Лабораторная работа № 3

Использование логистической регрессии для решения задачи множественной классификации – распознавания рукописных цифр от 0 до 9

В этой работе вам нужно решить задачу классификации рукописных цифр от 0 до 9 с помощью множественной логистической регрессии, а точнее метода *one-vs-all.*

Идея этого метода очень проста. Если у нас есть объекты К классов, мы строим К различных бинарных классификатора, которые объекты определенного класса отделяют от всех остальных. Т.е. первый классификатор отделяет объекты первого класса от всех прочих ("не первого" класса), второй – второго и т.д. Теперь, когда у нас есть К таких классификаторов, для любого нового объекта мы можем вычислить вероятность его принадлежности к каждому из этих классов и выбрать тот класс, для которого это значение оказалось наибольшим.

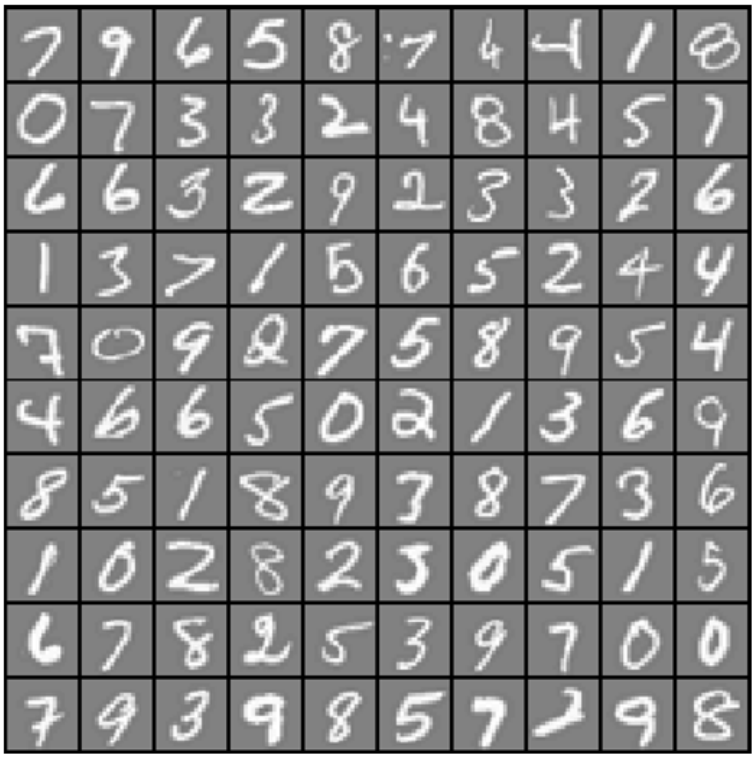
В этой работе набор данных содержит объекты 10 классов – это рукописные цифры от 0 до 9. Поэтому Вам предстоит обучать 10 различных бинарных классификаторов.

Загрузите данные из файла ex3data.txt

Первые 400 столбцов – это "цифры" X, последний столбец – метки классов y. Отделите их.

Набор данных содержит 5000 рукописных цифр. Каждая цифра была изначально gray scale картинкой 20x20 пикселей, которую затем "развернули" в строку из 400 элементов со значениями, характеризующими интенсивность данного пикселя.

Вот так выглядят некоторые из цифр нашего набора.



Все цифры теперь хранятся у Вас в массиве X размером 5000x400, а правильные ответы в y размером 5000x1.

Добавьте, как обычно, к массиву X столбец из 1.

Далее нам нужны все те же самые функции, что и в прошлой лабораторной работе: **sigmoid**, **costFunction**, **gradientFunc**, которые мы можем просто взять оттуда.

Мы также будем использовать **fmin_tnc** функцию, которая дает оптимальные значения θ при данных X и y. Помимо этих входных значений, она требует значения целевой функции и значения ее производной.

Постройте 10 бинарных классификаторов, решающих задачи "цифра k" - не "цифра k ", для $k = 0, \dots, 9$. У Вас должно получиться 10 наборов оптимальных значений θ, которые разумно хранить в одном массиве (размером 10x401.)

Осталось найти выходные значения каждого из 10 классификаторов

```
h=sigmoid(np.dot(X,theta.T))                # size (5000,10)
```

и выбрать класс с максимальным значением вероятности принадлежности объекта к нему

```
h_argmax = np.argmax(h, axis=1)
```

Тем самым, мы можем

предсказать, к какому классу относится наш объект. Найдите долю правильных ответов Вашей модели.

Итак, мы применили простой алгоритм логистической регрессии и метода one-vs-all к достаточно сложной задаче распознавания рукописных цифр и получили, как видите, очень неплохой результат.