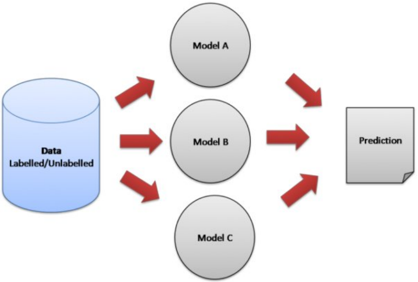
Композиции алгоритмов

Виктор Китов

victorkitov.github.io

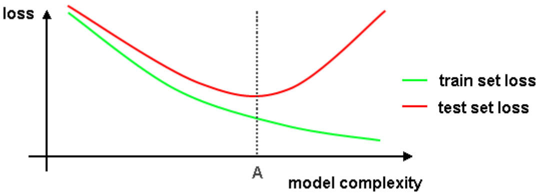


Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Содержание

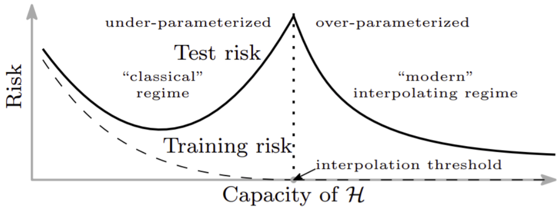
Средние потери в зависимости от сложности модели



Комментарии:

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

Дальнейшее усложнение модели



- Некоторые эмпирические наблюдения свидетельствуют, что для слишком сложных моделей (многослойные нейросети) качество выше ожиданий.
- Феномен double descent risk curve[1].

[1]Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off.

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 $(X, Y) = \{(x_n, y_n),\ n = 1, 2...N\}$.
- 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

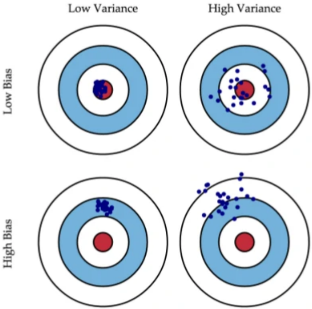
Разложение на смещение и разброс
(bias-variance decomposition):

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

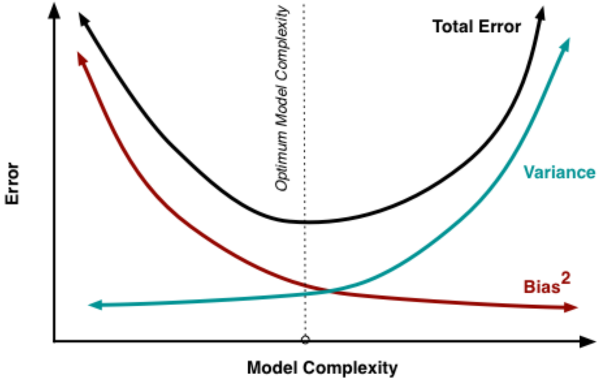
- Интуиция:
  $MSE =$ смещение$^2 +$ дисперсия $+$ неснижаемая ошибка
  - смещение (bias) - степень недообученности
  - дисперсия (variance) - степень переобученности
  - неснижаемая ошибка (irreducible error) - определяется шумом в данных

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

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

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 (meta-model)
- 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)

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

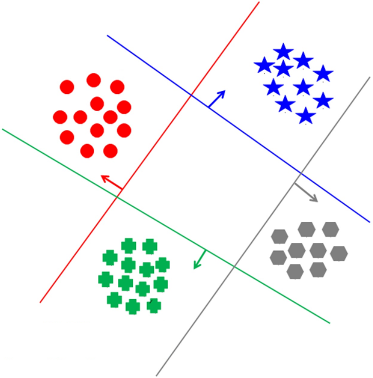
$$w_m=\begin{cases}0, & m\neq i(x)\\ 1 & m=i(x)\end{cases}\qquad \text{либо}\qquad w=\text{SoftMax}(\cdots)$$

- построение $\widehat{y}(x)$ декомпозируется на решение подзадач $f_1(x),...f_M(x)$
- ускорение обучения
  - например, усреднение ядерных SVM на подвыборках

Содержание

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 и недообучением

Ансамбли моделей используются для борьбы

- с переобучением базовых моделей
  - голосование по большинству, усреднение, бэггинг, метод случайных подпространств, случайный лес, особо случайные деревья.
- с недообучением базовых моделей
  - AND, OR, K-out-of-M, стэкинг, блендинг, бустинг.

Содержание

Борьба с переобучением

- Предположим $f_1(x), ...f_M(x)$ - слишком простые модели.
- Можем повысить сложность, применяя сложную мета-модель:

$$\widehat{y}(x) = G(f_1(x), ...f_M(x))$$

- Примеры:
    - $G\left(f_1(x), f_2(x)\right) = f_1(x)$ AND $f_2(x)$;
      $G\left(f_1(x), f_2(x)\right) = f_1(x)$ OR $f_2(x)$;

Борьба с переобучением

- Предположим $f_1(x), ...f_M(x)$ - слишком простые модели.
- Можем повысить сложность, применяя сложную мета-модель:

$$\widehat{y}(x) = G(f_1(x), ...f_M(x))$$

- Примеры:
  - $G\left(f_1(x), f_2(x)\right) = f_1(x)$ AND $f_2(x)$;
    $G\left(f_1(x), f_2(x)\right) = f_1(x)$ OR $f_2(x)$;
  - Стэкинг, блендинг: $G(\cdot)$ настраиваемая функция.
    - $f_1(x), ...f_M(x)$ - сложные, $G(f_1(x), ...f_M(x))$ - ещё сложнее!

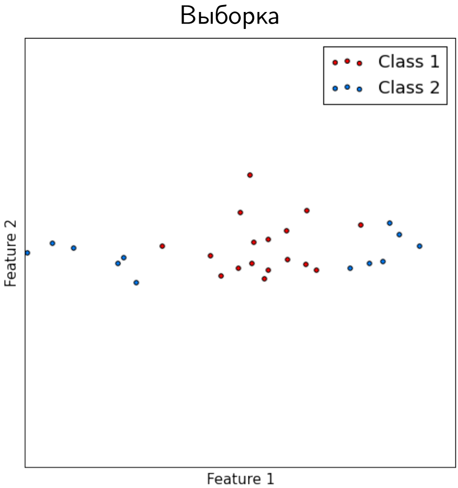
Борьба с переобучением

- Предположим $f_1(x), ...f_M(x)$ - слишком простые модели.
- Можем повысить сложность, применяя сложную мета-модель:

$$\widehat{y}(x) = G(f_1(x), ...f_M(x))$$

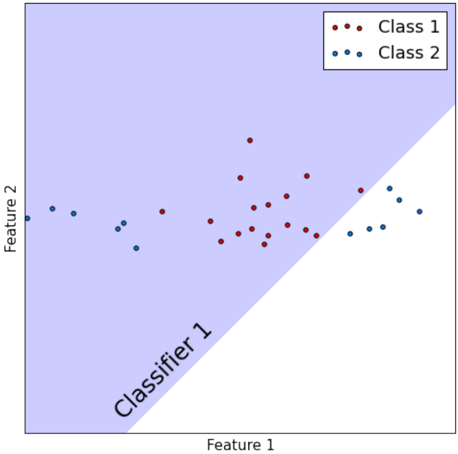
- Примеры:
    - $G\left(f_1(x), f_2(x)\right) = f_1(x)$ AND $f_2(x)$;
      $G\left(f_1(x), f_2(x)\right) = f_1(x)$ OR $f_2(x)$;
    - Стэкинг, блендинг: $G(\cdot)$ настраиваемая функция.
        - $f_1(x), ...f_M(x)$ - сложные, $G(f_1(x), ...f_M(x))$ - ещё сложнее!
    - Бустинг: $G_i(x) = \varepsilon f_1(x) + \varepsilon f_2(x) + ... + \varepsilon f_i(x)$, $\varepsilon \in (0, 1]$.
        - $f_1(x)$ учится предсказывать y
        - $f_2(x)$ исправляет ошибки $G_1(x)$
        - $f_3(x)$ исправляет ошибки $G_2(x)$
        - ...

Идея композиции AND



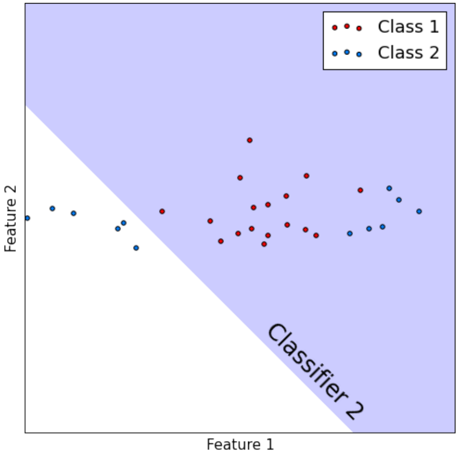
Идея композиции AND

Классификатор 1



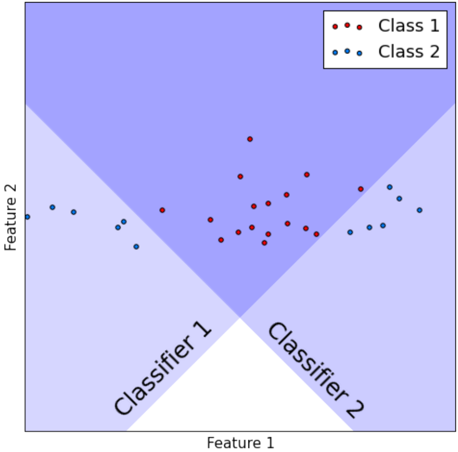
Идея композиции AND

Классификатор 2



Идея композиции AND

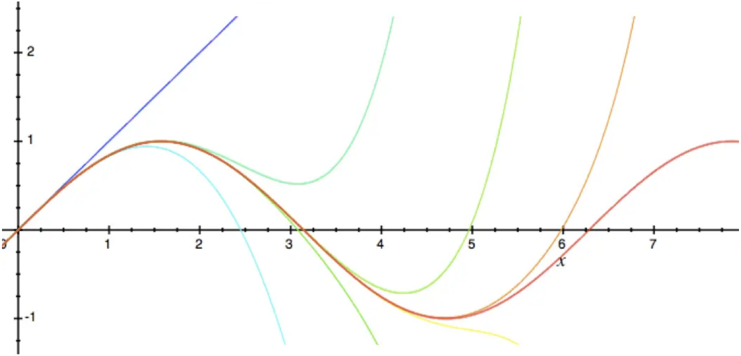
(Классификатор 1) AND (классификатор 2)



Идея стэкинга

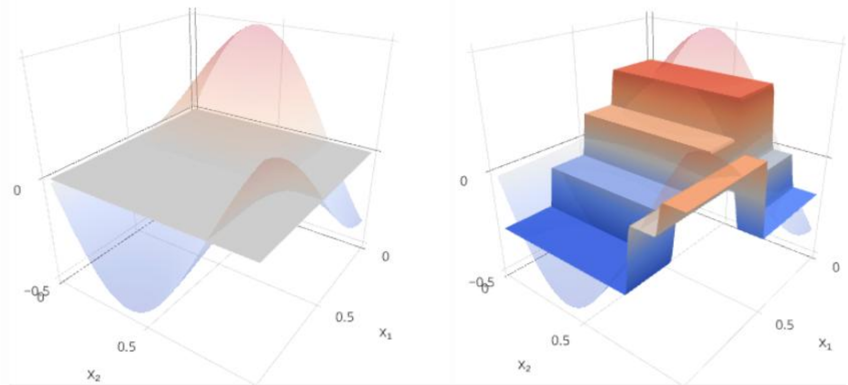
- $f_1(x) = x,\ f_2(x) = (x)^2,\ \dots\ f_M(x) = (x)^M.$
- $G\left(f_1(x), \dots f(x)_M\right) = w_0 + w_1 f_1(x) + \dots w_M f_M(x) = w_0 + w_1 x + w_2 (x)^2 + \dots w_M (x)^M$

Последовательное приближение $y = sin(x)$:



Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

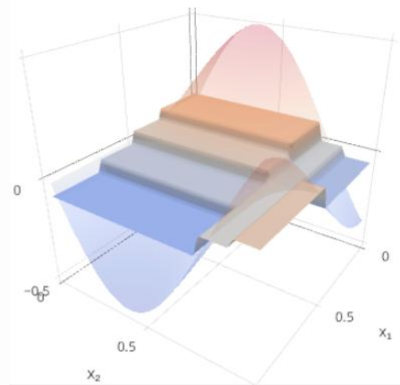
Слева:$y(x)$, $G_0(x)$. Справа: отклонение и $f_1(x)$

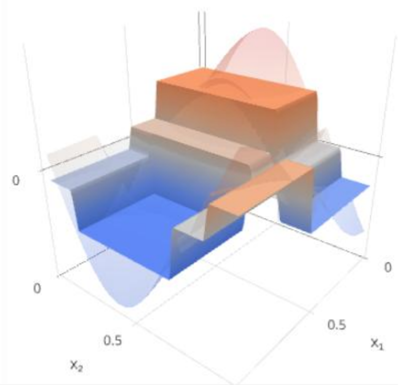


[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

Слева:$y(x)$, $G_1(x)$. Справа: отклонение,$f_2(x)$

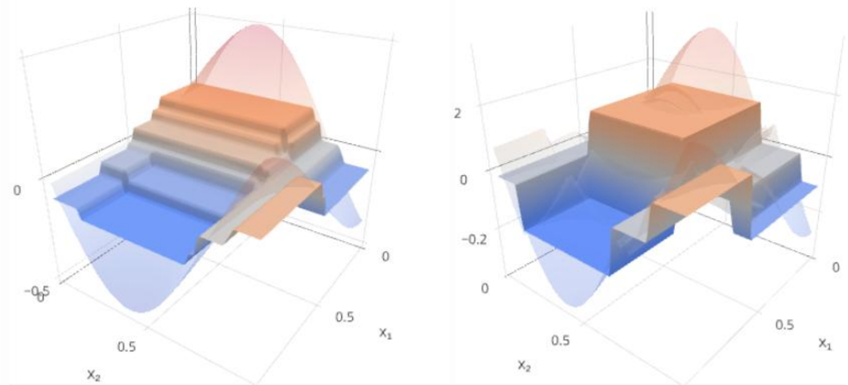




[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

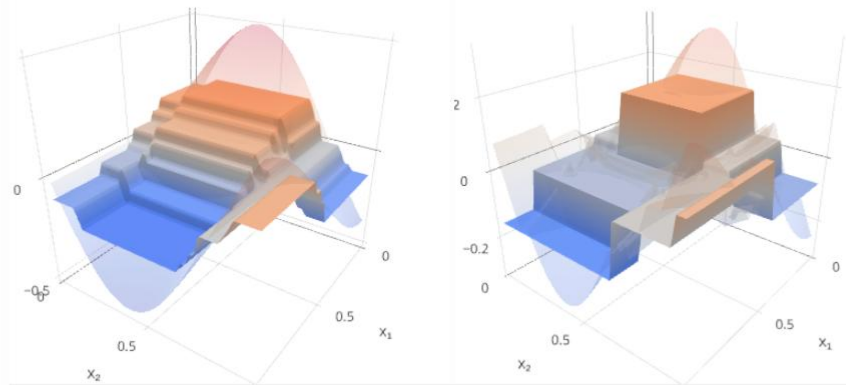
Слева:$y(x)$, $G_2(x)$. Справа: отклонение,$f_3(x)$



[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

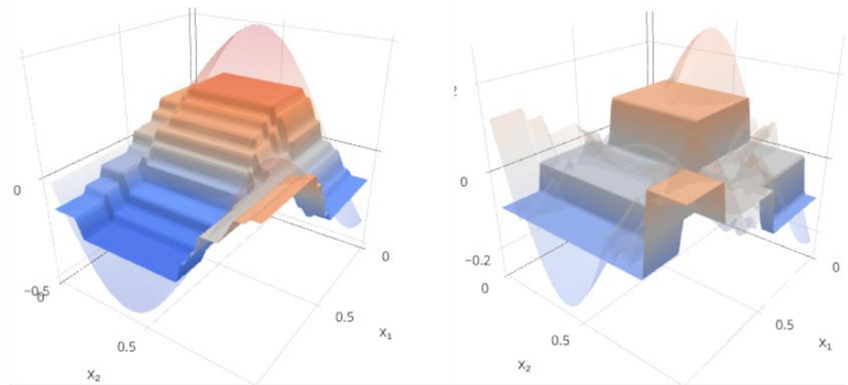
Слева: $y(x)$, $G_3(x)$. Справа: отклонение, $f_4(x)$



[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

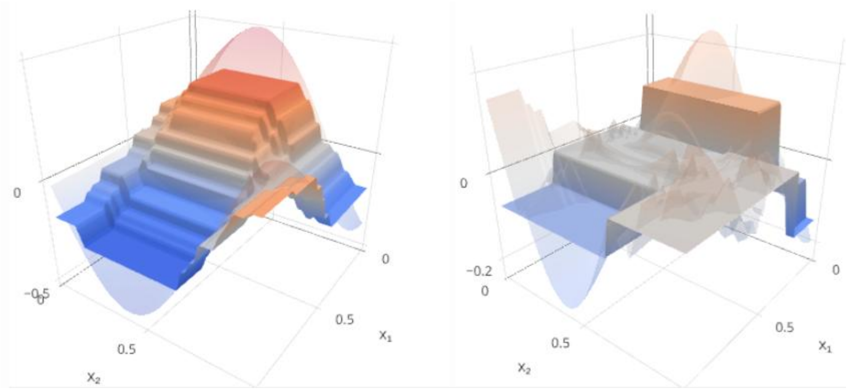
Слева: $y(x)$, $G_4(x)$. Справа: отклонение, $f_5(x)$



[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

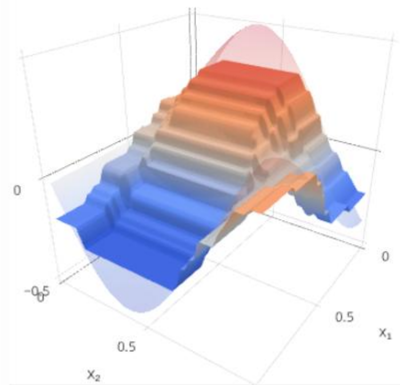
Слева:$y(x)$, $G_5(x)$. Справа: отклонение, $f_6(x)$

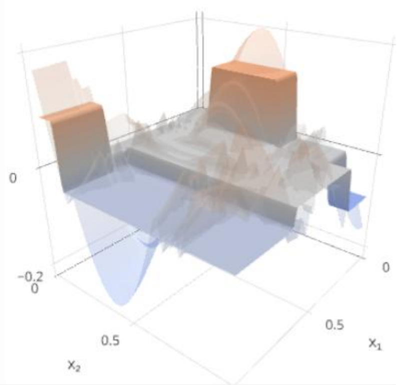


[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

Слева: $y(x)$, $G_6(x)$. Справа: отклонение, $f_7(x)$

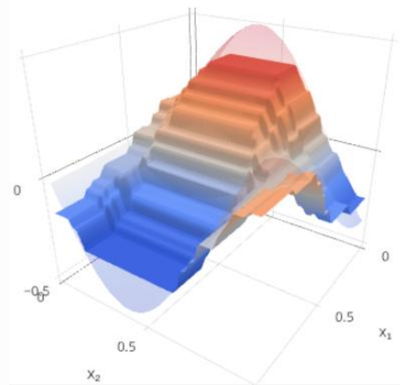


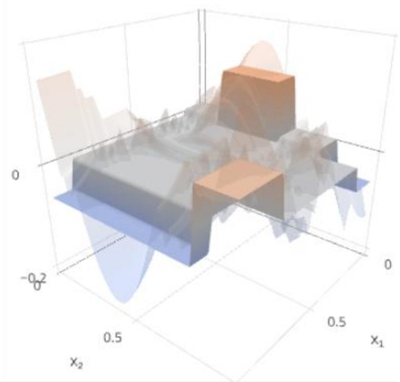


[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

Слева: $y(x)$, $G_7(x)$. Справа: отклонение, $f_8(x)$

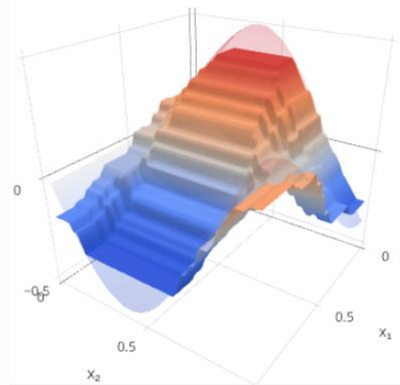


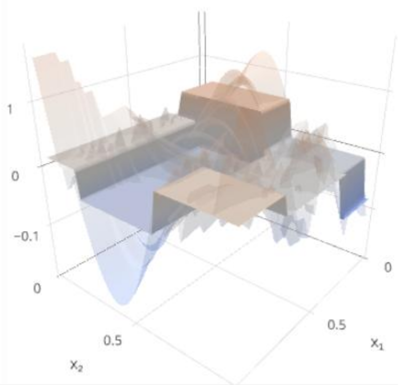


[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

Слева: $y(x)$, $G_8(x)$. Справа: отклонение, $f_9(x)$

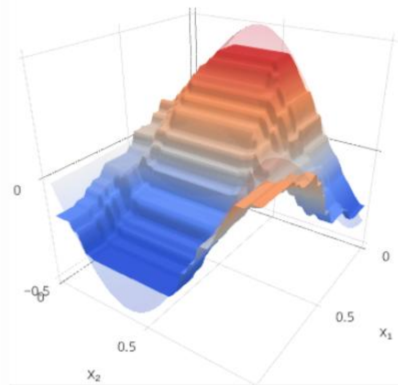


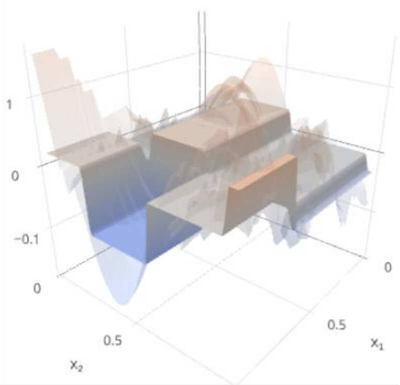


[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

Слева: $y(x)$, $G_9(x)$. Справа: отклонение, $f_{10}(x)$

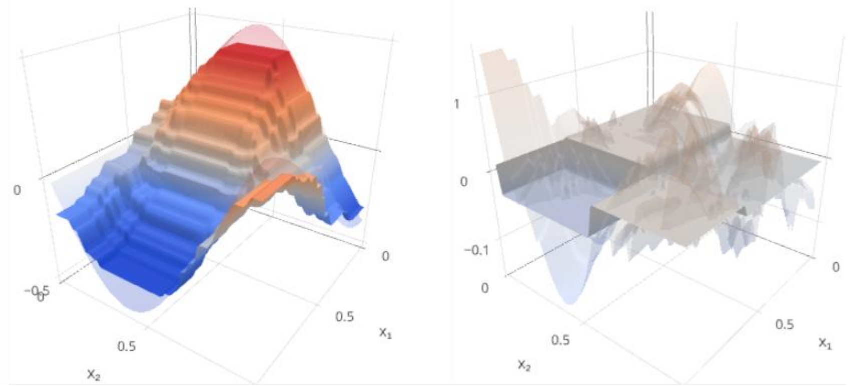




[2] Интерактивная иллюстрация Алексея Рогожникова.

Идея бустинга (на деревьях глубины 3, $\varepsilon = 0.3$)[2]

Слева: $y(x)$, $G_{10}(x)$. Справа: отклонение, $f_{11}(x)$



[2] Интерактивная иллюстрация Алексея Рогожникова.

Содержание

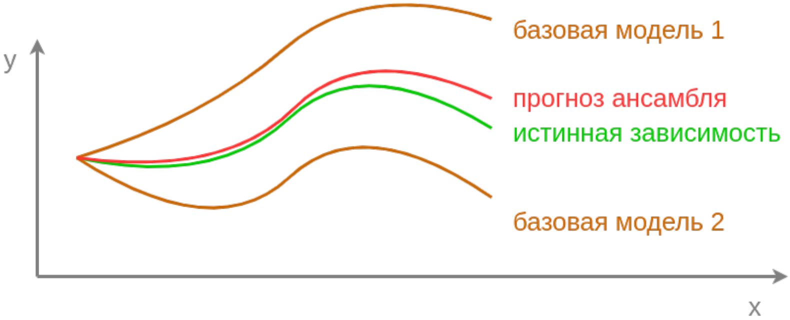
Борьба с переобучением

- $f_1(x), ...f_M(x)$ 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

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

- Классификация: прогноз самым частым классом среди $\{f_1(x), ...f_M(x)\}$.
  - голосование по по большинству, majority voting

Регрессия: переобучение



Усреднение ошибок

Рассмотрим задачу регрессии с усредняющим ансамблем:

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

Пусть

$$\varepsilon_i = f_i(x) - y(x), \quad \mathbb{E}\varepsilon_i = 0, \quad \mathbb{D}\varepsilon_i = \mathbb{E}\left\{(\varepsilon_i - \mathbb{E}\varepsilon_i)^2\right\} = \mathbb{E}\varepsilon_i^2 = \sigma^2$$

$$\text{corr}\{\varepsilon_i, \varepsilon_j\} = \rho = \text{cov}\{\varepsilon_i, \varepsilon_j\}/\sqrt{\mathbb{D}\varepsilon_i \cdot \mathbb{D}\varepsilon_j}$$

$$\text{cov}\{\varepsilon_i, \varepsilon_j\} = \mathbb{E}\left\{(\varepsilon_i - \mathbb{E}\varepsilon_i)(\varepsilon_j - \mathbb{E}\varepsilon_j)\right\} = \mathbb{E}\left\{\varepsilon_i\varepsilon_j\right\} = \rho\sigma^2$$

$$\mathbb{E}\left\{(f_i(x) - y(x))^2\right\} = \mathbb{E}\varepsilon_i^2 = \sigma^2, \quad \mathbb{E}\left\{(G(x) - y(x))^2\right\} - ?$$

Усреднение ошибок

Ожидаемый квадрат ошибки усредняющего ансамбля:

Усреднение ошибок

Ожидаемый квадрат ошибки усредняющего ансамбля:

$$\mathbb{E}\left\{(G(x)-y(x))^2\right\} = \mathbb{E}\left\{\left(\frac{\sum_{m=1}^{M}(f_m(x)-y(x))}{M}\right)^2\right\}$$

$$= \frac{1}{M^2}\mathbb{E}\left\{\left(\sum_{m=1}^{M}\varepsilon_m\right)^2\right\} = \frac{1}{M^2}\mathbb{E}\left\{\sum_{m=1}^{M}\varepsilon_m^2 + \sum_{i\neq j}\varepsilon_i\varepsilon_j\right\}$$

$$= \frac{M}{M^2}\sigma^2 + \frac{M^2-M}{M^2}\rho\sigma^2 = \frac{\sigma^2}{M} + \left(1-\frac{1}{M}\right)\rho\sigma^2$$

- При $\rho = 1$: σ^2; при $\rho = 0$: σ^2/M
  - может быть еще меньше при $\rho < 0$.

Усреднение ошибок

Ожидаемый квадрат ошибки усредняющего ансамбля:

$$\mathbb{E}\left\{(G(x)-y(x))^2\right\}=\mathbb{E}\left\{\left(\frac{\sum_{m=1}^{M}\left(f_m(x)-y(x)\right)}{M}\right)^2\right\}$$

$$=\frac{1}{M^2}\mathbb{E}\left\{\left(\sum_{m=1}^{M}\varepsilon_m\right)^2\right\}=\frac{1}{M^2}\mathbb{E}\left\{\sum_{m=1}^{M}\varepsilon_m^2+\sum_{i\neq j}\varepsilon_i\varepsilon_j\right\}$$

$$=\frac{M}{M^2}\sigma^2+\frac{M^2-M}{M^2}\rho\sigma^2=\frac{\sigma^2}{M}+\left(1-\frac{1}{M}\right)\rho\sigma^2$$

- При $\rho=1$: σ^2; при $\rho=0$: σ^2/M
  - может быть еще меньше при $\rho<0$.
- На практике $\rho>0$, т.к. модели прогнозируют одинаковый отклик по одинаковой выборке.

Голосование большинства (против переобучения)

- Рассмотрим M классификаторов $f_1(x), ...f_M(x)$.
- Пусть $p(f_m(x) \neq y) = p < 0.5\,\forall m$
    - например, $p = 0.49$.
- Пусть модели ошибаются независимо друг от друга.
- Пусть $G(x)$ - выбор самого частого класса.
- Тогда $p(G(x) \neq y) \to 0$ при $M \to \infty$

Доказательство

Доказательство

Рассмотрим сл. вел.: $\xi_m = \begin{cases} +1, & f_m(x) = y \\ 0 & f_m(x) \neq y \end{cases}$ $\quad \eta = \frac{\xi_1 + ... + \xi_M}{M}$

$$G(x) = y \; \Leftrightarrow \; \eta > 0.5; \quad P(G(x) = y) = P(\eta > 0.5)$$

$$\frac{\sum_{i=1}^{M}[\xi_i - \mathbb{E}\xi_i]}{\sqrt{M\mathbb{D}(\xi_1)}} = \frac{\sum_{i=1}^{M}[\xi_i - p]}{\sqrt{Mp(1-p)}} = \frac{\sqrt{M}}{\sqrt{p(1-p)}} \frac{\sum_{i=1}^{M}\xi_i - Mp}{M}$$

$$= \frac{\sqrt{M}}{\sqrt{p(1-p)}}(\eta - p) \to \mathcal{N}(0,1) \text{ по ЦПТ.}$$

$$\eta \to \mathcal{N}(p, \frac{p(1-p)}{M}) \text{ при } M \to \infty$$

$$\mathbb{E}\eta = p \text{ и } \mathbb{D}\eta = \frac{p(1-p)}{M} \to 0 \text{ при } M \to \infty \;\Rightarrow$$

$P(\eta > 0.5) \to 1$ при $M \to \infty$ по неравенству Чебышева.

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

$$\underbrace{(G(x) - y)^2}_{\text{ошибка ансамбля}} = \underbrace{\sum_m w_m \left(f_m(x) - y\right)^2}_{\text{ошибки базовых моделей}} - \underbrace{\sum_m w_m \left(f_m(x) - G(x)\right)^2}_{\text{неоднозначность}}$$

- 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 разложения неоднозначности

Доказательство:

$$\sum_m w_m \left(f_m(x) - G(x)\right)^2 = \sum_m w_m \left(f_m(x) - y + y - G(x)\right)^2$$

$$= \sum_m w_m \left(f_m(x) - y\right)^2 + \sum_m w_m \left(y - G(x)\right)^2 + 2\sum_m w_m \left(f_m(x) - y\right)\left(y - G(x)\right)$$

$$= \sum_m w_m \left(f_m(x) - y\right)^2 + \left(G(x) - y\right)^2 + 2\left(y - G(x)\right)\sum_m w_m \left(f_m(x) - y\right)$$

$$= \sum_m w_m \left(f_m(x) - y\right)^2 + \left(G(x) - y\right)^2 + 2\left(y - G(x)\right)\left(G(x) - y\right)$$

$$= \sum_m w_m \left(f_m(x) - y\right)^2 + \left(G(x) - y\right)^2 - 2\left(G(x) - y\right)^2$$

$$= \sum_m w_m \left(f_m(x) - y\right)^2 - \left(G(x) - y\right)^2$$

Выпуклые потери

Выпуклые потери поощряют использование взвешенных прогнозов вместо индивидуальных.

- Рассмотрим регрессию с выпуклой ф-цией потерь $\mathcal{L}(\widehat{y}-y)$.
- Учитываем $f_1(x), ...f_M(x)$ с весами $w_1, ...w_M \geq 0$, $\sum_i w_i = 1$.

Для фикс. x какая стратегия выгоднее?

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

Выпуклые потери

Выпуклые потери поощряют использование взвешенных прогнозов вместо индивидуальных.

- Рассмотрим регрессию с выпуклой ф-цией потерь $\mathcal{L}(\widehat{y}-y)$.
- Учитываем $f_1(x),...f_M(x)$ с весами $w_1,...w_M \geq 0$, $\sum_i w_i = 1$.

Для фикс. x какая стратегия выгоднее?

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

выпуклость функции: $\mathcal{L}(\alpha\varepsilon_1 + (1-\alpha)\varepsilon_2) \leq \alpha\mathcal{L}(\varepsilon_1) + (1-\alpha)\mathcal{L}(\varepsilon_2)$

по индукции можно показать, что:

$$\mathcal{L}\left(w_1\varepsilon_1 + w_2\varepsilon_2 + ... + w_M\varepsilon_M\right) \leq w_1\mathcal{L}(\varepsilon_1) + w_2\mathcal{L}(\varepsilon_2) + ... + w_M\mathcal{L}(\varepsilon_M)$$