

Ансамбль моделей (Ensemble of models)

Синонимы: Комитет моделей, Ensemble learning

В статистике и машинном обучении под ансамблем моделей понимают комбинацию нескольких алгоритмов обучения, которые, работая вместе, позволяют построить модель более эффективную и точную, чем любая из моделей, построенная с помощью отдельного алгоритма. Модель, построенную на основе ансамбля, часто называют «метамоделью».

Алгоритмы обучения с учителем можно рассматривать как выполняющие задачу поиска в пространстве гипотез с целью найти ту из них, которая будет соответствовать решению конкретной задачи. Даже если такая гипотеза существует, то найти ее достаточно сложно. Ансамбли объединяют несколько гипотез, чтобы сформировать из них лучшую.

Ансамбль моделей требует больше вычислительных затрат, чем отдельная модель, поэтому работу ансамбля можно рассматривать как способ компенсации «плохих» алгоритмов обучения путем дополнительных вычислений.

Что касается количества моделей, используемых в ансамбле, которое позволяет получить наиболее точные результаты, то четких критериев здесь пока не существует. Однако при построении ансамблей классификаторов во многих случаях наилучшим оказывается число моделей, равное числу классов.

Существует несколько методов объединения моделей в ансамбли:

- **Оптимальный байесовский классификатор** — ансамбль, состоящий из простых классификаторов Байеса, взвешенных их апостериорными вероятностями.
- **Бэггинг** — ансамбль моделей, обучающихся параллельно, на различных случайных выборках одного и того же обучающего множества. Определение конечного результата определяется голосованием классификаторов ансамбля — выбирается класс, который предсказало большинство классификаторов.
- **Бустинг** — ансамбль моделей, обучающихся последовательно, при этом каждый последующий алгоритм обучается на примерах, в которых предыдущий классификатор допустил ошибку. Считается, что бустинг дает более точные результаты, чем бэггинг, но при этом склонен к переобучению.
- **Стекинг** — обучающее множество разбивается на N блоков, и на $N - 1$ из них обучается столько же базовых моделей. Затем N -я модель обучается на оставшемся блоке, но в качестве целевой переменной используются выходы базовых классификаторов, образующие так называемый метапризнак. Недостатком стекинга является то, что он начинает значительно улучшать результаты базовых

классификаторов при относительно большом числе обучающих примеров (несколько десятков тысяч).

Приведенные выше методы построения ансамблей могут работать с различными видами моделей — регрессией, нейронными сетями, деревьями решений, алгоритмами кластеризации. Однако есть ансамбли, которые разработаны специально для моделей одного типа. Например, случайный лес специально предназначен для ансамблей из деревьев решений.