Задание №1.4. Логистическая регрессия и гауссовский дискриминантный анализ

В этом упражнении мы рассмотрим два вероятностных линейных классификатора. Во-первых, дискриминативный линейный классификатор: логистическую регрессию. Во-вторых, генеративный линейный классификатор: гауссовский дискриминантный анализ (ГДА). Оба алгоритма находят линейную решающую границу, которая разделяет входные данные на два класса, правда модели делают при этом разные предположения. Наша цель – получить более глубокое представление о сходстве и различиях (а также, сильных и слабых сторонах) этих двух алгоритмов.

Мы будем работать с двумя обучающими выборками, а также стартовым кодом, предоставленными в следующих файлах:

- ds1_{train, valid}.csv
- ds2_{train, valid}.csv
- logreg.py
- gda.py

Каждый датасет содержит n обучающих примеров, по одному $(x^{(i)}, y^{(i)})$ в строке. В частности, i-тая строка содержит значения $x_0^{(i)} \in \mathbb{R}$, $x_1^{(i)} \in \mathbb{R}$ и $y^{(i)} \in \{0,1\}$. В подзадачах, приведенных ниже, мы исследуем поведение логистической регрессии и гауссовского дискриминантного анализа на этих двух обучающих выборках.

Вопрос №1 [2 балла]

Мы помним, что функция потери для логистической регрессии есть

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\Big(y^{(i)}\log\Big(h_\theta\big(x^{(i)}\big)\Big) + \big(1-y^{(i)}\big)\log\Big(1-h_\theta\big(x^{(i)}\big)\Big)\Big),$$

где $y^{(i)} \in \{0,1\}$, $h_\theta(x) = g(\theta^T x)$, а $g(z) = \frac{1}{1+e^{-z}}$.

Найдите гессиан H этой функции, и покажите, что для любого вектора z выполняется неравенство

$$z^T H z \geq 0.$$

Подсказка: Вы можете начать с того, чтобы показать, что $\sum_i \sum_j z_i x_i x_j z_j = (x^T z)^2 \geq 0$. Также напомним, что $g'(z) = g(z)\big(1-g(z)\big)$.

Замечание: Это один из стандартных способов показать, что матрица H является положительной полуопределенной, что записывается как «$H \preccurlyeq 0$». Вы также можете воспользоваться другим способом доказать приведенное утверждение.

Вопрос №2. Задача на программирование [1 балл]

Следуя инструкциям в файле logreg.py, обучите логистическую регрессию с помощью метода Ньютона. Начиная со значения $\theta = \vec{0}$, повторяйте алгоритм Ньютона до тех пор, пока обновления вектора параметров не станут совсем незначительными: более конкретно, проводите обучение до первой итерации k такой, что $\|\theta_k - \theta_{k-1}\|_1 < \varepsilon$ для $\varepsilon = 1 \times 10^{-5}$. Убедитесь, что предсказания модели на валидационной выборке записаны в файл, указанный в коде.

Визуализируйте значения **валидационной** выборки на графике, в котором по горизонтальной оси идут значения x_1, а по вертикальной – x_2. Используйте разные символы, чтобы обозначать примеры

разных классов. На этом же графике отобразите решающую границу, найденную логистической регрессией (т.е. график прямой, соответствующей $p(y|x) = 0.5$).

Вопрос №3 [1 балл]

Напомним, что в ГДА мы моделируем совместное распределение вероятностей (x, y) с помощью следующих уравнений:

$$p(y) = \begin{cases} \phi, \text{если } y = 1, \\ 1 - \phi, \text{если } y = 0, \end{cases}$$
$$p(x|y = 0) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu_0)^T\Sigma^{-1}(x - \mu_0)\right),$$
$$p(x|y = 1) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu_1)^T\Sigma^{-1}(x - \mu_1)\right),$$

где ϕ, μ_0, μ_1 и Σ являются параметрами модели.

Предположим, мы уже подобрали ϕ, μ_0, μ_1 и Σ, и хотим предсказать значение y по новой точке x. Для того чтобы показать, что ГДА приводит к классификатору с линейной решающей границей, покажите, что апостериорное распределение вероятностей может быть записано в виде

$$\mathrm{p}(y = 1|x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp\big(-(\theta^T x + \theta_0)\big)},$$

где $\theta \in \mathbb{R}^d$, а $\theta_0 \in \mathbb{R}$ – подходящие функции ϕ, μ_0, μ_1 и Σ.

Вопрос №4 [1.4 балла]

Мы утверждаем, что для заданной обучающей выборки оценки максимального правдоподобия параметров задаются следующим образом:

$$\phi = \frac{1}{n}\sum_{i=1}^{n} 1\{y^{(i)} = 1\},$$
$$\mu_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}},$$
$$\mu_1 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}},$$
$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T.$$

Лог-правдоподобие обучающей выборки по параметрам есть

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log\prod_{i=1}^{n} p\big(x^{(i)}, y^{(i)};\ \phi, \mu_0, \mu_1, \Sigma\big) = \log\prod_{i=1}^{n} p\big(x^{(i)}|\ y^{(i)}; \mu_0, \mu_1, \Sigma\big)p\big(y^{(i)}; \phi\big).$$

Максимизируя ℓ по указанным четырем параметрам (приравняв частные производные к нулю и выразив параметры), докажите корректность вышеприведенных оценок. (Вы можете предполагать, что есть хотя бы один положительный и один отрицательный примеры, так что знаменатели в формулах для μ_0 и μ_1 будут ненулевыми.)

Вопрос №5. Задача на программирование [1 балл]

В файле-заготовке gda.py допишите код для вычисления ϕ, μ_0, μ_1 и Σ, затем на их основе вычислите θ, а потом, используя полученную модель, классифицируйте примеры из валидационной выборки. Убедитесь, что предсказания модели на валидационной выборке записаны в файл, указанный в коде.

Визуализируйте значения **валидационной** выборки на графике, в котором по горизонтальной оси идут значения x_1, а по вертикальной – x_2. Используйте разные символы, чтобы обозначать примеры разных классов. На этом же графике отобразите решающую границу, найденную ГДА (т.е. график прямой, соответствующей $p(y|x) = 0.5$).

Вопрос №6 [0.4 балла]

Сравните графики, полученные при выполнении подзадач 2 и 5 – для логистической регрессии и ГДА, соответственно. Несколькими предложениями прокомментируйте ваши наблюдения.

Вопрос №7 [1 балл]

Повторите обучение моделей и построение графиков из подзадач 2 и 5 для второй обучающей выборки из ds2_train.csv. Постройте аналогичные графики на валидационном множестве ds2_valid.csv. На каком датасете ГДА справляется хуже? Почему это может быть так?

Вопрос №8 [0.2 балла]

Какую трансформацию признаков можно было бы применить к исходным данным, чтобы улучшить результаты ГДА на той выборке, на которой она показывает неудовлетворительные результаты работы?