

Домашнее задание №8 по курсу «Машинное обучение»: SVM

Колесов Алексей

1 декабря 2018 г.

1 Задания

1. В этом задании вам предложено применить алгоритм SVM для задачи классификации спама. Данные будут использоваться из этого датасета: <http://archive.ics.uci.edu/ml/datasets/Spambase>. В задании будет дана ссылка на перемешанный датасет.

- скачайте и установите библиотеку `libsvm` с <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- скачайте перемешанную версию датасета отсюда <http://www.cs.nyu.edu/~mohri/yml/spambase.data.shuffled>
- используйте `libsvm scaling tool` для масштабирования данных. Первые 3450 объектов используйте в качестве тренировочной выборки, последние 1151 — для тестовой. Параметры масштабирования должны быть вычислены только на тренировочной выборке и лишь потом применены к тестовой
- решите задачу бинарной классификации на этом датасете с помощью SVM с полиномиальными ядрами.

Для этого, разделите тренировочную выборку на 10 равных непересекающихся частей. Для каждого значения степени полинома $d = 1, 2, 3, 4$ постройте график среднего значения ошибки на кросс-валидации плюс-минус стандартное отклонение как функции от C (другие параметры полиномиального ядра можете оставить по умолчанию в `libsvm`). Перебирайте C по степеням двойки от 2^{-k} до 2^k . k выберите таким образом, чтоб видеть изменение empirical risk (от очень большого до низкого). Чем больше значение C , тем дольше работает `libsvm`, это нормально
- пусть (C^*, d^*) — лучшая пара из предыдущего пункта. Зафиксируйте $C = C^*$ и постройте графики ошибки на кросс-валидации и тестовой выборки как функции от d . Постройте график количество опорных векторов, как функции от d . Как много опорных векторов в вашем решении? Какие выводы вы можете сделать?
- **Бонус:** какое ядро подходит лучше всего для этой задачи? Почему?

2. Примените SVM для решения задачи <https://archive.ics.uci.edu/ml/datasets/Gisette>. Ваша задача получить модель с какой можно более хорошим true risk. Проведите эксперименты и опишите ваш подход. Особое внимание в отчёте уделите тому, как вы используете train/dev/test разделение, а также как вы справились с тем, что в данном датасете большое количество и объектов, и признаков.