Домашнее задание 7

Авласов Владислав

Gisette dataset

В гайде по свм, приложение С, есть указание о том, что в случае большого количества объектов и фич стоит использовать линейное ядро: svm-guide.
Для этого скачали библиотеку liblinear и поместили её в текущую директорию в папку liblinear_.
Используем кросс-валидацию, как и в первом задании, но поставим параметр 5, т.к. датасет довольно объёмный. Запускаем на значениях C от 2^{-15} до 2^2.
Данные по всем запускам можно найти в gisette_errors.txt. Первый параметр – значение C, второй – значение ошибки, третий – стандартное отклонение.
Наименьшая ошибка получилась с $C = 2^{-7}$, accuracy= 98.86%. Для $C = 2^{-8}$ среднее значение ошибки почти такое же, но стандартное отклонение меньше.
Точность модели проверим для этих двух значений C на наборе dev. Т.к. у нас кросс-валидация, т.е. не нужно было дополнительно откладывать отдельно данные для валидации, то test-set у нас будут объекты из gisette-valid.
Получили ответ:
log2(C) = -8: Test accuracy = 97.7%, error = 0.02300000000000002
log2(C) = -7: Test accuracy = 97.8%, error = 0.02200000000000002
true-risk всего на 1% выше empirical. Значит, с высокой вероятностью переобучения не произошло. Лучшиий вариант $C = 2^{-7}$, на нём и остановимся.