Zaawansowane uczenie maszynowe:
wykład 9

Paweł Cichosz

Tworzenie modeli bazowych

Różne zbiory trenujące: zastosowanie tego samego algorytmu do różnych zbiorów przykładów z tej samej dziedziny (lub tego samego zbioru przykładów z różnymi wagami).

Różne algorytmy: zastosowanie różnych algorytmów do tego samego zbioru przykładów,

Różne ustawienia parametrów: zastosowanie tego samego algorytmu z różnymi ustawieniami parametrów do tego samego zbioru przykładów.

Randomizacja algorytmu: wielokrotne niezależne uruchomienia algorytmu z elementami losowymi w zastosowaniu do tego samego zbioru przykładów.

Łączenie predykcji

Głosowanie/uśrednianie: klasa/wartość funkcji docelowej uzyskiwana przez zwykłe głosowanie/uśrednianie predykcji modeli bazowych.

Ważone głosowanie/uśrednianie: klasa/wartość funkcji docelowej uzyskiwana przez ważone głosowanie/uśrednianie predykcji modeli bazowych.

Użycie jako atrybutów: klasa/wartość funkcji docelowej wyznaczana przez model, dla którego predykcje modeli bazowych pełnią funkcję atrybutów.

Bagging

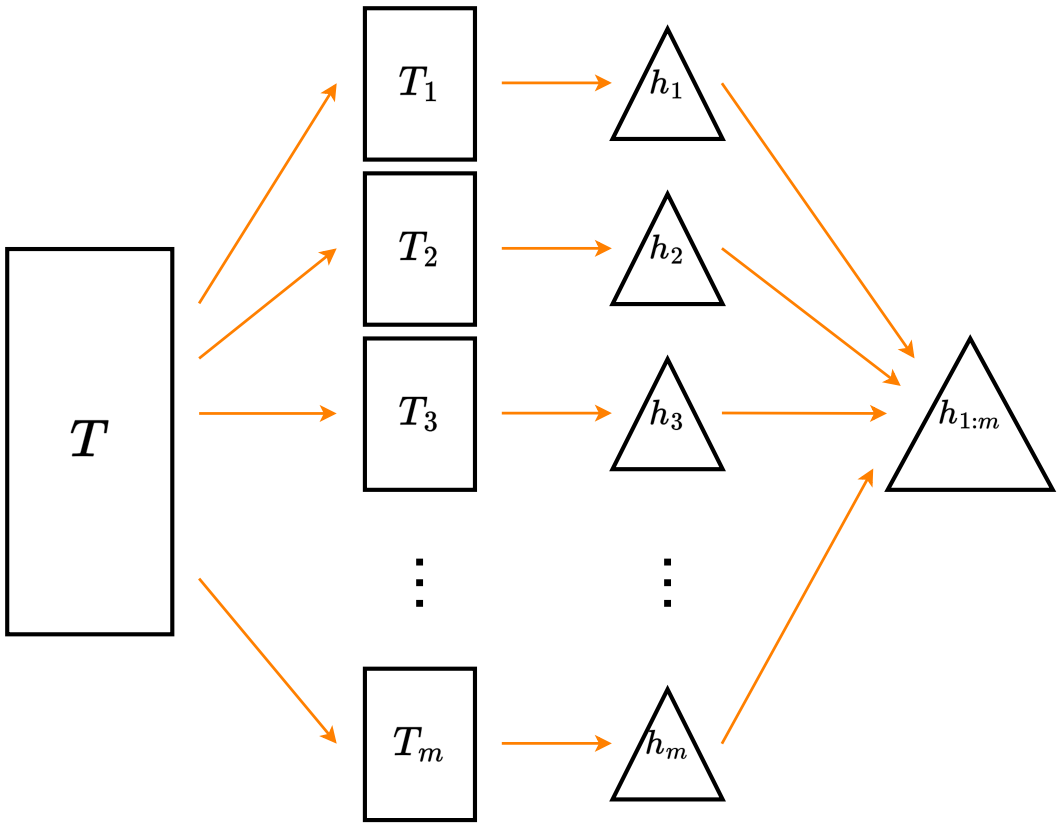
Modele bazowe: tworzone przez ten sam algorytm stosowany do *bootstrapowych prób* (losowanych ze zwracaniem) ze zbioru trenującego:

bag: ok. 63.2% przykładów $(1 - (1 - \frac{1}{N})^N \approx 1 - 1/e \approx 0.632)$.

out of bag: ok. 36.8% przykładów.

Łączenie predykcji: zwykłe głosowanie/uśrednianie.

Bagging



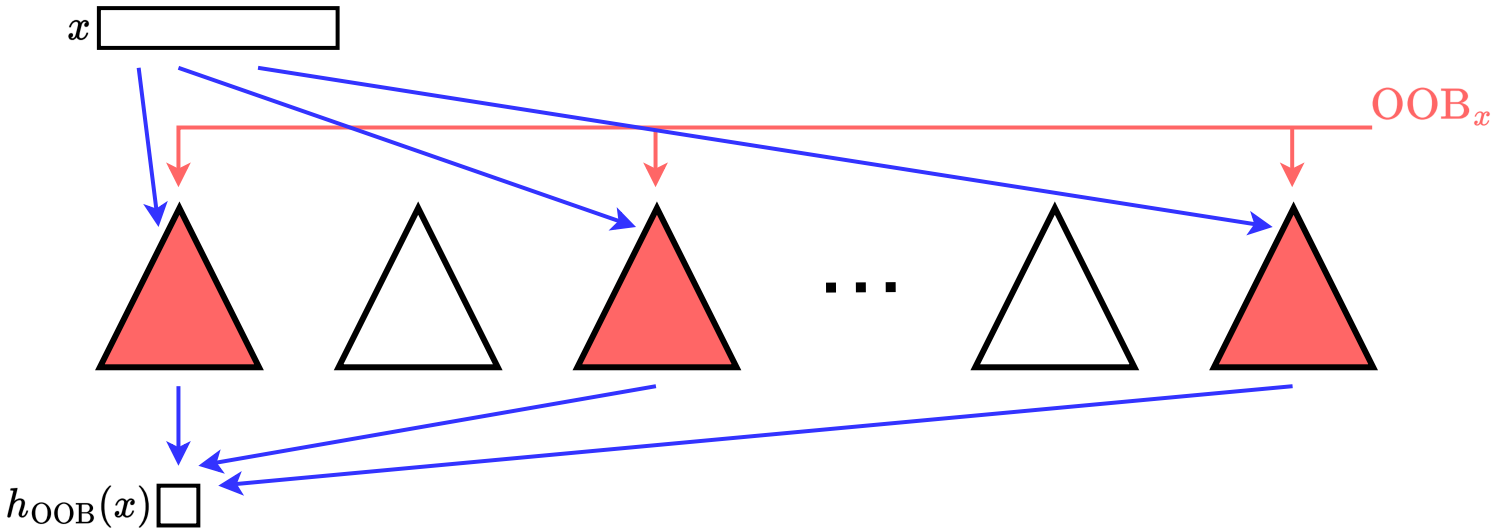
Właściwości: umiarkowana poprawa jeśli do tworzenia modeli bazowych jest stosowany niestabilny algorytm – wrażliwy na zaburzenia zbioru trenującego (zwykle drzewa decyzyjne/drzewa regresji).

Ocena jakości OOB

- $T_i' = T - T_i$ – zbiór przykładów *niewylosowanych* do próby bootstrapowej T_i (*out-of-bag*).
- Dla każdego $x \in T$:

$$\mathrm{OOB}_x = \left\{ i \in \{1, 2, \ldots, m\} \mid x \in T_i' \right\}$$

- Predykcja $h_{\mathrm{OOB}}(x)$ uzyskiwana przez głosowanie modeli h_i dla $i \in \mathrm{OOB}_x$ i porównywana z $c(x)$.



Boosting

Modele bazowe: tworzone przez ten sam algorytm stosowany tak, aby kolejny model poprawiał niedoskonałości poprzednich, np.:

- zmiany wag przykładów dla klasyfikacji: zwiększanie wag przykładów niepoprawnie klasyfikowanych przy tworzeniu kolejnego modelu.
- kompensacja residuów dla regresji: kolejny model przewiduje residua zespołu dotychczasowych modeli.

Łączenie predykcji: ważone głosowanie/uśrednianie (sumowanie).

Właściwości: duża poprawa jakości predykcji, dobre efekty z użyciem prostych modeli bazowych (zwykle drzewa decyzyjne o małej liczbie poziomów).

Algorytm *xgboost*

Modele bazowe: drzewa drzewa regresji z binarnymi podziałami – w liściach wartości funkcji docelowej (klasyfikacja możliwa np. przez zastosowanie zewnętrznej funkcji logistycznej).

Tworzenie modeli bazowych: optymalizacja miary jakości obejmującej:

- składnik straty reprezentujący poziom dopasowania do danych trenujących,
- składnik regularyzacji reprezentujący złożoność.

Łączenie predykcji: sumowanie.

Jakość modelu

Miara jakości modelu:

$$Q(h_{1:m}) = \sum_{x \in T} \mathcal{L}(f(x), h_{1:m}(x)) + \sum_{i=1}^{m} \Gamma(h_i)$$

gdzie:

- h_i – i-ty model bazowy (drzewo),
- $h_{1:m}$ – model zespołowy,
- $h_{1:m}(x)$ – predykcja zespołu drzew $h_1, \ldots, h_m$ dla przykładu x:

$$h_{1:m}(x) = \sum_{i=1}^{m} h_i(x)$$

- $f(x)$ – prawdziwa wartość atrybutu docelowego dla przykładu x (także klasa $0/1$ w przypadku klasyfikacji),
- $\mathcal{L}$ – funkcja straty – miara jakości predykcji.
- $\Gamma(h_i)$ – składnik regularyzacji dla drzewa h_i – kara za złożoność w celu ograniczenia ryzyka nadmiernego dopasowania.

Tworzenie modeli bazowych

Pierwszy model: h_1 – stałe 0 (jeden liść z wartością 0).

Kolejny model: h_i dla $i = 2, \ldots, m$: tworzony tak aby optymalizować $Q(h_{1:i})$, gdzie $h_{1:i}$ jest zespołem złożonym z modeli $h_1, \ldots, h_i$.

Funkcja celu optymalizacji: suma wartości funkcji straty i składnika regularyzacji dla i-tego modelu:

$$Q(h_i) = \sum_{x \in T} \mathcal{L}(f(x), h_{1:i-1}(x) + h_i(x)) + \Gamma(h_i)$$

Rozwinięcie Taylora

$$Q(h_i) = \sum_{x \in T} \mathcal{L}(f(x), h_{1:i-1}(x) + h_i(x)) + \Gamma(h_i)$$

Szereg Taylora (przypomnienie):

$$\phi(v) = \phi(b) + \phi'(b)(v - b) + \frac{1}{2}\phi''(b)(v - b)^2 + \dots$$

przy czym u nas:

$$\begin{aligned} \phi(v) &= \mathcal{L}(f(x), v) \\ v &= h_{1:i-1}(x) + h_i(x) \\ b &= h_{1:i-1}(x)) \\ v - b &= h_i(x) \end{aligned}$$

Realizacja optymalizacji

Składnik straty: przybliżenie za pomocą początkowych wyrazów rozwinięcia Taylora:

$$\mathcal{L}(f(x), h_{1:i-1}(x) + h_i(x))$$
$$\approx \mathcal{L}(f(x), h_{1:i-1}(x)) + \mathcal{G}_x h_i(x) + \frac{1}{2}\mathcal{H}_x h_i^2(x)$$

gdzie

$$\mathcal{L}(f(x), h_{1:i-1}(x)) = \text{const} \quad \text{(nie zależy od } h_i(x)\text{)}$$
$$\mathcal{G}_x = \frac{\partial \mathcal{L}(f(x), h_{1:i-1}(x))}{\partial h_{1:i-1}(x)}$$
$$\mathcal{H}_x = \frac{\partial^2 \mathcal{L}(f(x), h_{1:i-1}(x))}{\partial h_{1:i-1}^2(x)}$$

(w literaturze zwykle stosowane oznaczenia g jak *gradient* i h jak *hesjan* zamiast $\mathcal{G}$ i $\mathcal{H}$, ale u nas te symbole były wcześniej wykorzystywane w innej roli).

Realizacja optymalizacji

Składnik regularyzacji:

$$\Gamma(h_i) = \gamma|\mathbf{L}(h_i)| + \frac{1}{2}\lambda \sum_{\mathbf{l}\in\mathbf{L}(h_i)} w_{\mathbf{l}}^2$$

- $\mathbf{L}(h_i)$ – zbiór liści drzewa h_i,
- $w_{\mathbf{l}}$ – wartość w liściu $\mathbf{l}$.

Budowa drzewa

Minimalizacja:

$$
\begin{aligned}
q(h_i) &= \sum_{x \in T} \left[\mathcal{G}_x h_i(x) + \frac{1}{2} \mathcal{H}_x h_i^2(x) \right] + \gamma |\mathbf{L}(h_i)| + \frac{1}{2} \lambda \sum_{\mathbf{l} \in \mathbf{L}(h_i)} w_{\mathbf{l}}^2 \\
&= \sum_{\mathbf{l} \in \mathbf{L}(h_i)} \left[\sum_{x \in T_{\mathbf{l}}} \mathcal{G}_x w_{\mathbf{l}} + \frac{1}{2} \left(\sum_{x \in T_{\mathbf{l}}} \mathcal{H}_x + \lambda \right) w_{\mathbf{l}}^2 \right] + \gamma |\mathbf{L}(h_i)| \\
&= \sum_{\mathbf{l} \in \mathbf{L}(h_i)} \left[\mathcal{G}_{\mathbf{l}} w_{\mathbf{l}} + \frac{1}{2} \left(\mathcal{H}_{\mathbf{l}} + \lambda \right) w_{\mathbf{l}}^2 \right] + \gamma |\mathbf{L}(h_i)|
\end{aligned}
$$

gdzie:

$$
\mathcal{G}_{\mathbf{l}} = \sum_{x \in T_{\mathbf{l}}} \mathcal{G}_x
$$

$$
\mathcal{H}_{\mathbf{l}} = \sum_{x \in T_{\mathbf{l}}} \mathcal{H}_x
$$

Budowa drzewa

Optymalizacja wartości w liściach:

$$\frac{\partial q(h_i)}{\partial w_{\mathbf{l}}} = 0$$
$$\mathcal{G}_{\mathbf{l}} + (\mathcal{H}_{\mathbf{l}} + \lambda) w_{\mathbf{l}} = 0$$
$$w_{\mathbf{l}} = -\frac{\mathcal{G}_{\mathbf{l}}}{\mathcal{H}_{\mathbf{l}} + \lambda}$$

Jakość drzewa po optymalizacji wartości w liściach:

$$q^*(h_i) = -\frac{1}{2} \sum_{\mathbf{l} \in \mathbf{L}(h_i)} \frac{\mathcal{G}_{\mathbf{l}}^2}{\mathcal{H}_{\mathbf{l}} + \lambda} + \gamma |\mathbf{L}(h_i)|$$

Budowa drzewa

Kryterium wyboru podziału dla $\mathbf{l}$: maksymalizacja różnicy jakości drzewa przed podziałem i po podziale:

$$\Delta q^*_{\mathbf{l}}(t) = \frac{1}{2}\left[\sum_{r \in R_t} \frac{\mathcal{G}^2_{\mathbf{l}_r}}{\mathcal{H}_{\mathbf{l}_r} + \lambda} - \frac{\mathcal{G}^2_{\mathbf{l}}}{\mathcal{H}_{\mathbf{l}} + \lambda}\right] - (|R_t| - 1)\gamma$$

gdzie:

- R_t – zbiór wyników podziału t,
- $\mathbf{l}_r$ – liść potomny odpowiadający wynikowi r.

Kryterium stopu dla $\mathbf{l}$: $\max_t \Delta q^*_{\mathbf{l}}(t) \leq 0$ – jakość drzewa po najlepszym podziale taka sama lub gorsza niż przed podziałem.

Konkretyzacje

Strata kwadratowa (minimalizacja MSE):

$$\begin{aligned}
\mathcal{L}(f(x), h_{1:m}(x)) &= (f(x) - h_{1:m}(x))^2 \\
\mathcal{L}(f(x), h_{1:i-1}(x)) &= (f(x) - h_{1:i-1}(x))^2 \\
\mathcal{G}_x &= 2(h_{1:i-1}(x) - f(x)) \\
\mathcal{H}_x &= 2
\end{aligned}$$

Strata logarytmiczna (maksymalizacja logarytmu wiarygodności):

$$\begin{aligned}
\mathcal{L}(f(x), h_{1:m}(x)) &= -f(x) \ln \pi_{1:m}(x) - (1 - f(x)) \ln(1 - \pi_{1:m}(x)) \\
\pi_{1:m}(x) &= \text{logit}^{-1}(h_{1:m}(x)) = \frac{e^{h_{1:m}(x)}}{e^{h_{1:m}(x)} + 1} \\
\mathcal{L}(f(x), h_{1:i-1}(x)) &= -f(x) \ln \pi_{1:i-1}(x) - (1 - f(x)) \ln(1 - \pi_{1:i-1}(x)) \\
\mathcal{G}_x &= \pi_{1:i-1}(x) - f(x) \\
\mathcal{H}_x &= \pi_{1:i-1}(x)(1 - \pi_{1:i-1}(x))
\end{aligned}$$

Podsumowanie

Właściwości:

- zazwyczaj bardzo wysoka jakość predykcji,
- liczba modeli bazowych wymaga dostrojenia w celu uniknięcia nadmiernego dopasowania (zwykle kilkanaście–kilkadziesiąt).

Pokrewne algorytmy: *LightGBM*, *CatBoost* – techniki przyspieszające, obsługa atrybutów dyskretnych.