

1. Для чего и в каких случаях полезны различные варианты усреднения для метрик качества классификации: micro, macro, weighted?

Усреднение для метрик качества классификации, как правило, используется в многомерной классификации.

1. Macro – Усреднение вычисляет метрики для каждого класса и находит их невзвешенное среднее. Всем классам, независимо от их размера, присваивается одинаковый вес.
2. Micro — Усреднение вычисляет общее количество ложно положительных примеров, ложно отрицательных примеров и истинно положительных примеров по всем классам, а затем вычисляет метрику с помощью этих показателей
3. Weighted - усреднение вычисляет метрику для каждого класса и находит их среднее, взвешенное по поддержке (количеству фактических примеров для каждого класса).

Если необходимо присвоить одинаковый вес каждому примеру, рекомендуется использовать микро-усреднение, если необходимо присвоить одинаковый вес каждому классу, рекомендуется использовать макро-усреднение

2. В чём разница между моделями xgboost, lightgbm и catboost или какие их основные особенности?

XGBoost:

- строит дерево по уровню
- перебирает все данные в фичах и строит дерево по наибольшему приросту

LGBM

- строит дерево по листьям
- использует алгоритм GOSS, который выбирает данные с наибольшим градиентом ошибки и дополняет их небольшой подвыборкой из данных с маленьким градиентом.
- использует метод смещения фич

CatBoost

- разделяет дерево по одной переменной
- обучает несколько разных моделей на разных данных, в то время, как XGBoost и LGBM строит модели на уже используемых данных
- перед построением дерева данные сильно перемешиваются
- умеет работать с категориальными признаками.