Московский Государственный Университет имени М.В.Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Курсовая работа на тему:

Обзор методов линейного регрессионного анализа

Остапец Андрей Александрович
группа 317

Москва, 2012

Оглавление

Введение

Термин **«регрессия»** был введён Фрэнсисом Гальтоном 1886 году.

В исследовании Ф. Гальтона был измерен рост 205 отцов и 928 их взрослых детей. При этом, если за Y взять рост ребенка, а за X рост родителя, уравнение регрессии, связывающее рост ребенка с ростом родителей, имеет вид:

$$Y_i = Y_{cp} - \frac{2}{3}(X_i - X_{cp}), \qquad (*)$$

где Y_{cp} , X_{cp} - средние по всей выборке испытуемых.

Таким образом, зная величины средних по всей выборке и рост одного из родителей X_i по формуле (*) можно подсчитать величину Y_i, т.е. рост ребенка.

Таблица 1: **Таблица Ф. Гальтона, иллюстрирующая наличие зависимости между ростом родителей и их детей**

	Рост детей								
Рост родителей	$\leqslant 60.7$	62.7	64.7	66.7	68.7	70.7	72.7	$\geqslant 74.7$	Всего
$\geqslant 74$							4		4
72			1	4	11	17	20	6	62
70	1	2	21	48	83	66	22	8	251
68	1	15	56	130	148	69	11		430
66	1	15	19	56	41	11	1		144
$\leqslant 64$	2	7	10	14	4				37
Всего	5	39	107	255	387	163	58	14	928

Гальтон назвал это явление **«регрессией к середине»**, т.е к среднему значению в популяции. В настоящее время термин, возникший в частной прикладной задаче, закрепился за широким классом методов восстановления зависимостей.

В этих методах предполагается, что полученные данные состоят из набора $X^l = (x_i, y_i)_{i=1}^l$, где y_i - **зависимая** переменная, а вектор x_i - вектор **независимых** переменных $x_i = (x_i^1, \ldots, x_i^p)$.

Регрессионный анализ - это моделирование данных и исследования их свойств. Коэффициенты $w = (w_1, ..., w_p)$, настраиваются таким образом, чтобы среднеквадратичная ошибка была минимальна:

$$S = \sum_{i=1}^{l} (f(w, x_i) - y_i)^2 \to min$$

Регрессионный анализ используется для прогнозирования, выявления скрытых взаимосвязей в данных, анализа временных рядов, тестирования гипотез.

Глава 1

Разложение средней ошибки на дисперсию и смещение

Задачу обучения по прецедентам при $\mathbb{Y} = \mathbb{R}$ принято называть задачей **восстановления регрессии**. Задано пространство объектов $\mathbb{X}$ и множество возможных ответов $\mathbb{Y}$. Существует неизвестная целевая зависимость $y^* : \mathbb{X} \to \mathbb{Y}$, значения которой известны только на объектах обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$, $y_i = y^*(x_i)$. Требуется построить алгоритм, который в данной задаче принято называть «функцией регрессии» $f : X \to Y$, аппроксимирующий целевую зависимость y^*.

Предположим существование неизвестного вероятностного распределения на множестве $X \times Y$ с плотностью $p(x, y)$, из которого случайно и независимо выбираются l наблюдений $X^l = (x_i, y_i)_{i=1}^l$. Такие выборки называются простыми или случайными одинаково распределёнными (independent identically distributed, i.i.d.).

Введём **функцию потерь (loss function)** $L(y; f(x))$, которая «наказывает за ошибки»; для поставленной задачи логично взять **квадратичную функцию потерь:**

$$L(y, f(x)) = (y - f(x))^2$$

Тогда каждой f можно сопоставить **ожидаемую ошибку предсказания (expected prediction error)**:

$$EPE(f) = E(y - f(x))^2 = \iint (y - f(x))^2 p(x, y)\, dxdy$$

И теперь самая хорошая функция предсказания $\hat{f}$ - это та, которая минимизирует EPE(f). Можно представить ожидаемую ошибку предсказания следующим образом:

$$EPE(f) = E_x E_{y|x}[(y - f(x))^2 | x],$$

и, значит, можно минимизировать EPE поточечно:

$$\hat{f}(x) = arg\min_c E_{y|x'}[(y - c)^2 | x' = x],$$

и, в итоге, получить следующий ответ:

$$\hat{f}(x) = E_{y|x'}(y | x' = x).$$

Это решение называется **функцией регрессии** и является наилучшим предсказанием y в любой точке x.

Подсчитаем ожидаемую ошибку и перепишем её в другой форме:

$$E[L] = E[(y - f(x))^2] = E[(y - E[y|x] + E[y|x] - f(x))^2] =$$
$$= \int (f(x) - E[y|x])^2 p(x)\, dx + \int (E[y|x] - y)^2 p(x, y)\, dxdy,$$

потому что

$$\int (f(x) - E[y|x])(E[y|x] - y) p(x, y)\, dxdy = 0.$$

Эта форма записи - разложение на **bias-variance** и **noise**:

$$E[L] = \underbrace{\int (f(x) - E[y|x])^2 p(x)\, dx}_{\text{bias-variance}} + \underbrace{\int (E[y|x] - y)^2 p(x, y)\, dxdy}_{\text{noise}}$$

Отсюда, кстати, тоже сразу видно, что от $f(x)$ зависит только первый член, и он минимизируется, когда $f(x) = \hat{f}(x) = E[y|x]$.

$\int (E[y|x] - y)^2 p(x, y)\, dxdy$ - это просто свойство данных, дисперсия шума.

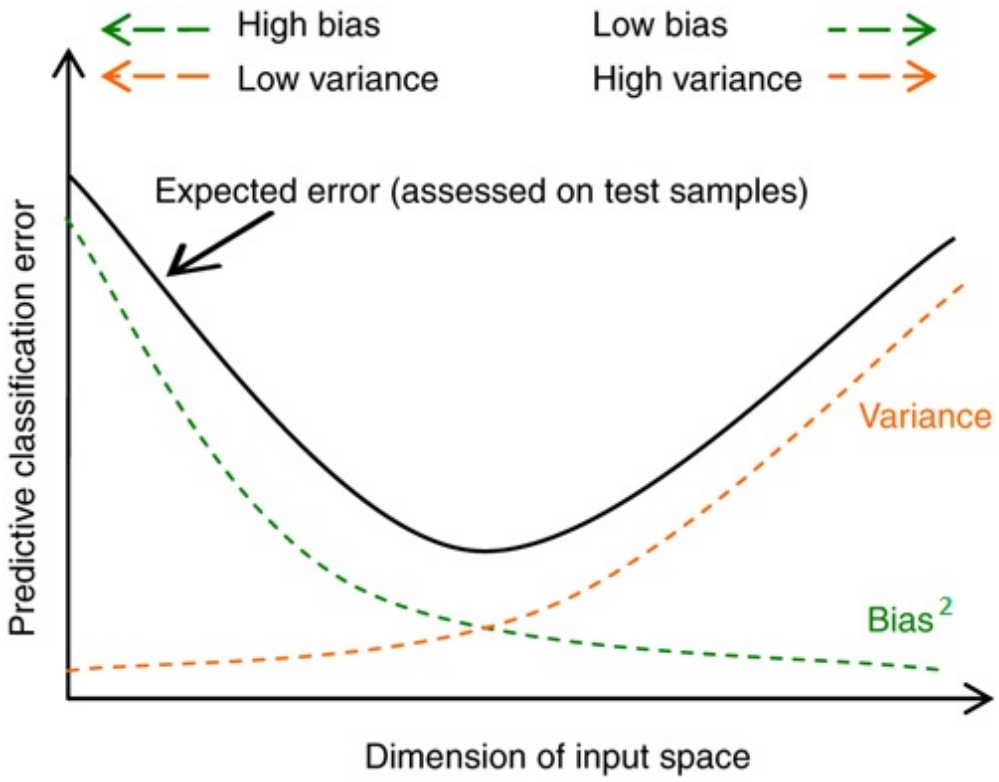


Рис. 1.1: Пример разложения ошибки на дисперсию и смещение

В реальности встречаются выборки только конечного размера, поэтому точно вычислить величину $\hat{f}(x) = E[y|x]$ невозможно.

Предположим, что выборка берётся по распределению $p(x; y)$ - т.е. фактически мы можем проводить эксперименты только такого вида:

- взяли выборку D из N точек по распределению $p(x; y)$;
- подсчитали нашу регрессию;
- получили новую функцию предсказания $f(x; D)$.

Разные выборки будут приводить к разным функциям предсказания. Попробуем усреднить наши функции по выборкам.

Наш первый член в ожидаемой ошибке выглядел как $(f(x) - \hat{f}(x))^2$, а теперь будет $(f(x; D) - \hat{f}(x))^2$, и его можно усреднить по D, применив такой трюк:

$$(f(x; D) - \hat{f}(x))^2 = (f(x; D) - E_D[f(x; D)] + E_D[f(x; D)] - \hat{f}(x))^2$$

и в ожидании получится

$$E_D[(f(x; D) - \hat{f}(x))^2] = E_D[(f(x; D) - E_D[f(x; D)])^2] + (E_D[f(x; D] - \hat{f}(x))^2.$$

Теперь мы можем представить ожиадемую потерю следующим образом:

$$\textbf{Expected loss} = \textbf{(bias)}^2 + \textbf{variance} + \textbf{noise}, \text{ где}$$

$(\text{bias})^2 = (E_D[f(x; D] - \hat{f}(x))^2,$

$\text{variance} = E_D[(f(x; D) - E_D[f(x; D)])^2],$

$\text{noise} = \int (E[y|x] - y)^2 p(x, y)\, dxdy.$

Глава 2

Линейная регрессионная модель и метод наименьших квадратов

Пусть объект x задается своим признаковым описанием

$$x = (x^0, x^1, x^2, ..., x^p) \in \mathbb{R}^{p+1}$$

Будем предполагать, что

$$\forall x : x^0 \equiv 1$$

(всегда можно пополнить признаковое пространство таким фиктивным признаком).
y - зависимая переменная, значение которой необходимо предсказать.

Линейная регрессионная модель имеет вид:

$$f(w, x_i) = w_0 + \sum_{j=1}^{p} x_i^j w_j = x_i^T w$$

Таким образом, по вектору входов $x = (x^0 \equiv 1, x^1, ..., x^p)$ мы будем предсказывать выход y как

$$\hat{y}(x) = \hat{w_0} + \sum_{j=1}^{p} x^j \hat{w_j} = x^T \hat{w}$$

Линейная регрессия предполагает, что функция f зависит от параметров w линейно. При этом линейная зависимость от свободной переменной x необязательна.

Пусть у нас есть тренировочная выборка $(x_1, y_1)...(x_N, y_N)$ и мы хотим оценить параметры w. Каждый вектор x_i есть вектор измеренных значений $x_i = (x_i^0 \equiv 1, x_i^1, x_i^2, ..., x_i^p)$
Самый популярный метод настройки параметров - это **метод наименьших квадратов**, где коэффициенты $w = (w_0, w_1, \ldots, w_p)$ подбираются так, чтобы минимизировать остаточную сумму квадратов(**residual sum of squares**):

$$RSS(w) = \sum_{i=1}^{N} (x_i^T w - y_i)^2 \tag{2.1}$$

Как же минимизировать это выражение?

Введём матричные обозначения:

$X = (x_i^j)_{N \times (p+1)}$ - матрица объекты–признаки;
$y = (y_1, \ldots, y_N)^T$ - целевой вектор;
$w = (w_0, w_1, \ldots, w_p)$ - вектор параметров.

В матричных обозначениях формула (2.1) принимает вид

$$RSS(w) = \|y - Xw\|^2$$

Запишем необходимое условие минимума функционала в матричном виде:

$$\frac{\partial RSS}{\partial w}(w) = 2X^T(y - Xw) = 0,$$

откуда следует, что $X^T Xw = X^T y$. Эта система линейных уравнений относительно w называется нормальной системой для задачи наименьших квадратов. Если матрица $X^T X$ невырождена, то решением нормальной системы является вектор

$$\hat{w} = (X^T X)^{-1} X^T y$$

Замечание: Матрица $X^+ = (X^T X)^{-1} X^T$ называется **псевдообратной матрицей Мура–Пенроуза** матрицы X; это обобщение понятия обратной матрицы на неквадратные матрицы

2.1 Сингулярное разложение

Произвольную $l \times n$-матрицу ранга n можно представить в виде **сингулярного разложения (singular value decomposition, SVD)**

$$X = VDU^T$$

обладающего рядом хороших свойств:

1. $n \times n$-матрица D диагональна, $D = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$, где $\lambda_1, ..., \lambda_n$ - общие ненулевые собственные значения матриц $X^T X$ и XX^T.
2. $l \times n$-матрица $V = (v_1, ..., v_n)$ ортогональна, $V^T V = I_n$, столбцы v_j являются собственными векторами матрицы XX^T, соответствующими $\lambda_1, ..., \lambda_n$.
3. $n \times n$-матрица $U = (u_1, ..., u_n)$ ортогональна, $U^T U = I_n$, столбцы u_j являются собственными векторами матрицы $X^T X$, соответствующими $\lambda_1, ..., \lambda_n$.

Имея сингулярное разложение, легко записать псевдообратную матрицу:

$$X^+ = (UDV^T VDU^T)^{-1} UDV^T = UD^{-1}V^T = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^T$$

МНК-решение:

$$\hat{w} = (X^T X)^{-1} X^T y = X^+ y = UD^{-1}V^T y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y)$$

вектор $X\hat{w}$ - МНК-аппроксимацию целевого вектора y:

$$X\hat{w} = (VDU^T)UD^{-1}V^T y = VV^T y = \sum_{j=1}^{n} v_j (v_j^T y) \tag{2.2}$$

и норму вектора коэффициентов:

$$\|\hat{w}\|^2 = y^T VD^{-1}Y^T UD^{-1}V^T y = y^T VD^{-2}V^T y = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^T y)^2 \tag{2.3}$$

Итак, если есть сингулярное разложение, то обращать матрицы уже не нужно. Однако вычисление сингулярного разложения практически столь же трудоёмко, как и обращение. Эффективные численные алгоритмы, вычисляющие SVD, реализованы во многих стандартных математических пакетах.

Для демонстрации работы данного метода была взята следующая модельная задача. Взято 200 точек $(x_i)_{i=1}^{200}$, которые делят отрезок [0,10] на равные части. Для каждого x_i соответствующее значение y_i вычисляется следующим образом:

$$y = x + 1.5 \cdot \sin(x) - 0.5 \cdot \cos(x) + h,$$

где случайная величина $h \in N(0, 0.5)$ является умышленно добавленным шумом

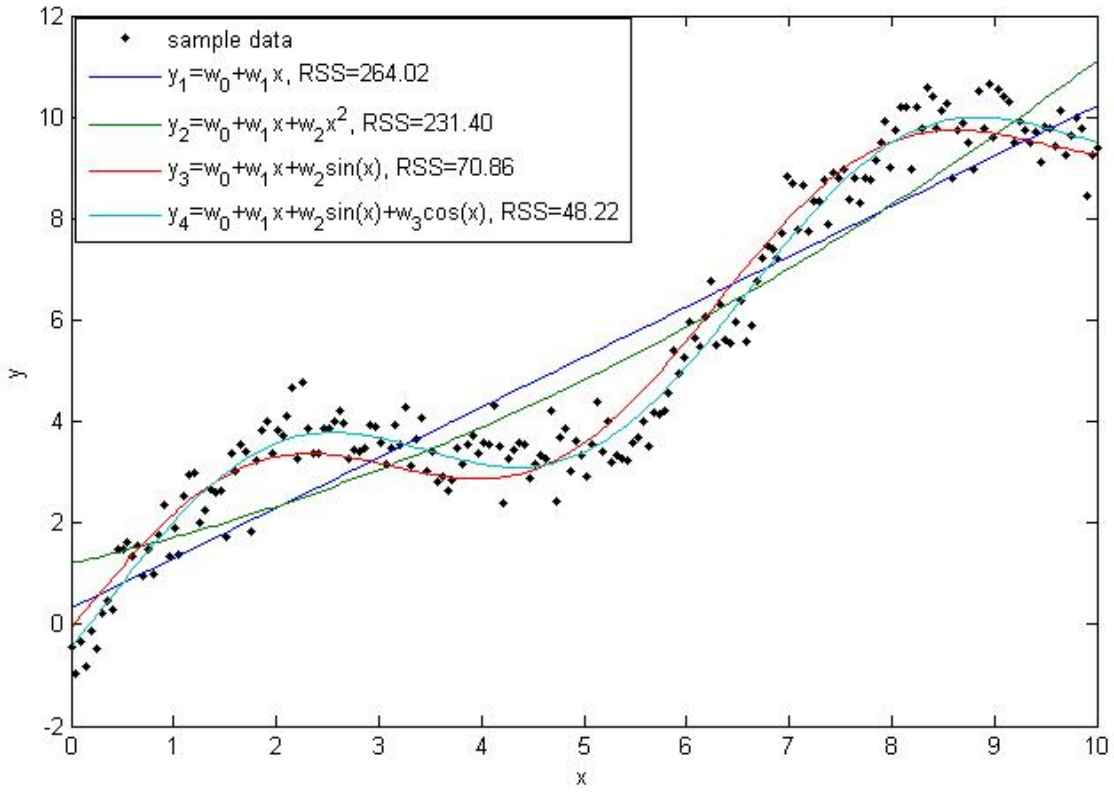


Рис. 2.1: Линейные регрессионные модели, построенные методом наименьших квадратов

2.2 Теорема Гаусса-Маркова

Рассматривается модель парной регрессии, в которой наблюдения y связаны с x следующей зависимостью:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

На основе выборки из n прецедентов оценивается уравнение регрессии: $\hat{y_i} = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Теорема Гаусса-Маркова гласит, что если данные обладают следующими свойствами:

- Регрессионная модель является линейной относительно параметров и корректно специфицирована(т.е, адекватна устройству данных).
- Все x_i детерминированы и не все равны между собой;
- Ошибки не носят систематического характера, то есть $E(\epsilon_i) = 0, \forall i$
- Дисперсия ошибок одинакова и равна некоторой σ^2 (гомоскедастичность).
- Ошибки независимы, то есть $Cov(\epsilon_i, \epsilon_i) = 0, \forall i, j$

Тогда оценка параметров модели, которая получена **методом наименьших квадратов**, является:

- Линейной
- Несмещенной(математическое ожидание параметра равно истинному значению параметра)
- Эффективной в классе линейных несмещенных оценок.(BLUE - best linear unbiased estimators)

Пояснение:

Эффективность является мерой отклонения от истинного значения. Для несмещенной оценки это эквивалентно вариации:

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \underbrace{var(\hat{\theta})}_{\text{дисперсия}} + \underbrace{(E(\hat{\theta}) - \theta)^2}_{\text{смещение}}$$

Теорема Гаусса-Маркова утверждает, что любая другая линейная несмещенная оценка будет иметь большую дисперсию, чем МНК-оценка.

Глава 3

Отбор подмножеств

Как бы не был прост и логичен метод наименьших квадратов, существует 2 причины, по которым его результаты не являются удовлетворительными:

- Первой является **плохая предсказательная сила**: МНК часто имеет низкое смещение, но большую дисперсию. Для повышения точности некоторые коэффициенты можно уменьшить или положить равными нулю. В результате, мы можем немного увеличить смещение, но зато значительно уменьшить дисперсию, и, в результате, точность увеличивается.
- Вторая причина - **интерпретируемость**. Когда в задаче существует большое число признаков, хотелось бы ограничиться маленьким подмножеством, которое дает наилучший результат. И мы готовы пожертвовать маленькими деталями, чтобы получить простую и ясную модель.

В методах отбора подмножеств в модели остается лишь какое-то подмножество признаков, а остальные исключаются из модели. Регрессией наименьших квадратов оцениваются параметры тех признаков, которые остались в модели. Ниже рассматриваются различные стратегии отбора подмножеств.

3.1 Отбор наилучшего подмножества

Пусть у нас есть тренировочная выборка $X^l = (x_i, y_i)_{i=1}^l$. Каждый вектор x_i есть вектор измеренных значений $x_i = (x_0 \equiv 1, x_i^1, x_i^2, ..., x_i^p)$

Отбор наилучшего подмножества(Best Subset) находит для каждого k $\in$ {1,2,. . . ,p+1} подмножество из k входных переменных, которые дает наименьшую остаточную сумму квадратов:

$$RSS(w, k) = \sum_{i=1}^{N}(y_i - \sum_{j=1}^{k} x_i^j w_j)^2$$

Эта процедура занимает слишком много времени, поскольку идет полный перебор всех возможных подмножеств. Ниже предемонстрируем работу данного алгоритма на примере.

Взято 100 точек $(x_i)_{i=1}^{100}$, где $x_i = (x_i^1, ..., x_i^{10})$, каждый из 10 признаков равномерно распределен на отрезке [0,1]. Для каждого x_i соответствующее значение y_i вычисляется по формуле:

$$y_i = x_i^1 - x_i^2 + 0.5x_i^3 - 0.5x_i^4 + 0.5 \cdot h,$$

где h является случайной величиной из стандартного нормального распределения.

Можно заметить, что необязательно лучшее подмножество размера 2 включает ту переменную, которая была лучшей в подмножестве размера 1.

Есть целый ряд критериев, которые используются для выбора подмножества. Обычно выбирается подмножество наименьшей мощности, остаточная сумма которого, не сильно отличается от остаточной суммы полного множества признаков.

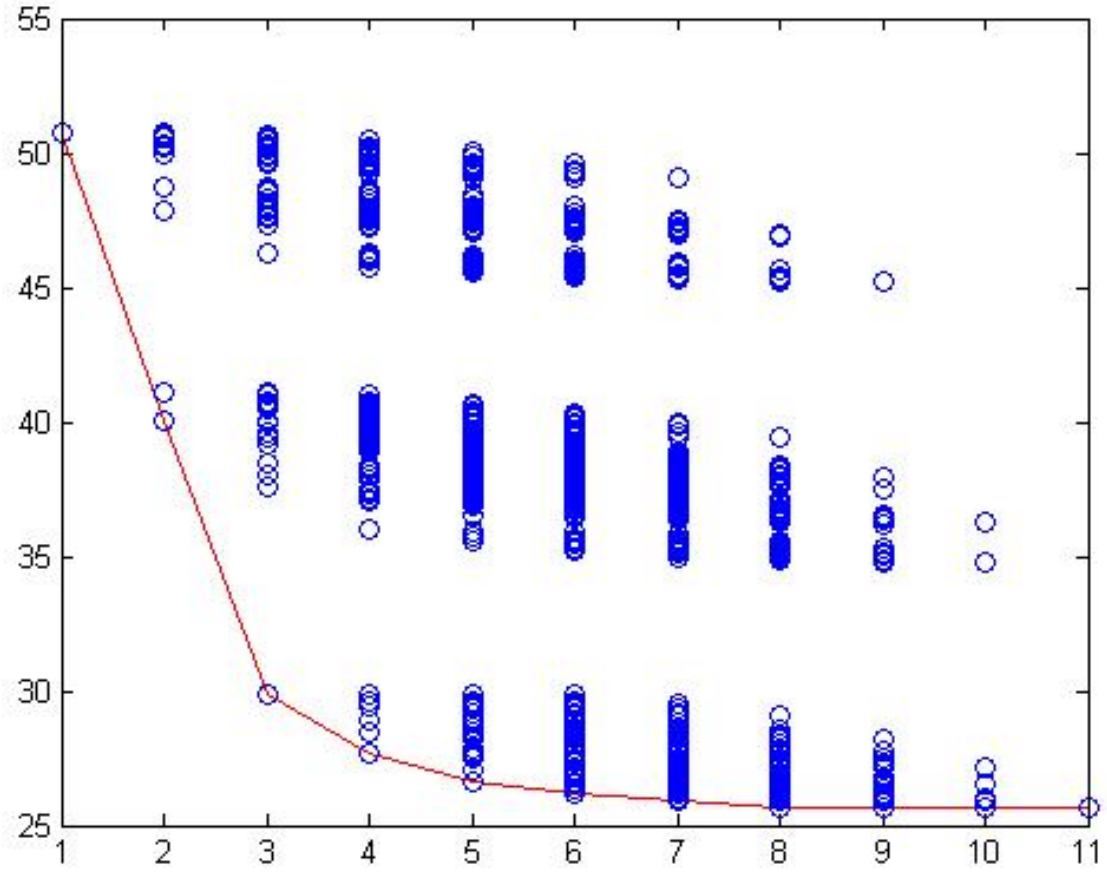


Рис. 3.1: Пример работы алгоритма Best Subset

3.2 Прямая и обратная шаговая регрессия

Существуют различные эвристические подходы к задаче выбора подмножества.
Одна из таких эвристик - **прямая шаговая регрессия(Forward Stepwise).**
В ней мы добавляет на каждом шаге предиктор, который максимально уменьшает ошибку.

Недостатки:

- Это жадный алгоритм, каждое предыдущее подмножество вложено в последующие подмножества.
- Его результат может оказаться хуже результата алгоритма отбора наилучшего подмножества.

Достоинства:

- Вычислимость. При большом количестве предикторов (больше 50), алгоритм отбора наилучшего подмножества должен перебрать $> 10^{15}$ подмножеств, что может занять много времени. Прямая шаговая регрессия перебирает не более чем $(p+1)^2$ подмножеств.

- При отборе наилучшего подмножества мы получаем алгоритм с низким смещением, но возможно большой дисперсией. Прямая шаговая регрессия обладает ограниченным поиском и будет иметь низкую дисперсию, но возможно большое смещение.

В **обратной шаговой регрессии(Backward Stepwise)** мы действуем наоборот: начинаем с полного набора признаков и на каждом шаге убираем предиктор, который оказывает меньше всего влияния на ошибку. Этот метод имеет абсолютно такие же достоинства и недостатки, что и предыдущий. Эти два метода практически всегда показывают схожие результаты.
Поэтому на практике используется гибридная стратегия, которая включает оба метода и называется **шаговой регрессией**.

На каждом шаге делаются оба действия: находится несколько кандидатов, которые включаются в модель, а затем несколько признаков удаляются.

На рисунке 3.2 представлены результаты работы всех вышеперечисленных алгоритмов.

Взято 100 точек $(x_i)_{i=1}^{200}$, где $x_i = (x_i^1, ..., x_i^{20})$, каждый из 20 признаков равномерно распределен на отрезке [0,1]. Для каждого x_i ответ y_i вычисляется по формуле:

$$y_i = x_i^1 - x_i^2 + 0.75x_i^3 - 0.75x_i^4 + 0.5x_i^5 - 0.5x_i^6 + 0.25x_i^7 - 0.25x_i^8 + 0.2 \cdot h,$$

где h является случайной величиной из стандартного нормального распределения.

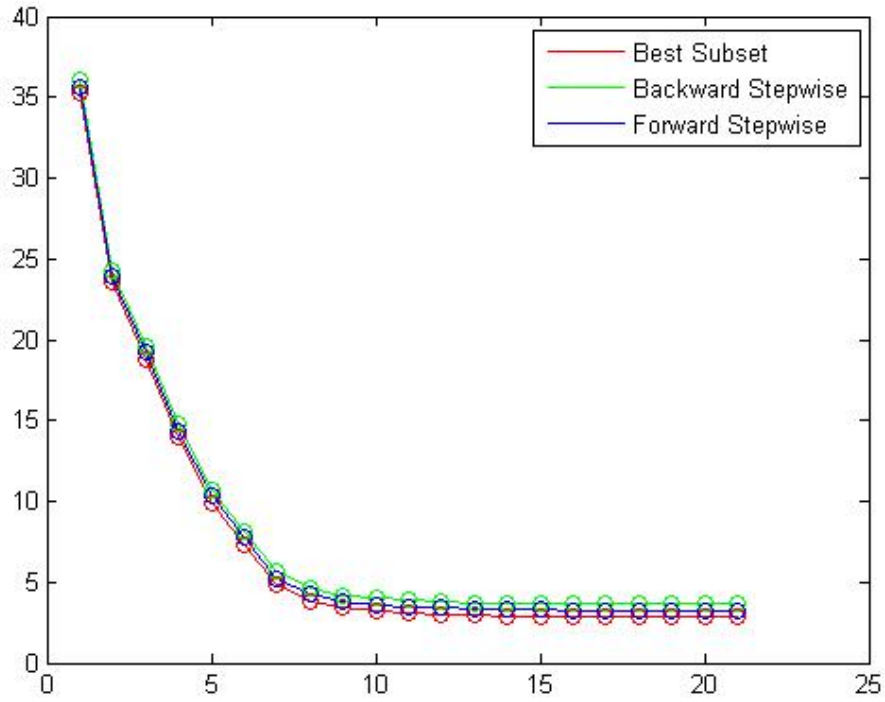


Рис. 3.2: Сравнение работы алгоритмов отбора признаков

Глава 4

Методы, основанные на сжатии коэффициентов

Используя в модели лишь какое-то подмножество признаков и отбрасывая остальные, методы отбора подмножества производят модели, которые являются интерпретируемыми и возможно имеют более низкую ошибку, чем полная модель.
Однако, поскольку данные методы являются дискретными - переменная либо включается в подмножество, либо отбрасывается - они часто обладают высокой дисперсией. Методы сжатия коэффициентов на каждом шаге корректируют веса, поэтому обладают более низкой дисперсией.

4.1 Гребневая регрессия

Гребневая регрессия производит сжатие коэффициентов путем добавления к (2.1) регуляризатора, штрафующего большие значения нормы вектора весов $\|w\|$. Коэффициенты гребневой регрессии подбираются решая задачу минимизации остаточной суммы квадратов:

$$RSS(w, \alpha) = \|y - Xw\|^2 + \alpha \|w\|^2 ,$$

где α - неотрицательный параметр регуляризации.

Приравнивая нулю производную $RSS(w, \alpha)$ по параметру α, находим:

$$\hat{w}_\alpha = (X^T X + \alpha I_n)^{-1} X^T y. \tag{4.1}$$

Таким образом, перед обращением матрицы к ней добавляется «гребень» - диагональная матрица αI_n. Отсюда и название метода - **гребневая регрессия (ridge regression).** После этого все её собственные значения становятся больше на величину α, а собственные векторы не изменяются. В результате матрица превращается в хорошо обусловленную, оставаясь «похожей» на исходную. Это и есть вторая причина появления гребневой регрессии - попытка решить проблему мультиколлинеарности.

4.1.1 Проблема мультиколлинеарности

Если ковариационная матрица $\Sigma = X^T X$ имеет неполный ранг, то её обращение невозможно. Тогда приходится выкидывать линейно зависимые признаки или применять другие методы. В реальности чаще встречается **проблема мультиколлинеарности** - когда матрица Σ имеет полный ранг, но близка к какой-то матрице неполного ранга. Тогда можно сказать, что Σ - матрица неполного псевдоранга или что она является плохо обусловленной. Признаком мультиколлинеарности является наличие у матрицы Σ собственных значений, близких к нулю.

Число обусловленности матрицы Σ есть

$$\mu(\Sigma) = \|\Sigma\| \left\|\Sigma^{-1}\right\| = \frac{\max\limits_{u:\|u\|=1} \|\Sigma u\|}{\min\limits_{u:\|u\|=1} \|\Sigma u\|} = \frac{\lambda_{max}}{\lambda_{min}},$$

где λ_{max} и λ_{min} - максимальное и минимальное собственные значения матрицы Σ, где все нормы евклидовы. Матрица считается плохо обусловленной, если $\mu(\Sigma) \gtrsim 10^2 ... 10^4$ Обращение такой матрицы неустойчиво.

При использовании метода наименьших квадратов в формуле (2.3) близкие к нулю собственные значения оказываются в знаменателе, в результате растет норма вектора коэффициентов $\hat{w}$, появляются большие по абсолютной величине положительные и отрицательные коэффициенты. МНК-решение становится неустойчивым - малые погрешности в измерении признаков или ответов на объектах обучающей выборки могут существенно изменить вектор решения $\hat{w}$. Мультиколлинеарность вызывает неустойчивость и переобучение, а также и неинтерпретируемость коэффициентов, так как по абсолютной величине коэффиецента w_j становится невозможно судить о степени важности j-го признака.

4.1.2 Гребневая регрессия и сингулярное разложение

Выразим решение (4.1) через сингулярное разложение:

$$\hat{w}_\alpha = (UD^2U^T + \alpha I_n)^{-1} UDV^T y = U(D^2 + \alpha I_n)^{-1} DV^T y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \alpha} u_j (v_j^T y).$$

Теперь найдём МНК-аппроксимацию целевого вектора y:

$$X\hat{w}_\alpha = VDU^T \hat{w}_\alpha = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \alpha} v_j (v_j^T y).$$

Как и прежде в (2.2), аппроксимация методом наименьших квадратов представляется в виде разложения вектора y по базису собственных векторов матрицы XX^T. Только теперь проекции на собственные векторы сокращаются, умножаясь на $\frac{\sqrt{\lambda_j}}{\lambda_j + \alpha} \in (0, 1)$. В сравнении с (2.3) уменьшается и норма вектора коэффициентов:

$$\|\hat{w}_\alpha\|^2 = \left\|D^2(D^2 + \alpha I_n)^{-1} D^{-1} V^T y\right\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j + \alpha} (v_j^T y)^2 < \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^T y)^2 = \|\hat{w}\|^2$$

Произведем **центрирование данных:**

- Каждое x_i^j заменяется на $x_i^j - \bar{x^j}$ $(i = 1, 2, ..., n; j = 1, 2, ..., p)$
- В качестве w_0 выбирается $\bar{y}$, где

$$\bar{x^j} = \frac{1}{n} \sum_{i=1}^{n} x_i^j, \quad \bar{y} = \sum_{i=1}^{n} y_i.$$

Пусть центрирование уже было осуществлено центрирование, следовательно, X имеет p(а не $p + 1$) столбцов.

Теперь рассмотрим как зависят от данных d_j.

Выборочная матрица ковариаций равна $S = \frac{X^T X}{n}$, откуда

$$S = \frac{1}{n} X^T X = \frac{1}{n} V D U^T U D V^T = V \cdot \frac{D^2}{n} \cdot V^T$$

Итак, столбцы $v_1, v_2, ..., v_p$ матрицы V представляют собой собственный базис матрицы ковариаций S,
$\frac{d_j^2}{n}$ - собственные числа этой матрицы. Векторы v_j называются также **главными компонентами (principal components)** для данных X.

Пусть $z_j = X v_j$.
Можно проверить, что

$$Var(z_j) = Var(X v_j) = \frac{d_j^2}{n}, \;\; z_j = X v_j = d_j u_j.$$

Первая главная компонента u_1 обладает тем свойством, что z_1 имеет максимальную дисперсию среди всех нормированных линейных комбинаций столбцов матрицы X.
Вектор u_j выбран среди всех векторов, ортогональных $u_1, ..., u_{j-1}$, так, что z_j имеет максимальную дисперсию.
Вектор z_p имеет минимальную дисперсию.
Таким образом, малые сингулярные числа d_j соответствуют направлениям v_j, для которых мала дисперсия величины z_j, и гребневая регрессия осуществляет наибольшее уменьшение соответствующих компонент.

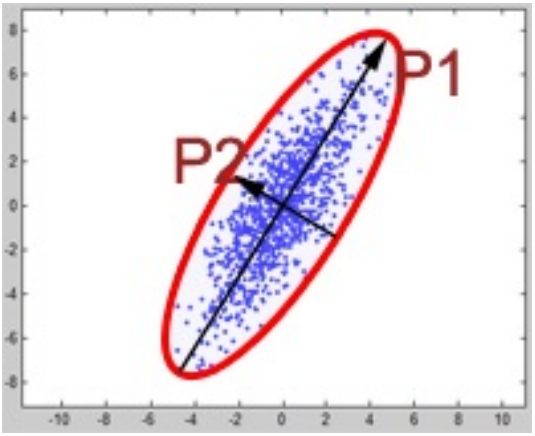


Рис. 4.1: Пример работы метода главных компонент

4.1.3 Понятие эффективной размерности

При увеличении параметра α вектор коэффициентов $\hat{w}_\alpha$ становится всё более устойчивым и жёстко определённым. В сущности, происходит понижение эффективной размерности решения - это второй смысл термина «сжатие». Доказано, что роль размерности играет след матрицы $X(X^T X)^{-1} X^T$. Тогда, в нерегуляризованном случае имеем

$$tr X(X^T X)^{-1} X^T = tr(X^T X)^{-1} X^T X = tr I_n = n$$

При использовании регуляризации эффективная размерность принимает значение от 0 до n, не всегда целое, и убывает при возрастании α :

$$n_{ef} = tr\, X(X^T X + \alpha I_n)^{-1} X^T = tr\, diag(\frac{\sqrt{\lambda_j}}{\lambda_j + \alpha}) = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \alpha} < n.$$

4.2 Лассо Тибширани

Ещё один метод регуляризации постановкой задачи схож с гребневой регрессией, но приводит к совершенно другому поведению вектора коэффициентов. Вместо добавления регуляризатора к функционалу качества вводится ограничение-неравенство, которое запрещает чрезмерно большие абсолютные значения коэффициентов:

$$\begin{cases} RSS(w) = \|Xw - y\|^2 \to \min\limits_w \\ \sum\limits_{j=1}^{n} |w_j| \leq \chi \end{cases}$$

где χ - параметр регуляризации. При больших значениях χ ограничение (1) становится строгим неравенством, и решение совпадает с МНК-решением. Чем меньше χ, тем больше коэффициентов w_j становятся нулевыми.
Происходит отбор (селекция) признаков, поэтому параметр χ называют ещё селективностью. Параметр χ ограничивает вектор коэффициентов, лишая его избыточных степеней свободы. Отсюда и название метода - **лассо (LASSO, least absolute shrinkage and selection operator)**.

Эту задачу можно представить немного в другом виде. Записываем w_j как

$$w_j = w_j^+ - w_j^-,$$

где $w_j^+ \geq 0$ и $w_j^- \geq 0$, $|w_j| = w_j^+ + w_j^-$

Получили новую задачу:

$$\begin{cases} RSS(w) = \|Xw - y\|^2 \to \min\limits_w \\ \sum\limits_{j=1}^{n} (w_j^+ + w_j^-) \leq \chi \\ w_j^+ \geq 0 \\ w_j^- \geq 0 \end{cases}$$

Чем меньше χ, тем больше ограничений обращаются в равенства $w_j^+ = w_j^- = 0$, что соответствует обнулению коэффициента w_j и исключению j-го признака.

4.2.1 Нахождение решения

Зафиксируем χ.
Задачу можно решать, последовательно вводя ограничения-неравенства и требуя от решения удовлетворения условий Куна-Такера.

Обозначим через $\delta_i,\ i = 1, ..., 2^n$ - n-мерные векторы вида $(\pm 1, ..., \pm 1)$. Тогда условия $\sum\limits_{j=1}^{n} |w_j| \leq \chi$ эквивалентны $\delta_i^T w \leq \chi$ для всех i.

Для заданного вектора w, пусть $E = \{i : \delta_i^T = t\}$, $I = \{i : \delta_i^T < t\}$,где E - набор индексов, соответствующих равенствам, I - набор индексов,для которых неравенство не выполняется. Выделим матрицу G_E , строками которой являются векторы δ_i , где $i \in E$. Пусть 1 - вектор из единиц длинной, равной числу строк в G_E.

1. Начальное приближение для алгоритма: $E = i_0$, где $\delta_{i_0} = sign(w^0)$, w^0 - оценка параметров методом наименьших квадратов без введения ограничений.

2. Нахождение w^{new}, минимизирующего $RSS(w)$ при $G_E w \leq 1\chi$.

3. Пока $\sum\limits_{j=1}^{n} |w_j^{new}| > \chi$.

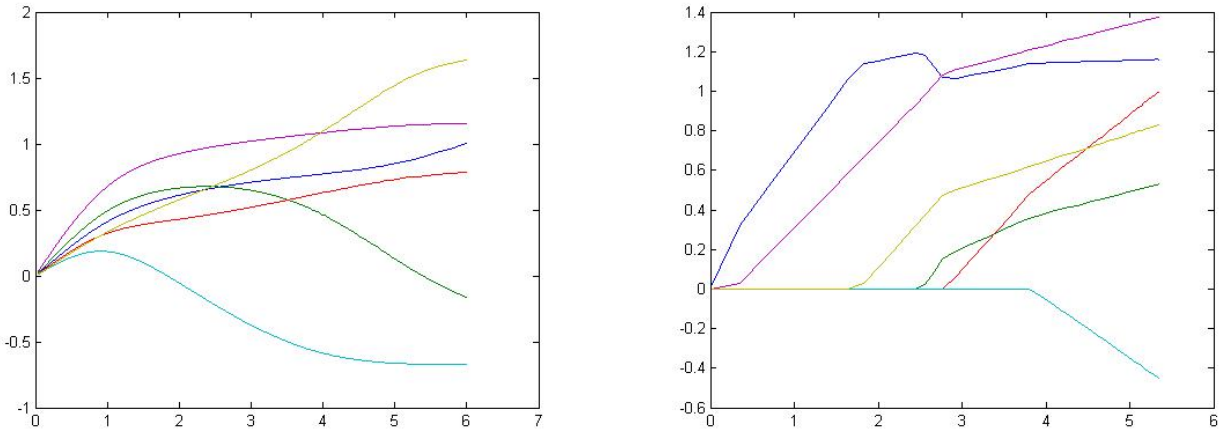
4. Добавить i в набор E. где $\delta_i = sign(w^{new})$.

5. Повторять шаги 2-4 пока процесс не сойдется.

Процедура сходится за конечное число шагов, так как на каждом шаге добавляется по одному элементу и число добавляемых элементов конечно.
Решение, получаемое на последнем шаге, является решением всей задачи, так как условия Куна-Такера соблюдены для наборов E и I.

4.3 Сравнение лассо и гребневой регрессии

Оба метода эффективно решают проблему мультиколлинеарности. Гребневая регрессия использует все признаки, стараясь максимально эффективно использовать имеющуюся информацию. Лассо производит отбор признаков, что эффективнее, если среди признаков есть шумовые. На Рис 4.2, левый график соответствует гребневой регрессии, правый - лассо. Ослабление регуляризации ведёт к уменьшению ошибки на обучении и увеличению нормы вектора коэффициентов. При этом ошибка на контроле в какой-то момент проходит через минимум, и далее только возрастает - это и есть переобучение.



Гребневая регрессия Лассо

Рис. 4.2: Зависимость коэффициентов линейной модели от параметра n_{ef} для гребневой регрессии и от параметра χ для лассо Тибширани

4.4 Метод наименьших углов

Метод наименьших углов (англ. **least angle regression, LARS**) является еще одним алгоритмом построения регрессионной модели. Алгоритм предложили Бредли Эфрон, Тревор Хасти и Роберт Тибширани.

LARS является улучшенной версией прямой шаговой регрессии. При большом количестве свободных переменных возникает проблема неустойчивого оценивания весов модели. LARS предлагает метод выбора и оценки весов такого набора свободных переменных, который имел бы наиболее значимую статистическую связь с зависимой переменной.

Прямая шаговая регрессия строит модель последовательно, добавляя по одной переменной. На каждом шаге алгоритм определяет лучшую переменную и включает ее в текущее подмножество, а затем методом наименьших квадратов определяет веса для отобранных переменных.

Метод наименьших углов, вместо последовательного добавления свободных переменных, на каждом шаге изменяет их веса. Веса увеличиваются так, чтобы доставить наибольшую корреляцию с вектором регрессионных остатков. Основным достоинством LARS является то, что он выполняется за число шагов, не превышающее число свободных переменных.

Пусть задана выборка - матрица X, столбцы которой соответствуют независимым переменным, а строки - элементам выборки и вектор y, содержащий элементы зависимой переменной. Обозначим множество столбцов матрицы X как $\{x^1, ..., x^p\}$. Будем строить следующую регрессионную модель

$$\mu(x_i, \beta) = \sum_{j=1}^{n} x_i^j \beta_j = x_i \beta,$$

Критерий качества - среднеквадратичная ошибка

$$S(X^l, \beta) = \sum_{i=1}^{l} (y - \mu(x_i, \beta))^2,$$

Обозначим через Ω текущий набор активных признаков. На j-м шаге только j входят в активный набор признаков Ω.

Алгоритм 4.1. Метод наименьших углов

1. Преобразование свободных переменные так, чтобы они имели нулевое среднее и единичную норму.

$$x^i = x^i - \bar{x^i}, \quad x^i = \frac{x^i}{\|x^i\|}, \ i = 1...p$$

Инициализировать: $r = y - \bar{y}, \quad \beta_1 = 0, ..., \beta_p = 0.$

2. На первом шаге находится свободная переменная x^j, которая имеет наибольшую корреляцию c_j с вектором r. Корреляция c_j вектора остатков r на некоторый признак x^j вычисляется как

$$c_j = x^j r$$

Эта переменная включается в набор активных признаков $\Omega = \{j\}$.

3. Далее каждый раз вычисляется единичный вектор u, лежащий на биссекторе уже выбранных признаков. Алгоритм смещает текущее приближение μ_Ω в направлении вектора u,

$$\mu_\Omega^{new} = \mu_\Omega + \gamma u,$$

где γ - коэффициент смещения, который определяется из условия, что корреляция нового вектора остатков $r^{new} = y - \mu_\Omega^{new}$ на некоторый неактивный признак x_k будет равна корреляции на все активные признаки, $\Omega^{new} = \Omega \cup \{k\}$ - новое активное множество признаков. Смещение в направлении вектора u обеспечивает равенство корреляций вектора остатков r^{new} на выбранные признаки, или иначе говоря, обеспечивает равенство углов между вектором остатков и выбранными признаками.

Через $(p-1)$ повторений этого шага мы получим такое же решение, что дает нам метод наименьших квадратов.

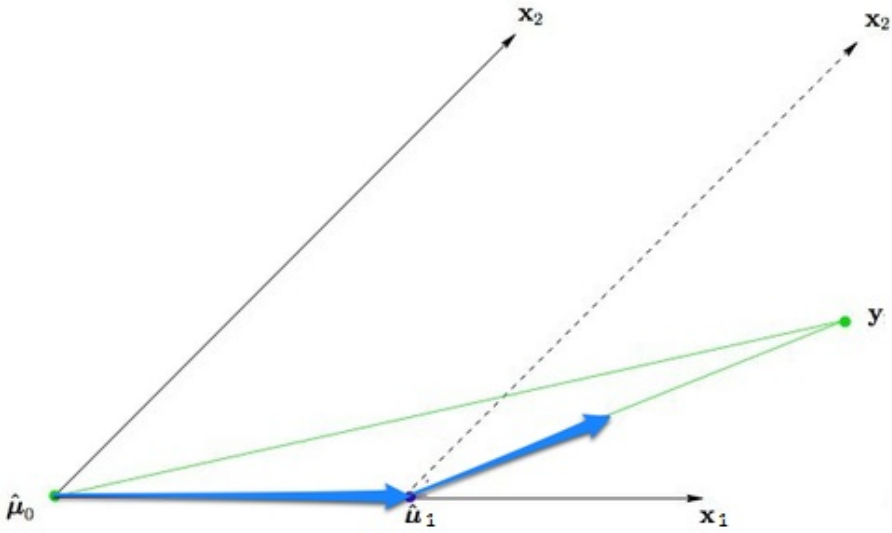


Рис. 4.3: Пример работы метода наименьших углов

На Рис. 4.3 продемонстирирована работа алгоритма LARS в случае двух признаков x_1 и x_2. Назначим начальное приближение $\mu_0 = 0$. Вектор регрессионных остатков $y - \mu_0$ имеет большую корреляцию с вектором x_1, чем с вектором x_2. Первый шаг заключается в оценке $\mu_1 = \mu_0 + \gamma_1 x_1$. Скаляр γ_1 выбирается так, чтобы вектор остатков $y - \mu_1$ делил пополам угол между векторами x_1 и x_2. Далее получаем значение $\mu_2 = \mu_1 + \gamma_2 u_2$, где u_2 - нормированный вектор, делящий этот угол пополам.

Глава 5

Методы сокращения размерности данных

Очень часто при большом количестве признаков, эти самые признаки бывают сильно взаимосвязаны. Методы, описанные в этом разделе, строят небольшое количество линейных комбинаций z_j, $j = 1, ..., m$, которые заменяют исходное множество признаков.Алгоритмы в этом разделе описывают построение таких линеных комбинаций.

5.1 Метод Главных Компонент

Метод главных компонент - это один из способов уменьшения размерности, который заключается в переходе к новому ортогональному базису, оси которого ориентированы по направлениям максимальной дисперсии набора входных данных. Вдоль первой оси нового базиса дисперсия максимальна, вторая ось максимизирует дисперсию при условии ортогональности первой оси, и т.д., последняя ось имеет минимальную дисперсию из всех возможных. Такое преобразование позволяет понижать размерность путем отбрасывания тех координат, которые соответствуют направлениям с минимальной дисперсией. Предполагается, что если нам надо отказаться от одного из базисных векторов, то лучше, если это будет тот вектор, вдоль которого набор входных данных меняется менее значительно.
Можно отметить, что в основе метода главных компонент лежат следующие предположения:

- Предположение о том, что размерность данных может быть эффективно понижена путем линейного преобразования.
- Предположение о том, что больше всего информации несут те направления, в которых дисперсия входных данных максимальна.

Можно заметить, что эти условия не всегда выполняются. Например, если точки входного множества располагаются на поверхности гиперсферы, то никакое линейное преобразование не сможет понизить размерность. Это недостаток в равной мере свойственен всем линейным алгоритмам и может быть преодолен за счет использования дополнительных фиктивных переменных, являющихся нелинейными функциями от элементов набора входных данных.

Второй недостаток метода главных компонент состоит в том, что направления, максимизирующие дисперсию, далеко не всегда максимизируют информативность. Можно рассмореть следующую задачу - переменная с максимальной дисперсией не несет почти никакой информации, в то время как переменная с минимальной дисперсией позволяет полностью разделить классы. Метод главных компонент в данном случае отдаст предпочтение первой (менее информативной) переменной. Этот недостаток тесно связан с тем, что метод главных компонент не осуществляет линейное разделение классов, линейную регрессию или иные подобные операции - он всего лишь позволяет оптимальным образом восстановить входной вектор на основе неполной информации о нем. Вся дополнительная информация, связанная с вектором игнорируется.

Описание метода

В пункте 4.1.2 уже вводилось определение главных компонент.

Рассмотрим регрессию y относительно $z_1, ..., z_m$, где $m \leq p$. Так как векторы $z_1, z_2, ..., z_m$ попарно ортогональны, то

$$y^{pcr} = \bar{y} + \sum_{j=1}^{m} \theta_m z_m, \;\; w^{pcr}(m) = \sum_{j=1}^{m} \theta_m v_m, \;\; \theta_m = \frac{< y, z_m >}{< z_m, z_m >}.$$

Регрессия методом главных компонент имеет много общего с гребневой регрессией.

- гребневая регрессия уменьшает коэффициенты главных компонент матрицы X, при этом чем меньше сингулярное значение, тем в большей степени уменьшается коээфициент.
- регрессия методом главных компонент просто отбрасывает компоненты, соответствующие меньшим $p - m$ сингулярным числам.

5.2 Частичные наименьшие квадраты

Метод главных компонент смотрит только на величину дисперсии x_j и не анализирует вектор y. Метод частичных наименьших квадратов находит направления, соответствующие большой дисперсии x_j и имеющие большую корреляцию с откликом y.

Пусть y центрированы и x_j нормализовано так, что имеет среднее 0, дисперсию 1.

1: $y^{(0)} := \bar{y}$
2: $x_j^{(0)} := x_j \; (j = 1, 2, ...p)$
3: **для** $m = 1, \ldots, p$
4: $\quad z_m := \sum_{j=1}^{p} \phi_{mj} x_j^(m-1)$, где $\phi_{mj} =< y, x_j^{m-1} >$
5: $\quad y^{(m)} := y^{(m-1)} + \theta_m z_m$, где $\theta_m = \frac{<y, z_m>}{<z_m, z_m>}$
6: $\quad x_j^{(m)} := x_j^{(m-1)} - \frac{<x_j^(m-1), z_m>}{<z_m, z_m>} z_m (j = 1, ..., p)$
7: Возвратить $y^{(m)} (m = 1, ...p)$ и $w_{jm}^{pls} = \sum_{l=1}^{m} \phi_{lj} \theta_l$

В итоге, на выходе алгоритма получаем решение $y^{(m)} = \sum_{j=1}^{p} w_{jm}^{pls} x_j (m = 1, .., p)$

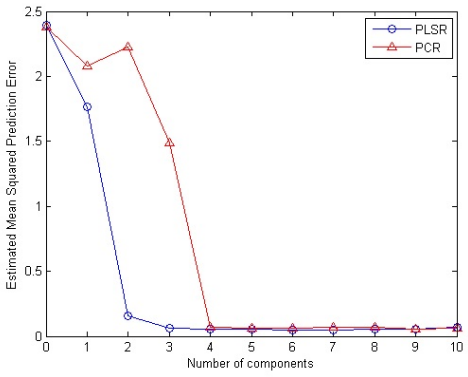


Рис. 5.1: Сравнение метода главных компонент и частичных наименьших квадратов

На рисунке 5.1 представлена работа 2-х методов преобразования данных. Можно заметить, что при маленьком числе компонент лучше работает метод частичных наименьших квадратов.

Заключение

Многие методы исходят из одних и тех же идей, но немного отличаются реализацией. Например, гребневая регрессия уменьшает коэффициенты главных компонент матрицы, при этом чем меньше сингулярное значение, тем в большей степени уменьшается соответствующий коэффициент. Регрессия методом главных компонент оставляет компоненты, соответствующие m максимальным сингулярным числам, а остальные отбрасывает. Интересно, что можно показать, что метод частичных наименьших квадратов также стремится уменьшить коээфициенты направлений с низкой дисперсией, но он может повысить коэффициенты при некоторых компонентах с высокой дисперсией. Это может сделать метод частичных наименьших квадратов немного неустойчивым, и ожидаемая ошибка у него немного выше, чем у гребневой регрессии. Полное исследование провели Франк и Фридман (1993). Авторы приходят к заключению, что для минимизации ошибки прогноза, гребневая регрессия, как правило, предпочтительнее методов отбора подмножеств, метода главных компонент и частичных наименьших квадратов.

Однако улучшение по сравнению с двумя последними методами совсем незначительно. Подводя итог, можно сказать, что методы гребневой регрессии, частичных наименьших квадратов и главных компонент, как правило, показывают практически одинаковые результаты. Гребневая регрессия может быть предпочтительней, так как уменьшается коэффициенты плавно, а не скачкообразно. Лассо Тибширани находится где-то между гребневой регрессией и наилучшим подмножеством, и включает в себя некоторые свойства каждого из них.

Литература

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning.
   http://www-stat.stanford.edu/~tibs/ElemStatLearn
2. Tibshirani R. J. Regression shrinkage and selection via the lasso.
   http://citeseer.ist.psu.edu/tibshirani94regression.html
3. Bishop C. M. Pattern Recognition and Machine Learning.
4. Воронцов К.В. Курс лекций по машинному обучению.
   http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
5. Золотых Н.Ю. Курс лекций по машинному обучению.
   http://www.uic.unn.ru/~zny/ml/Lectures/ml_pres.pdf
6. Hastie T., Tibshirani R., Johnstone J. Efron B. Least Angle Regression
   http://www.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf
7. Frank, I. and Friedman, J. A statistical view of some chemometrics regression tools (with discussion).
8. Lawson C., Hansen R. Solving Least Squares Problems.