Ensemble Learning

Uczenie Złożone

Aleksander Czeszejko-Sochacki

07.12.2017

Plan

- Definicja modeli złożonych
- Budowa modeli złożonych
- Zastosowania modeli złożonych
- Wady i zalety modeli złożonych

Idea uczenia złożonego

Problem

Mamy dane N modeli, chcemy na ich podstawie skonstruować jeden model.

Idea uczenia złożonego

Problem

Mamy dane N modeli, chcemy na ich podstawie skonstruować jeden model.

Rozwiązania naiwne (niekoniecznie złe)

- Dla wektora z danych testowych uśredniamy wyniki modeli.
- Dla wektora z danych testowych zwracamy wynik, który zwróciłoby z osobna najwięcej z naszych N modeli.

Idea uczenia złożonego

Problem

Mamy dane N modeli, chcemy na ich podstawie skonstruować jeden model.

Rozwiązania naiwne (niekoniecznie złe)

- Dla wektora z danych testowych uśredniamy wyniki modeli.
- Dla wektora z danych testowych zwracamy wynik, który zwróciłoby z osobna najwięcej z naszych N modeli.

Rozwiązanie mądre

Opracowujemy algorytm, który pozwoli stwierdzić, jak bardzo możemy ufać każdemu z wybranych modeli.

Przykład rozwiązania naiwnego

Klasyfikacja słów kodowych na podstawie zero-jedynkowych biklasyfikatorów

TABLE 16.1. *Part of a 15-bit error-correcting coding matrix* **C** *for the 10-class digit classification problem. Each column defines a two-class classification problem.*

Digit	C_1	C_2	C_3	C_4	C_5	C_6	$\cdots$	C_{15}
0	1	1	0	0	0	0	$\cdots$	1
1	0	0	1	1	1	1	$\cdots$	0
2	1	0	0	1	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
8	1	1	0	1	0	1	$\cdots$	1
9	0	1	1	1	0	0	$\cdots$	0

ECOC - error correcting output codes

Ważne

Z własności kodów korekcyjnych wynika jednoznaczność klasyfikacji każdego nowego słowa kodowego przez model złożony.

Regresja z penalizacją jako sposób na budowanie modelu złożonego

Niech $\{\Gamma_k\}_{k \in 1,2,\ldots,K}$ - pewien zbiór modeli. Funkcją przyporządkowania wektora x definiujemy następująco:

$$f(x) = \sum_{k=1}^{K} \alpha_k T_k$$

gdzie K $=$ card(Γ)

Regresja z penalizacją jako sposób na budowanie modelu złożonego

Chcemy minimalizować wartość wyrażenia względem wektora współczynników α:

$$\min \left\{ \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} \alpha_k T_k(x_i) \right)^2 + \lambda * J(\alpha) \right\}$$

gdzie J(α) - funkcja penalizacji (λ ustalona), na przykład

$J(\alpha) = \sum_{k=1}^{k} (|\alpha_k|)^2$ regresja grzbietowa
$J(\alpha) = \sum_{k=1}^{k} |\alpha_k|$ lasso

Regresja z penalizacją jako sposób na budowanie modelu złożonego

Jak to zrealizować?

- Metoda brutalna
- Algorytm Forward Stagewise Linear Regression

Algorithm 16.1 *Forward Stagewise Linear Regression.*

1. Initialize $\check{\alpha}_k = 0,\ k = 1, \dots, K$. Set $\varepsilon > 0$ to some small constant, and M large.
2. For $m = 1$ to M:
   (a) $(\beta^*, k^*) = \arg\min_{\beta,k} \sum_{i=1}^{N} \left(y_i - \sum_{l=1}^{K} \check{\alpha}_l T_l(x_i) - \beta T_k(x_i) \right)^2$.
   (b) $\check{\alpha}_{k^*} \leftarrow \check{\alpha}_{k^*} + \varepsilon \cdot \operatorname{sign}(\beta^*)$.
3. Output $f_M(x) = \sum_{k=1}^{K} \check{\alpha}_k T_k(x)$.

Forward Stagewise Linear Regression

Algorytm ten posiada ważną własność: wiele ze współczynników α będzie zerami.

Porównanie Lasso i Forward Stagewise Linear Regression

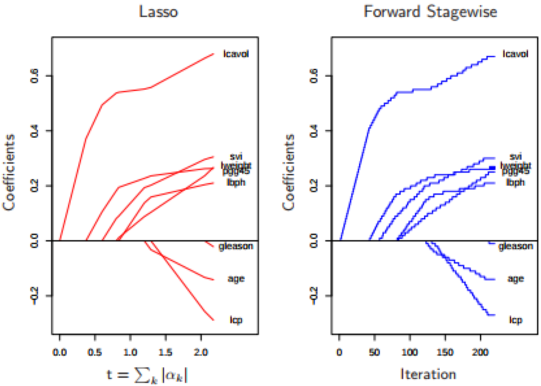


FIGURE 16.1. *Profiles of estimated coefficients from linear regression, for the prostate data studied in Chapter 3. The left panel shows the results from the lasso, for different values of the bound parameter $t = \sum_k |\alpha_k|$. The right panel shows the results of the stagewise linear regression Algorithm 16.1, using $M = 220$ consecutive steps of size $\varepsilon = .01$.*

Zasada "The Bet on Sparsity"(założenie rzadkości)

Wybór funkcji penalizacji ma zazwyczaj znaczący wpływ na dokładność modelu.

Zasada "The Bet on Sparsity"(założenie rzadkości)

Wybór funkcji penalizacji ma zazwyczaj znaczący wpływ na dokładność modelu.

Przykład

Jeśli przypuszczamy, że zbiór wartości współczynników α będzie zbliżony do zbioru losowych wartości z rozkładu gaussowskiego, to najlepszą predykcją wykaże się model zbudowany przy użyciu regresji grzbietowej (ridge regression). (Mało współczynników o dużych wartościach bezwzględnych)

Zasada “The Bet on Sparsity”(założenie rzadkości)

Wybór funkcji penalizacji ma zazwyczaj znaczący wpływ na dokładność modelu.

Przykład

Jeśli przypuszczamy, że zbiór wartości współczynników α będzie zbliżony do zbioru losowych wartości z rozkładu gaussowskiego, to najlepszą predykcją wykaże się model zbudowany przy użyciu regresji grzbietowej (ridge regression). (Mało współczynników o dużych wartościach bezwzględnych)

Rzadkie i gęste zbiory współczynników

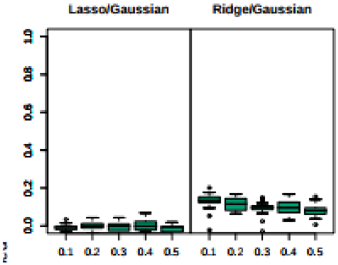
Naszym celem jest to, aby zbiór współczynników był rzadki. Wtedy mniejszym błędem będzie obarczony model zbudowany przy użyciu regresji lasso.

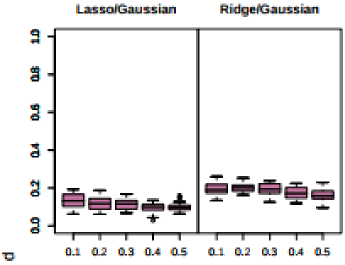
Sformułowanie zasady "The Bet on Sparsity"

The Bet on Sparsity

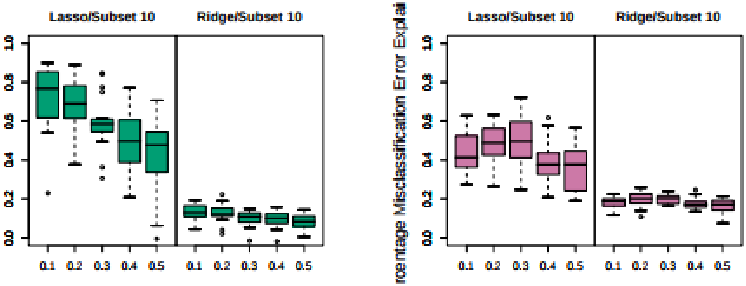
Do zbudowania modelu złożonego użyj metody, która sprawdza się dla zbiorów rzadkich, ponieważ nie ma dostatecznie dobrej metody dla zbiorów gęstych.

Porównanie działania modeli opartych na regresji grzbietowej i regresji lasso na różnych zbiorach danych

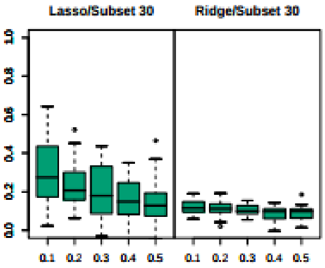


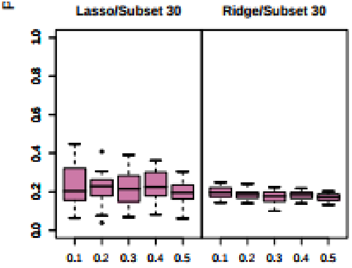


Porównanie działania modeli opartych na regresji grzbietowej i regresji lasso na różnych zbiorach danych



Porównanie działania modeli opartych na regresji grzbietowej i regresji lasso na różnych zbiorach danych





Porównanie rezultatów regresji lasso i algorytmu Forward Stagewise

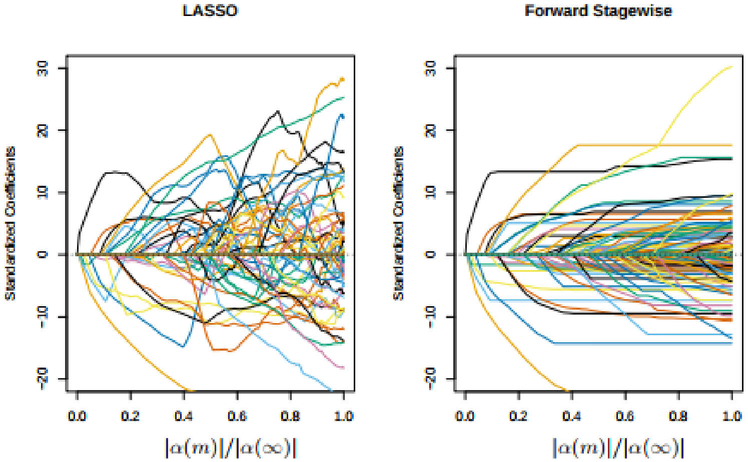


FIGURE 16.3. *Comparison of lasso and infinitesimal forward stagewise paths*

Czym jest przetwarzanie wtórne (post-processing)?

Do metod złożonych zaliczamy m.in. Random Forest czy drzewa addytywne. Można je poddać tzw. post-processingowi, tzn, użyć pewnego rodzaju regresji do zbudowania w oparciu o ich wyniki bardziej wiarygodnego modelu. Przypomijmy:

$$f(x) = a_0 + \sum_{T_k \in T} \alpha_k T_k(x)$$

gdzie T - zbiór modeli, card(T) - duże.

Czym jest przetwarzanie wtórne (post-processing)?

Do metod złożonych zaliczamy m.in. Random Forest czy drzewa addytywne. Można je poddać tzw. post-processingowi, tzn, użyć pewnego rodzaju regresji do zbudowania w oparciu o ich wyniki bardziej wiarygodnego modelu. Przypomijmy:

$$f(x) = a_0 + \sum_{T_k \in T} \alpha_k T_k(x)$$

gdzie T - zbiór modeli, card(T) - duże.

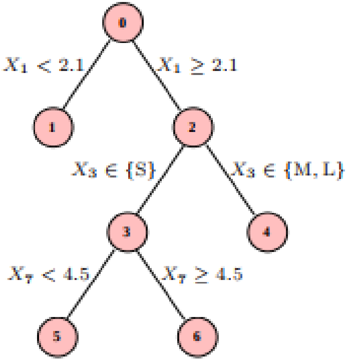
Post-processing (Friedman and Popescu, 2003)

- Indukujemy zbiory $T_L = \{T_1(x), T_2(x), \ldots, T_M(x)\}$ po T.
- Używamy regresji lasso w następujący sposób (minimalizacja po α):

$$\alpha(\lambda) = argmin \sum_{i=1}^{N} L[y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m(x_i)] + \lambda \sum_{m=1}^{M} |\alpha|$$

Modele złożone reguł (Rule ensembles)

Mamy pewne drzewo klasyfikacyjne (dla uproszczenia przykładu nieduże). Dla każdego węzła i liścia będziemy chcieli znaleźć zbiór reguł go opisujący.



Modele złożone reguł (Rule ensembles)

Zbiory te wyglądają następująco:

$$
\begin{aligned}
R_1(X) &= I(X_1 < 2.1) \\
R_2(X) &= I(X_1 \geq 2.1) \\
R_3(X) &= I(X_1 \geq 2.1) * I(X_3 \in \{S\}) \\
R_4(X) &= I(X_1 \geq 2.1) * I(X_3 \in \{M, L\}) \\
R_5(X) &= I(X_1 \geq 2.1) * I(X_3 \in \{S\}) * I(X_7 < 4.5) \\
R_6(X) &= I(X_1 \geq 2.1) * I(X_3 \in \{S\}) * I(X_7 \geq 4.5)
\end{aligned}
$$

Budowa modelu złożonego opartego na regułach

1. Niech T - zbiór drzew, card(T) = M. Dla każdego drzewa T^m konstruujemy zbiór reguł T^m_{RULE} $(m \in \{1, 2, \dots, M\})$
2. Model złożony budowany na zbiorze T wyraża się następująco:

$$T_{RULE} = \bigcup_{m=1}^{M} T^m_{RULE}$$

3. (Opcjonalnie) Dokonujemy regularyzacji powyższego zbioru.

Porównanie jakości modelu złożonego opartego na regułach przed użyciem regresji liniowej i po użyciu

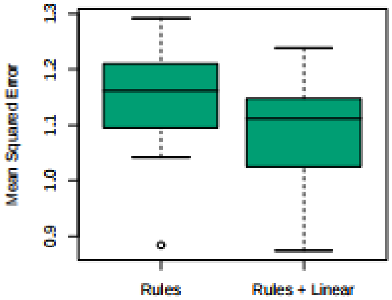
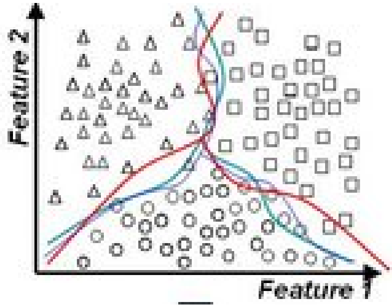


FIGURE 16.10. *Mean squared error for rule ensembles, using 20 realizations of the simulation example (16.13).*

Modele złożone a granice decyzyjne

Mamy zbiór modeli wyznaczających granice decyzyjne. Możliwe są:

- Wyznaczanie obszarów "wątpliwych"
- Aproksymacja rzeczywistych granic decyzyjnych
- Interpolacja granicy wielomianem przechodzącym przez punkty wspólne dla granic z różnych modeli



Bibliografia

- Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning - Data Mining, Interference, and Prediction, Springer, 2013
- dr Robi Polikar, $http://www.scholarpedia.org/article/Ensemble_learning$