Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н., н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

Ансамбли моделей

Михаил Криницкий

к.т.н., с.н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

Ансамбли моделей

деревья решений **сильно склонны к переобучению.** **НИКОГДА НЕ ПРИМЕНЯЙТЕ** деревья решений как таковые!
Способ борьбы с этой особенностью – использование ансамблей алгоритмов.

“Ансамбль искусственных нейронов” by Alexander Gavrikov

Ансамбли моделей

деревья решений **сильно склонны к переобучению. НИКОГДА НЕ ПРИМЕНЯЙТЕ** деревья решений как таковые!
Способ борьбы с этой особенностью – ансамблирование моделей.

Виды ансамблей:

- **Weighted averaging** (взвешенное осреднение): обучить K различных методов («базовых алгоритмов») на одних и тех же данных; результат взвешенно осреднять (в случае регрессии) или получать взвешенным голосованием (в случае классификации):

$$\widehat{y_i} = \frac{1}{\sum_i w_i} \sum_{k=1}^{K} w_k y_i^{(k)}$$

$$\hat{c}_i = \underset{c \in \mathbb{Y}}{\operatorname{argmax}} \sum_{k=1}^{K} w_k * [c_i^{(k)} == c]$$

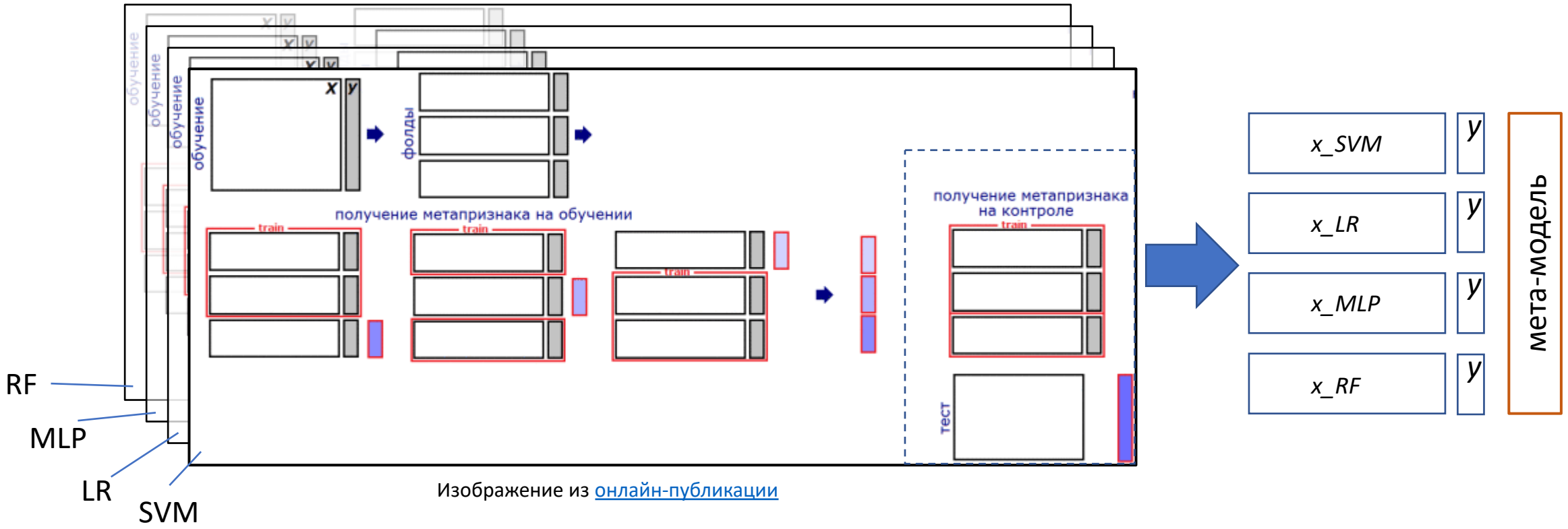
если все веса w_k равны 1, получим простое голосование/осреднение; в случае, когда веса зависят от x (а значит обучаемые) – получим т.н. «смесь экспертов» (blending)

- **Stacking** (стекинг): обучить К различных базовых алгоритмов на одних и тех же данных; вывод каждого из алгоритмов (значения параметров целевого распределения μ_i или p_i) использовать как новые признаки для новой мета-модели (обычно довольно простого, напр., любой из вариантов GLM/GAM: линейная регрессия в случае регрессии или логистическая регрессия в случае классификации);
- Обучать один и тот же базовый алгоритм на K полностью различных тренировочных выборках; результаты взвешенно осреднять (см. выше) или использовать эти K моделей в подходе стекинга. При этом рассчитывать на то, что каждой из этих выборок достаточно для обучения модели; все они порождены из одного и того же распределения. Однако набирать две или больше достаточно объемные тренировочные выборки – дорого и долго;
- **Random Subspace Method** (метод случайных подпространств): обучать К базовых алгоритмов (различных или одинаковых)
- **Bagging** (**B**ootstrap **Agg**regat**ing**, аггрегирование в подходе бутстрэп) – см. далее;
- **Boosting** («бустинг») – см. далее.

Ансамбли моделей: stacking

Идея:

1. обучить несколько базовых алгоритмов, каждый из которых где-то хорошо работает, а где-то систематически ошибается, получать этими моделями т.н. «метапризнаки»;
2. агрегировать результаты еще одной (тоже обучаемой) моделью – «метамоделью».



Изображение из [онлайн-публикации]

Ансамбли моделей: bagging

Bagging (**B**ootstrap **Agg**regat**ing**, аггрегирование в подходе бутстрэп)

Идея:

1. обучить множество базовых алгоритмов, склонных к переобучению, на подвыборках, гарантированно порожденных одним и тем же распределением (identically distributed, “i.d.”);
2. агрегировать результаты в подходе простого голосования/осреднения.

Сэмплирование из тренировочной выборки в подходе Bootstrap гарантирует* идентичность порождающего распределения**. В случае ограниченного количества выборок Bootstrap предоставляет лучшее из доступных приближений***.

Размер каждой выборки bootstrap:

- в случае сильно ограниченного размера тренировочной выборки – берут размером с тренировочную;
- в случае большого тренировочного набора данных размер bootstrap-выборки – гиперпараметр, подбирается по качеству на валидационной выборке.

Почему вообще ансамблирование одинаковых переобучающихся алгоритмов может работать, если сам алгоритм «плохой»?

* в пределе бесконечного количества выборок

** в смысле статистик, оцениваемых эмпирически

*** Efron B. Bootstrap Methods: Another Look at the Jackknife Springer Series in Statistics / под ред. S. Kotz, N.L. Johnson, New York, NY: Springer, 1992. 569–593 с.

Ансамбли моделей: bagging

Bagging (**B**ootstrap **Agg**regat**ing**, аггрегирование в подходе бутстрэп)

Почему вообще ансамблирование K одинаковых переобучающихся алгоритмов может работать, если сам алгоритм «плохой»?

Оценка целевой переменной (точнее, какого-то параметра распределения $P(y|x)$, например, $\mu(x)$) – тоже случайная величина (обозначим T), с определенным распределением $P(T)$.

Обычно (в предположении, что $T_1, T_2, \dots T_K$ – оценки K разными алгоритмами, i.i.d.* случайные величины):

$$Var(T_1) = Var(T_2) = \cdots = Var(T_n) = \sigma^2$$

тогда

$$Var(\bar{T}) = Var\left(\frac{1}{K}\sum_{k=1}^{K} T_k\right) = \frac{\sigma^2}{K}$$

Представим, что эти случайные величины – не независимы (в случае обучения K одинаковых алгоритмов на bootstrap-выборках уже нельзя говорить о независимости результатов). Например (простейший вариант), попарные корреляции между ними одинаковы и составляют ρ:

$$\rho = \frac{Cov(T_j, T_i)}{\sigma_{T_i}\sigma_{T_j}}$$

$$Cov(T_i, T_i) = Var(T_i) = \sigma^2$$

Тогда:

$$\boxed{Var(\bar{T})} = Var\left(\frac{1}{K}\sum_{k} T_k\right) = \frac{1}{K^2}\sum_{i,j} Cov(T_i, T_j) = K\frac{\sigma^2}{K^2} + \frac{K(K-1)}{K^2}\rho\sigma^2 = \boxed{\rho\sigma^2 + \frac{1-\rho}{K}\sigma^2}$$

*i.i.d. – independent, idetically distributed

Ансамбли моделей: bagging

Bagging (**B**ootstrap **Agg**regat**ing**, аггрегирование в подходе бутстрэп)

$$Var(\bar{T}) = \rho\sigma^2 + \frac{1-\rho}{K}\sigma^2$$

где T – случайная переменная оценки параметра условного распределения $P(y|x)$ целевой переменной (μ для регрессии, p для классификации)
$Var(\bar{T})$ - дисперсия средней оценки этого параметра при ансамблировании K одинаковых алгоритмов при смягчении предположения о независимости их ответов (например, при обучении на пересекающихся bootstrap-выборках)

Выводы: для снижения дисперсии (неопределенности) ответов ансамбля

- следует повышать K – количество членов ансамбля
- следует снижать ρ, характеризующую степень их скоррелированности – делать результаты базовых алгоритмов как можно менее похожими

Bagging эксплуатирует подход обучения большого количества ($K \gg 1$) моделей, склонных к переобучению (σ^2 - существенна, но ρ сильно меньше единицы, алгоритмы раскоррелированы за счет склонности к переобучению и за счет обучения на различающихся подвыборках).

Способ применения в случае решающих деревьев: обучить очень много довольно решающих деревьев до конца (не ограничивая их глубину, без регуляризаций); обучать на bootstrap-выборках, агрегировать результаты по принципу простого голосования (в случае классификции) или простого осреднения (в случае регрессии).