Задание №1.1. Регрессия Пуассона

В этом задании мы построим обобщенную линейную модель (Generalized Linear Model) иного класса, чем мы строили до этого, а именно *регрессию Пуассона*. В ОЛМ выбор распределения из экспоненциального семейства основан на типе рассматриваемой проблемы. Если мы решаем задачу классификации, то мы используем распределение из экспоненциального семейства, определенное на множестве дискретных классов (например, Бернулли или категориальное). Аналогично, если выход модели имеет вещественное значение, мы можем использовать нормальное распределение или распределение Лапласа, которые тоже относятся к экспоненциальному семейству. Иногда в задаче требуется прогнозировать некоторое *количество* чего-либо, например, количество электронных писем, ожидаемых в течение дня, или количество покупателей, которые могут войти в магазин в течение следующего часа, и т.д., на основе входных признаков (также называемых ковариатами). Как вы, наверное, помните, распределение вероятностей, заданное на множестве целых чисел и подходящее для подобного рода задач, есть распределение Пуассона, которое тоже относится к экспоненциальному семейству.

В следующих подзадачах мы начнем с того, что покажем, что распределение Пуассона относится к экспоненциальному семейству, выведем функциональную форму гипотезы, выведем правило обновления параметров для итерационного метода обучения модели и, наконец, используя предоставленную обучающую выборку, обучим реальную модель и сделаем прогнозы на тестовой выборке.

Вопрос №1

Рассмотрим распределение Пуассона с параметром λ:

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}.$$

Здесь y может принимать целые положительные значения. Покажите, что распределение Пуассона принадлежит семейству экспоненциальных распределений, явно записав выражения для $b(y)$, η, $T(y)$ и $a(\eta)$.

Вопрос №2

Предположим, что мы рассматриваем регрессию на основе ОЛМ модели с переменной отклика[1], моделируемой распределением Пуассона. Как будет выглядеть каноническая функция отклика[2] для этого класса распределений? Вы можете использовать тот факт, что математическое ожидание случайной величины, распределенной по Пуассону с параметром λ, есть λ.

Вопрос №3

Для обучающей выборки $\{(x^{(i)}, y^{(i)}); i = 1, \dots, m\}$ функция лог-правдоподобия для одного примера есть $\log \mathrm{p}(y^{(i)}|x^{(i)}; \theta)$. Путем нахождения частной производной этой функции по θ_j выпишите правило для обновления данного параметра в алгоритме *стохастического градиентного подъема*, который в дальнейшем может быть использован для обучения ОЛМ модели с переменной отклика y, распределенной по Пуассону.

---

[1] В терминологии ОЛМ *переменной отклика* (response variable) называется переменная, значение которой мы пытаемся спрогнозировать, т.е. в наших обозначениях это y.

[2] В терминологии ОЛМ *канонической функцией отклика* (canonical response function) называется отображение g, выражающее ожидаемое значение распределения как функцию от естественного параметра η (т.е. $g(\eta) = E[T(y); \eta]$). В частности, для нормального класса распределений это будет функция тождественности, а для распределения Бернулли – логистическая функция.

Вопрос №4 (задача на программирование)

Предположим, мы хотим спрогнозировать среднюю за день величину трафика некоторого веб-сайта. Владельцы данного веб-сайта собрали статистику его посещения за некоторый период времени, включающую в себя некоторые признаки, которые, по их мнению, будут полезны для прогнозирования количества посетителей в день. Выборка разбита на обучающий и тестовый наборы, а начальный код доступен в файле poisson.py.

Мы применим регрессию Пуассона для моделирования количества посетителей в день. Обратите внимание, что применение регрессии Пуассона, в частности, предполагает, что данные соответствуют распределению Пуассона, естественным параметром которого является линейная комбинация входных признаков (т.е. $\eta = \theta^T x$). Реализуйте регрессию Пуассона и используйте метод градиентного восхождения, чтобы максимизировать лог-правдоподобие θ. В качестве критерия остановки проверяйте, что норма вектора изменения параметров стала меньше какого-нибудь маленького порога, например 10^{-5}.

Далее, используя обученную модель, спрогнозируйте ожидаемые значения количества посещений для тестовой выборки и создайте точечную диаграмму, отображающую истинные и прогнозируемые значения переменной отклика. На диаграмме пусть на оси X будут представлены истинные количества посещений, а на оси y – соответствующие им прогнозы. Обратите внимание, что истинные значения являются целыми, в то время как ожидаемые значения в общем случае будут действительными числами.