Листок по теме «Линейная регрессия»

Кантор Виктор

16 февраля 2018 г.

1 Формулировка задачи

Пусть, для начала, есть векторы признаков $x_1, \cdots, x_l$ для l объектов. На них же известны значения прогнозируемой величины $y_1, \cdots, y_l$. Будем пытаться аппроксимировать зависимость этой величины от признаков линейной: $y \approx \hat{y} =< w, x >= w^T x$, где w - вектор параметров, который мы хотим настроить на обучающей выборке. Здесь мы также имеем ввиду, что могли бы рассматривать более широкий класс зависимостей $\hat{y} =< w, x > +w_0$, но мысленно добавив к x фиктивный признак $x_0 = 1$ (равный единице для всех объектов), мы всегда сведем задачу к предыдущей.

Пусть, далее, мы хотим минимизировать суммарную потерю на обучающей выборке:

$$Q(w) = \sum_{i=1}^{l} L(y_i, \hat{y}_i) \to \min_w$$

И пусть, кроме того, функция потерь квадратичная: $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$.

Наконец, пусть X – матрица признаков, строки которой соответствуют признаковым описаниям объектов $x_1, \cdots, x_l$ (т.е. в матрице l строк), вектором y будем обозначать вектор длины l с координатами y_i, вектором $\hat{y}$ – вектор длины l с координатами $\hat{y}_i$.

1. Как тогда выразить вектор $\hat{y}$ через w и X?
2. Как записать функционал $Q(w)$ через $\hat{y}$ и y? Что получится, если подставить выражение для $\hat{y}$?
3. Как выражается $\hat{y}$ через столбцы матрицы X? В каком линейном пространстве лежит $\hat{y}$ (в терминах линейных обочек набора векторов)?

2 Нормальное уравнение (normal equation)

1. Покажите справедливость следующего выражения:
$$\frac{\partial}{\partial x}(Ax + b)^T (Ax + b) = 2A^T (Ax + b)$$
2. Приравняв производную $Q(w)$ по вектору w к нулю, выразите вектор весов w в точке минимума через X и y. Полученное выражение носит название $NormalEquation$. Запомните, как оно получается — пригодится на экзамене.

3 Геометрическая интерпретация

1. Вспомните, что у вас получалось в последнем вопросе из раздела "формулировка задачи". Мы хотим минимизировать сумму квадратов отклонений, значит мы хотим для заданного вектора y найти в этом линейном пространстве точку $\hat{y}$, которая будет самой близкой к вектору y. Нетрудно сообразить, что такой точкой будет проекция y на линейное пространство, в котором должен жить $\hat{y}$. Чему в этом случае равно скалярное произведение $< x^{(j)}, \hat{y} - y >$, где $x^{(j)}$ - j-ый столбец матрицы X?

2. Исходя из предыдущего пункта, чему будет равно произведение

$$X^T(\hat{y} - y) \ ?$$

3. Закончите геометрический вывод нормального уравнения, подставив сюда выражение для $\hat{y}$ и выразив w.

4 Вероятностная интерпретация

4. Вспомните формулу плотности одномерного нормального распределения и как оцениваются параметры распределения методом максимального правдоподобия.

5. Предположим, что для каждого объекта x_i наблюдаемое значение величины y_i распределено нормально с матожиданием $\hat{y}_i$ (некоторым "истинным значением"в рамках линейной модели, которое мы хотим оценить) и дисперсией σ^2, одинаковой для всех i. Напишите минимизируемый функционал в задаче оценки $\hat{y}_i$ по методу максимального правдоподобия.

6. Покажите, что метод максимумального правдоподобия в этом случае приводит к минимизации суммы квадратов отклонений.

7. Теперь вы видите, что есть связь между функцией потерь и нашим преставлением о распределении "правильных ответов". А какое распределение привело бы к минимизации суммы модулей отклонений?

5 $l2$-регуляризация: гребневая регрессия (ridge regression)

8. Давайте теперь добавим к функционалу $Q(w)$ штрафное слагаемое $\tau||w||^2$. Дифференцируя по вектору w получите новое выражение для его оптимального значения.

9. Покажите, что это это дает тот же результат, что и добавление к матрице X^TX единичной матрицы I, умноженной на τ (с последующим применением стандартной формулы).

10. Модификация матрицы из предыдущего пункта изменяет собственные числа матрицы, меняя число обусловленности и позволяя обратить $X^T X$, если до модификации обращение было неустойчиво. Но при этом сохраняются собственные векторы матрицы. Покажите, как меняются минимальное и максимальное собственное число и число обусловленности матрицы. А затем, — что собственные векторы остаются теми же.

6 $l1$-регуляризация: лассо Тибширани (LASSO)

11. Применив теорему Куна-Таккера, покажите, что добавление ограничения $||w||_{l1} < r$ равносильно добавлению штрафного слагаемого $\tau||w||_{l1}$ в функционал $Q(w)$ для некоторого τ.

12. Попробуйте с помощью теоремы или из геометрических соображений объяснить, почему минимум попадает в "угловые точки"шаров $l1$-нормы (т.е. почему зануляются некоторые коэффициенты w), и почему с увеличением τ нулей будет становиться больше.

7 Дополнительные вопросы

13. Покажите, что если предварительно центрировать выборку, параметр сдвига w_0 получится равным нулю.

14. Представьте теперь, что вы пытаетесь восстановить пряму по известным ее точкам на изображении. Здесь вам уже захочется минимизировать не сумму квадратов отклонений по y, а сумму квадратов расстояний от известных точек до прямой. Как в этом случае будут выглядеть формулы для коэффициентов искомой прямой $w_x x + w_y y + w_0 = 0$?

15. Придумайте изящный ответ на предыдущий вопрос с помощью метода главных компонент.