

Для чего может добавляться шум к выходным меткам?

Добавление шума в принципе можно использовать в случаях небольших датасетов. Обучение на маленькой выборке может привести к переобучению, добавление шумов позволяет “увеличить” в некотором смысле объем тренировочных данных и построить достаточно общую модель.

Какие методы увеличения данных Вы знаете?

В случае классификации изображений можно применять следующую хитрость: если у нас имелось, допустим, фото с собакой, то можно добавить к датасету то же самое изображение, только повернутое на некоторый угол или отраженное с той же самой меткой. Можно масштабировать изображение или применить изменение насыщенности цветов. CNN в принципе и так неплохо справляются с обнаружением одного объекта в разных позициях, но такая методика иногда тоже может быть полезна.

В категориальном анализе рассматривают, например, таблицы сопряженности. Если в какой-то из ячеек число наблюдений менее 5, то проводить, например, хи-квадрат тест не рекомендовано. В R в таких ситуациях можно использовать симулирование данных, например, методами Монте-Карло.

Почему были выбраны именно такие модели?

Первая причина достаточно очевидна – нам их заботливо подсказали. Во-вторых, модели являются достаточно разными по конфигурации для объединения их в ансамбль. Изменение параметров каждой модели либо ухудшало, либо незначительно изменяло итоговую точность. Поскольку не существует правила, позволяющего определить идеальные параметры модели, то текущая конфигурация была принята как оптимальная, так как справлялась с поставленной задачей.