OzonMasters

Теоретическое задание 2
Оптимизация и регуляризация линейных моделей

курс «Машинное обучение 1», 2020

Базовая стоимость задания — 12 баллов, максимальная — 14 баллов. Для выполнения базовой части достаточно решить любое число задач, дающих в сумме 12 баллов.

Задачи

1. Продифференцируйте следующие функции:
   (a) (1 балл) $f(\mathbf{x}) = \log(1 + \mathbf{x}^\intercal A\mathbf{x})$
   (b) (1 балл) $f(\mathbf{x}) = \dfrac{-1}{1 + \mathbf{x}^\intercal \mathbf{x}}$
   (c) (2 балла) $f(\mathbf{x}) = \mathbf{x}^\intercal A[\mathbf{x}]^2, \quad$ где $[\mathbf{x}]^2 = (x_1^2 \dots x_n^2)^\intercal$

2. (2 балла) Для функции

$$f(x) = \sin(x)\sin(2x), \quad x \in \left[-\frac{3\pi}{2} - a; \frac{3\pi}{2} + a\right], \quad a = \arcsin\sqrt{\frac{2}{3}}$$

   применяется метод градиентного спуска с убывающим шагом. Начальный шаг градиентного спуска достаточно мал. Критерий останова: $\nabla f(x) = 0$.
   После того, как очередной спуск завершился, выбирается новая точка, и метод применяется заново. Точки для запуска метода градиентного спуска выбираются равномерно из интервала $\left[-\frac{3\pi}{2} - a; \frac{3\pi}{2} + a\right]$.
   Вблизи каких точек будет останавливаться алгоритм? Какая доля спусков сойдется к каждой из точек останова, если точек для запуска спуска было выбрано достаточно много?

3. (1 балл за пункт) Найдите субдифференциалы во всех точках для следующих функций:
   (a) $f(x) = \max(0, 1 - ax)$, a — $const$
   (b) $f(x) = \sqrt{|x|}$

4. (2 балла) Представим, что в решаемой задаче регрессии признаки можно разбить на G непересекающихся групп (например, каждая группа — one-hot кодирование признака до обработки). Будем использовать для решения линейную модель и введём новый вид регуляризации: $\|w\|_{\frac{1}{2}} = \sum_{g \in G} \|w_g\|_2$, где $\|w_g\|_2$ — $l2$-норма весов всех признаков внутри группы g. Например, для $g = \{w_1, w_2\}$, $\|w_g\|_2 = \sqrt{w_1^2 + w_2^2}$. Утверждается, что такой способ регуляризации позволяет отбирать признаки не по отдельности, а рамках групп (т.е. одновременно занулять веса признаков всей группы).

   Рассмотрите случай, когда в задаче всего три признака (w_1, w_2, w_3), $G = \{g_1, g_2\}$, $g_1 = \{w_1, w_2\}$, $g_2 = \{w_3\}$. Нарисуйте линии уровня функции $\|w\|_{\frac{1}{2}}$. Интерпретируйте полученную картинку аналогично случаю $l1$-регуляризации. Поясните, почему при такой регуляризации будет иметь место отбор групп признаков, указанный выше.

5. (2 балла) Покажите, что регуляризация из предыдущей задачи позволяет отбирать группы признаков, с помощью аппарата субдифференциалов. Используйте аппроксимацию основного функционала с помощью ряда Тейлора.

   Замечание 1. Для решения может быть полезно показать, что субдифференциал функции $f(x) = \|x\|_2$ в точке 0 равен $\overline{B}_2(0, 1) = \{x : \|x\|_2 \leqslant 1\}$.

   Замечание 2. В отличие от семинара не требуется выводить зависимость решения регуляризованной задачи от нерегуляризованной во всех случаях, достаточно проанализировать только случаи зануления признаков.

6. (2 балла) Докажите, что в случае линейно разделимой выборки не существует вектора параметров (весов), который бы максимизировал правдоподобие модели логистической регрессии в задаче двухклассовой классификации. Покажите, как можно модифицировать модель, чтобы оптимум достигался.