# YSDA ML HW #3

Тихон Волжин

9 октября 2023 г.

## Задача 1

### 1)

$$p_\theta(x) = \frac{\theta^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\theta x} \qquad \theta > 0, \beta > 0, x > 0$$

Функция правдоподобия (из-за независимости $X_1, \ldots, X_N$) имеет вид:

$$\mathscr{L}_{X_1,\ldots,X_N}(\theta) = \mathbb{P}(X_1, \ldots, X_N|\theta) = \prod_{i=1}^{N} P(X_i|\theta) = \prod_{i=1}^{N} \frac{\theta^\beta}{\Gamma(\beta)} X_i^{\beta-1} e^{-\theta X_i}$$

Оценка максимального правдоподобия получается из максимизации данной функции:

$$\hat{\theta} = \operatorname*{argmax}_\theta \mathscr{L}_{\langle X_i \rangle}(\theta) = \operatorname*{argmax}_\theta \log \mathscr{L}_{\langle X_i \rangle}(\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^{N} \log \left( \frac{\theta^\beta}{\Gamma(\beta)} X_i^{\beta-1} e^{-\theta X_i} \right) =$$

$$= \operatorname*{argmax}_\theta \sum_{i=1}^{N} \left[ const + \log \left( \theta^\beta e^{-\theta X_i} \right) \right] = \operatorname*{argmax}_\theta \sum_{i=1}^{N} \left[ \beta \log \theta - \theta X_i \right] = \operatorname*{argmax}_\theta \left[ N\beta \log \theta - \theta \sum_{i=1}^{N} X_i \right]$$

Находим производную логарифма функции правдоподобия:

$$\frac{\partial \log \mathscr{L}_{\langle X_i \rangle}(\theta)}{\partial \theta} = N\beta \frac{1}{\theta} - \sum_{i=1}^{N} X_i$$

Производная имеет один корень, и очевидно, что это максимум (при $\theta > 0, \beta > 0, X_i > 0$ до $\hat{\theta}$ производная положительна, после - отрицательна), поэтому ответ:

$$\hat{\theta} = \frac{N\beta}{\sum_{i=1}^{N} X_i} = \frac{\beta}{\bar{X}}$$

### 2)

$$\mathrm{P}_\theta(X_i = k) = \frac{\theta^k}{k!} e^{-\theta} \qquad \theta > 0, k \in \mathbb{Z}_+ = \{0, 1, 2, ...\}.$$

Аналогично:

$$\hat{\theta} = \operatorname*{argmax}_\theta \log \mathscr{L}_{\langle X_i \rangle}(\theta) = \operatorname*{argmax}_\theta \sum_{i=1}^{N} \left[ const + X_i \log \theta - \theta \right] = \operatorname*{argmax}_\theta \left( -N\theta + \log \theta \sum_{i=1}^{N} X_i \right)$$

$$\frac{\partial \log \mathscr{L}_{\langle X_i \rangle}(\theta)}{\partial \theta} = -N + \frac{1}{\theta} \sum_{i=1}^{N} X_i = 0$$

Аналогично: корень производной - единственный максимум. Поэтому ответ (оценки не существует, если все $X_i = 0$):

$$\hat{\theta} = \frac{\sum_{i=1}^{N} X_i}{N} = \bar{X}$$

# Задача 2

## 1)

$$\mathcal{F}(\theta) = -\ell_Y(\theta) + \lambda||\theta||^2 \longrightarrow \min_{\theta \in \mathrm{R}^d}$$

$$\mathcal{F}(\theta) = -\sum_{j=1}^{N} \left[ Y^j \log \sigma(\theta^T X^j) + \left(1 - Y^j\right) \log(1 - \sigma(\theta^T X^j)\right] + \lambda \theta^T \theta$$

Пользуясь формулой производной сложной функции и

$$\sigma' = \sigma(1-\sigma) \quad (\log \sigma)' = (1-\sigma) \quad (\log(1-\sigma))' = -\sigma$$

Получаем:

$$\nabla_\theta F = -\sum_{j=1}^{N} \left[ Y^j (1 - \sigma(\theta^T X^j)) - \left(1 - Y^j\right) \sigma(\theta^T X^j)\right] X^j + 2\lambda\theta = -\sum_{j=1}^{N} \left[ Y^j - \sigma(\theta^T X^j)\right] X^j + 2\lambda\theta$$

Формула для GD с шагом $\alpha$ (добавим еще в градиенте $\frac{1}{N}$, чтобы на это можно было смотреть как на оценку матожидания градиента по выборке + регуляризация):

$$\theta_t = \theta_{t-1} - \alpha \nabla_\theta F = \theta_{t-1} + \alpha \left( \frac{1}{N} \sum_{j=1}^{N} \left[ Y^j - \sigma(\theta^T X^j)\right] X^j - 2\lambda\theta \right)$$

Формула для SGD $\left(\sum_{j=1}^{N} \sim \frac{N}{k} \sum_{j_1,\ldots,j_k}\right)$:

$$\theta_t = \theta_{t-1} + \alpha \left( \frac{1}{k} \sum_{j=1}^{N} \left[ Y^j - \sigma(\theta^T X^j)\right] X^j - 2\lambda\theta \right)$$

## 2)

Запишем приращение первого порядка в терминах скалярного произведения градиента с вектором приращений (обозначения идентичны обозначениям из учебника и обозначениям на лекции ($S(\theta)$ - вектор сигм $\sigma(\theta^T X^j)$ и.т.п):

$$[D_{\theta_0} F(\theta)](h_1) = \langle \nabla_\theta F, h_1 \rangle = \langle -X^T (Y - S(\theta)) + 2\lambda\theta, h_1 \rangle$$

Запишем второй дифференциал как дифференциал первого дифференциала и загоним его через все линейные отображения, которые встретим на своем пути:

$$[D_{\theta_0}[D_{\theta_0} F(\theta)](h_1)](h_2) = \langle -X^T (Y - [D_{\theta_0} S(\theta)](h_2)) + 2\lambda [D_{\theta_0}\theta](h_2), h_1 \rangle$$

Второй дифференциал применяется построчно и в форме $\langle ..., h \rangle$ записывается как линейное преобразование $h_2$ с матрицей в строках которой градиент каждой компоненты вектор функции от $h_2$. В случае вектор функции $\theta$ это очевидно $(\delta_{ij})$, т.е единичная матрица. В случае $S(\theta)$ мы в строку с $\sigma_k = \sigma(\theta^T X^k)$ добавляем ее градиент (верхний индекс - индекс объекта, нижний - индекс признака):

$$\nabla_{\theta_0} \sigma_k = \langle \sigma_k(1-\sigma_k) X_1^k, \ldots, \sigma_k(1-\sigma_k) X_d^k \rangle$$

$$[D_{\theta_0} \sigma_k] = \langle \nabla_{\theta_0} \sigma_k, h_2 \rangle$$

$$[D_{\theta_0} S(\theta)](h_2) = \begin{pmatrix} \sigma_1(1-\sigma_1)X_1^1 & \sigma_1(1-\sigma_1)X_2^1 & \cdots & \sigma_1(1-\sigma_1)X_d^1 \\ \sigma_2(1-\sigma_2)X_1^2 & \sigma_2(1-\sigma_2)X_2^2 & \cdots & \sigma_2(1-\sigma_2)X_d^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_N(1-\sigma_N)X_1^N & \sigma_N(1-\sigma_N)X_2^N & \cdots & \sigma_N(1-\sigma_N)X_d^N \end{pmatrix} \times h_2$$

Так как каждая строка умножена на одно и то же $\sigma'_k$, а без этих коэффициентов это в точности матрица X, то эту матрицу можно представить как произведение:

$$\Sigma_\sigma = diag\langle\sigma_0(1-\sigma_0),\ldots,\sigma_N(1-\sigma_N)\rangle$$

$$[D_{\theta_0}S(\theta)](h_2) = (\Sigma_\sigma \times X)\, h_2$$

Тогда приращение второго порядка:

$$[D^2_\theta F(\theta)](h_1,h_2) = h_2^T\left(X^T \times \Sigma_\sigma \times X + 2\lambda\mathbb{I}\right) h_1$$

Тк производная сигмоиды $\sigma_k(1-\sigma_k) > 0$, то можем взять корень из матрицы $\Sigma_\sigma$ и ввести новую матрицу:

$$X^* = \sqrt{\Sigma_\sigma} \times X$$

Finally, гессиан имеет вид:

$$Hess(F(\theta)) = X^{*T}X^* + 2\lambda\mathbb{I}$$

Положительная определенность (а следовательно выпуклость и единственный минимум в экстремуме) доказывается очевидным образом ($\langle Xh, Xh\rangle \geq 0$, $2\lambda\langle h,h\rangle > 0$ при $h \neq 0$ ):

$$h^T(X^{*T}X^* + 2\lambda\mathbb{I})h = (X^*h)^T(X^*h) + 2\lambda h^T h = \langle Xh, Xh\rangle + 2\lambda\langle h,h\rangle > 0 \quad \text{при } h \neq 0$$

Положительная определенность гессиана доказана, а значит и единственность и существование экстремума (глобального минимума).

## 3)

Сначала покажем очевидную вещь, связанную со значениями функции $\sigma(\theta^T x)$, $x \in \mathbb{R}^d$, где d - размерность пространства признаков:

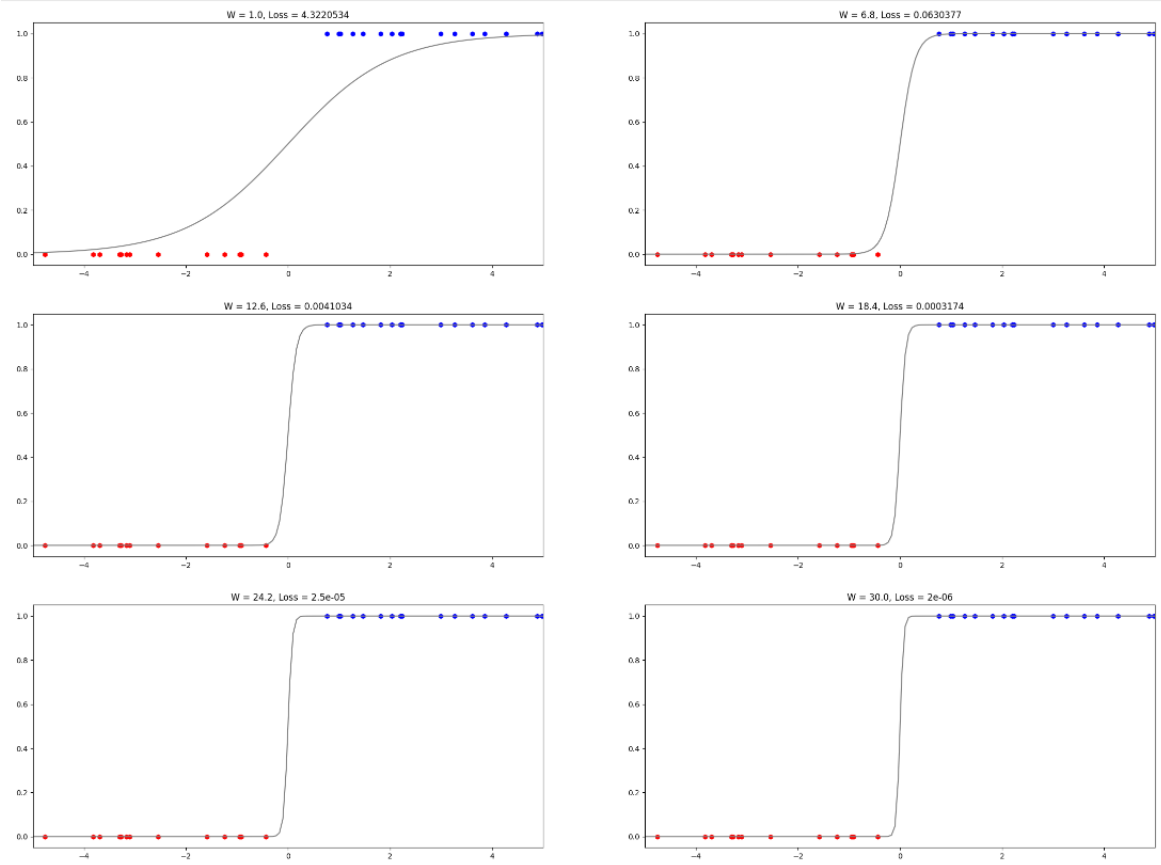
$$\sigma(\theta^T x) = \sigma(\theta_0 + x_1\theta_1 + \cdots + x_n\theta_n)$$

Подсигмоидное выражение - линейная функция, которая при $\theta_0 + x_1\theta_1 + \cdots + x_n\theta_n = C$ задает гиперплоскость (при C = 0 она разделяющая). То есть линии уровня - это гиперплоскости, параллельные разделяющей плоскости, так как нормальный вектор для всех плоскостей одинаков. Если вместо нормального вектора $\bar\theta = (\theta_1,\ldots,\theta_d)$ взять $\bar\theta^* = |W| * (\theta_1,\ldots,\theta_d)$, то есть увеличить/уменьшить его в |W|>0 раз, то положение разделяющей гиперплоскости и направление роста сигмоиды (а меняется она, как мы уже показали только вдоль нормали к разделяющей гиперплоскости) не изменится. Из за того, что функция меняется только вдоль одного направления то можно для наглядности рассмотреть как раз редукцию исходной задачи к одномерному случаю: мы будем рассматривать по оси X - C определенное выше, и можно отмечать $C_1,\ldots,C_N$ (N - число объектов). Теперь рассмотрим две ситуации:

1) Выпуклые оболочки двух классов не пересекаются, то есть существует разделяющая гиперплоскость, которая полностью делит эти два множества
2) Выпуклые оболочки пересекаются

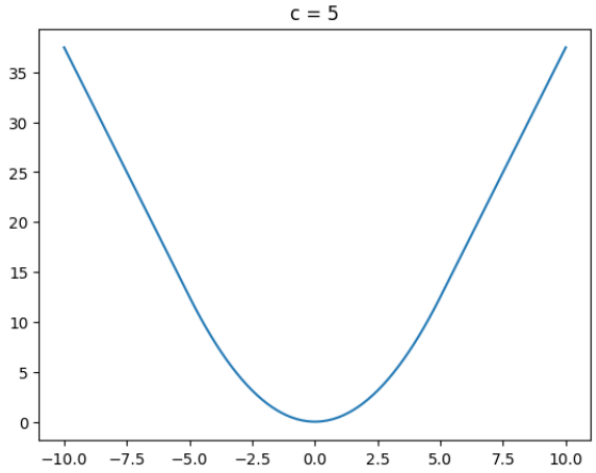
Проблемы есть с первым случаем, тк по втором случае для каждой разделяющей плоскости при увеличении нормы весов (без сдвига), величина точки находящейся не по ту сторону будет приводить к близкой к нулю величине под логарифмом, и функция потерь будет расти. НО в случае с разделяемыми классами сигмоида будет неограниченно прижиматься к меткам классов за счет увелечения нормы вектора весов (опять же сдвиг не трогаем). То есть разделяющая гиперплоскость все та же, модель предсказывает метки правильно, но без регуляризации норма вектора весов неограниченно увеличивается, при этом функция потерь будет слева стримиться к нулю. А значит минимума в данном случае не будет - он будет на бесконечности, а следовательно модуль весов без сдвига будет неограниченно расти. Снизу приведена иллюстрация уже в проекции на одномерный случай (для простоты без ограничения общности сдвиг тут нулевой, и классы разделены относительно нуля). Лосс стремиться к нулю при $W \longrightarrow \infty$ и модуль весов

растет неограниченно, минимизируя функцию потерь. Регуляризация препятствует росту нормы, поэтому с ней такой проблемы не будет.




# Задача 3

График:




Эта функция более устойчива к выбросам, то есть градиент не сильно меняется при наличии слишком больших отклонений, чтобы при SGD модель не содержала огромных вкладов выбросов в градиент в ущерб оптимизации менее отклонившихся экземпляров. Также она менее концентрируется на малых отклонениях.

$$R(x) = \frac{x^2}{2} I\{|x| \le c\} + c\left(|x| - \frac{c}{2}\right) I\{|x| > c\}.$$

$$R'(x) = xI\{|x| \le c\} + c\,sign(x) I\{|x| > c\}.$$

$$\nabla_\theta F(\theta) = \sum_{k=1}^{N} R'(Y^k + \theta^T X^k) X^k$$

$$\nabla_\theta F(\theta) = \sum_{k=1}^{N} ((Y^k + \theta^T X^k) I\{|Y^k + \theta^T X^k| \le c\} + c\,sign(Y^k + \theta^T X^k) I\{|Y^k + \theta^T X^k| > c\}) X^k$$

GD:

$$\theta_t = \theta_{t-1} - \alpha \nabla_{\theta_{t-1}} F(\theta)$$

SGD (K - размер батча):

$$\theta_t = \theta_{t-1} - \alpha \frac{N}{K} \nabla_{\theta_{t-1}} F(\theta)_{batch}$$