Лекция 6.

БББББ
(Бутстрепинг, бутстреп, бэггинг и бустинг).

Павел Владимирович
Слипенчук
PavelMSTU@stego.su
ИУ-8

Москва
14.10.2016

Бутстрепинг (статистика)

- Генерация выборки размера **N** из подвыборки размера **n<<N** путем выбора с повторением.
- Сколько вариантов выбора?

Бутстрепинг (статистика)

- Генерация выборки размера **N** из подвыборки размера **n<<N** путем выбора с повторением.
- Вариантов выбора: $\overline{C_N^n} = C_{N+n-1}^n$
- В бутстреппинге выборка выбирается большое количество раз (например 1000, 10000)
- На каждой выборке высчитывается мат.ожидание, дисперсия.
- Для **N** мат.ожидание и дисперсия принимаются как среднее арифметическое этих величин на совокупности выборок.

Бутстрепинг (статистика)

- **В каких случаях бутстрепинг применим, а когда нет?**

Бутстрепинг (статистика)

- **Центральная Предельная Теорема.**

$x_1, \ldots, x_n$ – **n** случайных велчин, ***независимых*** и ***одинакого Распределенных***, которые ***имеют*** конечное мат.ожидание и дисперсию.

μ, σ – математическое ожидание и дисперсия

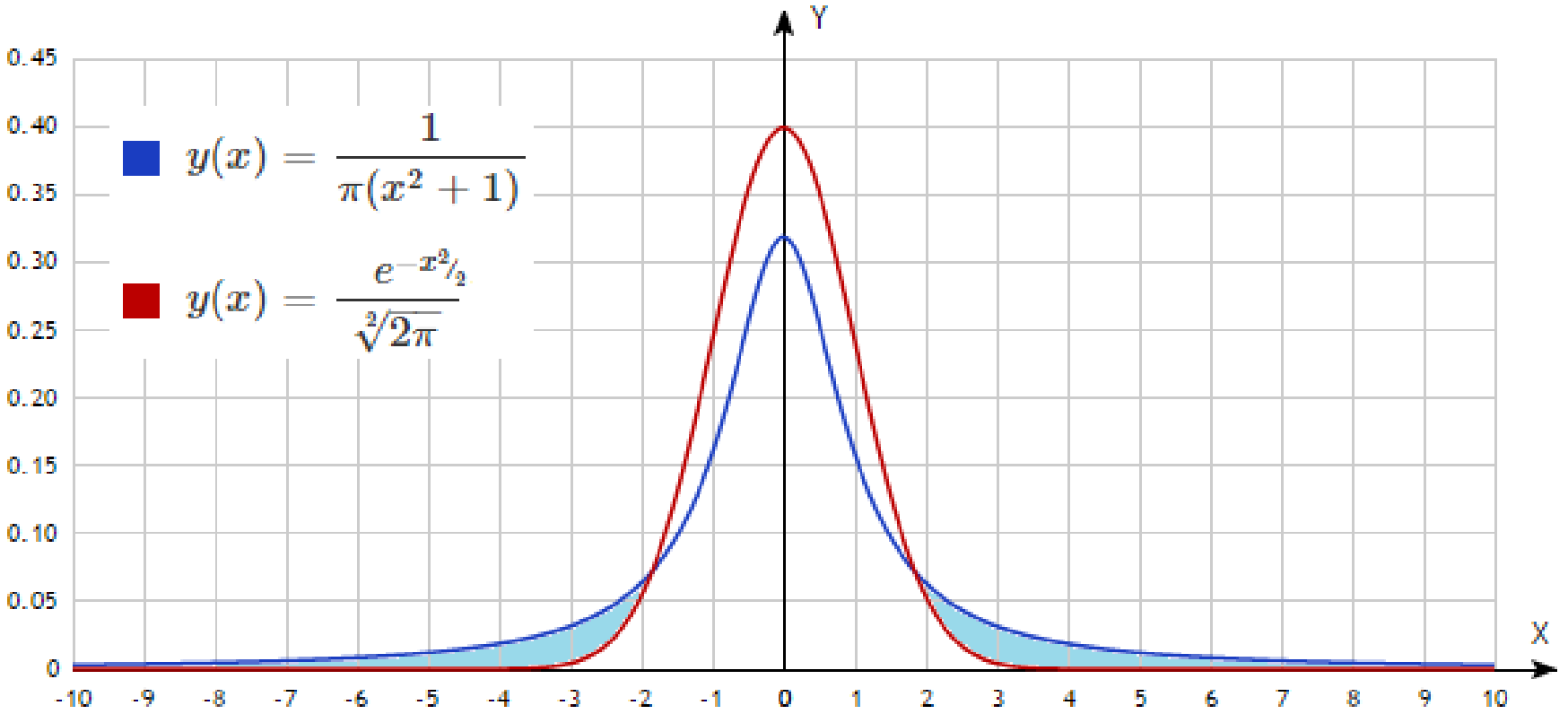
$X_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} x_i$ – сумма n величин

$$\lim_{n \to \infty} \frac{X_n - \mu \cdot n}{\sigma \cdot \sqrt{n}} = N(0,1)$$

- Если не конечное мат.ожидание или дисперсия – бутстрепинг не поможет.

Бутстрепинг (статистика)

- **Пример:** распределение Коши.



- это называют «проблемой тяжелых хвостов»

См статью «Тяжелые хвосты» блог А.Дьяконова Анализ Малых Данных. 2016.03.04

Бутсрепинг (Machine Learning)

- **Идеи?**

Бутсрепинг (Machine Learning)

- Можно использовать бутстреп для генерации нескольких выборок из одной определенной выборки.
- **Зачем**?

Бутсрепинг (Machine Learning)

- Можно использовать бутстреп для генерации нескольких выборок из одной определенной выборки.
- **Примеры.**
  1) Создание различных выборок для классификаторов ансамбля (например для деревьев случайного леса) – **бутстреп (bootstrap)**.
  2) Проверка качества модели (создается несколько тестовых выборок)
  3) Требуется **n>>N**. Например для решения проблемы разномощности классов
  (*см. след.слайд*)

Разномощные классы
(~ «несбалансированные классы»)

- Типичная ситуация: количество элементов в одном классе существенно меньше количества элементов в другом классе.
- **Примеры**:
  1) фрода существенно меньше, чем легитимных транзакций
  2) больных раком меньше, чем здоровых
  3) порядочных водителей больше, чем «лихачей»
  4) ….
- **Методы решения?**

Разномощные классы

- **Методы решения**

1) Из большего класса взять случайным образом одну подвыборку размерности меньшего класса («классика жанра»)

2) Продолжать наблюдения и получать новые события, определять их классы. Затем п.1

3) **Бутстрепинг** для меньшего класса (n>>N).

4) «экспертная генерация выборки»

Бутстреп-агрегация или **бэггинг** (**Bootstrap aggregating** or **bagging**)

- По сути это ансамбль над классификатором.
- Бутстреп-агрегация – это обобщение одгого и того же классификатора на различных подвыборках или случайных параметрах, которые различны при каждой новой подгонки (fitting).
- Если итоговый классификатор – случайный лес, то беггинг – это усреднение этого леса при **m>1** подгонках.

Бустинг (boosting)

- **«Общее определение»**
  Бустинг – это итеративное перевзвешивание наблюдений обучающей выборки. Т.е. это «последовательный» ансамбль, каждый следующий член которого корректирует предыдущий.
- + можно добавлять веса каждому ансамблю в соответствии с ошибками которые он допустил

Бустинг (boosting)

- **Простой вариант**.
  Обучающая выборка подается на вход классификатору №1. Из контрольной выборки выбираются объекты для которых классификатор №1 ошибся. Это – обучающая выборка для классификатора №2. … …

Бустинг (boosting)

ДЗ. Посмотреть.

- Градиент, градиентный спуск
- AnyBoost
- AdaBoost
- BrownBoost

Резюме.
Who is who?

- **Бутстрепинг**
- **Бутстреп**
- **Бутстреп аггрегация**
- **Бэггинг**
- **Бустинг**