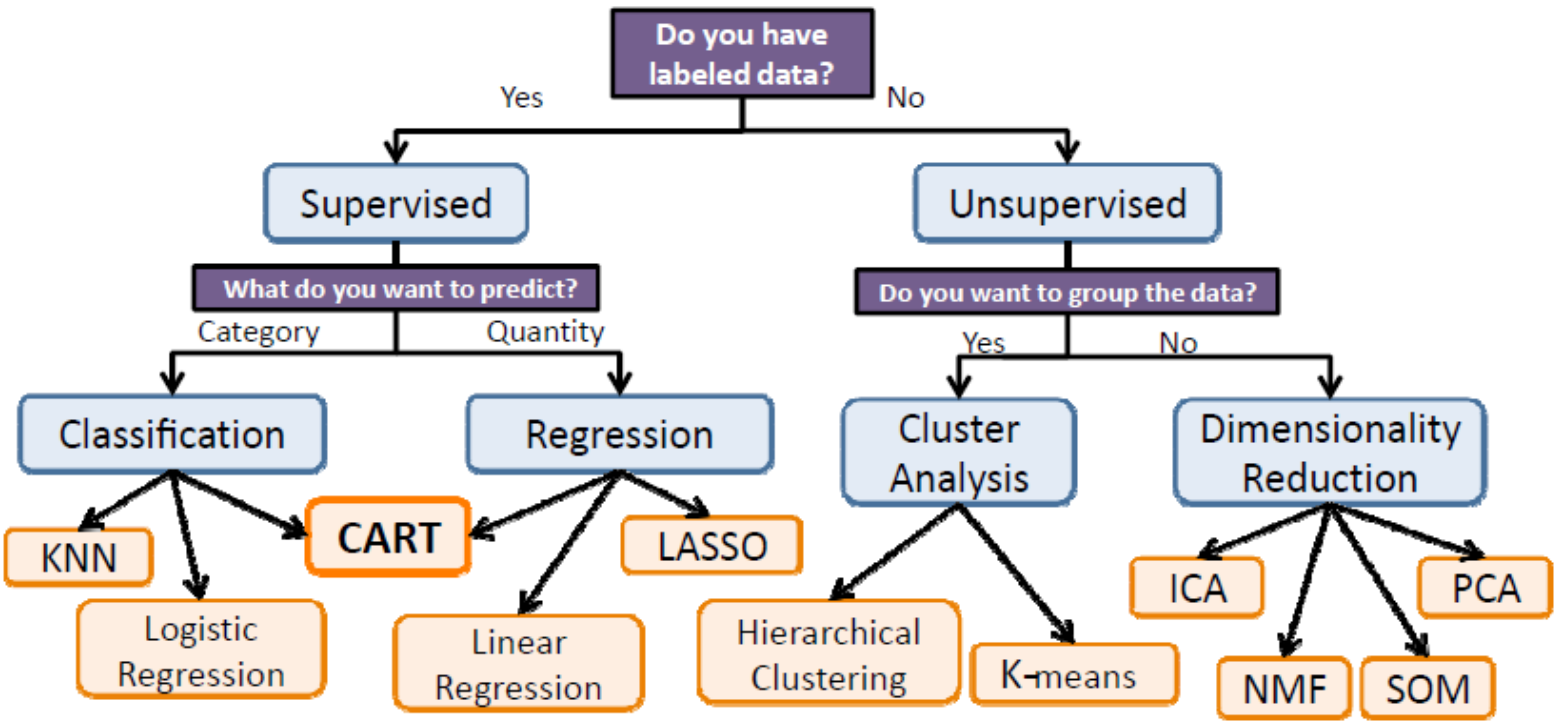
Ngày 16 tháng 1 năm 2019

Machine learning/Data mining là gì?

- Một nhóm các công cụ và thuật toán để tối đa hóa khả năng dự báo của mô hình.
- Khác biệt với tiếp cận kinh tế học, ML không cần thiết phải giả định về cấu trúc của mô hình.
- Nguồn gốc của ML là khoa học thống kê, tuy nhiên khả năng áp dụng trong kinh tế, kinh doanh, và xã hội rất lớn.

Một số phương pháp của machine learning

Machine Learning = Algorithm + Data



Source: Nguyễn Thanh Tùng, CSE445.

Một số ứng dụng của ML/DM trong phân tích kinh tế và kinh doanh

- Dự báo (prediction)
- Phân loại (classification)
- Phân cụm dữ liệu (clustering)
- Giảm chiều dữ liệu (dimension reduction)

Sử dụng ML để cải thiện mô hình hồi quy

- Khi chúng ta quan tâm đến khả năng dự báo của mô hình (prediction) thay vì hàm ý quan hệ nhân quả (causal relation)
- Có thể xây dựng mô hình để dự báo cho mẫu dữ liệu phân tích, nhưng khả năng dự báo ngoài mẫu (out-of-sample prediction) rất kém
- Các thủ thuật kiểm chứng chéo (cross-validation) có thể được sử dụng để giảm vấn đề ước lượng quá khớp (overfitting)

Ví dụ hiện tượng ước lượng quá khớp (overfitting)

- Sử dụng lại bộ dữ liệu VHLSS 2010 và ước lượng hàm tỷ suất thu nhập của đi học.
- Tạo ra các biến dummies đại diện cho từng tỉnh, huyện, xã, và số hộ gia đình.
- Ước lượng mô hình với lần lượt các biến dummies kể trên. So sánh sự thay đổi của R^2.
- Nhận xét khả năng dự báo của mô hình cho nhóm hộ không nằm trong mẫu dữ liệu?

Phương pháp kiểm chứng chéo (cross-validation)

Dùng mô hình để dự báo cho quan sát ngoài mẫu (out-of-sample prediction). Mô hình ước lượng quá khớp với dữ liệu ước lượng sẽ có sai số dự báo lớn với quan sát ngoài mẫu. Lựa chọn mô hình tối ưu sao cho sai số dự báo MSE là nhỏ nhất.

$$MSE = E[(y - \hat{y})^2]$$

Các thuật ngữ trong ML

- Supervised learning (học máy có giám sát)
  - Biến phụ thuộc liên tục: ML = Hồi quy
  - Biến phụ thuộc định tính: ML = Phân loại (classification)
- Unsupervised learning (học máy không giám sát)
  - Không có biến phụ thuộc
  - Phân nhóm dữ liệu tùy thuộc vào đặc tính của các biến giải thích
- Training data: Dữ liệu ước lượng
- Test data (validation data, hold-out sample): Dữ liệu kiểm chứng

Thuật giải của phương pháp kiểm chứng chéo

- Chia bộ dữ liệu ngẫu nhiên thành hai phần là bộ dữ liệu ước lượng (training data) và bộ dữ liệu kiểm chứng (validation data)
- Ước lượng mô hình đối với bộ dữ liệu ước lượng.
- Sử dụng mô hình của dữ liệu ước lượng để ước tính *MSE* cho dữ liệu kiểm chứng.
- Lựa chọn mô hình sao cho *MSE* là tối thiểu.

Các hình thức kiểm chứng chéo

- Leave-one-out Cross Validation (LOOCV)
  - Lần lượt chia bộ dữ liệu n quan sát thành training data với $(n-1)$ quan sát và test data với 1 quan sát.
  - Ước lượng giá trị dự báo $\hat{y}_{(-i)}$ đối với lần lượt các quan sát bị tách làm nhóm kiểm chứng.
  - Ước tính LOOCV như sau:

$$CV_(n) = \frac{1}{n}\sum_{i=1}^{n} MSE_{(-i)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(-i)})^2$$

- k-fold Cross Validation
  - Chia bộ dữ liệu thành K nhóm với số quan sát bằng nhau. Lấy nhóm 1 được sử dụng làm test data, $K-1$ nhóm sử dụng làm training data.
  - Ước lượng mô hình với training data, ước tính MSE cho nhóm 1.
  - Lặp lại K lần cho nhóm 2, 3,...
  - Ước tính

$$CV_{(K)} = \frac{1}{K}\sum_{j=1}^{K} MSE_{(j)}$$

- LOOCV là trường hợp khi $K = n$. Thông thường $K = 5$ hoặc $K = 10$.

Đánh đổi giữa độ chệch và phương sai (Bias-Variance Trade-off)

Giả sử chúng ta ước lượng mô hình từ training data:

$$y = f(x) + \varepsilon$$

và ước lượng *MSE* cho test data (x_0, y_0):

$$\begin{aligned} MSE &= E[(y_0 - \hat{f}(x_0))^2] \\ &= Var[\hat{f}(x_0)] + \{Bias(\hat{f}(x_0))\}^2 + Var(\varepsilon) \end{aligned}$$

Thông thường các mô hình càng linh động (flexible function) thì bias càng thấp nhưng phương sai của ước lượng càng cao.

Thực hành

- Thực hành với bộ dữ liệu mô phỏng.
- Thực hành với các tình huống dự báo khác.

Shrinkage Estimators

Phương pháp làm giảm độ phức tạp của mô hình bằng cách giảm SSR (tương tự như OLS), tuy nhiên có điều chỉnh cho kích cỡ (số biến giải thích) của mô hình (giống như sử dụng R^2 điều chỉnh để chọn biến giải thích).

Ridge Regression

Tối thiểu hóa *SSR* và *Penalty* lên kích cỡ của mô hình bằng β^2 và một hệ số λ:

$$\underbrace{\sum_{i=1}^{n}(y_i - X_i\beta)^2}_{SSR} + \underbrace{\lambda \sum_{j=1}^{K} \beta_j^2}_{Penalty}$$

- Tăng số biến giải thích trong mô hình (tăng K) thì *SSR* giảm nhưng *Penalty* có thể tăng.
- λ được gọi là tham số điều chỉnh (tuning parameter).

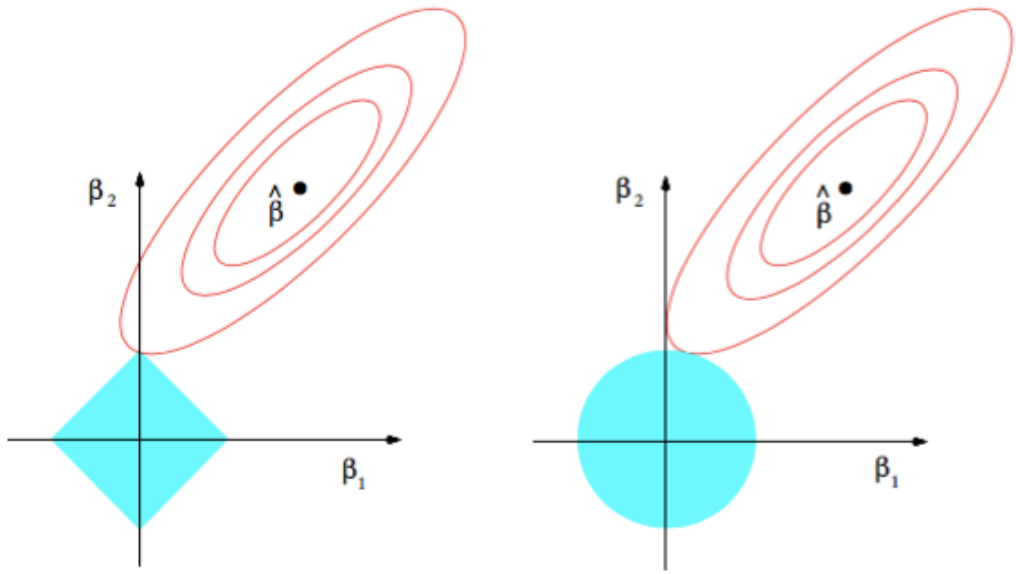
Least Absolute Shrinkage and Selection - LASSO

Tương tự như Ridge Regression, tuy nhiên *Penalty* được tính bằng $|\beta|$

$$\underbrace{\sum_{i=1}^{n}(y_i - X_i\beta)^2}_{SSR} + \underbrace{\lambda \sum_{j=1}^{K} |\beta_j|}_{Penalty}$$

Diễn giải phương pháp Ridge và LASSO

- Do β bị ảnh hưởng bởi đơn vị (scaling) của dữ liệu nên các biến giải thích được chuẩn hóa ($x_i^* = \frac{x_i - \bar{x}_i}{se(x_i)}$) trước khi ước lượng.
- Các phương pháp đều làm giảm β xuống ("shrink" an estimator) theo hướng bằng 0.
- Kết quả tối ưu khi mô hình chỉ có một vài $\beta_j \neq 0$ trong số các biến giải thích đưa vào mô hình (Lasso, trái) hay các tham số β_j nhỏ đi (Ridge, phải).



Thực hành

1. Chuẩn hóa bộ dữ liệu
2. Ước lượng mô hình với Lasso và Ridge
3. So sánh và lựa chọn mô hình tối ưu