Домашнее задание №5 по курсу «Машинное обучение»:

Кукуев Михаил

1. Параметр k влияет на значение получаемой минимизированной оценки L_V и на качество выбора классификатора. При сильно большом k увеличивается вероятность получить сильно зашумленные значения L_{Vi} на поднаборах и в итоге неадекватное значение L_V, т.к. части валидационного набора могут оказаться слишком малыми и нерепрезентативными (в случае небольшого количества объектов в общей выборке).

При слишком малом k мы получим уже надежную оценку L_V, но обучение будет недостаточным и L_V слишком большим по сравнению с L_D, т.к. тренировочная выборка будет намного меньше настоящей, используемой впоследствии для решения задачи.

Данные рассуждения применимы для случаев, когда данных для обучения и валидации не так уж много. Если же объем выборки действительно очень большой, то различные значения k по идее слабо повлияют на значение оценки L_V.

Поэтому для больших наборов данных можно изпользовать малые значения k и значительно сэкономить время проведения кросс-валидации. А для небольших наборов нужно искать баланс, чтобы и тренировочная часть была репрезентативной, и тестовая, поэтому для таких случаев значения k рекомендуется обычно выбирать так, чтобы размер валидационной части при разбиении выборки составлял 10-20%.

2. На валидационном наборе мы подбираем параметры модели либо модификации алгоритма, стараясь минимизировать L_V. На тестовом же мы применяем выбранную модель, чтобы оценить насколько она подходит для решения задачи на новых данных, получая оценку L_D и вычисляя $|L_D - L_V|$. Если тестировать модель на валидационной выборке, то L_D будет равна L_V, которую мы уже заранее минимизировали во время валидации, т.е. мы получим неправдоподобную заниженную оценку L_D и не сможем определить насколько подходит модель для решения задачи в будущем на реальных данных.

3. Недостатки leave-all-out:

1. Применим только для выборок малого размера, т.к. при количестве разбиений 2^m на средних и больших наборах алгоритм будет работать слишком долго.
2. Из-за местами очень малого размера тренировочных и валидационных наборов могут быть получены совсем неадекватные оценки, которые затем наравне с остальными будут учтены при расчете L_D.

Преимущества:

1. Получаемая оценка не зависит от выбора k.
2. В алгоритме участвуют подвыборки различных размеров, а не фиксированного, что может быть полезно для определения склонности алгоритма к переобучению.