Домашнее задание №4.

Предлагается потестировать некоторые изученные методы первого порядка на задаче логистической регрессии (вспомнить/познакомиться с тем, что такое логистическая регрессия, можно, например, на хабре или прочитать в википедии; для выполнения задания это будет необязательно). Пусть $A \in \mathbb{R}^{n\times d}$ — матрица признаков, $y \in \{-1,1\}^n$ — вектор ответов. Здесь n — число объектов в датасете, d — число признаков. Функцию потерь в логистической регрессии можно записать в следующем виде:

$$Loss_{logreg}(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-y_i\cdot(Ax)_i\right)\right), \tag{1}$$

где $x \in \mathbb{R}^d$ — вектор параметров (переменные, по которым нужно оптимизировать), $(Ax)_i$ — i-я компонента вектора Ax. Как правило, на практике задачу минимизации функции (1) не решают в чистом виде.[1], а решают регуляризованную задачу

$$Loss_{logreg}(x) + R(x) \to \min_{x\in\mathbb{R}^d}, \tag{2}$$

где функция $R(x)$ — выпуклая и замкнутая. В этом задании предлагается потестировать некоторые методы оптимизации на задаче (2). Везде далее в этом проекте предлагается рассматривать регуляризатор вида $R(x) = \frac{\mu}{2}\|x\|_2^2$. В таком случае задачу можно переписать в следующем виде

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\log\left(1+\exp\left(-y_i\cdot(Ax)_i\right)\right) + \frac{\mu}{2}\|x\|_2^2\right)}_{f_i(x)} \to \min_{x\in\mathbb{R}^d}. \tag{3}$$

Задание.

1. Загрузите ваш любимый датасет для бинарной классификации (датасеты можно загружать из LIBSVM; в sklearn есть удобная для этих целей функция `load_svmlight_file`; к заданию прикладывается датасет `mushrooms`, можно использовать его). Обратите внимание, что вектор меток y должен принимать значения из множества $\{-1,1\}$ (если это изначально не так, то преобразуйте вектор y к нужному виду).

2. Докажите, что дважды непрерывно дифференцируемая функция $f(x)$ является L-гладкой тогда и только тогда, когда $\|\nabla^2 f(x)\| \le L$ для всех $x \in \mathbb{R}^d$. Покажите отсюда, что константу L для функции f можно взять равной $\mu + \frac{1}{4}\max\limits_{i=1,\ldots,n}\|A_i\|_2^2$, где A_i — i-я строчка матрицы A.

3. Докажите, что константа сильной выпуклости функции $f(x)$ равна μ.

4. Имплементируйте следующие методы:

   - Градиентный спуск.
   - Стохастический градиентный спуск.
   - Ускоренный метод Нестерова.

   Запустите методы на задаче (3). Для справедливого сравнения методов считайте не число итераций, а так называемое число эпох. Для полноградиентных методов число эпох равняется числу итераций, а для стохастического градиентного спуска число эпох равняется числу итераций, поделённому на n. В таком случае методы сравниваются по числу вычислений градиентов функций f_i (слагаемых в (3)). По оси ординат откладывайте значение функции $f(x)$ и норму градиента $\|\nabla f(x)\|_2$.

   Рассмотрите все изученные на семинарах правила выбора шагов для градиентного спуска и для стохастического градиентного спуска.

   Запустите методы с разными μ. Рассмотрите случаи с разным числом обусловленности, то есть возьмите $\mu = \frac{L}{10000}, \frac{L}{1000}, \frac{L}{100}, \frac{L}{2}$ и сравните методы в каждом случае.

[1] Кстати говоря, неплохое разминочное упражнение – подумайте, почему эту задачу не решают в чистом виде. При всех ли A данная задача имеет решение?