

Прикладная статистика в машинном обучении

Лекция 6

Непараметрическое оценивание

И. К. Козлов
(Мехмат МГУ)

2022

Непараметрическая регрессия

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

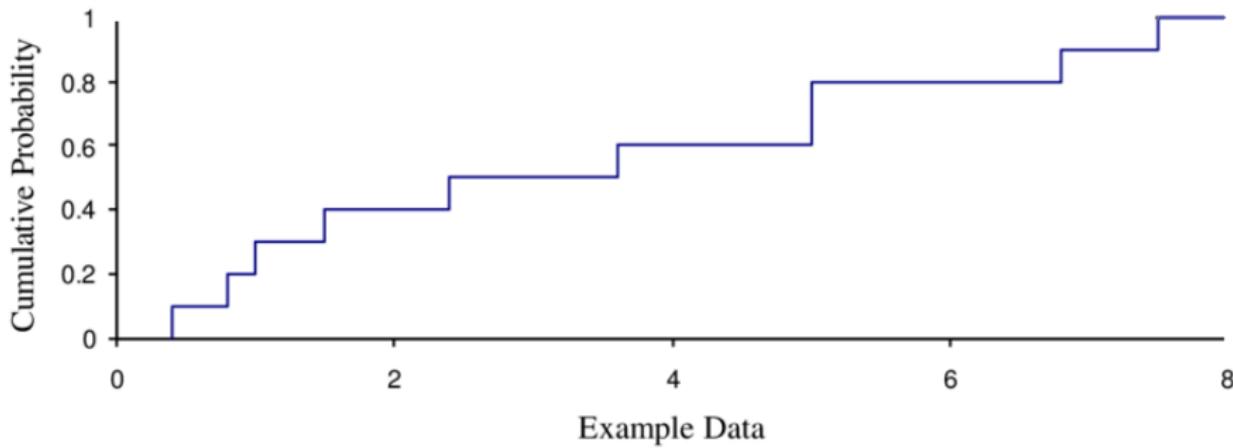
5 ML. Bias and Variance

- Bagging

Эмпирическая функция распределения

Q: Как аппроксимировать функцию распределения $F(x)$?

A: Конечно, эмпирической функцией распределения $\hat{F}(x)!$



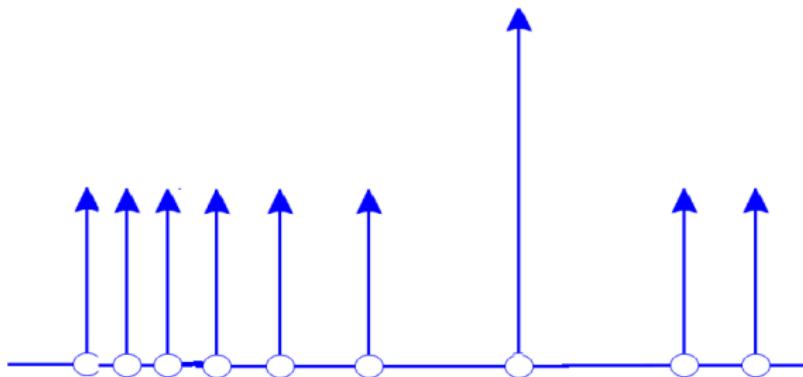
Плотность — сумма дельта-функций

Q: А какая у $\hat{F}(x)$ плотность?

Плотность — сумма дельта-функций

Q: А какая у $\hat{F}(x)$ плотность?

A: Сумма дельта-функций.

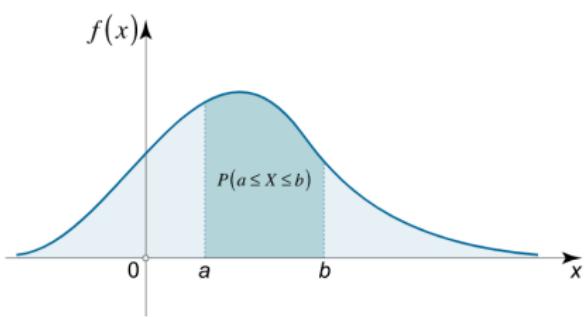


Почти всюду - ноль! Не гладко!

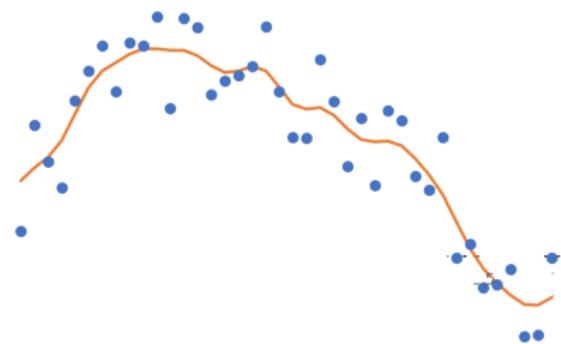
2 задачи

Рассмотрим 2 задачи.

Восстановление плотности $f(x)$

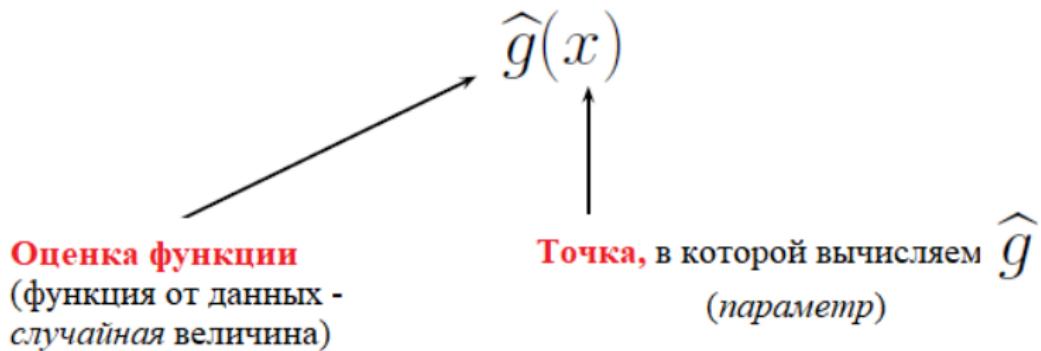


Нахождение регрессора $r(x) = E(Y | X = x)$



Оценка функции

Мы изучаем оценку $\hat{g}(x)$ функции $g(x)$.



Риск

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML, Bias and Variance

- Bagging

Функция потерь

Q: Как оценить — насколько хорошо приближение?

Функция потерь

Q: Как оценить — насколько хорошо приближение?

A: Функция потерь — интегральный квадрат ошибки (ISE):

$$L(g, \hat{g}_n) = \int (g(u) - \hat{g}(u))^2 du.$$

Так будут самые лучшие формулы.

Функция потерь

Q: Как оценить — насколько хорошо приближение?

A: Функция потерь — интегральный квадрат ошибки (ISE):

$$L(g, \hat{g}_n) = \int (g(u) - \hat{g}(u))^2 du.$$

Так будут самые лучшие формулы.

Мы хотим минимизировать риск = средний интегральный квадрат ошибки (MISE)

$$R(g, \hat{g}_n(x)) = \mathbb{E}(L(g, \hat{g}_n)).$$

Смещение и дисперсия

Теорема. Риск может быть записан в виде

$$R(g, \hat{g}_n) = \int b^2(x)dx + \int v(x)dx$$

- ① 1ое слагаемое — это смещение $\hat{g}(x)$ в точке x :

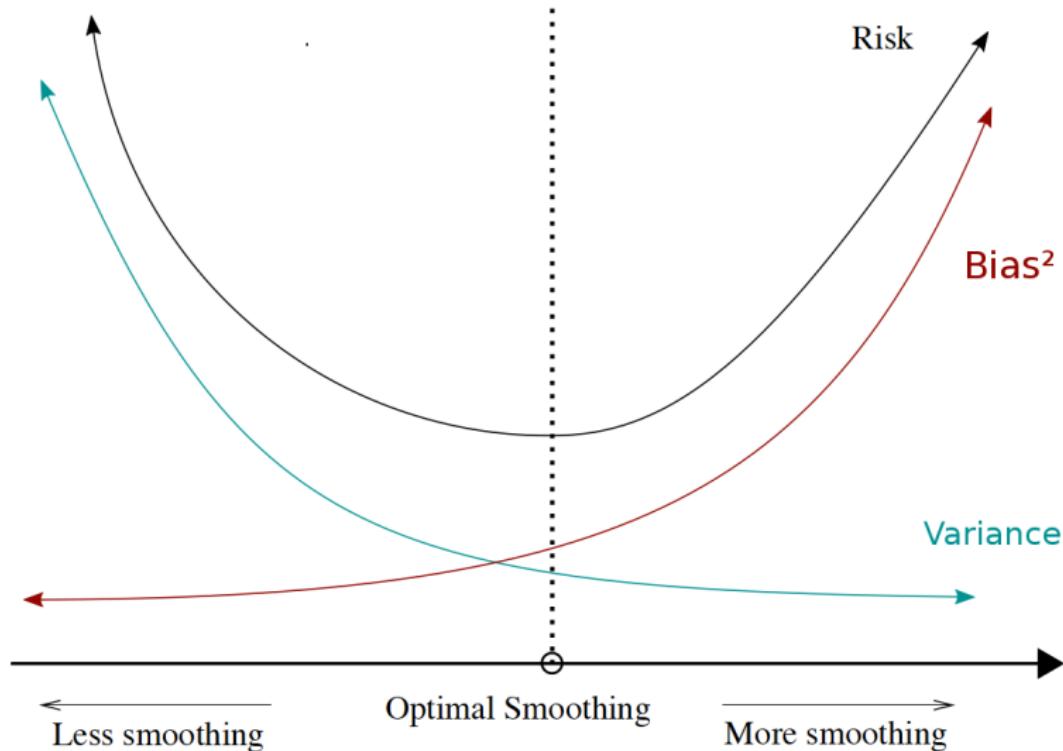
$$b(x) = \mathbb{E}[\hat{g}_n(x)] - g(x).$$

- ② 2ое слагаемое — это дисперсия $\hat{g}(x)$ в точке x :

$$v(x) = \mathbb{V}[\hat{g}_n(x)] = \mathbb{E}[\hat{g}_n(x) - \mathbb{E}(\hat{g}_n(x))]^2.$$

Дilemma смещения-дисперсия

Это всё тот же Bias-Variance Tradeoff



$$\text{Risk} = \text{Bias}^2 + \text{Variance}$$

Гистограммы

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML. Bias and Variance

- Bagging

Демонстрация ЦПТ

Начнём с восстановления плотности $f(x)$.

Какое самое важное распределение? Нормальное $\mathcal{N}(0, 1)$.

Q: Где возникает $\mathcal{N}(0, 1)$?

Демонстрация ЦПТ

Начнём с восстановления плотности $f(x)$.

Какое самое важное распределение? Нормальное $\mathcal{N}(0, 1)$.

Q: Где возникает $\mathcal{N}(0, 1)$?

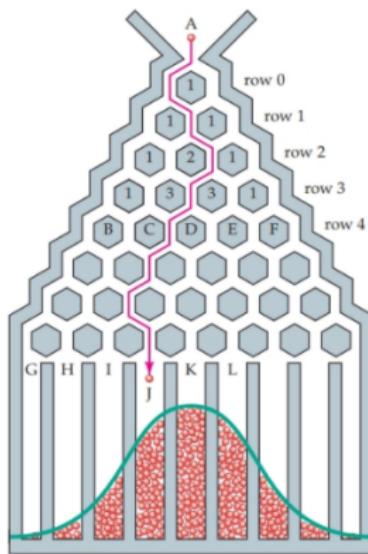
A: В ЦПТ.

Q-2: Сталкивались ли Вы с устройствами для демонстраций ЦПТ?

Доска Гальтона

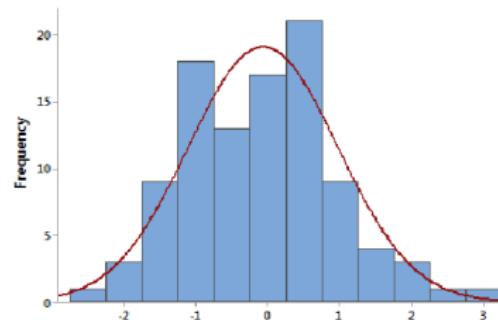
Доска Гальтона. Падающие шарики реализуют плотность $\mathcal{N}(0, 1)$.

Q: Можно ли аппроксимировать столбиками произвольную плотность?

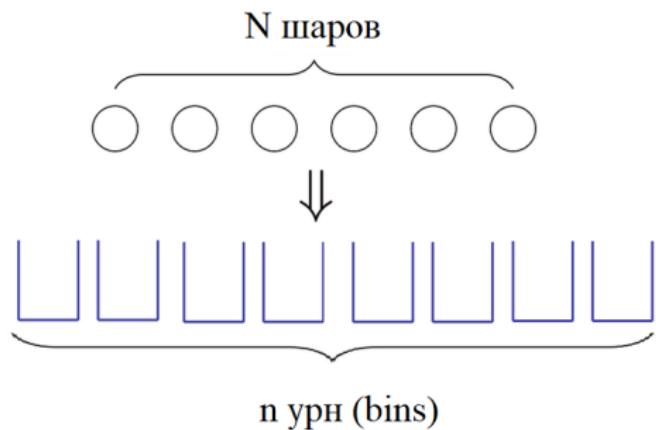


Гистограмма

Строим гистограмму.



Гистограмма



Гистограмма

- Пусть X_1, \dots, X_n — i.i.d. выборка из $[0, 1]$ с плотностью $f(x)$.

- Разбиваем отрезок на $m \in \mathbb{N}$ ячеек:

$$B_1 = \left[0, \frac{1}{m}\right), \quad B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \quad \dots, \quad B_m = \left[\frac{m-1}{m}, 1\right].$$

- Ширина (binwidth) $h = \frac{1}{m}$.

Гистограмма

- Пусть в B_j попадает ν_j элементов. Полагаем $\hat{p}_j = \frac{\nu_j}{n}$.

Гистограмма (оценка плотности):

$$\hat{f}_n(x) = \sum_{j=1}^n \frac{\hat{p}_j}{h} I(x \in B_j).$$

То есть $f(x) = \frac{\hat{p}_j}{h}$, если $x \in B_j$.

Оценка плотности

Аргументируем формулу

$$\hat{f}_n(x) = \sum_{j=1}^n \frac{\hat{p}_j}{h} I(x \in B_j).$$

Положим $p_j = \int_{B_j} f(u) du$.

Q: Пусть $x \in B_j$. Чему равно $\mathbb{E}[\hat{f}_n(x)]$?

Оценка плотности

Аргументируем формулу

$$\hat{f}_n(x) = \sum_{j=1}^n \frac{\hat{p}_j}{h} I(x \in B_j).$$

Положим $p_j = \int_{B_j} f(u) du$.

Q: Пусть $x \in B_j$. Чему равно $\mathbb{E}[\hat{f}_n(x)]$?

A: По линейности матожидания

$$\mathbb{E}[\hat{f}_n(x)] = \mathbb{E}\left[\frac{\hat{p}_j}{h}\right] = \frac{p_j}{h}.$$

Если h мало, то

$$\frac{p_j}{h} = \frac{\int_{B_j} f(u) du}{h} \approx \frac{f(x)h}{h} = f(x).$$

Выбор ширины

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

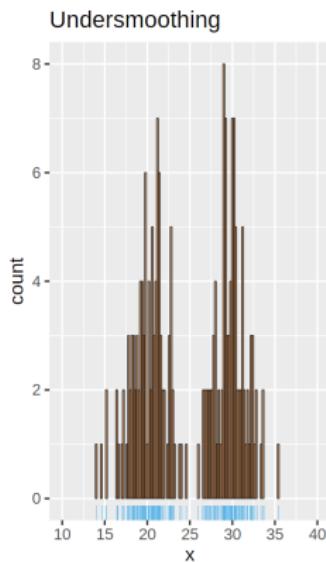
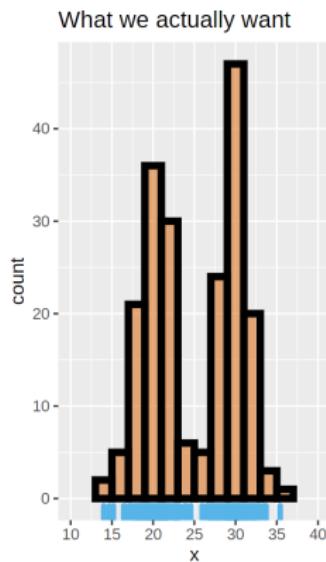
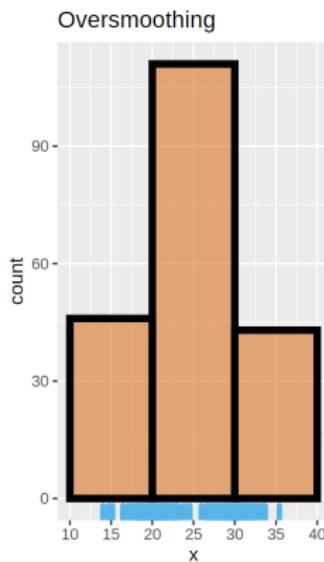
4 Формула Надаля-Ватсона

5 ML. Bias and Variance

- Bagging

Оптимальная ширина

Главный вопрос — как подобрать оптимальную ширину h ?



Q: Как оценить “оптимальность” h ?

Risk management

Q: Как оценить “оптимальность” h ?

A: Мы минимизируем риск $R(\hat{f}_n, f)$.



Рис.: Занятие по минимизации рисков

Risk, Bias and Variance

Оценим риск R .

Разложим плотность $f(x)$ в ряд Тейлора в каждой ячейке B_j .

Фиксируем $x \in B_j$. Тогда для $u \in B_j$

$$f(u) \approx f(x) + (u - x)f'(x).$$

Risk, Bias and Variance

Оценим риск R .

Разложим плотность $f(x)$ в ряд Тейлора в каждой ячейке B_j .

Фиксируем $x \in B_j$. Тогда для $u \in B_j$

$$f(u) \approx f(x) + (u - x)f'(x).$$

Теорема

Пусть $\int (f'(u))^2 du < \infty$. Тогда

$$R(\hat{f}_n, f) \approx \frac{h^2}{12} \int (f'(u))^2 du + \frac{1}{nh}.$$

Опускаем технические детали

Выкладки можно найти в Главе 20



Wasserman L.
All of Statistics.

Мы опускаем подобную технику, т.к. это страница простеньких разложений в ряд Тейлора.

Можно полюбоваться на эту страницу на следующем слайде.

Хотя лучше её **пропустить** 😊.

Бассерман. Глава 20

Let's take a closer look at the bias-variance tradeoff using equation (20.9). Consider some $x \in B_j$. For any other $u \in B_j$,

$$f(u) \approx f(x) + (u - x)f'(x)$$

and so

$$\begin{aligned} p_j = \int_{B_j} f(u) du &\approx \int_{B_j} \left(f(x) + (u - x)f'(x) \right) du \\ &= f(x)h + hf'(x) \left(h \left(j - \frac{1}{2} \right) - x \right). \end{aligned}$$

Therefore, the bias $b(x)$ is

$$\begin{aligned} b(x) &= \mathbb{E}(\widehat{f}_n(x)) - f(x) = \frac{p_j}{h} - f(x) \\ &\approx \frac{f(x)h + hf'(x) \left(h \left(j - \frac{1}{2} \right) - x \right)}{h} - f(x) \\ &= f'(x) \left(h \left(j - \frac{1}{2} \right) - x \right). \end{aligned}$$

If \bar{x}_j is the center of the bin, then

$$\begin{aligned} \int_{B_j} b^2(x) dx &\approx \int_{B_j} (f'(x))^2 \left(h \left(j - \frac{1}{2} \right) - x \right)^2 dx \\ &\approx (f'(\bar{x}_j))^2 \int_{B_j} \left(h \left(j - \frac{1}{2} \right) - x \right)^2 dx \\ &= (f'(\bar{x}_j))^2 \frac{h^3}{12}. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_0^1 b^2(x) dx &= \sum_{j=1}^m \int_{B_j} b^2(x) dx \approx \sum_{j=1}^m (f'(\bar{x}_j))^2 \frac{h^3}{12} \\ &= \frac{h^2}{12} \sum_{j=1}^m h (f'(\bar{x}_j))^2 \approx \frac{h^2}{12} \int_0^1 (f'(x))^2 dx. \end{aligned}$$

Note that this increases as a function of h . Now consider the variance. For h small, $1 - p_j \approx 1$, so

$$\begin{aligned} v(x) &\approx \frac{p_j}{nh^2} \\ &= \frac{f(x)h + hf'(x) \left(h \left(j - \frac{1}{2} \right) - x \right)}{nh^2} \\ &\approx \frac{f(x)}{nh} \end{aligned}$$

Риск

Итак, риск имеет вид

$$R(h) \approx \boxed{h^2 C_0} + \boxed{\frac{1}{nh}}$$

Risk Bias Variance

Q: При каком h указанный риск минимален?

Оптимальная ширина в теории

A: Оптимальная ширина

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int (f'(u))^2 du} \right)^{1/3} \sim \frac{1}{n^{1/3}}.$$

При такой ширине

$$R(\hat{f}_n, f) \approx \frac{C}{n^{2/3}}.$$

Q: Можем ли мы на практике вычислить h^* ?

Оптимальная ширина в теории

A: Оптимальная ширина

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int (f'(u))^2 du} \right)^{1/3} \sim \frac{1}{n^{1/3}}.$$

При такой ширине

$$R(\hat{f}_n, f) \approx \frac{C}{n^{2/3}}.$$

Q: Можем ли мы *на практике* вычислить h^* ?

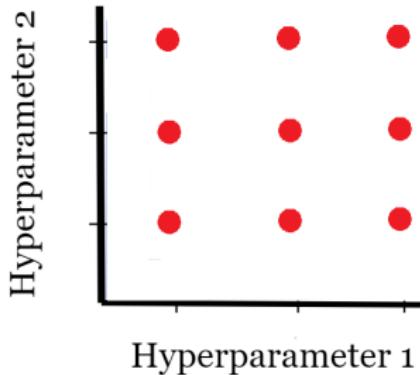
A: Нет, мы не знаем $f'(x)$.

Q: Что же делать? В ML как находятся гиперпараметры?

Q: Что же делать? В ML как находятся гиперпараметры?

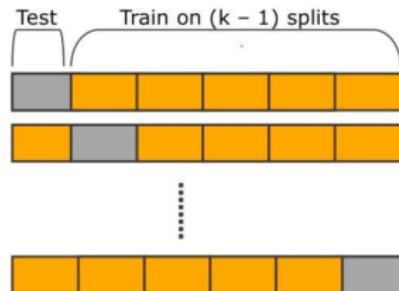
GridSearch

"Перебираем параметры по сеточке"



Cross-Validation

"Обучаемся на $n-1$ элементе, проверяем на оставшемся"



Опускаем технические детали

Без доказательства приведём следующие факты из Главы 20



Wasserman L.
All of Statistics.

Запутываем

Изменим лосс: раскроем скобки и отбросим последнее слагаемое-константу.

$$\begin{aligned} L(\hat{f}_n, f) &= \int (\hat{f}_n(x) - f(x))^2 dx = \\ &= \underbrace{\int (\hat{f}_n(x))^2 dx - 2 \int \hat{f}_n(x)f(x)dx + \int (f(x))^2 dx}_{J(h)} \end{aligned}$$

Теперь **риском** будем называть

$$R(h) = \mathbb{E}[J(h)].$$

Кросс-валидационная оценка риска

На практике:

- ➊ Лосс $J(h)$ оценивается через кросс-валидационную оценку риска

$$\hat{J}(h) = \int (\hat{f}_n(x))^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i),$$

где \hat{f}_{-i} — гистограммная оценка с опущенным i -тым наблюдением.

Кросс-валидационная оценка риска

На практике:

- Лосс $J(h)$ оценивается через кросс-валидационную оценку риска

$$\hat{J}(h) = \int (\hat{f}_n(x))^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i),$$

где \hat{f}_{-i} — гистограммная оценка с опущенным i -тым наблюдением.

- Важно — есть удобная формула для вычислений

$$\hat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{n-1} \sum_{j=1}^m \hat{\rho}_j^2.$$

Перерыв

Перерыв

Доверительные интервалы

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

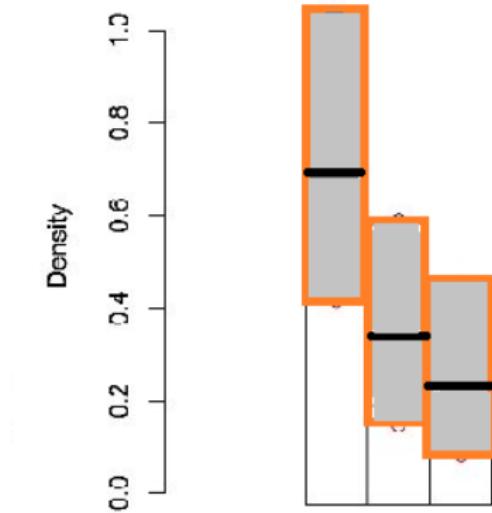
5 ML, Bias and Variance

- Bagging

Доверительные полосы

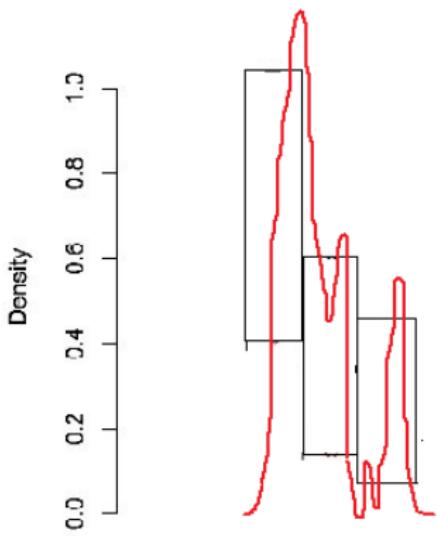
Также можно построить доверительные полосы

Confidence band for histogram

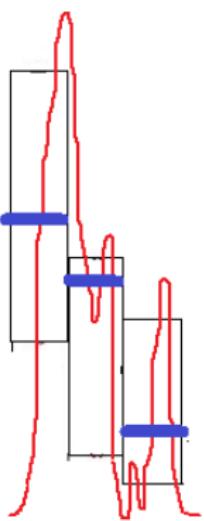


Нужно усреднить

Можно ли вписать
сильноколеблющуюся функцию
плотности?



Нет! Нужно усреднять по ячейкам.



Нужно усреднить

- “Гистограммная” версия функции $f(x)$:

$$\bar{f}_n(x) = \frac{p_j}{h}, \quad p_j = \int_{B_j} f(u) du, \quad \text{для любого } x \in B_j.$$

- Пара функций $(\ell_n(x), u_n(x))$ — это $1 - \alpha$ доверительная полоса, если

$$\mathbb{P}(\ell_n(x) \leq \bar{f}(x) \leq u_n(x), \quad \text{для всех } x) \geq 1 - \alpha.$$

Доверительные полосы

Для справки приведём ответ из



Wasserman L.
All of Statistics.

Теорема

Пусть $m = m(n)$ — количество бинов. Условия на ширину:

$$m(n) \rightarrow 0, \quad \frac{m(n) \log n}{n} \rightarrow 0 \quad \text{при } n \rightarrow \infty$$

Тогда $1 - \alpha$ доверительная полоса:

$$\ell_n(x) = \left(\max \left\{ \sqrt{\hat{f}_n(x)} - c, 0 \right\} \right)^2, \quad u_n(x) = \left(\sqrt{\hat{f}_n(x)} + c \right)^2,$$

где

$$c = \frac{z_{\alpha/2m}}{2} \sqrt{\frac{m}{n}}.$$

Нужно смириться

Q: Нас не смущает, что оценка для $\bar{f}(x)$ — “гистограммной” версии $f(x)$?

Нужно смириться

Q: Нас не смущает, что оценка для $\bar{f}(x)$ — “гистограммной” версии $f(x)$?

A: Нет, от ошибки в аппроксимации $f(x) - \bar{f}(x)$ избавиться нельзя.



Рис.: С этим нужно смириться

Ядерное сглаживание

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML. Bias and Variance

- Bagging

Всё должно быть гладко

Наши оценки плотности — негладкие! Исправим это.

Аппроксимируем дельта-функции локальными “шапочками”.



Дельта-функция

Достаточно гладкая функция

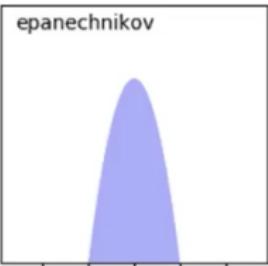
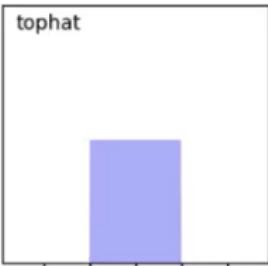
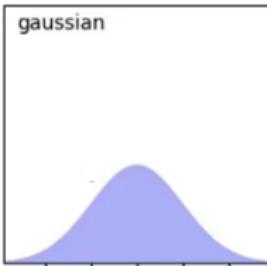
Ядро

Ядром $K(x)$ мы будем называть симметричную функцию плотности

$$K(x) \geq 0, \quad \int K(x) dx = 1, \quad K(x) = K(-x).$$

Ядро

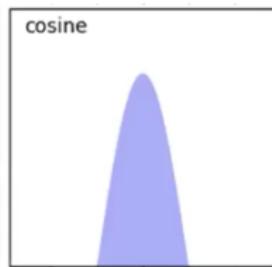
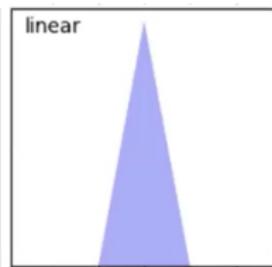
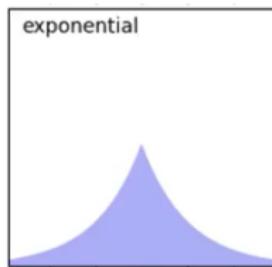
Есть много разных ядер. Запоминать их не нужно.



$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

$$K(u) = \frac{1}{2}$$

$$K(u) = \frac{3}{4}(1 - u^2)$$

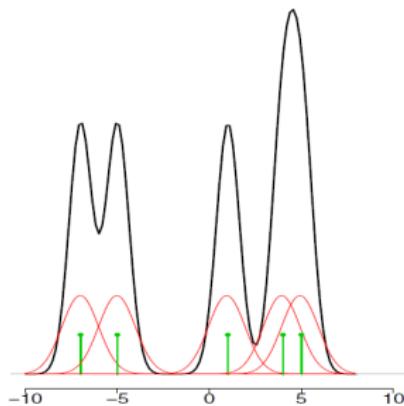


$$K(u) = \frac{1}{2} e^{-|u|}$$

$$K(u) = (1 - |u|)$$

$$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right)$$

Ядерная оценка плотности



Пусть X_1, \dots, X_n — наблюдаемые данные.

Фиксируем ядро K и ширину полосы h .

Ядерная оценка плотности:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

Выбор ядра

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML, Bias and Variance

- Bagging

Поиск лучшего ядра

Естественный вопрос — как найти “самое лучшее ядро”
(и нужно ли это делать)?



Рис.: Все ли ядра — чистый изумруд?

Оптимальность ядра Епанечникова

В Главе 22 Лагутина объясняется, в каком смысле
ядро Епанечникова является “**оптимальным**”.



М. Б. Лагутин,
Наглядная математическая статистика.

Это теоретический результат, не особо важный на практике.

В правом столбце таблицы — эффективность ядер по отношению к оптимальному ядру Епанечникова.

N	Ядро	$q(y)$	$E(q)$
1	Епанечникова	$(3/4)(1 - y^2) I_{\{ y \leq 1\}}$	1
2	Квартическое	$(15/16)(1 - y^2)^2 I_{\{ y \leq 1\}}$	0,995
3	Треугольное	$(1 - y) I_{\{ y \leq 1\}}$	0,989
4	Гаусса	$(2\pi)^{-1/2} \exp\{-y^2/2\}$	0,961
5	Прямоугольное	$(1/2) I_{\{ y \leq 1\}}$	0,943

Дистанции
огромного
размера!

Рис.: Таблица из главы 22 Лагутина

Важна ширина ядра, а не его тип

“Народная мудрость”:



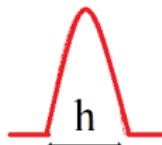
Тип ядра



Не важно



Ширина ядра



Важно

Утверждения аналогичны

Для ядерных оценок есть результаты, похожие на случай гистограмм.

Подробнее — Глава 20



Wasserman L.

All of Statistics.

или Глава 22



М. Б. Лагутин,

Наглядная математическая статистика.

Теорема Колмогорова

- Оптимальная ширина окна порядка $\frac{1}{n^{1/5}}$. Риск

$$R(f, \hat{f}_n) \approx \frac{C}{n^{4/5}}.$$

Теорема Колмогорова

- Оптимальная ширина окна порядка $\frac{1}{n^{1/5}}$. Риск

$$R(f, \hat{f}_n) \approx \frac{C}{n^{4/5}}.$$

- По теореме Колмогорова (для непр. $F(x)$)

$$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{d} K$$

скорость сходимости эмпирической функции распределения порядка $\frac{1}{n}$.

- Q: Почему ядерная оценка медленнее сходится?

Теорема Колмогорова

- Оптимальная ширина окна порядка $\frac{1}{n^{1/5}}$. Риск

$$R(f, \hat{f}_n) \approx \frac{C}{n^{4/5}}.$$

- По теореме Колмогорова (для непр. $F(x)$)

$$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{d} K$$

скорость сходимости эмпирической функции распределения порядка $\frac{1}{n}$.

- **Q:** Почему ядерная оценка медленнее сходится?

A: Она “видит” не все точки. (Зато гладкая. ☺)

Формула Надаля-Ватсона

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML. Bias and Variance

- Bagging

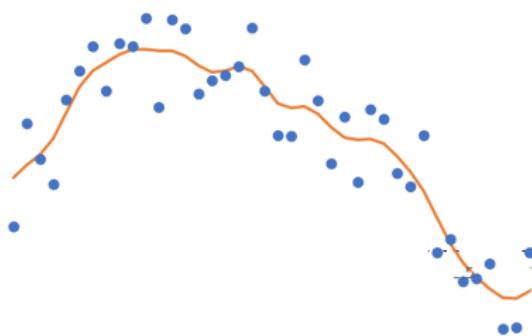
Регрессия

Регрессия. Даны пары $(x_1, Y_1), \dots, (x_n, Y_n)$, где

$$Y_i = r(x_i) + \varepsilon_i, \quad \text{и} \quad \mathbb{E}\varepsilon_i = 0.$$

Мы ищем регрессор

$$r(x) = \mathbb{E}(Y|X=x)$$



Взвешенные средние

Регрессор $r(x)$ — функция от Y_i .

Q: Логично, что чем x_i ближе, тем вклад Y_i должен быть больше?

Взвешенные средние

Регрессор $r(x)$ — функция от Y_i .

Q: Логично, что чем x_i ближе, тем вклад Y_i должен быть больше?

A: Поэтому многие оценки — **взвешенные средние**

$$\sum_{i=1}^n \frac{w_i(x)}{\sum_j w_j(x)} Y_i.$$

Оценка Надарадя–Ватсона

Q: Что взять за $w_i(x)$? Мы уже сталкивались сегодня с локальными функциями?

Оценка Надарадя–Ватсона

Q: Что взять за $w_i(x)$? Мы уже сталкивались сегодня с локальными функциями?

A: Оценка Надарадя–Ватсона

$$\hat{r}(x) = \sum_{i=1}^n \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x - x_j}{h}\right)} Y_i.$$

Оценки

При должных условиях для оценки Надарая-Ватсона схожи с ядерными:

- оптимальная ширина $\sim \frac{1}{n^{1/5}}$.
- риск $\sim \frac{1}{n^{4/5}}$.



Wasserman L.
All of Statistics.



М. Б. Лагутин,
Наглядная математическая статистика.

Подведём итоги

- ① При непараметрической регрессии главное — подбор **ширины окна**.
- ② При подборе ширины есть **Bias-Variance Tradeoff**.
- ③ Оценки сглаживаются при помощи ядер $K(x)$. Ширина ядра важна, тип - нет.
- ④ Есть формулы для доверительных интервалов.

Нужно помнить — они центрированы вокруг “усреднений” $\bar{f}(x)$, а не $f(x)$!

Перерыв

Перерыв

ML. Bias and Variance

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML. Bias and Variance

- Bagging

Смещение и дисперсия

Хотим точную оценку $\hat{\theta} = \theta$. Возможны 2 проблемы:

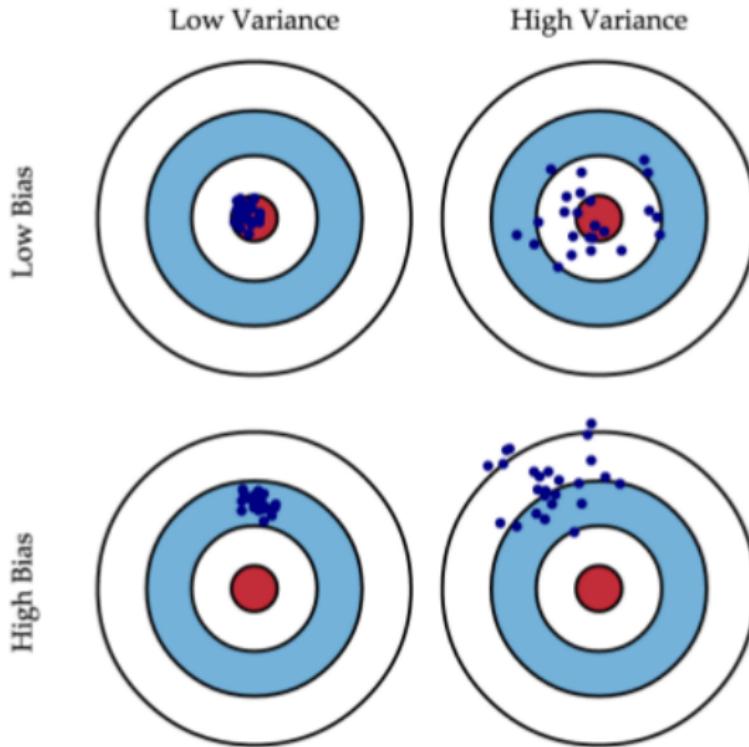
- ① Большое смещение

$$\text{Bias} = \mathbb{E}[\hat{\theta}] - \theta.$$

- ② Большая дисперсия

$$\mathbb{V}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})^2].$$

Bias and Variance



ML. Bias and Variance

Bias-Variance Tradeoff повсюду в ML:

- ① L_1 и L_2 -регуляризация линейных моделей.

Добавляется Bias, уменьшается Variance. (Лекция 2)

- ② Feature selection.

Больше признаков уменьшает Bias и увеличивает Variance. (Лекция 5)

ML. Bias and Variance

Bias-Variance Tradeoff повсюду в ML:

- ① L_1 и L_2 -регуляризация линейных моделей.

Добавляется Bias, уменьшается Variance. (Лекция 2)

- ② Feature selection.

Больше признаков уменьшает Bias и увеличивает Variance. (Лекция 5)

- ③ Метод k-ближайших соседей.

Большее k увеличивает Bias и уменьшает Variance. (Далее)

- ④ Решающие деревья.

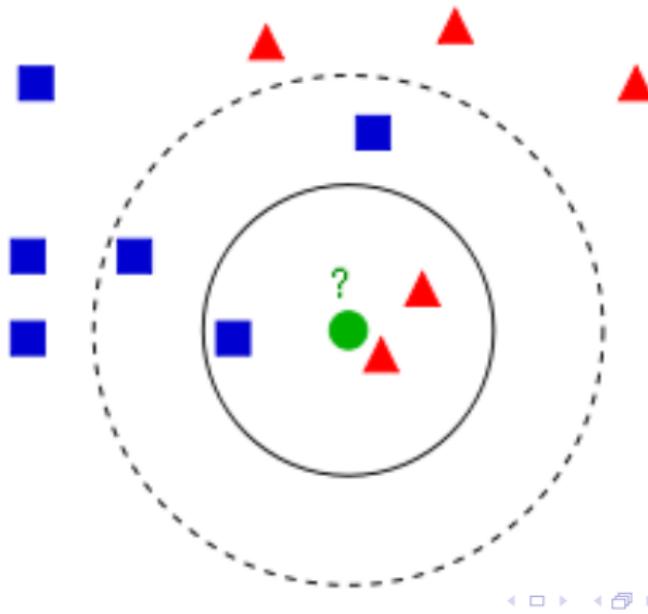
Чем больше глубина дерева, тем больше Variance. (Далее)

KNN

Метод k -ближайших соседей (KNN).

Классификация: метка элемента — самая частая у k ближайших соседей.

Регрессия: предсказание — среднее значение по k ближайшим объектам.



KNN

Для KNN-регрессии можно написать точную формулу Bias-Variance Tradeoff:

$$\mathbb{E}[(y - \hat{f}(x))^2 | X = x] = \left(f(x) - \frac{1}{k} \sum_{i=1}^k f(N_i(x)) \right)^2 + \frac{\sigma^2}{k} + \sigma^2$$

Bias Variance

Растёт с ростом k Уменьшается с ростом k

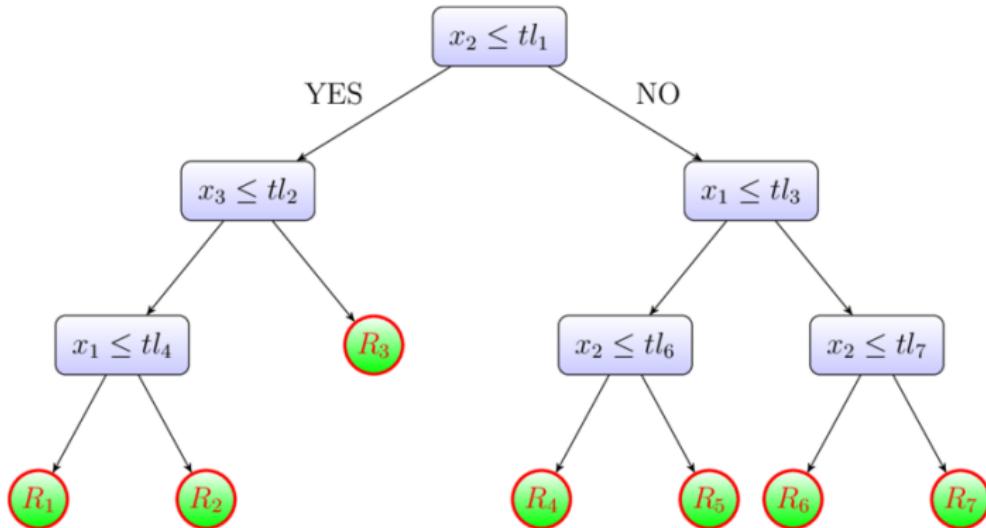


T. Hastie; R. Tibshirani; J. H. Friedman,
The Elements of Statistical Learning.

Decision Tree

Решающее дерево (Decision Tree).

Известная схема принятия решений ☺.



Decision Tree

Q: Что будет, если натравить О-О-ОЧЕНЬ большое дерево на датасет?

Decision Tree

Q: Что будет, если натравить О-О-ОЧЕНЬ большое дерево на датасет?

Модель выучит каждое событие (каждому элементу — по листу).

Если теперь чуть-чуть поменять данные — ответ резко изменится.

Поэтому у глубоких деревьев больший **Variance**.

Нужно смириться

Не всё коту Масленица.



Нужно смириться

Не всё коту Масленица.



Или нет?

Bagging

1 Непараметрическая регрессия

- Риск

2 Гистограммы

- Выбор ширины
- Доверительные интервалы

3 Ядерное сглаживание

- Выбор ядра

4 Формула Надаля-Ватсона

5 ML. Bias and Variance

- Bagging

Демократия нам поможет

Доверимся демократии!

Парламентаризм — решение принимается **большинством голосов**.

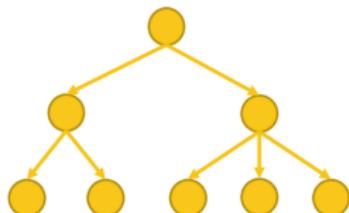


Нужно смириться

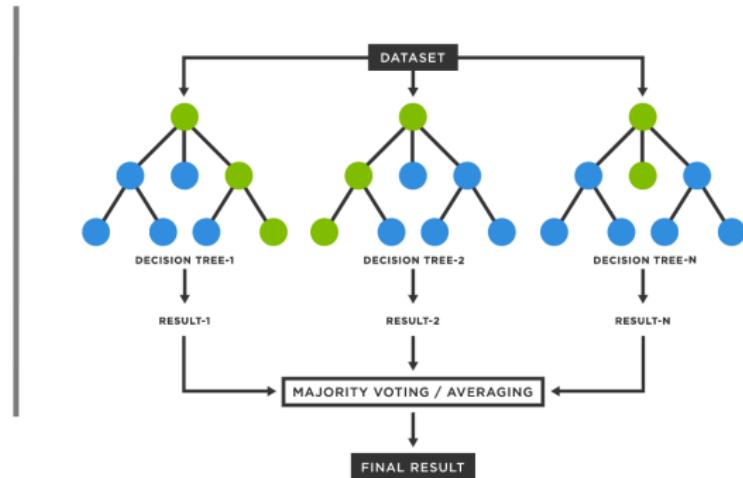
Как учит нас математика, **лес** — это набор **деревьев**.

Q: Что эффективней?

Decision Tree



Random Forest



Decision Tree

Простые, быстрые

Интерпретируемые

Random Forest

Точнее

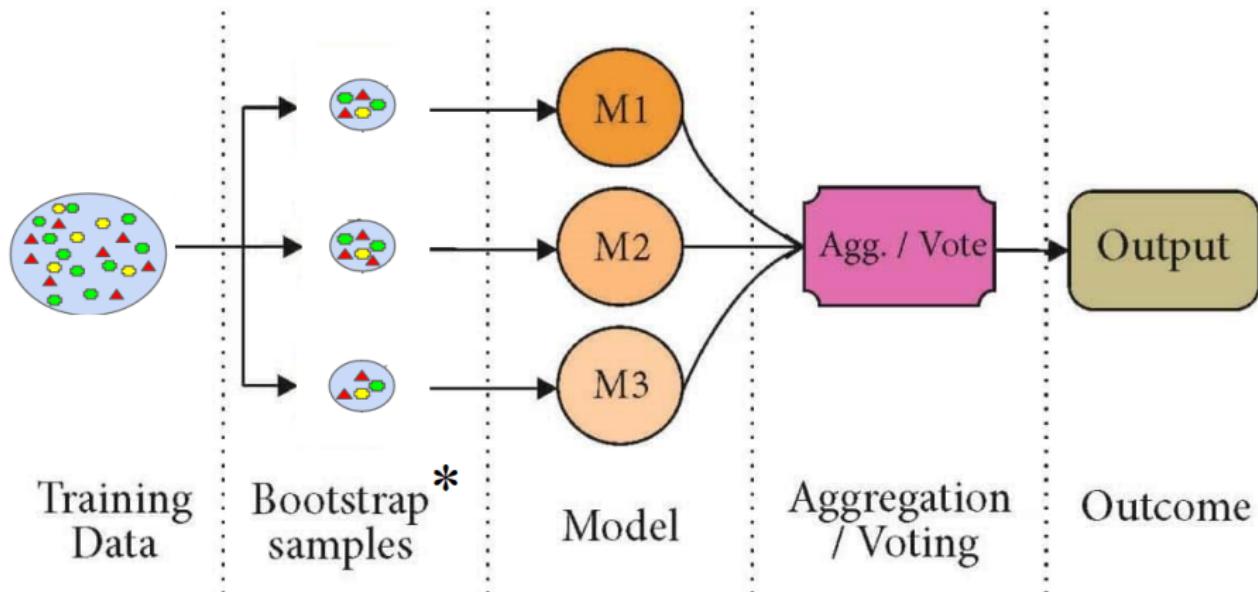
Робастные (меньше влияют выбросы)

Меньше переобучаются

Легко распараллеливаются

Q: А почему мы усредняем ответы именно деревьев?

BAGGING



Бэггинг. Дисперсия

Q: Что происходит с дисперсией при бэггинге?

Простое соображение: пусть X_1, \dots, X_n — i.i.d. выборка. Чему равны

$$\mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = ? \quad \mathbb{V}\left(\frac{X_1 + \dots + X_n}{n}\right) = ?$$

Бэггинг. Дисперсия

Q: Что происходит с дисперсией при бэггинге?

Простое соображение: пусть X_1, \dots, X_n — i.i.d. выборка. Чему равны

$$\mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = ? \quad \mathbb{V}\left(\frac{X_1 + \dots + X_n}{n}\right) = ?$$

Матожидание усредняется, а **дисперсия уменьшается**

$$\mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \mathbb{E}X_1, \quad \mathbb{V}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n}\mathbb{V}X_1.$$

Bagging decreases Variance

Bagging — способ уменьшить дисперсию
для “сильных” моделей (Low Bias).



T. Hastie; R. Tibshirani; J. H. Friedman,
The Elements of Statistical Learning.

Замечание. Способ уменьшить дисперсию для “слабых” моделей — бустинг¹.

¹ Нейронки и градиентный бустинг — сейчас самые эффективные модели.

Не недооценивайте силу демократии 😊.

И никогда не сдавайтесь!



To Be Continued ━━━━