**Учебный материал 2**

**Регрессия**

С помощью линейной регрессии мы стремимся подогнать линию (или гиперплоскость в более высоких измерениях) для разброса данных. В этом разделе мы опишем основные понятия, лежащие в основе этой процедуры.
Данные для задач регрессии поступают в виде обучающего набора P пар ввода / вывода наблюдений:

$$\{\left(x\_{1},y\_{1}\right),\left(x\_{2},y\_{2}\right),…, (x\_{p},y\_{p}) \}$$

Во многих случаях регрессии входные данные для задач регрессии являются скалярными (выход всегда будет считаться скалярным), и, следовательно, задача линейной регрессии выполняет подгонку линии к соответствующему разбросу точек данных в двумерном пространстве. В общем, однако, каждый входной xp может быть вектором столбца длины N.

В этом случае задача линейной регрессии аналогична задаче подгонки гиперплоскости к разбросу точек в N+1-мерном пространстве. В случае скалярного ввода, для подгонки линии к данным требуется определить наклон w и смещение (или «y-пересечение») b, чтобы между входными / выходными данными сохранялась приблизительная линейная зависимость,

 

Обратите внимание, что мы использовали знак приблизительно, потому что мы не можем быть уверены, что все данные полностью лежат в одной строке. В более общем случае, когда входное измерение равно N ≥ 1, тогда мы имеем смещение и N связанных весов, чтобы правильно настроиться, чтобы соответствовать гиперплоскости. Аналогичным образом, линейная зависимость в более общем случае определяется как

Элементы входного вектора xp называются входными признаками задачи регрессии. Например, описанные данные о студенческом долге имеют только одну особенность: год. И наоборот, в данных о темпах роста ВВП первый элемент вектора входных признаков может содержать уровень безработицы признаков.



**Функция стоимости наименьших квадратов для линейной регрессии**
Чтобы найти параметры гиперплоскости, которая наилучшим образом соответствует набору данных регрессии, обычно сначала формируют функцию стоимости наименьших квадратов. Для данного набора параметров (b, w) эта функция стоимости вычисляет общую квадратичную ошибку между связанной гиперплоскостью и данными, давая хорошую меру того, насколько хорошо конкретная линейная модель соответствует набору данных. Естественно, тогда лучшая подгонка гиперплоскости — это та, параметры которой сводят эту ошибку к минимуму. Поскольку мы стремимся сохранить систему уравнений как можно лучше, чтобы сформировать желаемую стоимость, мы просто возводим в квадрат разницу (или ошибку) между линейной моделью $b+x\_{p}^{T}\*w$ и соответствующим выходным значением yp по всему набору данных. Это дает функцию стоимости наименьших квадратов.



Имитация двумерного обучающего набора данных вместе с линией (в пурпурном цвете), подходящей к данным, с использованием структуры наименьших квадратов, которая направлена на восстановление линии, которая минимизирует общую квадратную длину штриховых полос ошибок.



Мы, конечно, хотим найти пару параметров (b, w), которая обеспечивает небольшое значение для g (b, w), поскольку чем больше это значение, тем больше квадрат ошибки между соответствующей линейной моделью и данными, и, следовательно, чем хуже мы представлять данные. Поэтому мы стремимся минимизировать g над вектором смещения и веса, чтобы восстановить лучшую пару (b, w), которая записана формально как

**

**Минимизация функции стоимости наименьших квадратов**
Для выполнения расчетов сначала будет удобно использовать следующие более компактные обозначения:



С помощью этих обозначений мы можем переписать функцию стоимости, показанную в единичном векторе параметров, как



Чтобы вычислить градиент этой стоимости, мы просто применяем правило цепочки из исчисления, которое дает



Используя это, мы можем выполнить градиентный спуск, чтобы минимизировать стоимость. Однако в этом (редком) случае мы можем фактически решить систему первого порядка напрямую, чтобы восстановить глобальный минимум. Установка градиента выше нуля и решение для w дает систему линейных уравнений



**Прогнозирование значения новых входных данных**
С помощью оптимальных параметров $(b^{\*},w^{\*})$, найденных путем минимизации стоимости наименьших квадратов, мы можем предсказать выход y новый для новой входной функции x new, просто подключив новый вход в настроенную линейную модель и оценив соответствующую вывод как



Это иллюстрировано в виде набора игрушечных данных для случая, когда N = 1





Набор данных моделируемой регрессии, где связь между входным элементом x и выходным y не является линейной. Однако, поскольку мы можем визуализировать этот набор данных, мы можем видеть, что существует четко структурированная нелинейная связь между его входом и выходом. Наши знания в этом случае, основанные на нашей способности визуализировать данные, позволяют нам разработать новую функцию для данных и сформулировать соответствующую функцию (показана здесь черным пунктиром), которая, по-видимому, генерирует данные.