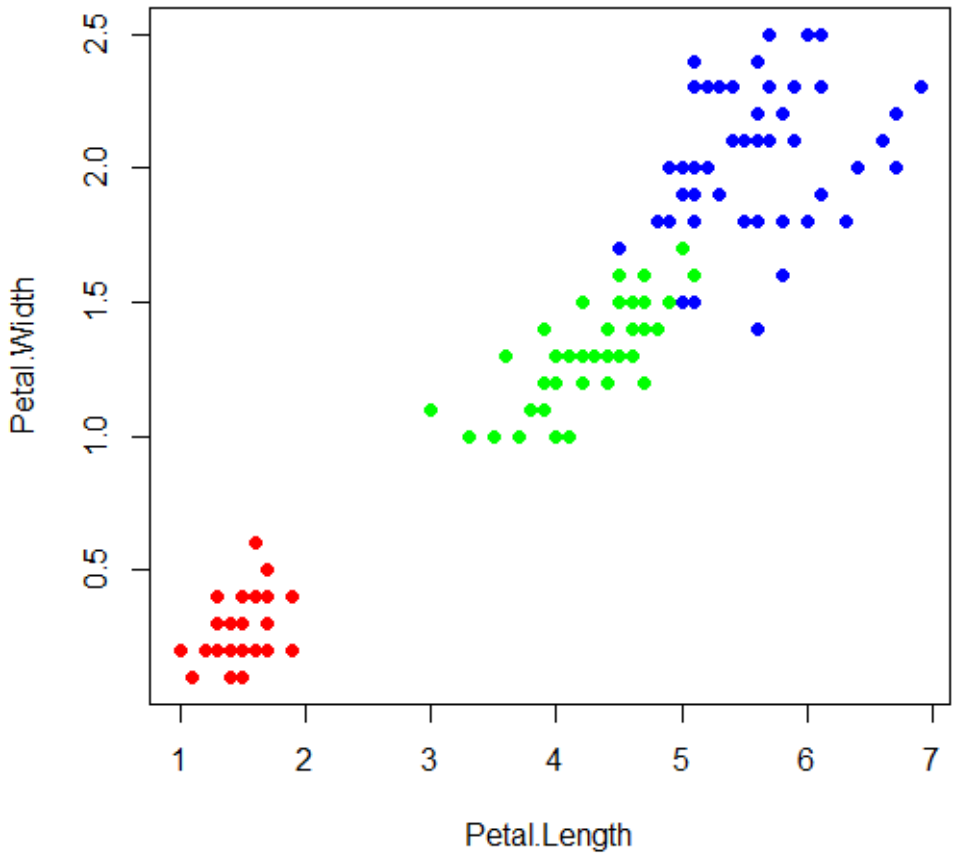
kNN

Метод k ближайших соседей. Для повышения надёжности классификации объект относится к тому классу, которому принадлежит большинство из его соседей — k ближайших к нему объектов обучающей выборки x_i. В задачах с двумя классами число соседей берут нечётным, чтобы не возникало ситуаций неоднозначности, когда одинаковое число соседей принадлежат разным классам.

Алгоритм:

$$a(u) = argmax_{y \epsilon Y} \sum_{i=1}^{m} [x_{i;u} = y]$$

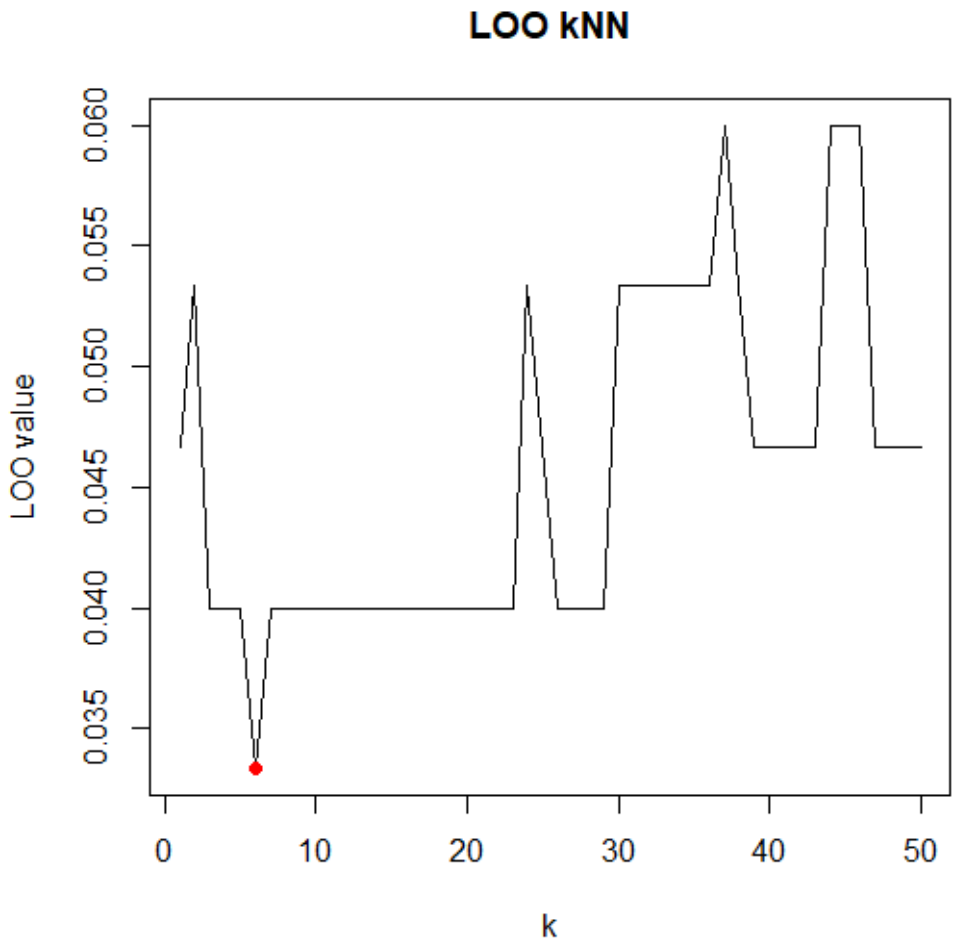
Возьмем датасет Ириса Фишера, график зависимостей признаков Petal.Length и Petal.Width, по ним же и будем классифицировать:



Возникает вопрос: какой k взять? Ответ - LOO(leave-one-out):

$$h = argmax_h \frac{1}{l} \sum_{i=1}^{l} (x_i; X^m / x_i)$$

Проверим для от $k = 1$ до $k = 50$. Лучший результат достигается при $k = 6$. Получим график:



Максимальная точность - 96%.