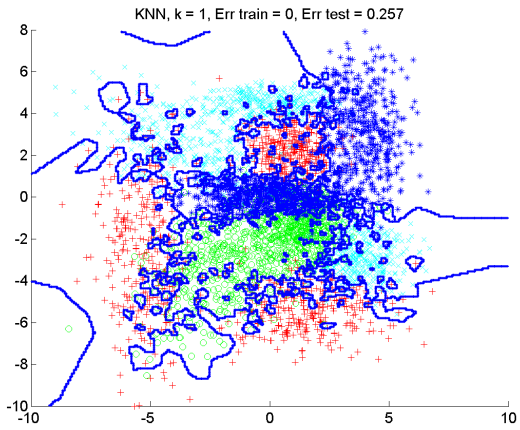
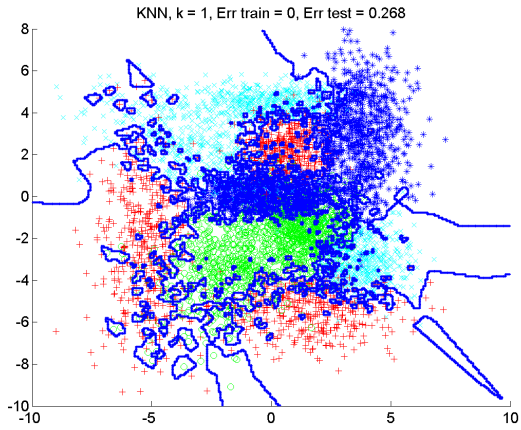
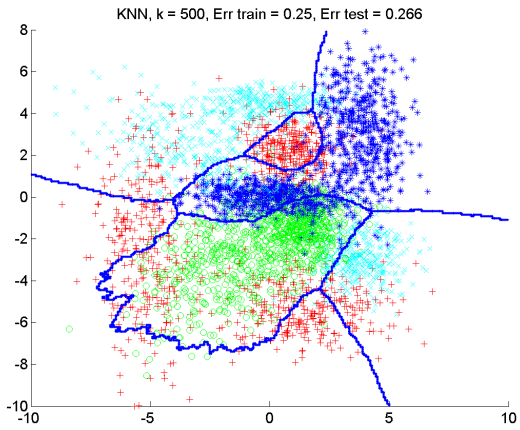
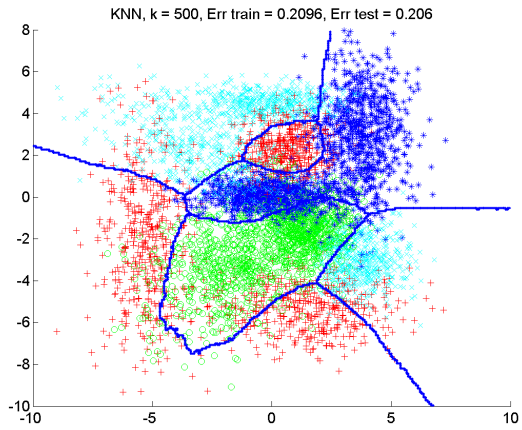
однако совершенно не приближается к оптимальному классификатору (уровень ошибки на тесте фиксирован на уровне 0.25). Увеличение объёма данных не приведёт к успеху в этой слишком простой модели, уровень ошибки существенно не уменьшится. Вот результат для 3000 и 5000 обучающих объектов:





Метод 500 ближайших соседей сильно недообучен, границы, выдаваемые им, слишком просты. Он обладает высоким *bias* (близкие кривые обучения на высоком уровне ошибки) при низком *variance*. При размере данных меньше 500 (50%) на тестовой выбоке наблюдается плато на уровне 75%: классификатор выдаёт в качестве ответа максимальный класс. На тестовой выборке при этом ошибка всё же ниже: сказывается случайный порядок объектов в обучающей выборке, соотношение классов не по 25%, и доминирующий класс состаляет чуть более 25%. Однако видна тенденция к падению уровня ошибки. Это происходит потому, что с увеличением объёма данных ослабляется эффект «доминирующего класса», и в этом случае добавление новых данных приведёт к значительному улучшению. Результат для 3000 обучающих объектов будет выглядеть гораздо более приемлемо, а для 5000 он даже приближается к оптимальному:





Налицо усложнение вида границ чисто за счёт большего объёма обучающей выборки. Это ещё раз наглядно иллюстрирует, насколько разный результат может давать одно и то же значение структурного параметра в зависимости от размера выборки.

3.2 SVM

Вот как выглядит зависимость ошибки SVM на обучении, валидации и тесте от каждого из параметров C и γ (другой параметр при этом фиксирован и равен 1; для параметров используется логарифмическая шкала):