Обнаружение аномалий в результате анализа

Лекция 9

Пищевая экспертиза

Метод k-ближайших соседей (k-Nearest Neighbors)

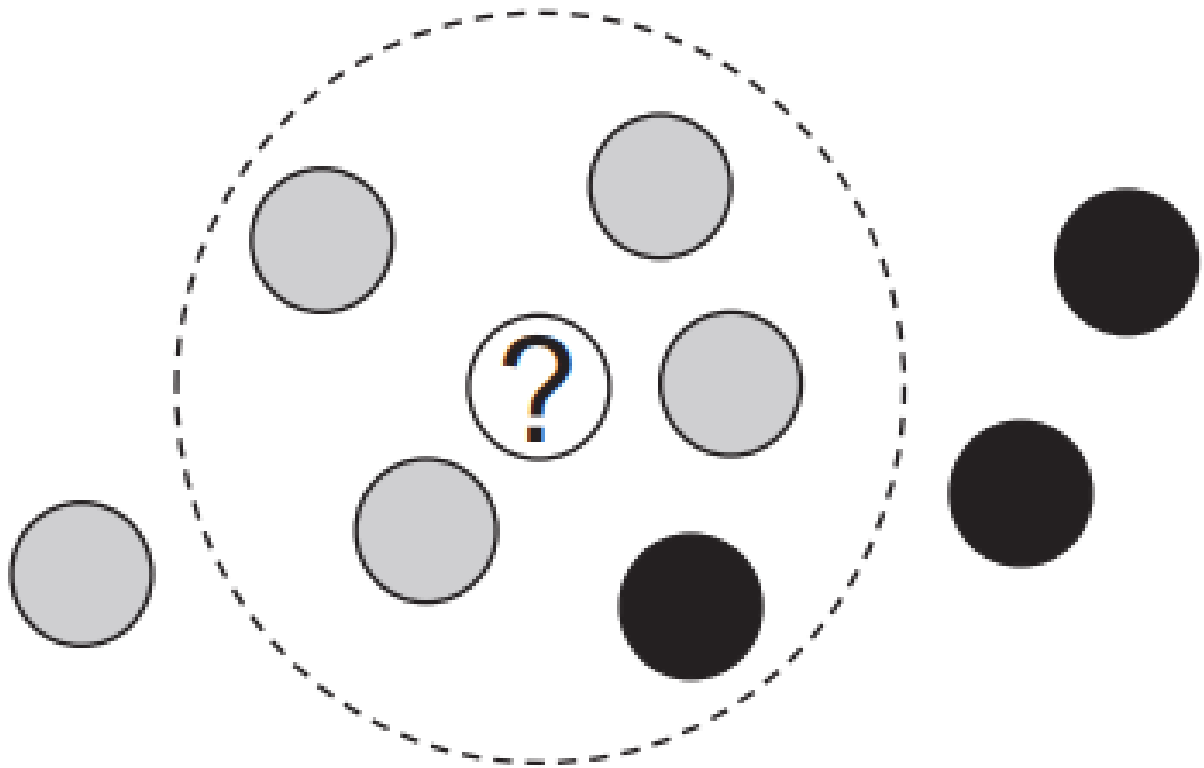


Рисунок 1

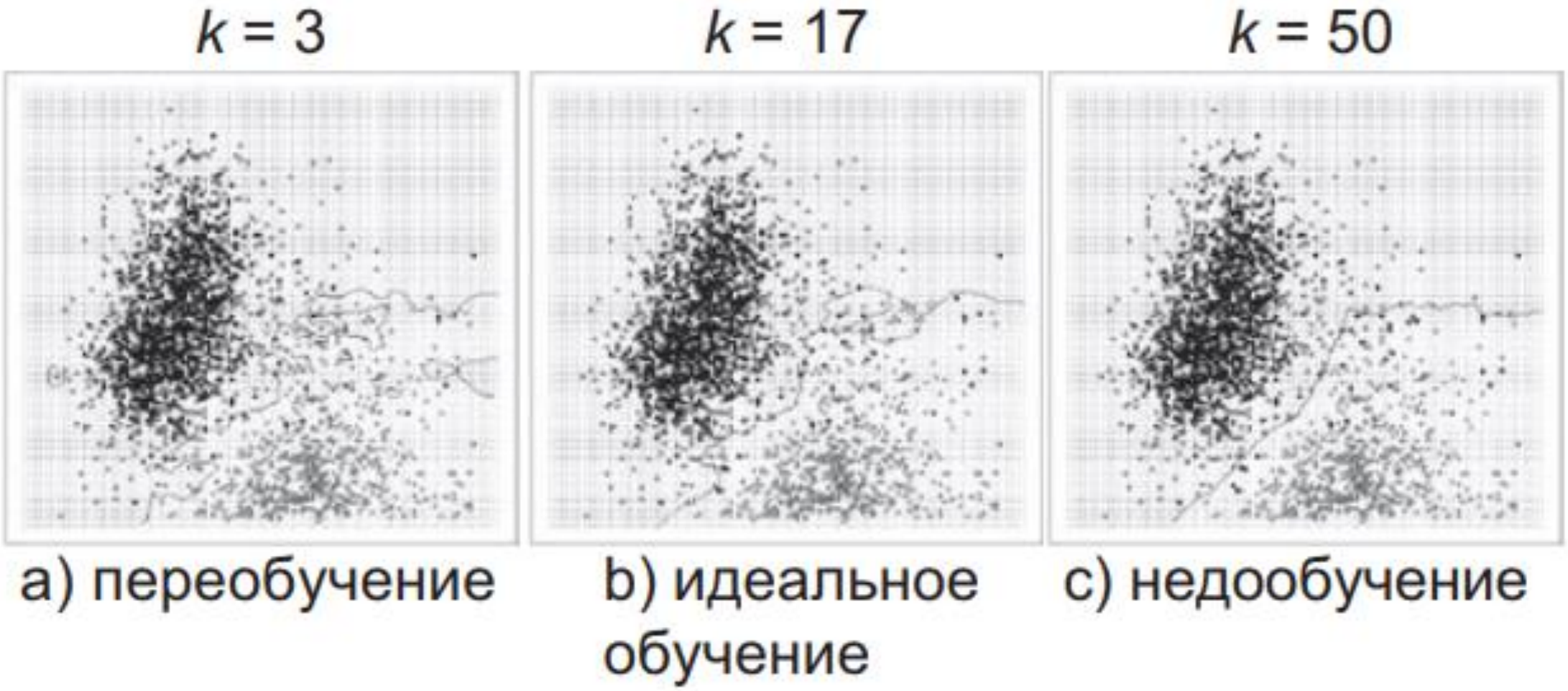
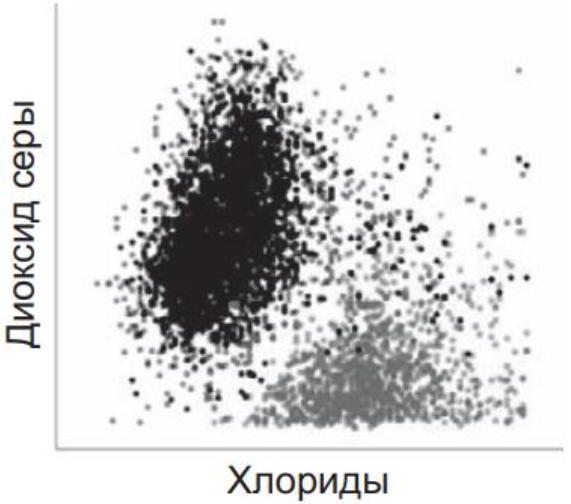


Рисунок 2. Сравнение моделей настройки при различных значениях k.

Пример: истинные различия в вине



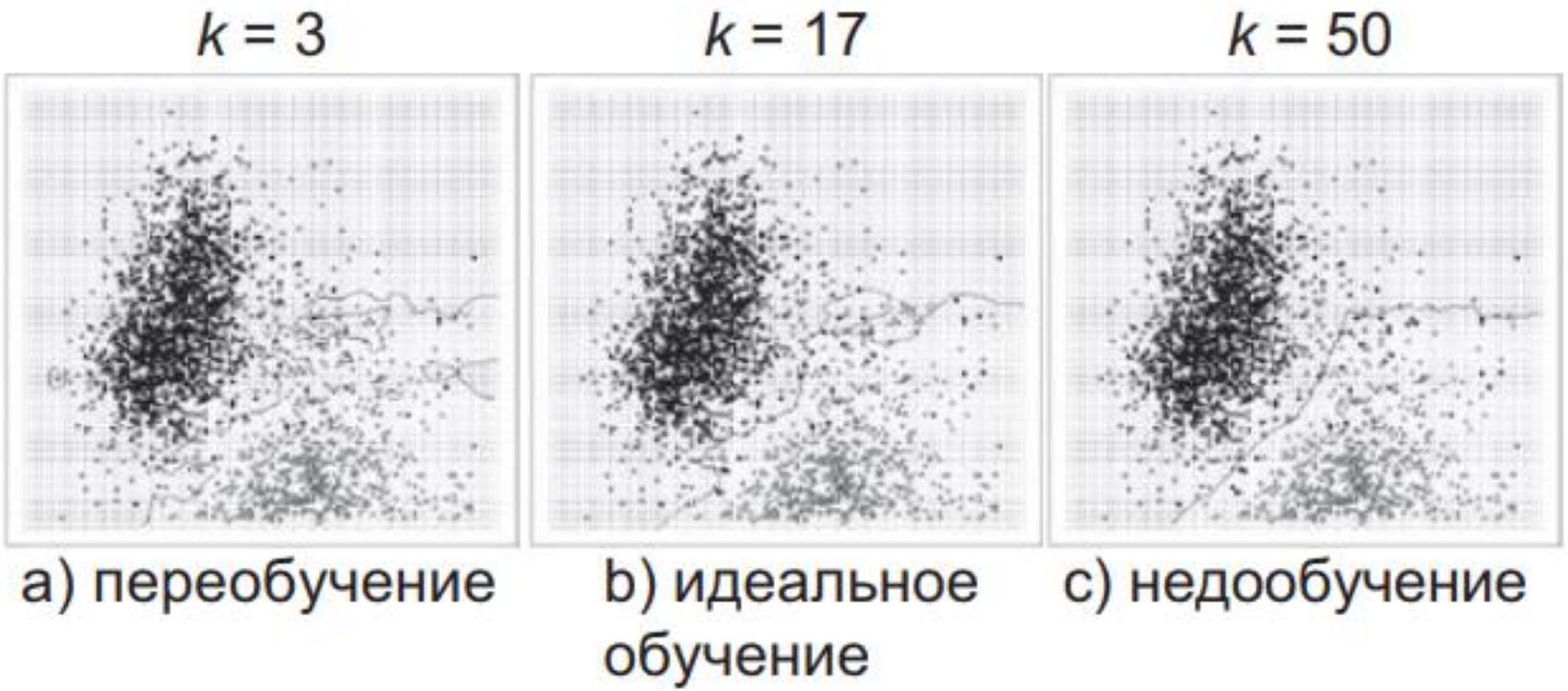
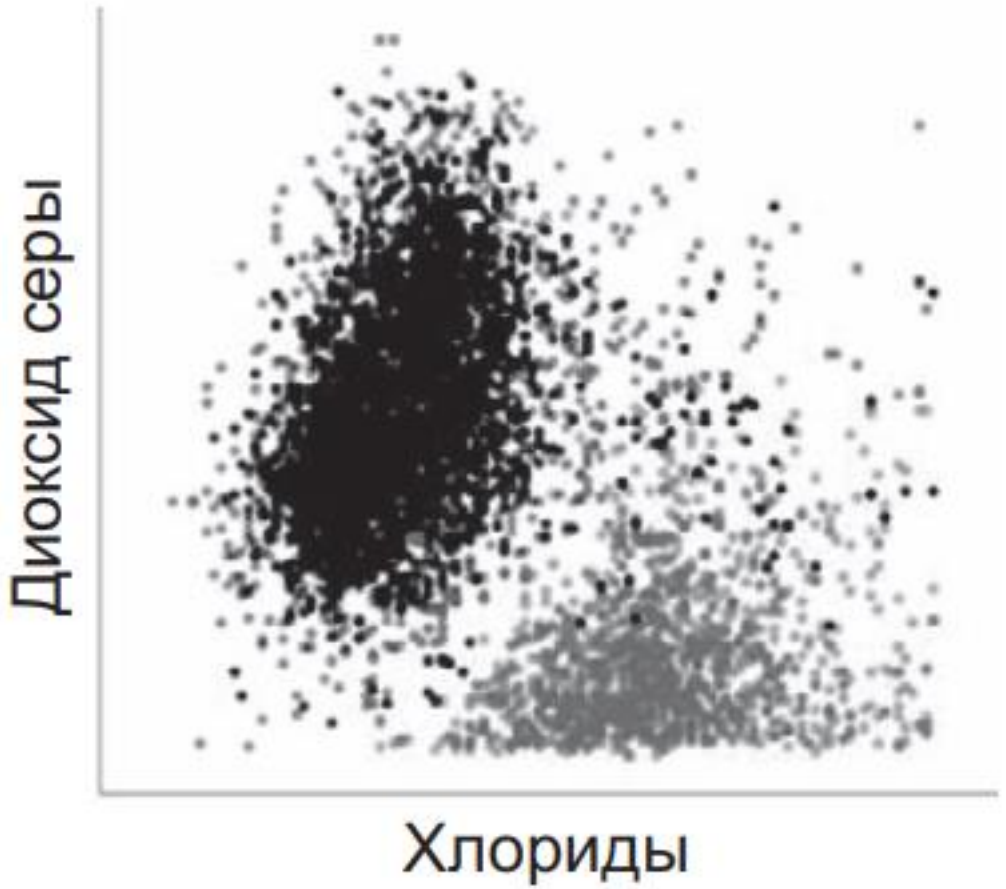


Рисунок 2. Сравнение моделей настройки при различных значениях k.

Обнаружение аномалий



Ограничения

- **Не классы.** Если имеется множество классов и эти классы существенно отличаются по размеру, то элементы данных, принадлежащие к самому небольшому из них, могут быть ошибочно включены в более крупные. Чтобы улучшить точность, можно и здесь использовать вместо равновесного вычисления весовые параметры, которые позволят больше ориентироваться на ближайшие элементы данных, а не на отдаленные.

- **Избыток предикторов.** Если предикторов слишком много, для определения ближайших соседей в нескольких измерениях могут потребоваться долгие вычисления. Более того, некоторые предикторы могут быть лишними и не улучшать точность прогноза. Чтобы исключить это, для выявления наиболее существенных предикторов для анализа можно воспользоваться уменьшением размерности

Краткие итоги

- Метод k-ближайших соседей представляет собой метод классификации элементов данных путем их соотнесения с ближайшими элементами.
- k — число таких ближайших элементов для расчета, которое определяется с помощью кросс-валидации.
- Лучше всего он работает при условиях, когда предикторов немного, а классы примерно одного размера. Неточные классификации могут служить верным признаком возможных аномалий.