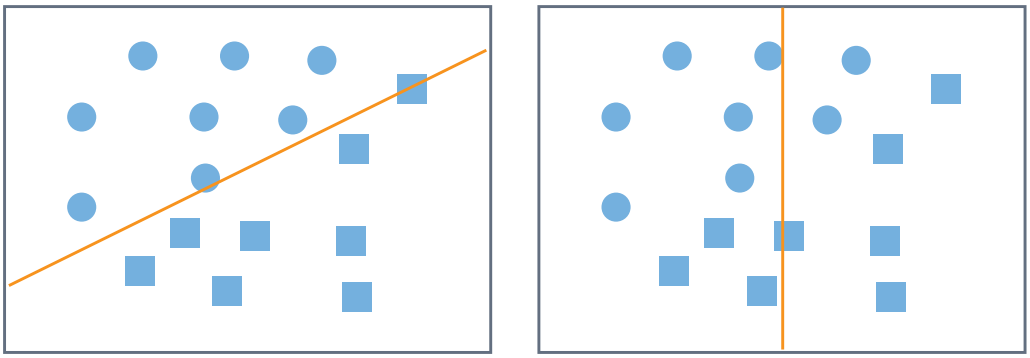
# Чистота множества (Purity)

В задачах классификации чистота — это показатель, характеризующий результаты разбиения множества данных на подмножества, ассоциированные с классами в классификационных моделях, таких как машины опорных векторов, алгоритмы покрытия, но чаще всего в деревьях решений.

Чем более однородным по классовому составу объектов является подмножество, тем выше его чистота, и наоборот. Чистота может служить критерием оптимальности разбиения: лучшим будет то, которое обеспечит наиболее чистые подмножества (в деревьях решений — **узлы** и **листья**).

Чистота подмножества может характеризоваться процентом «примеси» объектов других классов. Например, если в узле дерева решений содержится 50 примеров класса, ассоциированного с узлом, и 10 примеров других классов, то примесь будет 20%, а чистота — 80%. Если процент примеси нулевой, то подмножество называют **чистым**. Если в дереве решений в результате разбиения будет получен чистый узел, то разбиения прекращаются, и он объявляется **листом**.



На рисунке слева представлено линейное разделение двух классов объектов, в результате которого были получены чистые подмножества. На рисунке справа показано разбиение, породившее подмножества с примесью.

Поскольку слева от линии разделения больше кружков, можно ассоциировать данное подмножество с этим классом. Подмножество содержит 6 кругов и 3 квадрата, получим, что из 9 объектов 3 представляют примесь, т.е. 33,3%. Подмножество справа от линии также не является чистым. Оно ассоциировано с классом квадратов, но содержит примесь из 29% кружков.

Таким образом, задачу выбора оптимального разделения классов можно рассматривать как максимизацию чистоты или минимизацию примеси подмножеств.