

# Репрезентативность (Representativity)

Синонимы: Представительность

Loginom: Разбиение на множества (обработчик), Сэмплинг (обработчик).

Под репрезентативностью в статистике и машинном обучении понимается соответствие структурных характеристик выборки характеристикам генеральной совокупности, из которой она извлечена. Репрезентативность определяет, насколько возможно обобщать результаты исследования, полученные на основе выборочных данных на всю исходную совокупность.

Иными словами, репрезентативная выборка — это такая выборка, в которой представлены все группы, важные для исследования. Также репрезентативность можно определить как свойство выборочной совокупности представлять параметры генеральной совокупности, значимые с точки зрения задач исследования.

Отсутствие репрезентативности выборки приводит к некорректным результатам анализа. Поэтому перед использованием выборки необходимо убедиться, что она репрезентативна.

Обеспечение репрезентативности выборки достигается правильным выбором используемого алгоритма сэмплинга. При этом размер выборки, хотя и является важным, сам по себе не гарантирует ее репрезентативности.

Выделяют качественную и количественную репрезентативность. Качественная репрезентативность предполагает, что выборка содержит представителей всех групп объектов совокупности и в соответствующей пропорции. Например, если совокупность содержит мужчин и женщин в пропорции 2/3, то и выборка, чтобы быть репрезентативной, должна содержать представителей обоих полов в отношении 2/3.

Количественная репрезентативность предполагает достаточное количество наблюдений выборки. Т.е. выборка может быть качественно репрезентативной, но при этом содержать слишком мало наблюдений для построения эффективной модели. Особенно в этом случае страдает обобщающая способность модели. Например, чтобы нейронная сеть обладала обобщающей способностью, необходимо, что число примеров в обучающей выборке было в 2-3 раза больше числа межнейронных связей. При недостаточном объеме выборки это условие может не соблюдаться.

Если в распоряжении аналитика оказалась нерепрезентативная выборка и при этом исходная совокупность для повторной выборки недоступна или повторное выборочное исследование невозможно (из-за высоких затрат или невозможности воспроизвести

условия, при которых выборка была получена), то аналитик может произвести так называемый ремонт выборки — комплекс мер повышения репрезентативности.