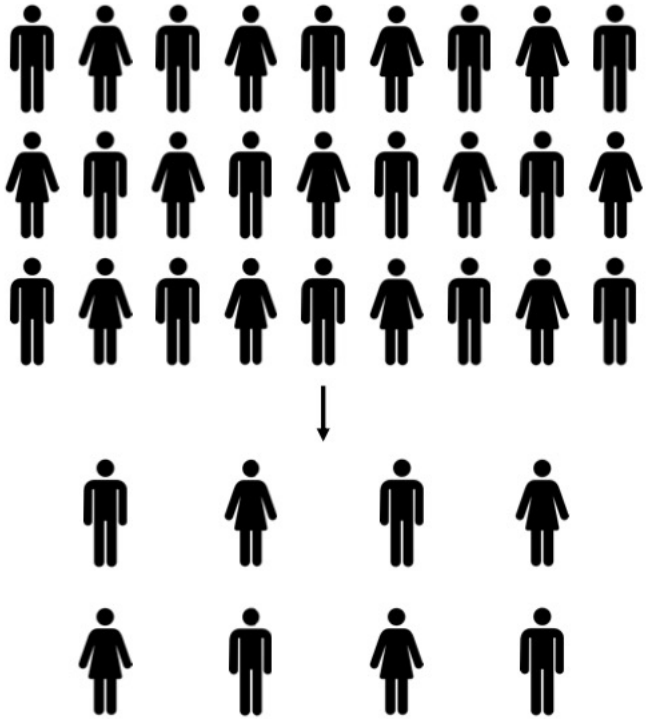
Методы семплирования

Генеральная совокупность и выборка

- **Генеральная совокупность** (population) – это совокупность всех объектов, обладающих общими признаками и относительно которых нам хочется делать какие-либо выводы при анализе некоторой конкретной задачи.
  Признаки, по которым мы разделяем объекты, могут быть абсолютно любыми: территориальное положение, возраст, уровень доходов и многие-многие другие.
- **Выборочная совокупность** или **выборка** (sample) – та часть генеральной совокупности, которую мы отбираем в рамках эксперимента и на основе которой мы будем описывать (характеризовать) генеральную совокупность.



- Некоторые объекты могут быть одновременно генеральной совокупностью и выборкой: например, все студенты ФПМИ.

Зачем нужно семплирование?

- Зачастую (практически всегда!) получить информацию и собрать данные о каждом или даже почти каждом элементе генеральной совокупности не представляется возможным. В связи с этим возникает необходимость в формировании выборки, то есть некоторой репрезентативной подгруппы генеральной совокупности.
- Такой процесс формирования выборки называется отбором или **семплированием** (sampling).

Методы семплирования

<table>
<tr><td colspan="2">Вероятностные методы
•каждый элемент г.с. имеет шанс быть выбранным
•возможность добиться репрезентативной выборки
•большое количество тонкостей реализации процедуры отбора</td><td colspan="2">Невероятностные методы
• неслучайный критерий отбора элементов генеральной совокупности
• простота процедуры отбора
• высокий риск возникновения выборочного смещения (sample bias)</td></tr>
<tr><td>Случайный отбор
(simple random sample)</td><td>Систематический отбор
(systematic sample)</td><td>Convenience sample</td><td>Voluntary response sample</td></tr>
<tr><td>Стратифицированный отбор
(stratified sample)</td><td>Кластерный отбор
(cluster sample)</td><td>Purposive sample</td><td>Snowball sample</td></tr>
</table>

Случайный отбор

- Каждый элемент имеет равные шансы быть отобранным
- Суть метода: каждый из $\boldsymbol{N}$ элементов генеральной совокупности нумеруется, и случайным образом отбираются $\boldsymbol{n}$ элементов из них

Пример

На заводе работает 500 сотрудников, нужно получить выборку из 5 сотрудников при помощи случайного отбора.

1. Нумеруем всех сотрудников от 1 до 500
2. Запускаем датчик случайных чисел 5 раз

Систематический отбор

- Каждый элемент имеет шанс быть отобранным
- Суть метода: каждый из $\boldsymbol{N}$ элементов генеральной совокупности нумеруется, вся г. с. делится на $\boldsymbol{n}$ равных (по возможности) интервалов. Из первого интервала случайно отбирается один элемент и с шагом, равным длине интервала, отбираются остальные элементы.

Пример

Рассмотрим тот же завод с 500 сотрудниками, но проведем систематический отбор.

1. Делим всех сотрудников на интервалы по 100 человек.
2. При помощи датчика случайных чисел выбираем число от 1 до 100 (например, получилось 81)
3. Далее с шагом 100 отбираем 5 сотрудников (81, 181, 281, 381, 481)

Стратифицированный отбор

- Каждый элемент имеет шанс быть отобранным
- Суть метода: генеральная совокупность разделяется на несколько подгрупп или **страт** (strata) в соответствии с каким-либо признаком (гендер, возраст или любой другой признак), а затем производится случайный отбор в каждой страте пропорционально доле данной страты от всей генеральной совокупности

Пример

Пусть известно, что из 500 сотрудников завода 300 мужчин и 200 женщин. Проведем стратифицированный отбор 5 человек.

1. Доля страты мужчин составляет 300/500 = 0,6 => отбираем 0,6*500 = 3 мужчины.
2. Доля страты женщин составляет 200/500 = 0,4 => отбираем 0,4*500 = 2 женщины.

Кластерный отбор

- Каждый элемент имеет шанс быть отобранным
- Суть метода: г.с. из N элементов разделяется на l кластеров таким образом, чтобы в каждом кластере было поровну (по возможности) элементов; случайным образом выбирается один кластер и из него при помощи случайного, систематического или стратифицированного отбора отбирается n элементов
- **Важно!** Кластеры должны быть организованы таким образом, чтобы элементы различных кластеров практически не различались своими характеристиками

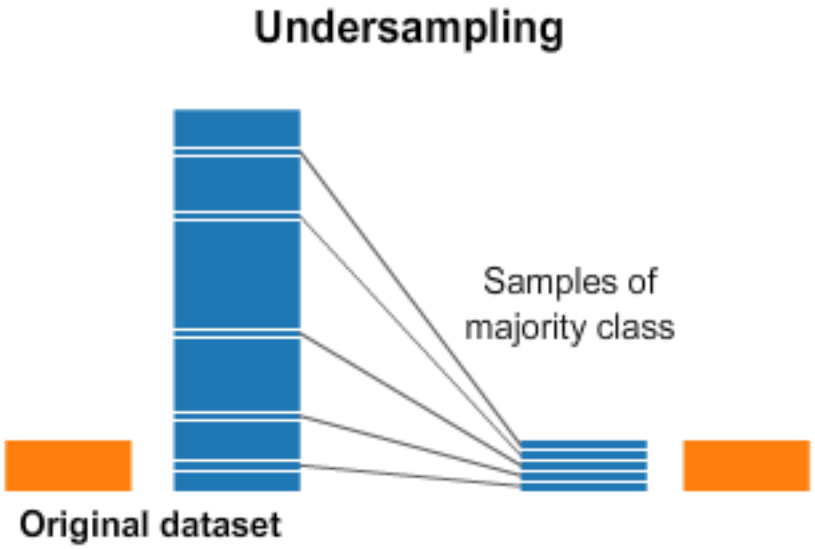
Пример

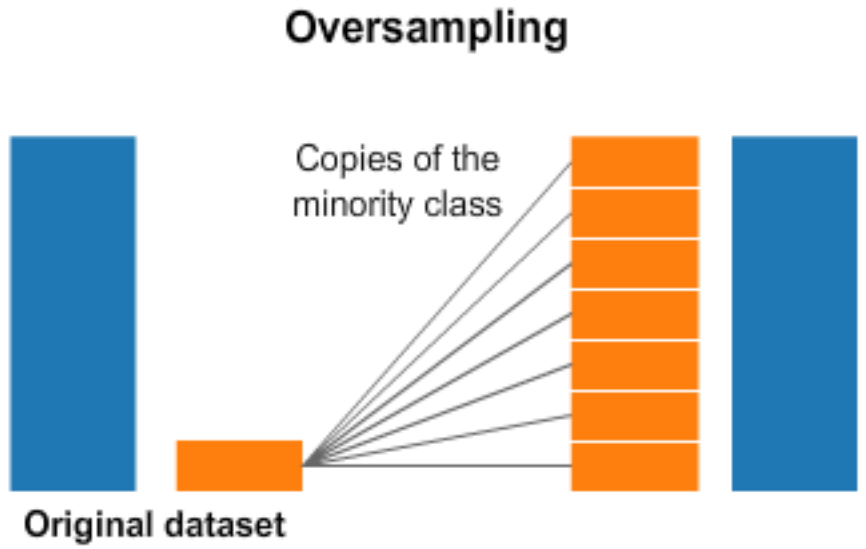
Уже знакомый нам завод разделили на три корпуса: A, B и C. Численность сотрудников в каждом из них составила 200, 150 и 150 человек, соответственно. Известно, что каждый из корпусов занимается схожими задачами, а сотрудники в них набирались случайным образом.

- предположим, датчик случайных чисел выбрал корпус C
- далее уже по одной из трех известных схем выбираем 5 сотрудников из корпуса C

Ресемплинг

- При работе с несбалансированными наборами данных обычно используется ресемплинг (передискретизация)
- Есть два вида ресемплинга: андерсемплинг - удаление элементов из слишком большого набора; оверсемплинг – добавление элементов в недостаточно большой набор.





Стратегии андерсемплинга

- Случайное удаление объектов большего класса
- Поиск связей Томека (Tomek Links)
- Правило сосредоточенного ближайшего соседа (Condensed Nearest Neighbor Rule)
- Односторонний семплинг (One-side sampling, one-sided selection)
- Правило «очищающего» соседа (Neighborhood cleaning rule)

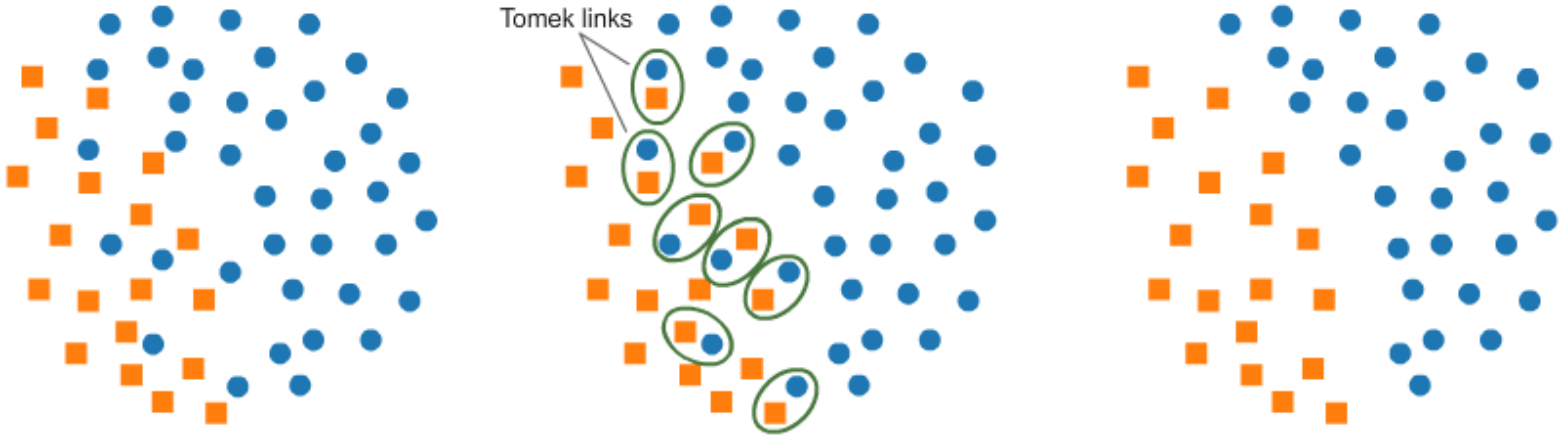
Подробнее по ссылке

Поиск связей Томека

- Пусть объекты E_i и E_j принадлежат к различным классам, $d(E_i, E_j)$ – расстояние между указанными примерами. Пара (E_i, E_j) называется связью Томека, если не найдется ни одного примера E_l, такого, что будет справедлива совокупность неравенств:

$$\begin{cases} d(E_i, E_l) < d(E_i, E_j), \\ d(E_j, E_l) < d(E_i, E_j) \end{cases}$$

- Согласно данному подходу, все мажоритарные записи, входящие в связи Томека, должны быть удалены из набора данных. Этот способ хорошо удаляет записи, которые можно рассматривать в качестве «зашумляющих».



Стратегии оверсемплинга

- Дублирование примеров миноритарного класса
- SMOTE (Synthetic Minority Oversampling Technique)
- ASMO (Adaptive Synthetic Minority Oversampling)

Подробнее по ссылке

SMOTE

- Этот алгоритм основан на идее генерации некоторого количества искусственных примеров, которые были бы похожи на имеющиеся в миноритарном классе, но при этом не дублировали их. Для создания новой записи находят разность $d=X_b–X_a$, где X_a,X_b – векторы признаков «соседних» объектов a и b из миноритарного класса. Их находят, используя алгоритм ближайшего соседа KNN.