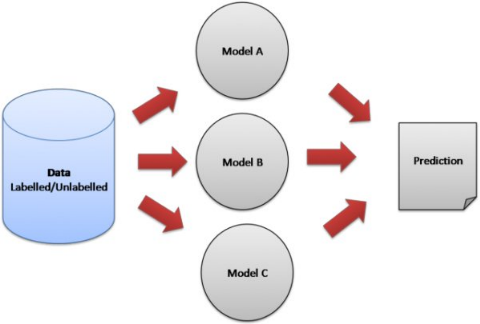
Композиции алгоритмов

Виктор Китов

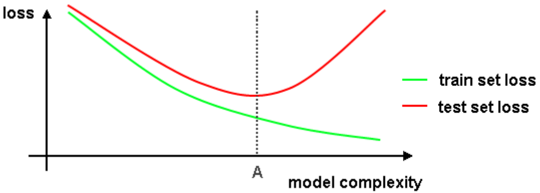
v.v.kitov@yandex.ru



Содержание

1. Разложение на смещение и разброс
2. Композиции алгоритмов
3. Фиксированная агрегирующая функция (против переобучения)
4. Стэкинг
5. Композиции на разных обучающих подвыборках (против переобучения)

Средние потери в зависимости от сложности модели



Комментарии:

- ожидаемые потери на тестовой выборке выше потерь на обучающей.
- слева от A: модель слишком простая, недообучение.
- справа от A: модель слишком сложная, переобучение

Разложение на смещение и разброс

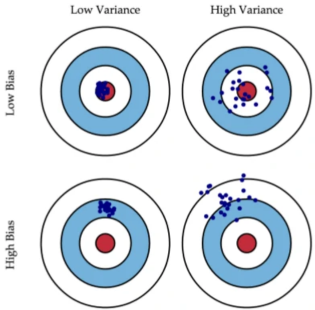
- Распределение реальных данных $y = f(x) + \varepsilon$
  - шум не зависит от x и старых наблюдений
  - хотим оценить $f(x)$
- Зависимость оценивается по $(X, Y) = \{(x_n, y_n),\ n = 1, 2 \ldots N\}$.
- Восстановленная зависимость $\widehat{f}(x)$.
- x - фикс. объект для прогноза.
- Шум ε не зависит от X, Y, $\mathbb{E}\varepsilon = 0$

Разложение на смещение и разброс (bias-variance decomposition)

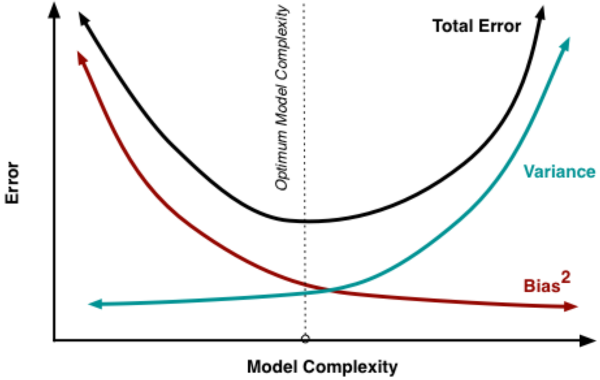
$$\mathbb{E}_{X,Y,\varepsilon}\{[\widehat{f}(x) - y(x)]^2\} = \left(\mathbb{E}_{X,Y}\{\widehat{f}(x)\} - f(x)\right)^2 + \mathbb{E}_{X,Y}\left\{[\widehat{f}(x) - \mathbb{E}_{X,Y}\widehat{f}(x)]^2\right\} + \mathbb{E}\varepsilon^2$$

Интуиция разложения

$$\mathbb{E}_{X,Y,\varepsilon}\{[\widehat{f}(x)-y(x)]^2\} = \left(\mathbb{E}_{X,Y}\{\widehat{f}(x)\}-f(x)\right)^2 + \mathbb{E}_{X,Y}\left\{[\widehat{f}(x)-\mathbb{E}_{X,Y}\widehat{f}(x)]^2\right\} + \mathbb{E}\varepsilon^2$$



Средние потери в зависимости от сложности модели



Доказательство разложения

Обозначим для краткости $f=f(x)$, $\widehat{f}=\widehat{f}(x)$, $\mathbb{E}=\mathbb{E}_{X,Y,\varepsilon}$.

$$\begin{aligned}\mathbb{E}\left(\widehat{f}-f\right)^2 &= \mathbb{E}\left(\widehat{f}-\mathbb{E}\widehat{f}+\mathbb{E}\widehat{f}-f\right)^2=\mathbb{E}\left(\widehat{f}-\mathbb{E}\widehat{f}\right)^2+\left(\mathbb{E}\widehat{f}-f\right)^2\\ &+2\mathbb{E}\left[(\widehat{f}-\mathbb{E}\widehat{f})(\mathbb{E}\widehat{f}-f)\right]\\ &=\mathbb{E}\left(\widehat{f}-\mathbb{E}\widehat{f}\right)^2+\left(\mathbb{E}\widehat{f}-f\right)^2\end{aligned}$$

т.к. $(\mathbb{E}\widehat{f}-f)$ - константа относительно X, Y,
$\mathbb{E}\left[(\widehat{f}-\mathbb{E}\widehat{f})(\mathbb{E}\widehat{f}-f)\right]=(\mathbb{E}\widehat{f}-f)\mathbb{E}(\widehat{f}-\mathbb{E}\widehat{f})=0.$

$$\begin{aligned}\mathbb{E}\left(\widehat{f}-y\right)^2 &= \mathbb{E}\left(\widehat{f}-f-\varepsilon\right)^2=\mathbb{E}\left(\widehat{f}-f\right)^2+\mathbb{E}\varepsilon^2-2\mathbb{E}\left[(\widehat{f}-f)\varepsilon\right]\\ &= \mathbb{E}\left(\widehat{f}-\mathbb{E}\widehat{f}\right)^2+\left(\mathbb{E}\widehat{f}-f\right)^2+\mathbb{E}\varepsilon^2\end{aligned}$$

$\mathbb{E}\left[(\widehat{f}-f)\varepsilon\right]=\mathbb{E}\left[(\widehat{f}-f)\right]\mathbb{E}\varepsilon=0$, поскольку ε не зависит от X, Y.

Содержание

1. Разложение на смещение и разброс
2. Композиции алгоритмов
   - Примеры использования композиций
3. Фиксированная агрегирующая функция (против переобучения)
4. Стэкинг
5. Композиции на разных обучающих подвыборках (против переобучения)

Композиции алгоритмов

- Композиция алгоритмов (ансамбль моделей, ensemble learning):
$$\widehat{y}(x) = G(f_1(x), ...f_M(x))$$
    - $f_1(x), ...f_M(x)$ - базовые модели=признаки для $G(\cdot)$
    - $G(\cdot)$ - агрегирующая модель, мета-модель
- Используется в
    - обучении с учителем (регрессия, классификация)
    - без учителя (кластеризация)

Мотивация композиций

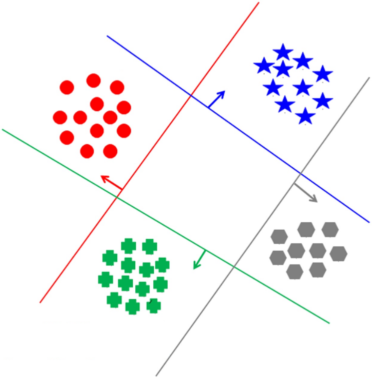
Мотивация:

- борьба с переобучением $f_1(x), ...f_M(x)$: простая $G(\cdot)$
- борьба с недообучением $f_1(x), ...f_M(x)$: сложная $G(\cdot)$
- каждая $f_1(x), ...f_M(x)$ отвечает за свою область признакового пространства (mixture of experts)
    - $G(\cdot)$ назначает одного из экспертов
- построение $\widehat{y}(x)$ декомпозируется на решение подзадач $f_1(x), ...f_M(x)$
- ускорение обучения
    - например, усреднение ядерных SVM на подвыборках

2. Композиции алгоритмов
   - Примеры использования композиций

Многоклассовая классификация

Многоклассовая классификация бинарными классификаторами (один против всех, один против одного, коды, исправляющие ошибки):



Последовательное решение, признаки разной природы

Последовательное решение

- Разделим классы: 1,2,"3+4"
  - если "3+4", применим модель, разделяющую 3 от 4.
- Прогнозирование стоимости квартир:
  - определяем тип покупки: для жилья/для инвестиций
  - для жилья: комфорт, индивидуальные вкусы и т.д.
  - для инвестиций: обменные курсы, процент по вкладам, рост рынка акций и т.д.
- Определение людей по фото:
  - определяем ракурс: фас/профиль
  - одна модель определяет людей по фото в фас
  - другая определяет людей по фото в профиль

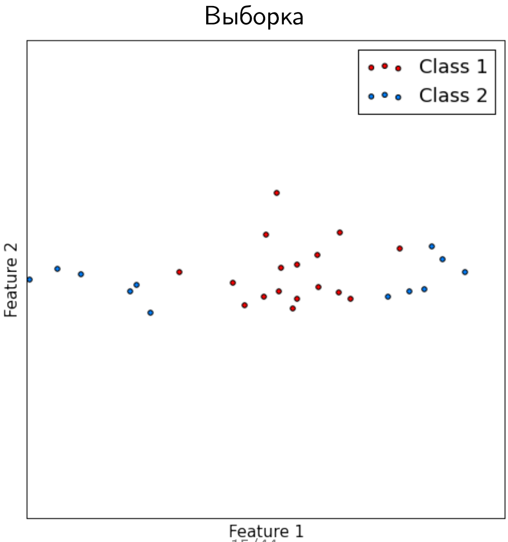
Признаки разной природы

- Идентификация человека по разнородной информации: по голосу, по лицу, по поведению, и т.д.

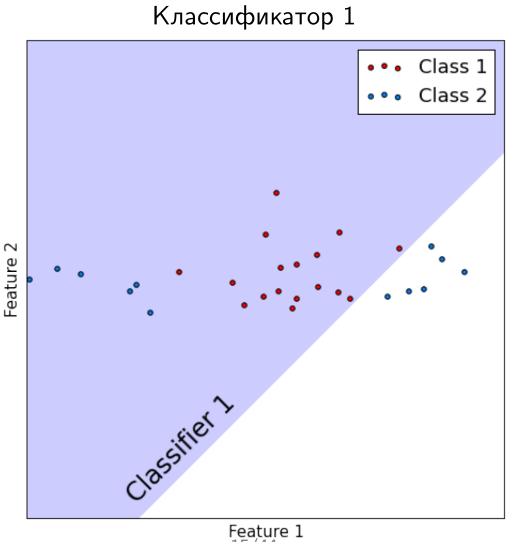
Борьба с переобучением

- Предположим $f_1(x), ...f_M(x)$ - слишком простые модели.
- Можем повысить сложность, применяя сложную мета-модель:
$$\widehat{y}(x) = G(f_1(x), ...f_M(x))$$

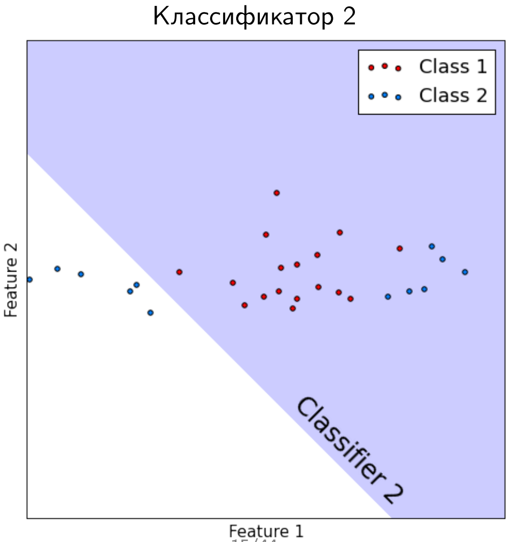
Пример



Пример

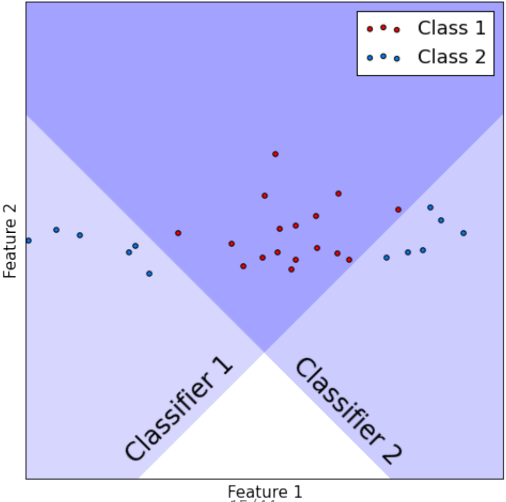

Классификатор 1
Class 1
Class 2
Classifier 1
Feature 1
Feature 2

Пример


Классификатор 2
Class 1
Class 2
Classifier 2
Feature 2
Feature 1

Пример

(Классификатор 1) AND (классификатор 2)



Борьба с переобучением

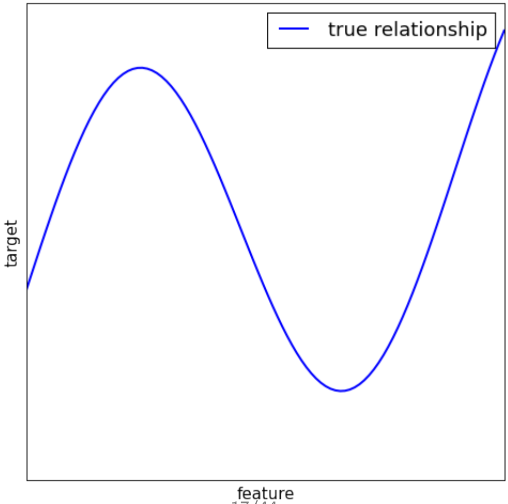
- $f_1(x), ...f_M(x)$ слишком сложные (переобученные модели)
    - решающие деревья большой глубины на разных подвыборках
    - глубокие нейросети
        - обученные из разных начальных приближений
        - разной архитектуры
- Регрессия: сделаем устойчивый прогноз за счет усреднения

$$\widehat{y}(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x)$$

- Классификация: прогноз=самый частый класс из $\{f_1(x), ...f_M(x)\}$.

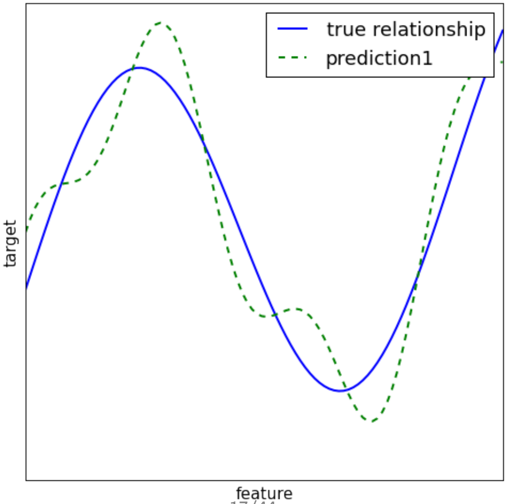
Регрессия: переобучение

Целевая зависимость



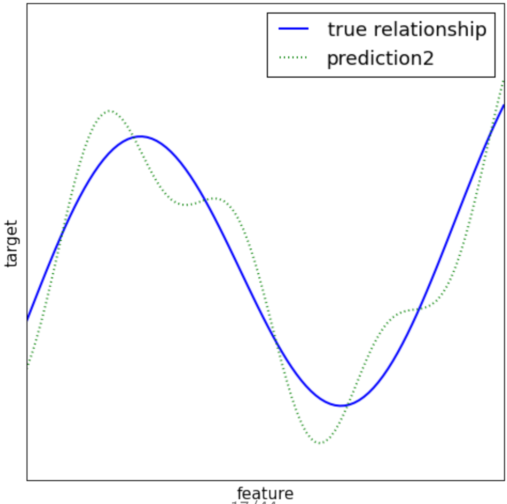
Регрессия: переобучение

Модель 1.



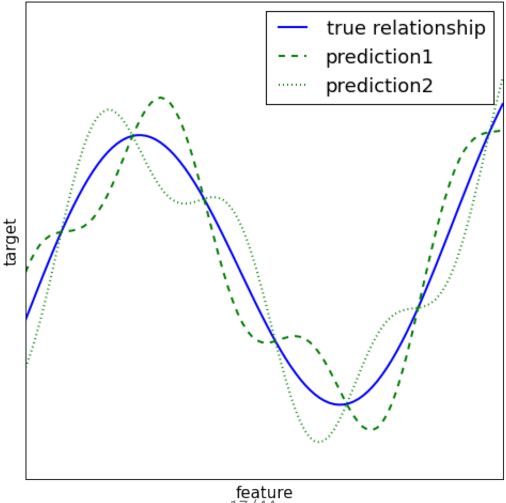
Регрессия: переобучение

Модель 2.



Регрессия: переобучение

Среднее модели 1 и 2 дает более точный прогноз.



Голосование большинства (против переобучения)

- Рассмотрим M бинарных классификаторов $f_1(x), ...f_M(x)$.
- Пусть $p(f_m(x) = y) = p < 0.5\,\forall m$
- Пусть модели ошибаются независимо друг от друга.
- Пусть $G(x)$ - выбор самого частого класса.
- Тогда $p(G(x) \neq y) \to 0$ при $M \to \infty$[1]

[1] Докажите это утверждение.

Взвешенное усреднение (против переобучения)

- **Разложение неоднозначности** (ambiguity decomposition): пусть (x, y) прогнозируется с помощью регрессии $G(x) = \sum_{m=1}^{M} w_m f_m(x)$, $w_m \geq 0$, $\sum_m w_m = 1$. Тогда

$$\underbrace{(G(x) - y)^2}_{\text{ensemble error}} = \underbrace{\sum_m w_m (f_m(x) - y)^2}_{\text{base learner error}} - \underbrace{\sum_m w_m (f_m(x) - G(x))^2}_{\text{ambiguity}}$$

- Композиция дает точные прогнозы когда:
    - $f_m(x)$ достаточно точны
    - индивидуальные прогнозы $\{f_m(x)\}_m$ сильно различаются
        - поэтому полезно усреднять по разным моделям

Доказательство разложения неоднозначности

Доказательство:

$$
\begin{aligned}
\sum_m w_m \left(f_m(x) - G(x)\right)^2 &= \sum_m w_m \left(f_m(x) - y + y - G(x)\right)^2 \\
&= \sum_m w_m \left(f_m(x) - y\right)^2 + \sum_m w_m \left(y - G(x)\right)^2 + 2\sum_m w_m \left(f_m(x) - y\right)\left(y - G(x)\right) \\
&= \sum_m w_m \left(f_m(x) - y\right)^2 + \left(G(x) - y\right)^2 + 2\left(y - G(x)\right)\sum_m w_m \left(f_m(x) - y\right) \\
&= \sum_m w_m \left(f_m(x) - y\right)^2 + \left(G(x) - y\right)^2 + 2\left(y - G(x)\right)\left(G(x) - y\right) \\
&= \sum_m w_m \left(f_m(x) - y\right)^2 + \left(G(x) - y\right)^2 - 2\left(G(x) - y\right)^2 \\
&= \sum_m w_m \left(f_m(x) - y\right)^2 - \left(G(x) - y\right)^2
\end{aligned}
$$

Выпуклые потери

Выпуклые потери поощряют использование взвешенных прогнозов вместо индивидуальных.

- Рассмотрим задачу регрессии с выпуклой ф-цией потерь $\mathcal{L}(\widehat{y} - y)$.
- Будем учитывать прогнозы моделей $f_1(x), ...f_M(x)$ с весами $w_1, ...w_M$.
- Для фиксированного x рассмотрим 2 стратегии прогнозирования:

1. сэмплировать $m \sim Categorical(w_1, ...w_M)$, $\widehat{y}(x) = f_m(x)$.
2. усреднять $\widehat{y}(x) = \sum_{m=1}^{M} w_m f_m(x)$

Какая стратегия в среднем по m будет давать меньшие ожидаемые потери?

Содержание

1. Разложение на смещение и разброс
2. Композиции алгоритмов
3. Фиксированная агрегирующая функция (против переобучения)
4. Стэкинг
5. Композиции на разных обучающих подвыборках (против переобучения)

Регрессия

$$\widehat{y}(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x)$$

$$\widehat{y}(x) = \frac{1}{\sum_{m=1}^{M} w_m}\sum_{m=1}^{M} w_m f_m(x)$$

- Взвешенное усреднение лучше, если модели сильно отличаются по точности.
- Веса $w_1 \geq 0, ...w_M \geq 0$ нужно настраивать на отдельной выборке (не той, на которой обучали $f_1(x), ...f_M(x)$)
- Альтернатива: медиана/взвешенная медиана

Классификаторы выдают **вероятности**

- Пусть $p_y^m(x)$ - вероятность класса y по мнению классификатора m.
- Равномерная агрегация:

$$p_y(x) = \frac{1}{M}\sum_{m=1}^{M} p_y^m(x)$$

- Взвешенная агрегация:

$$p_y(x) = \frac{1}{\sum_{m=1}^{M} w_m}\sum_{m=1}^{M} w_m p_y^m(x)$$

- Взвешенное усреднение лучше, если модели сильно отличаются по точности.
- Веса $w_1 \geq 0, ...w_M \geq 0$ нужно настраивать на отдельной выборке (не той, на которой обучали $f_1(x), ...f_M(x)$)

Классификаторы выдают **метки классов**

- Голосование по большинству (majority vote)
    - возможен взвешенный учет классификаторов
- Бинарная классификация: $\widehat{y} = +1$ <=>
    - $\geq k$ классификаторов выдают +1 (k-out-of-N)
        - возможен взвешенный учет классификаторов
    - все классификаторы выдают +1 (AND, N-out-of-N)
    - хотя бы один выдает +1 (OR, 1-out-of-N)

Классификаторы выдают **рейтинги**

- Пусть $g_y^m(x)$ - рейтинг класса y в модели m.
- Проблема: рейтинги несравнимы для разных моделей.
- Решение (Brier scores):
    1. Стандартизованный рейтинг - #классов ниже по рейтингу:

$$s_y^m(x) = \sum_{y \neq i} \mathbb{I}[g_y^m(x) > g_i^m(x)]$$
$$s_y^m(x) \in \{1, 2, ... C - 1\}$$

    2. Предскажем класс с максимальным усредненным по моделям рейтингом:

$$\widehat{y}(x) = \arg\max_y \frac{1}{M} \sum_{m=1}^{M} s_y^m(x)$$

Содержание

1. Разложение на смещение и разброс
2. Композиции алгоритмов
3. Фиксированная агрегирующая функция (против переобучения)
4. Стэкинг
5. Композиции на разных обучающих подвыборках (против переобучения)

Алгоритм стэкинга

- Рассмотрим обучающую выборку , базовые модели $f_1(x), ...f_M(x)$ и $G(\cdot)$.

Алгоритм стэкинга

- Рассмотрим обучающую выборку , базовые модели $f_1(x), ...f_M(x)$ и $G(\cdot)$.
- Обучение $f_1(x), ...f_M(x)$ и $G(\cdot)$ на одинаковой выборке вызывает переобучение.

Алгоритм стэкинга:

1. Инициализируем обучающую выборки $G(\cdot)$: $T' = \{\}$
2. Разобьем обучающую выборку $T = \{(x_n, y_n)\}_{n=1}^N$ на K блоков: T_1, $T_2, ...T_K$.
3. для $k = 1, 2, ...K$:
   обучим $f_1(x), ...f_M(x)$ на $T \backslash T_k$
   для каждого $(x, y) \in T_k$:
      дополним T' объектом $([f_1(x), ...f_M(x)], y)$
4. Обучим $G(\cdot)$ на T'.
5. Перенастроим $f_1(x), ...f_M(x)$ на всей T.

Расширения стэкинга

Кроме прогнозов $f_1(x), ...f_M(x)$ агрегирующая функция может зависеть от:

- исходных признаков x
  - в разных частях признакового пространства-разная агрегация
- внутренних представлений f_m (дискриминантных функций, вероятностей).

Линейный стэкинг (блендинг)

- Линейный стэкинг (блендинг, blending) :

$$f(x) = \sum_{m=1}^{M} w_m f_m(x)$$

$$\sum_{n=1}^{N} \left(\sum_{m=1}^{M} w_m f_m(x_n) - y_n \right)^2 \to \min_{\mathbf{w}}$$

- $f_1(x), \dots f_M(x)$ зависимы (предсказывают один и тот же y) => нестабильная оценка.
- Более устойчивая оценка:

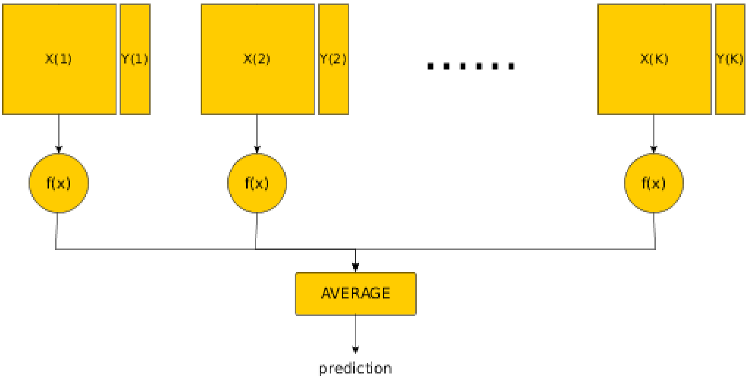
$$\begin{cases} \sum_{n=1}^{N} \left(\sum_{m=1}^{M} w_m f_m(x_n) - y_n \right)^2 + \lambda \sum_{m=1}^{M} \left(w_m - \frac{1}{M} \right)^2 \to \min_{\mathbf{w}} \\ w_1 \geq 0, \dots w_M \geq 0 \end{cases}$$

Содержание

1. Разложение на смещение и разброс
2. Композиции алгоритмов
3. Фиксированная агрегирующая функция (против переобучения)
4. Стэкинг
5. Композиции на разных обучающих подвыборках (против переобучения)

Усреднение по выборкам

Усреднение по выборкам



Усреднение по выборкам

Усреднение по выборкам: если модель переобучается на выборке (X,Y) можно усреднять множество моделей, обученных на разных реализациях обучающих выборок $(X_k,Y_k),\ k=1,2,...M$.

$$bias_{X,Y}\left[\frac{1}{M}\sum_{k=1}^{M}f(x,X_k,Y_k)\right]=y(x)-\mathbb{E}_{X,Y}\left[\frac{1}{M}\sum_{k=1}^{M}f(x,X_k,Y_k)\right]$$
$$=y(x)-\frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_{X,Y}f(x,X_k,Y_k)=y(x)-\frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_{X,Y}f(x,X,Y)$$
$$=y(x)-\mathbb{E}_{X,Y}f(x,X,Y)$$

т.е. смещение=смещению 1-го алгоритма.

Усреднение по выборкам: дисперсия

$$\mathrm{Var}_{X,Y}\left[\frac{1}{M}\sum_{k=1}^{M}f(x,X_k,Y_k)\right]=\mathbb{E}_{X,Y}\left[\frac{1}{M}\sum_{k=1}^{M}f(x,X_k,Y_k)-\mathbb{E}_{X,Y}\left[\frac{1}{M}\sum_{k=1}^{M}f(x,X_k,Y_k)\right]\right]^2$$

$$=\mathbb{E}_{X,Y}\left[\frac{1}{M}\sum_{k=1}^{M}\left(f(x,X_k,Y_k)-\mathbb{E}_{X,Y}f(x,X_k,Y_k)\right)\right]^2=$$

$$=\frac{1}{M^2}\mathbb{E}_{X,Y}\left[\sum_{k=1}^{M}\left(f(x,X_k,Y_k)-\mathbb{E}_{X,Y}f(x,X_k,Y_k)\right)\right]^2=$$

$$=\frac{1}{M^2}\sum_{k=1}^{M}\mathbb{E}_{X,Y}\left[f(x,X_k,Y_k)-\mathbb{E}_{X,Y}f(x,X_k,Y_k)\right]^2+$$

$$+\frac{1}{M^2}\sum_{k_1\neq k_2}\mathbb{E}_{X,Y}\left[\left(f(x,X_{k_1},Y_{k_1})-\mathbb{E}_{X,Y}f(x,X_{k_1},Y_{k_1})\right)\left(f(x,X_{k_2},Y_{k_2})-\mathbb{E}_{X,Y}f(x,X_{k_2},Y_{k_2})\right)\right]$$

$$=\frac{1}{M^2}\sum_{k=1}^{M}\mathrm{Var}_{X,Y}\left[f(x,X_k,Y_k)\right]+\frac{1}{M^2}\sum_{k_1\neq k_2}\mathrm{cov}\left[f(x,X_{k_1},Y_{k_1}),f(x,X_{k_2},Y_{k_2})\right]$$

Усреднение по выборкам: дисперсия

При нескоррелированных $f(x, X_k, Y_k)$:

$$\frac{1}{M^2}\sum_{k=1}^{K}\mathrm{Var}_{X,Y}\left[f(x, X_k, Y_k)\right]+$$
$$+\frac{1}{M^2}\sum_{k_1\neq k_2}\mathrm{cov}\left[f(x, X_{k_1}, Y_{k_1}), f(x, X_{k_2}, Y_{k_2})\right]=$$
$$=\frac{1}{M}\mathrm{Var}_{X,Y}\left[f(x, X_k, Y_k)\right]$$

Дисперсия в M раз меньше.

Усреднение по выборкам: дисперсия

При частичной скоррелированности, дисперсия тоже уменьшается (используем $\operatorname{cov}(x,y) \le \sqrt{\operatorname{Var}(x)}\sqrt{\operatorname{Var}(y)}$):

$$\frac{1}{M^2}\sum_{k=1}^{K}\operatorname{Var}_{X,Y}\left[f(x,X_k,Y_k)\right]+$$

$$+\frac{1}{M^2}\sum_{k_1\ne k_2}\operatorname{cov}\left[f(x,X_{k_1},Y_{k_1}),f(x,X_{k_2},Y_{k_2})\right]=$$

$$=\frac{1}{M^2}\sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\operatorname{cov}\left[f(x,X_{k_1},Y_{k_1}),f(x,X_{k_2},Y_{k_2})\right]\le$$

$$\le\frac{1}{M^2}\sum_{k_1=1}^{M}\sum_{k_2=1}^{M}\sqrt{\operatorname{Var}_{X,Y}\left[f(x,X_{k_1},Y_{k_1})\right]}\sqrt{\operatorname{Var}_{X,Y}\left[f(x,X_{k_2},Y_{k_2})\right]}\le$$

$$\le\operatorname{Var}_{X,Y}\left[f(x,X,Y)\right]$$

Бэггинг & метод случайных подпространств

На практике дана единственная (X, Y). Как генерировать $(X_1, Y_1), ...(X_M, Y_M)$?

[2] Какова вероятность того, что некоторый объект не попадет в подвыборку? Чему равен предел этой вероятности при $N \to \infty$?

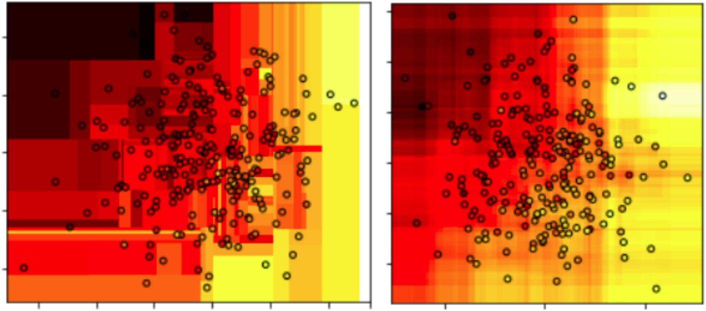
Бэггинг & метод случайных подпространств

На практике дана единственная (X, Y). Как генерировать $(X_1, Y_1), ...(X_M, Y_M)$?

- **Бэггинг (bagging)**:
  - случайный выбор N объектов (с возвращением)[2]
- **Метод случайных подпространств (random subspaces)**:
  - случайный выбор K признаков (без возвращения, $K < N$)
- Можно применять комбинацию методов.

[2] Какова вероятность того, что некоторый объект не попадет в подвыборку? Чему равен предел этой вероятности при $N \to \infty$?

Бэггинг деревьев



Регрессия: одно дерево и бэггинг над деревьями

Бэггинг деревьев

Настройка решающего правила в узле CART:

$$\widehat{f}, \widehat{h} = \arg\max_{f,h \in P(t)} \Delta\phi(t)$$

$P(t)$ для стандартных решающих деревьев:

$$
\begin{array}{l}
P = \{\} \\
\texttt{for each } f \texttt{ in } \{1, ..., D\} \\
\quad \texttt{for each } h \texttt{ in } \mathit{unique}\left\{x_n^f\right\}_{n: x_{n \in t}} \\
\qquad P := P \cup (f, h)
\end{array}
$$

Бэггинг над решающими деревьями успешно борется с их переобучением.

Случайный лес, особо случайные деревья

$P(t)$ для случайного леса (random forest, RF):

$P = \{\}, \quad K = \alpha D$
sample $d_1, ...d_K$ randomly from $\{1, ..., D\}$ # без возвращения
for each f in $d_1, ...d_K$
    for each h in $unique\left\{x_n^f\right\}_{n: x_n \in t}$
        $P := P \cup (f, h)$

$S(t)$ для особо случайных деревьев (extra random trees, ERT):

$S = \{\}, \quad K = \alpha D$
sample $d_1, ...d_K$ randomly from $\{1, ..., D\}$ # с возвращением
for each f in $d_1, ...d_K$
    sample h randomly from $unique\left\{x_n^f\right\}_{n: x_n \in t}$
    $P := P \cup (f, h)$

Вневыборочная оценка

Оценка по обучающей выборке - оптимистическая оценка сверху:

$$L=\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(\frac{1}{M}\sum_{m=1}^{M}f_m(x_n),y_n\right)$$

Вневыборочная оценка

Оценка по обучающей выборке - оптимистическая оценка сверху:

$$L = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(\frac{1}{M}\sum_{m=1}^{M} f_m(x_n), y_n\right)$$

Вневыборочная оценка (out-of-bag estimate) - если с бэггингом

$$L_{OOB} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(\frac{1}{|I_n|}\sum_{m\in I_n} f_m(x_n), y_n\right)$$

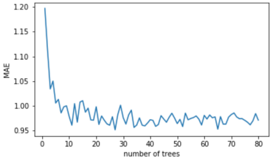
- $I_n \subset \{1, 2, ...M\}$- набор моделей, не использовавших (x_n, y_n) для обучения.
- Не требуется дополнительная валидационная выборка.
- Немного пессимистическая оценка потерь снизу.

Комментарии

- Бэггинг, случайный лес, особо случайные деревья:
    - легко распараллеливаются
    - базовые модели не учатся исправлять ошибки друг друга
- Деревья случайный лес, особо случайные деревья могут строиться на одинаковой обучающей выборке
    - bootstrap=False в sklearn
    - за счет случайности $P(t)$ модели все равно будут получаться разные

Число базовых моделей в ансамбле

- Пусть M=#базовых моделей.
- Типичная зависимость потерь от M:



- Нет переобучения с $\uparrow M$: просто избыточное усреднение по однотипным моделям.
- Настройка: подбор всех параметров с малым M, затем $\uparrow M$.

Заключение

- Разложение на смещение и разброс:
  - простые модели: высокое смещение
  - сложные модели: высокая дисперсия
- Разложение неопределенности:
  - выгодно усреднять разнородные модели
- Композиции:
  - простая агрегирующая модель: борьба с переобучением
  - сложная агрегирующая модель: борьба с недообучением
- Стэкинг: агрегирующая модель и базовые должны обучаться на разных выборках
- Борьба с переобучением: бэггинг, метод случайных подпространств, случайный лет, особо случайные деревья.