

Прикладная статистика в машинном обучении

Лекция 3

Бутстреп

И. К. Козлов
(Мехмат МГУ)

2022

Вопрос дня

Нужны данные о 100 котятах. У нас есть 6 котят.

Что делать?



Рис.: Каждый котёнок — индивидуальность!

Не CatBoost, а Bootstrap!

Q: Как получить больше информации, если мы не можем больше сэмплировать?

A: Изучать имеющиеся объекты по несколько раз!

Доверительные интервалы

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Доверительный интервал

Доверительный интервал с доверительной вероятностью $1 - \alpha$ для параметра θ — это интервал

$$C_n = (a, b), \quad a = a(X_1, \dots, X_n), \quad b = b(X_1, \dots, X_n),$$

для которого

$$P_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{для всех } \theta \in \Theta. \quad (1)$$

Q: Что такое 95% доверительный интервал? Что означает “вероятность 95%”?

Фриквентистский подход

Q2: C_n и θ — что здесь случайные величины?

Фриквентистский подход

Q2: C_n и θ — что здесь случайные величины?

- θ — (неизвестный) фиксированный параметр.
- C_n — случайная величина.

Фриквентистский подход

Фриквентистская (частотная) интерпретация вероятности:

- Проводим (бесконечную) серию экспериментов.

Вероятность события — предел частоты этого события в экспериментах:

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N}.$$

Фриквентистский подход

Фриквентистская (частотная) интерпретация вероятности:

- Проводим (бесконечную) серию экспериментов.

Вероятность события — предел частоты этого события в экспериментах:

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N}.$$

Пример. Подбрасываем монетку.

Вероятность выпадения орла $p = \lim_{N \rightarrow \infty} \frac{\text{число орлов}}{\text{число бросков}}$.

Фриквентистский подход

Интерпретация доверительного интервала-1.

Проводим один и тот же эксперимент. θ — не меняется, фиксированный параметр.

В 95% случаев (в пределе) θ попадаем в построенный интервал C_n .

Q: Слабое место этой интерпретации?

A: Нужно проводить один и тот же эксперимент.

Фриквентистский подход

Интерпретация доверительного интервала-2.

Проводим серию экспериментов.

- ① В 1ый день строим 95% доверительный интервал для параметра θ_1 .
- ② Во 2ой день строим 95% доверительный интервал для параметра θ_2 .
- ③ И так далее...

В 95% случаев (в пределах) θ_i попадаем в построенный интервал.

Пример. Покупаем каждый день газету. Там рейтинг кандидатов ± пара пунктов.

Обычно это 95% интервал.



Чёрный лебедь

Только в 95% случаев(!)



Рис.: Жизнь не идеальна¹

¹ Российскую социологию даже обсуждать не будем

Асимптотический доверительный интервал

Как строить доверительный интервал?

Точно построить сложно, пытаемся сделать это асимптотически:

- Поточечный асимптотический доверительный интервал

$$\liminf_{n \rightarrow \infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{для всех } \theta \in \Theta.$$

Асимптотическая нормальность и доверительные интервалы

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Асимптотический доверительный интервал

В начале рассмотрим случай, когда оценка асимптотически нормальна:

$$\frac{\hat{\theta} - \theta}{\sigma} \rightarrow \mathcal{N}(0, 1).$$

Для краткости будем писать

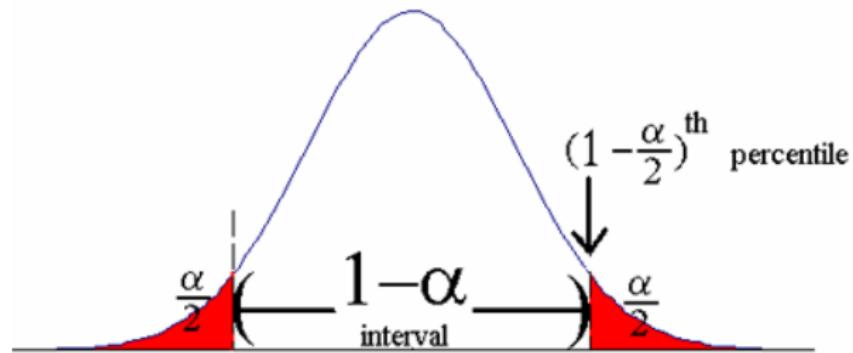
$$\hat{\theta} \approx \mathcal{N}(\theta, \sigma^2).$$

Типичный доверительный интервал

Q: Как построить $(1 - \alpha)$ доверительный интервал вокруг моды нормального распределения?

A: Неформально, “отрубить $\frac{\alpha}{2}$ хвосты” с обеих сторон.

z_α — стандартное обозначение квантиля $\Phi^{-1}(\alpha)$.



Типичный доверительный интервал

Часто оценка параметров асимптотически нормальна:

$$\hat{\theta}_n \approx \mathcal{N}(\theta, \hat{s}\hat{e}^2).$$

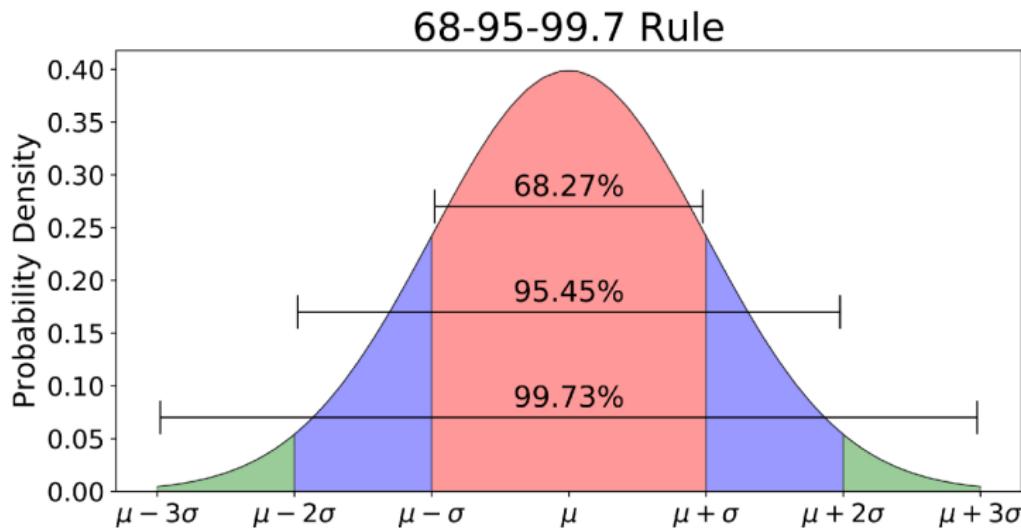
Асимптотический доверительный интервал

$$(\hat{\theta}_n - z_{\alpha/2} \hat{s}\hat{e}, \quad \hat{\theta}_n + z_{\alpha/2} \hat{s}\hat{e}).$$

Правило 3 сигм

$z_{0,025} = 1.96 \approx 2$, поэтому 95% данных лежит в интервале $\hat{\theta} \pm 2 \hat{s}_e$.

“Правило 3 сигм”: а 99% данных — в интервале $\hat{\theta} \pm 3 \hat{s}_e$.



Центральная предельная теорема.

Если $\mathbb{V}X_i = \sigma^2 < \infty$, то

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Дельта-метод

Дельта-метод. Пусть g дифференцируема в μ и $g'(\mu) \neq 0$, тогда

$$\sqrt{n}[X_n - \mu] \xrightarrow{D} \mathcal{N}(0, \sigma^2) \quad \Rightarrow \quad \sqrt{n}[g(X_n) - g(\mu)] \xrightarrow{D} \mathcal{N}(0, \sigma^2 \cdot [g'(\mu)]^2)$$

Дельта-метод

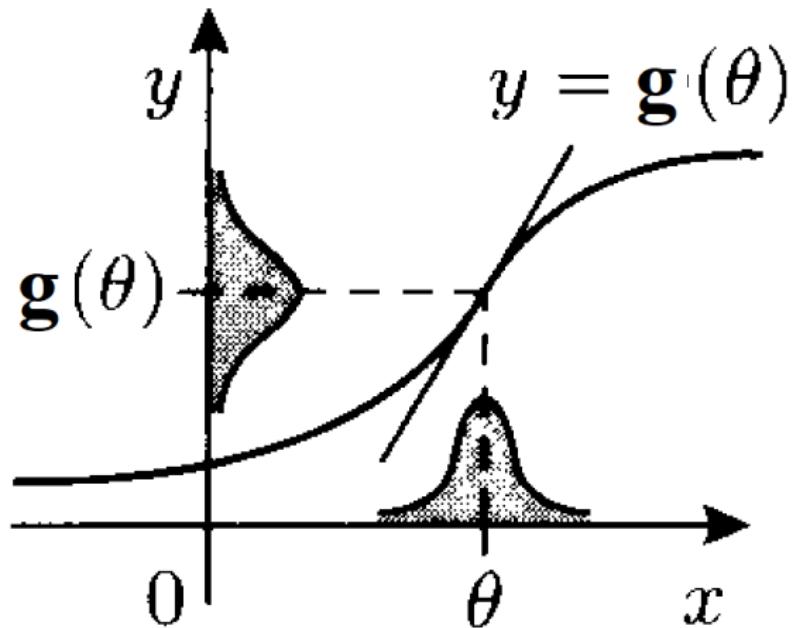


Рис.: Образ “нормальной шапочки” — “нормальная шапочка”

Дельта-метод

Доказательство — см. Лемму 1, Глава 7, §3,

 М. Б. Лагутин,
Наглядная математическая статистика.

По сути — разлагаем в ряд Тейлора и применяем свойства сходимости.

Дельта-метод

Дельта-метод (многомерный случай).

Пусть g дифференцируема.

Обозначим через ∇_μ градиент g в точке μ .

Если $\nabla_\mu \neq 0$, то

$$\sqrt{n}[X_n - \mu] \xrightarrow{D} \mathcal{N}(0, \Sigma) \quad \Rightarrow \quad \sqrt{n}[g(X_n) - g(\mu)] \xrightarrow{D} \mathcal{N}(0, \nabla_\mu^T \Sigma \nabla_\mu).$$

Бутстреп

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Бутстреп



Рис.: Bootstrap

Оценка дисперсии

$T = g(X_1, \dots, X_n)$ — статистика.

На практике мы часто хотим оценить разброс/дисперсию $\mathbb{V}_F(T)$.

Оценка дисперсии

$T = g(X_1, \dots, X_n)$ — статистика.

На практике мы часто хотим оценить разброс/дисперсию $\mathbb{V}_F(T)$.

Есть проблемы:

- ① Мы не знаем распределение на T .
- ② Мы не знаем F — распределение на X_i .

Оценка распределения

Более простой вопрос:

Q: Чем заменить распределение F ?

Оценка распределения

Более простой вопрос:

Q: Чем заменить распределение F ?

A: Эмпирической функцией распределения \hat{F} .

Оценка матожидания функции

Осталось понять — как считать дисперсию $\mathbb{V}_{\hat{F}}(T)$?

Более общий **вопрос**: как считать дисперсию

$$\mathbb{V}(Y) = \mathbb{E}(Y - \bar{Y})^2?$$

Оценка матожидания функции

Осталось понять — как считать дисперсию $\mathbb{V}_{\hat{F}}(T)$?

Более общий **вопрос**: как считать дисперсию

$$\mathbb{V}(Y) = \mathbb{E}(Y - \bar{Y})^2?$$

Дисперсия — матожидание от конкретной функции.

Ещё более общий **вопрос**: как вычислить матожидание

$$\mathbb{E}_G(h(Y)) = \int_{-\infty}^{\infty} h(y)dG(y)?$$

Метод Монте-Карло

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Метод Монте-Карло

Важная идея — [метод Монте-Карло](#).

Монте-Карло знаменито своими казино.



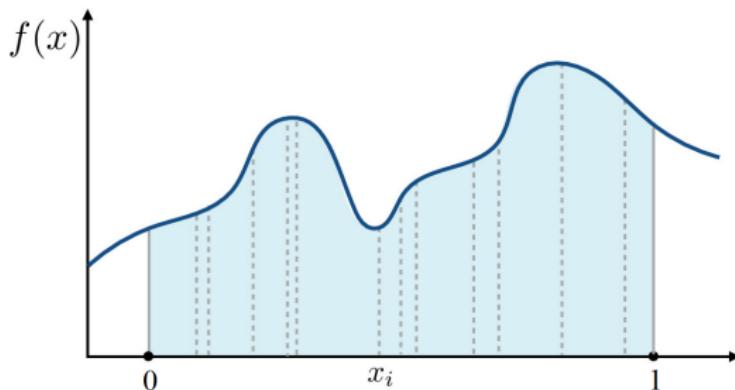
[Рис.](#) Рулетка — генератор случайных чисел

Метод Монте-Карло

Простая идея. Как посчитать $\int_0^1 f(x)dx$?

Интеграл = площадь под графиком = среднее значение $f(x)$ на отрезке $[0,1]$.

Оценка интеграла — берём $f(x_i)$ в случайных точках и усредняем.



Метод Монте-Карло

Как вычислить матожидание

$$\mathbb{E}_G(h(Y)) = \int_{-\infty}^{\infty} h(y) dG(y)?$$

Пусть мы умеем сэмплировать из распределения G .

Метод Монте-Карло:

- Берём Y_1, \dots, Y_B — i.i.d. из распределения G .
- Оценка матожидания: берём значения в этих точках и усредняем.

По (У)ЗБЧ

$$\frac{1}{B} \sum_{j=1}^B h(Y_j) \xrightarrow{a.s.} \mathbb{E}(h(Y_1)) = \int h(y) dG(y).$$

Дисперсия. Монте-Карло

В интересующем нас случае:

$\mathbb{V}(Y)$ можно аппроксимировать выборочной дисперсией $\frac{1}{B} \sum_{j=1}^B (Y_j - \bar{Y})^2$.

Метод Монте-Карло

Сделаем пару замечаний о методе Монте-Карло. Подробнее — см. Глава 3



М. Б. Лагутин,

Наглядная математическая статистика.

Q: Плюсы и минусы метода Монте-Карло?

Минусы МС

Недостатки метода Монте-Карло:

- Это вероятностный метод.
- Сходимость порядка² $\frac{1}{\sqrt{n}}$.
- Для оценки сходимости нужно считать дисперсию.
- Проклятия размерности.

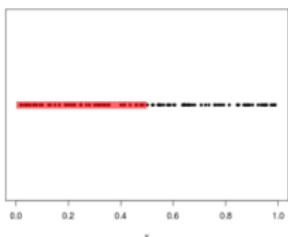
²Типичный порядок сходимости в статистике

Проклятие размерности

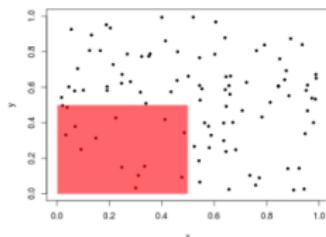
Проклятие размерности

Пусть данные в кубе $[0, 1]^D$. Какая доля данных попадёт в $[0, \frac{1}{2}]^D$?

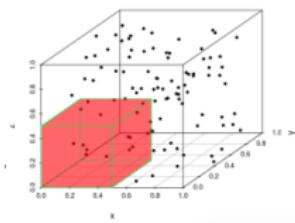
1-D: 42% попало



2-D: 14% попало



3-D: 7% попало



4-D: 3% попало

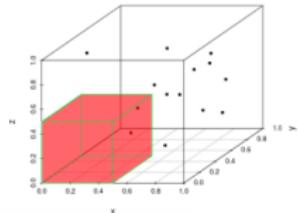


Рис.: Чем больше размерность, тем более разреженные данные.

Плюсы МС

Достоинства метода Монте-Карло:

- Применим и к многомерным интегралам.
- Универсальная и далеко идущая идея.

Схема Бутстрепа

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Идея бутстрепа

Идея бутстрепа:

- ① Оценить $\mathbb{V}_F(T)$ при помощи $\mathbb{V}_{\hat{F}_n}(T)$.
- ② Апроксимируем $\mathbb{V}_{\hat{F}_n}(T)$, используя моделирование Монте-Карло.

Замечание. Если мы можем явно посчитать $\mathbb{V}_{\hat{F}_n}(T)$, 2ой шаг проводить не надо.

Собираем воедино

Остаётся собрать всё воедино.



Схема бутстрепа

Оценка дисперсии с помощью бутстрепа.

- ① Сэмплируем $X_1^*, \dots, X_n^* \sim \hat{F}_n$.
- ② Вычисляем $T_n^* = g(X_1^*, \dots, X_n^*)$.
- ③ Повторяем шаги 1 и 2 B раз, получаем $T_{n,1}^*, \dots, T_{n,B}^*$.
- ④ Оценка дисперсии

$$v_{boot} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{j=1}^B T_{n,j}^* \right)^2.$$

Выборки

Контрольный вопрос. Чем такое сэмплировать из \hat{F}_n ?

Выборки

Контрольный вопрос. Что такое сэмплировать из \hat{F}_n ?

А: Равновероятно взять одно из X_1, \dots, X_n .

Бутстреп \rightarrow выборка с возвращением.

Аппроксимации

Аппроксимации в бутстрепе:

$$V_F(T_n) \overset{O(1/\sqrt{n})}{\approx} V_{\hat{F}_n}(T_n) \overset{O(1/\sqrt{B})}{\approx} v_{boot}.$$

Перерыв

Перерыв

Доверительные интервалы

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Нормальный интервал

Обсудим 3 способа построить доверительные интервалы с помощью бутстрепа.

Обозначим $\theta = T(F)$ — то, что предсказываем. $\hat{\theta}_n = T(\hat{F}_n)$ — наша оценка.

① Нормальный интервал.

$$\hat{\theta}_n \pm z_{\alpha/2} \hat{s}_{\text{boot}},$$

где $\hat{s}_{\text{boot}} = \sqrt{v_{\text{boot}}}$ — оценка дисперсии при помощи бутстрепа.

Не очень эффективная оценка, если данные не нормально распределены.

Терпение и труд всё перетрут

В следующих 2 способах будет немного формул.



Центральный интервал

② Центральный интервал.

Положим $R_n = \hat{\theta}_n - \theta$;

$H(r)$ — распределение величины R_n , т.е.

$$H(r) = \mathbb{P}(R_n \leq r).$$

Центральный интервал

Покажем, что **точный** $(1 - \alpha)$ **доверительный интервал** суть $C_n^* = (a, b)$, где

$$a = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right), \quad b = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right).$$

Доказательство:

Центральный интервал

Покажем, что **точный** $(1 - \alpha)$ доверительный интервал суть $C_n^* = (a, b)$, где

$$a = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right), \quad b = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right).$$

Доказательство:

$$\begin{aligned} \mathbb{P}(a \leq \theta \leq b) &= \mathbb{P}(a - \hat{\theta}_n \leq \theta - \hat{\theta}_n \leq b - \hat{\theta}_n) \\ &= \mathbb{P}(\hat{\theta}_n - b \leq \hat{\theta}_n - \theta \leq \hat{\theta}_n - a) \\ &= \mathbb{P}(\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a) \\ &= H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\ &= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \end{aligned}$$

Центральный интервал

Q: Мы не знаем распределение H . Что делать?

Центральный интервал

Q: Мы не знаем распределение H . Что делать?

A: Правильно, используем бутстреп .

$\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*$ — повторная выборка $\hat{\theta}_n = T(\hat{F}_n)$ на основе бутстрапа.

Полагаем $\hat{R}_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$ и строим эмпирическую функцию распределения

$$\hat{H}(r) = \frac{1}{B} \sum_{b=1}^B I(\hat{R}_{n,b}^* \leq r)$$

Центральный интервал

Границы интервала:

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1} \left(1 - \frac{\alpha}{2} \right) = 2\hat{\theta}_n - \hat{\theta}_{1-\frac{\alpha}{2}}^*,$$

$$b = \hat{\theta}_n - \hat{H}^{-1} \left(\frac{\alpha}{2} \right) = 2\hat{\theta}_n - \hat{\theta}_{\frac{\alpha}{2}}^*.$$

Центральный интервал

Границы интервала:

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = 2\hat{\theta}_n - \hat{\theta}_{1-\frac{\alpha}{2}}^*,$$

$$b = \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = 2\hat{\theta}_n - \hat{\theta}_{\frac{\alpha}{2}}^*.$$

В итоге — **центральный доверительный интервал** суть

$$\left(2\hat{\theta}_n - \hat{\theta}_{1-\frac{\alpha}{2}}^*, \quad 2\hat{\theta}_n - \hat{\theta}_{\frac{\alpha}{2}}^*\right).$$

Это асимптотический $1 - \alpha$ доверительный интервал.

Интервал на основе процентиелей

② Интервал на основе процентиелей.

$$\left(\hat{\theta}_{\frac{\alpha}{2}}^*, \quad \hat{\theta}_{1-\frac{\alpha}{2}}^* \right).$$

Почему это доверительный интервал?

Интервал на основе процентиелей

Пусть **существует** монотонное преобразование $U = m(T)$, что $U \sim \mathcal{N}(\phi, c^2)$.

Монотонное преобразование сохраняет квантили, поэтому $m(\theta_{\alpha/2}^*) = u_{\alpha/2}^*$ — соотв. квантиль нормального распределения $\mathcal{N}(\phi, c^2)$, т.е. $\phi - z_{\alpha/2} c$.

Доказательство, что доверительный интервал:

Интервал на основе процентиелей

Пусть существует монотонное преобразование $U = m(T)$, что $U \sim \mathcal{N}(\phi, c^2)$.

Монотонное преобразование сохраняет квантили, поэтому $m(\theta_{\alpha/2}^*) = u_{\alpha/2}^*$ — соотв. квантиль нормального распределения $\mathcal{N}(\phi, c^2)$, т.е. $\phi - z_{\alpha/2} c$.

Доказательство, что доверительный интервал:

$$\begin{aligned}\mathbb{P}(\theta_{\alpha/2}^* \leq \theta \leq \theta_{1-\alpha/2}^*) &= \mathbb{P}(m(\theta_{\alpha/2}^*) \leq m(\theta) \leq m(\theta_{1-\alpha/2}^*)) \\ &= \mathbb{P}(u_{\alpha/2}^* \leq \phi \leq u_{1-\alpha/2}^*) \\ &= \mathbb{P}(U - cz_{\alpha/2} \leq \phi \leq U + cz_{\alpha/2}) \\ &= \mathbb{P}(-z_{\alpha/2} \leq \frac{U - \phi}{c} \leq z_{\alpha/2}) \\ &= 1 - \alpha.\end{aligned}$$

Интервал на основе процентиляй

Нам не нужно **знать** монотонное преобразование $U = m(T)$, достаточно его существования.

Более того, на практике достаточно **существования** его аппроксимации.

Литература

На практике описанные выше интервалы могут быть близки.

Более точные обобщения этих бутстрепных интервалов описаны в



Wasserman L.

All of Nonparametric Statistics.

Доверительные интервалы

Итак, 3 способа построить доверительные интервалы:

① Нормальный интервал

$$(\hat{\theta}_n + z_{\alpha/2} \hat{s}e_{\text{boot}}, \quad \hat{\theta}_n - z_{\alpha/2} \hat{s}e_{\text{boot}})$$

где $\hat{s}e_{\text{boot}} = \sqrt{v_{\text{boot}}}$.

② Центральный интервал

$$\left(2\hat{\theta}_n - \hat{\theta}_{1-\frac{\alpha}{2}}^*, \quad 2\hat{\theta}_n - \hat{\theta}_{\frac{\alpha}{2}}^* \right).$$

③ Интервал на основе процентиелей

$$\left(\hat{\theta}_{\frac{\alpha}{2}}^*, \quad \hat{\theta}_{1-\frac{\alpha}{2}}^* \right).$$

Self-Supervision

1 Доверительные интервалы

2 Асимптотическая нормальность и доверительные интервалы

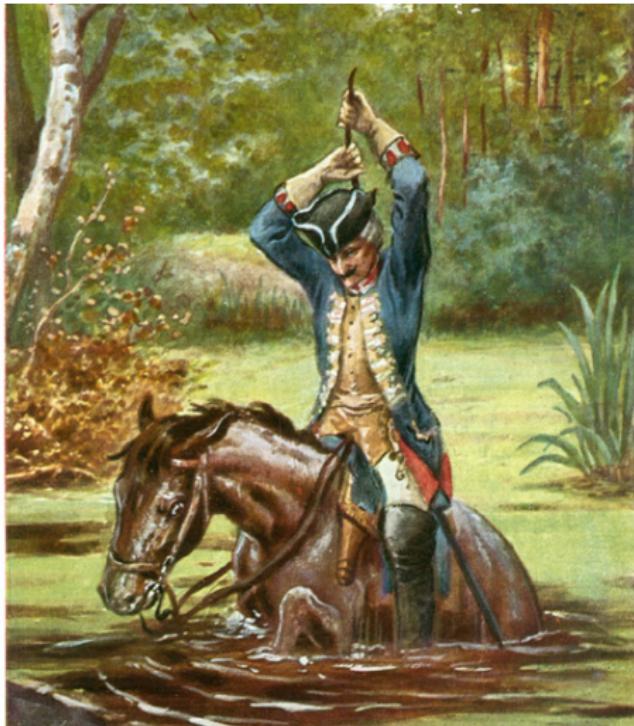
3 Бутстреп

- Метод Монте-Карло
- Схема Бутстрепа

4 Доверительные интервалы

5 Self-Supervision

Self-supervision



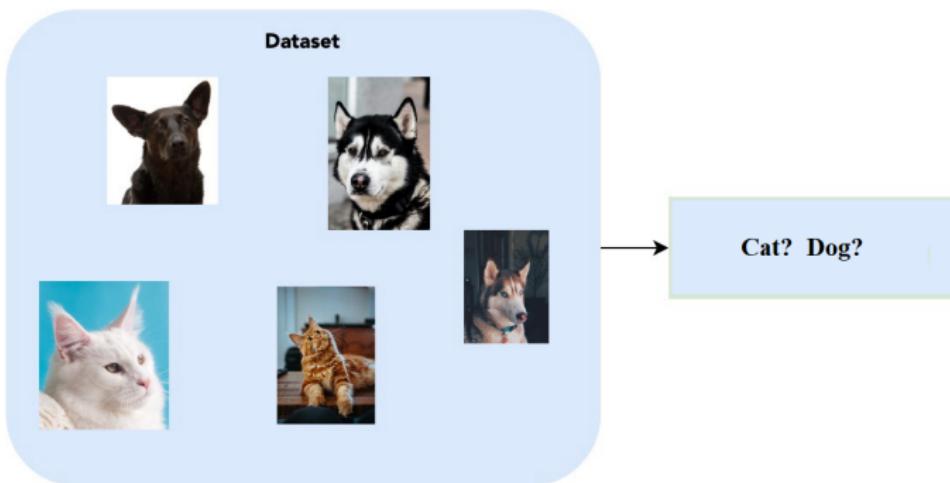
Self-supervision
~~Bootstrap~~

Self-supervision

Самообучение в ML (Self-supervision) идейно похоже на бутстреп.

Пусть у нас даны неразмеченные данные (картинки).

Откуда взять информацию для их классификации?

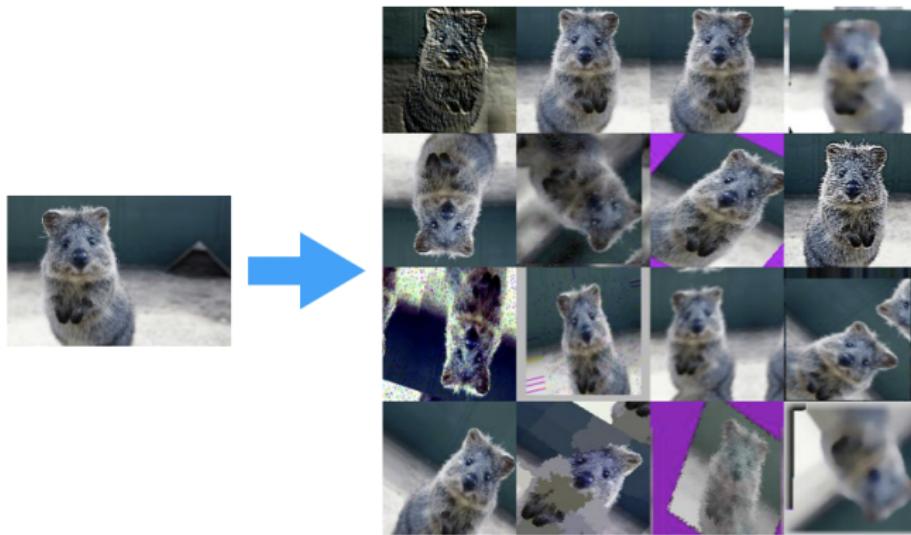


Меток нет!

Аугментация

Стандартная техника в DL для расширения датасета — [аугментация](#) данных.

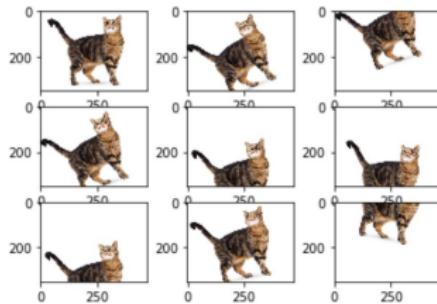
Картинки можно поворачивать, увеличивать, отражать, менять цвета и т.д.



Что спросить нейронку?

Итак, мы не знаем — котик или пёсик, но теперь у нас куча картинок с ним.

Q: Догадываетесь, что можно заставить выучить нейронку?

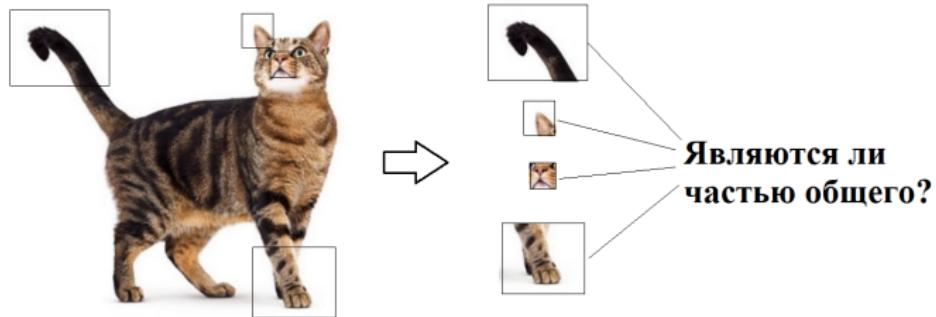


Идея самообучения

Как заставить нейронку понять, что такое кошка?

То есть — как построить функцию, принимающую одно значение на всех кошках?

Ключевая идея! Можно спросить — являются ли две аугментации производными одной и той же картинки?



Эквивариантность

- По сути мы строим инвариантные³ отображения.
- Если отображение F т., ч.

$$F(\text{хвост кошки}) = F(\text{ухо кошки}) = F(\text{лапы кошки}),$$

то отображение F зависит от общего, т.е. от “кошки”.

³ в смысле одинаковые на классах эквивалентности

Философский взгляд на вещи

"Вся европейская философия на самом деле - ряд примечаний к Платону". ☺

Платоновский "мир идей"



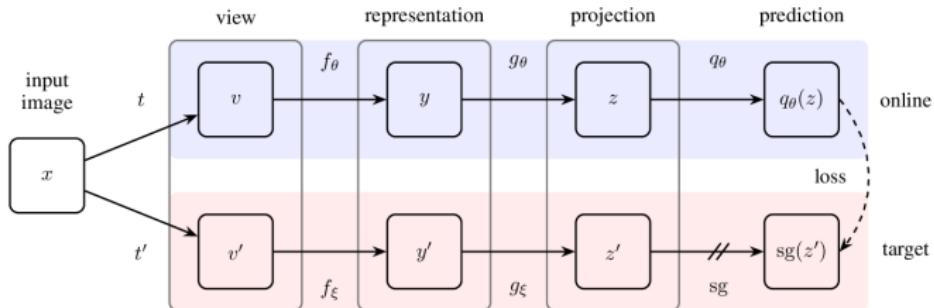
Что делает котика котиком? Идея "кошачности" - она есть в каждом котике.

Естественно, в ML и Self-Supervision полно разных идей.

Одна из известных статей про Self-supervision называется

Bootstrap your own latent: A new approach to self-supervised Learning

BYOL



To be continued

Неужели сэмплировать числа — предел наших генеративных способностей?

Обсудим генеративные модели **на 12 лекции**.



← To Be Continued →