Бэггинг, ансамбли моделей и случайный лес

Основные свойства решающих деревьев

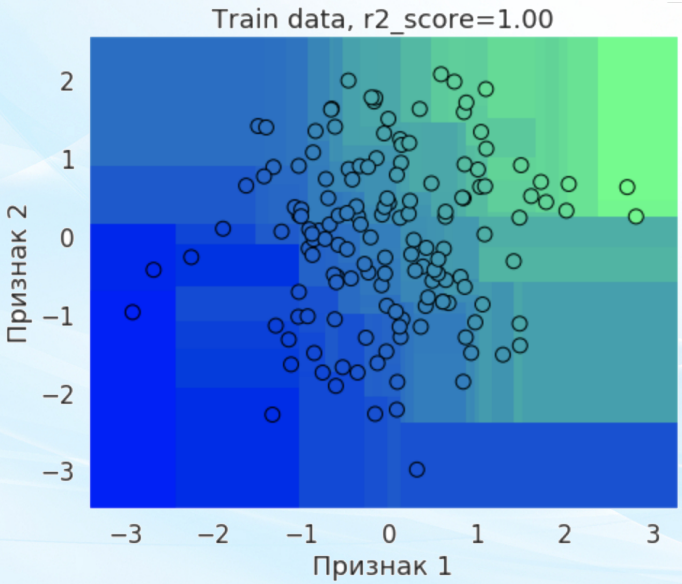
Плюсы

- Восстанавливают сложные закономерности

Минусы

- Очень легко переобучаются.

  Неустойчивы к малейшим изменениям в данных.
- Восстанавливаемая зависимость довольно ужасна.



$\Rightarrow$ Сами деревья не очень хороши.

Идея

Один в поле не воин...

Идея

Лес - много деревьев

Идея

А есть ли смысл брать деревья одинаковыми?

Нужны разные деревья

"Танцующий лес", нац. парк Куршская коса, Калининградская обл.

Идея

Возьмем композицию вида:

$$f = \frac{1}{T} \sum_{t=1}^{T} b_t$$

где b_t — решающее дерево.

Чтобы сделать деревья b_t разными:

- b_t обучаем на некоторой подвыборке.
- b_t обучаем на случайном подпространстве признаков.

Беггинг

Bagging = Boostrap Aggregating

Пусть есть выборка (X, Y).

Сгенерируем T бутстрепных подвыборок из нее.

На каждой из них обучим отдельную модель $\hat{y}_t = \mu_t(X_t^*, Y_t^*)$.

Итоговая модель строится как композиция:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$$

Модели $\hat{y}_t$ не обязаны быть моделями из одного вида моделей
Например, $\hat{y}_1$ может быть линейной моделью, а $\hat{y}_2$ - деревом.

Случайный лес

Возьмем в качестве базовых моделей решающие деревья.

Как сделать деревья разными?

- По объектам: Каждое дерево обучается на бутстрепной выборке.
- По признакам: Деревья в лесу являются *рандомизированными*.

  При каждом разбиении вершины выбираютсся случайные признаки для перебора.

Простая практика

Ансамбли моделей и случайные леса

Bias-variance tradeoff

Беггинг

Случайный лес

Важность признаков

Bias-variance tradeoff

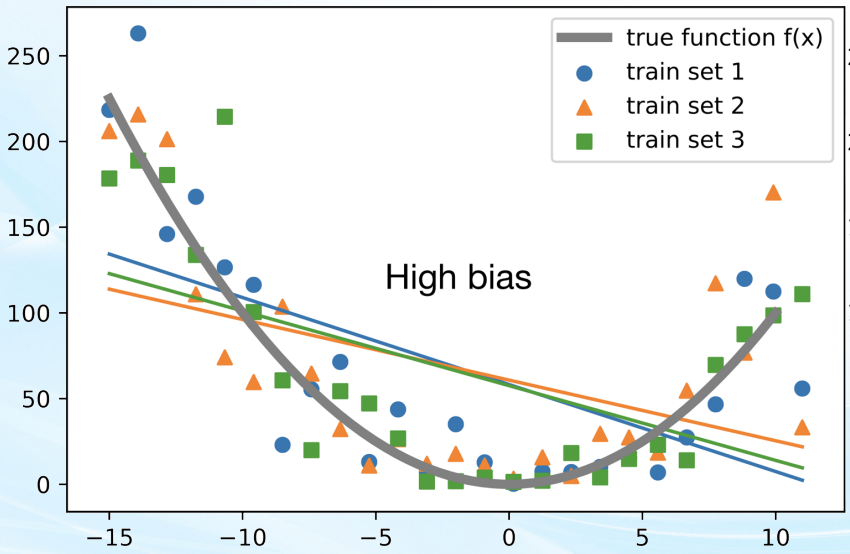
Bias-variance tradeoff

Шум = шум в данных.

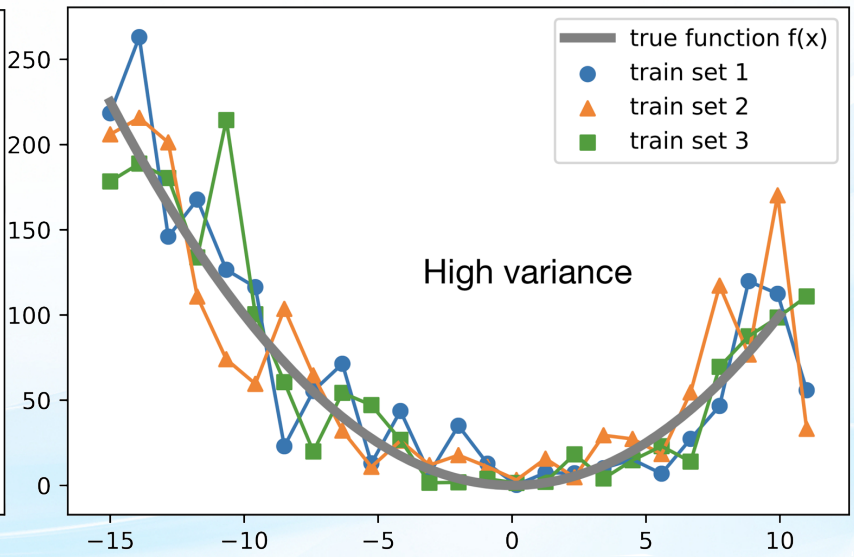
Смещение (bias) = среднее отклонение модели от истинной зависимости.

Разброс (variance) = среднеквадратичный разброс ответов обученных моделей относительно среднего ответа.

Разброс показывает, насколько сильно может измениться предсказание обученной модели в зависимости от разных реализаций выборки.



Большое смещение,
маленький разброс



Маленькое смещение,
большой разброс

Bias-variance tradeoff

Общий случай

Есть более общие формулы этого разложения для других функций, состоящие из трех компонент с похожим смыслом.

Т.е. для многих распространненых функций потерь ошибка метода обучения может быть разложена на шум, смещение и дисперсию.

Подробнее про общий вид разложения можно прочитать тут:
Domingos, Pedro (2000). A Unified Bias-Variance Decomposition and its Applications

Беггинг: вывод

Если базовые модели

- слабо коррелированы
- имеют низкое смещение
- имеют высокий разброс

то беггинг-композиция имеет низкое смещение и низкий разброс.

Когда модели менее коррелированы?

Когда они достаточно разные.

Как сделать модели разными?

- Использовать разные виды моделей и разные гиперпараметры.
- Обучать модели на разных признаках.
- Делать разную предобработку данных.

Ансамбли моделей и случайные леса

Bias-variance tradeoff

Беггинг

Случайный лес

Связь с метрическими моделями

Важность признаков

Случайный лес

Возьмем в качестве базовых моделей решающие деревья.

Свойства решающего дерева с большой глубиной:

- bias - низкий
- variance - высокий

  Напоминание: Деревья могут быть сильно разными даже при небольшом изменении выборки.

Случайный лес

- Деревья глубокие $\Rightarrow$ низкое смещение.
- Каждое дерево обучается на бутстрепной выборке $\Rightarrow$ разные.
- При разбиении признаки выбираются случайно $\Rightarrow$ разные.

Деревья разные $\Rightarrow$ малая корреляция, при объединении получим хорошую композицию.

Случайный лес

Пусть d — количество признаков, d_0 — количество случайно выбираемых признаков при разбиении.

Рекомендации:

- В задаче классификации

  Взять $d_0 = \left\lfloor \sqrt{d} \right\rfloor$.

  Строить каждое дерево до тех пор,

  пока в каждом листе не окажется по 1 объекту.

- В задаче регрессии

  Взять $d_0 = \left\lfloor d/3 \right\rfloor$.

  Строить каждое дерево до тех пор,

  пока в каждом листе не окажется по 5 объектов.

Ансамбли моделей и случайные леса

Bias-variance tradeoff

Беггинг

Случайный лес

Важность признаков

Mean Decrease in Impurity (MDI)

Случай одного дерева.

При разбиении одной вершины на две решаем задачу:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r) \to \min_{j,t}$$

Подобрав оптимальные j и t имеем уменьшение в критерии ошибки, равное $H(X_m) - \frac{|X_l|}{|X_m|} \cdot H(X_l) - \frac{|X_r|}{|X_m|} \cdot H(X_r)$.

Для задачи регрессии — это величина уменьшения MSE.

Это уменьшение является относительным по отношению к вершине m, а хотим посчитать общее уменьшение:

$$\Delta I_j^m = \frac{|X_m|}{|X|} H(X_m) - \frac{|X_l|}{|X|} \cdot H(X_l) - \frac{|X_r|}{|X|} \cdot H(X_r)$$

ΔI_j^m — уменьшение критерия ошибки на этапе разбиения вершины m по признаку j и оптимальному порогу t.

Mean Decrease in Impurity (MDI)

$\Rightarrow$ При построении дерева можем посчитать какой вклад каждый признак вносит в уменьшение ошибки:

$$\Delta I_j = \sum_{\substack{\text{m : разбиение в вершине } m \\ \text{происходит по признаку } j}} \Delta I_j^m$$

Отнормируем данные значения:

$$\widetilde{\Delta I_j} = \frac{\Delta I_j}{\sum\limits_i \Delta I_i}$$

Случай леса.

Для получения важности признаков для случайного леса усредним важности признаков, полученные от каждого дерева.

Пусть $\mathscr{T}$ — набор деревьев в лесу.

$\Delta I_j(T)$ — важность признака j для дерева T.

$$\Delta I_j = \frac{1}{|\mathscr{T}|} \sum_{T \in \mathscr{T}} \Delta I_j(T)$$

Mean Decrease in Impurity (MDI)

Плюсы:

- В sklearn RandomForest переменная feature_importances_ — важности признаков, посчитанные этим методом.
- Быстро считается, обучение происходит один раз.

Минусы:

- Важность признаков смещена в сторону признаков с большим количеством значений.

  Bias in information-based measures in decision tree induction, 1994
- Считается при использовании лишь обучающей выборки. Не смотрит на полезность признака при предсказании теста.