Линейная регрессия

Зинина Анастасия

Содержание

Оценка качества модели

- Подобрали модель, но как понять, хороша ли она?
- Для этого нужны новые данные
- Идея: отложенная выборка

Отложенная выборка

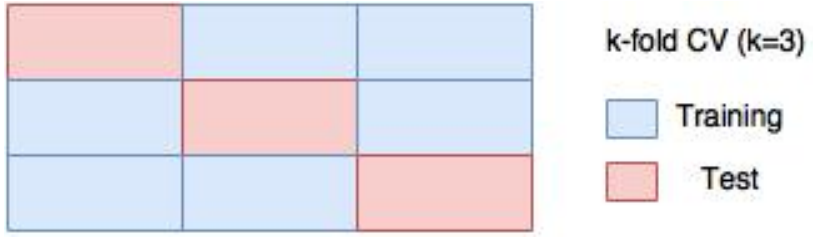
- Разделим выборку на две части: обучающую и тестовую
- Подберём параметры, минимизируя функционал ошибки на обучающей части
- Затем получим ответы алгоритма на тестовой части и оценим величину функционала ошибки на ней
- Как выбрать величину тестовой выборки?
  - Чем больше тестовая выборка, тем меньше данных для обучения
  - Чем меньше тестовая выборка, тем ненадёжнее оценка на ней
  - Ищем компромисс.Обычно ориентируются на объём данных и выбирают размер тестовой выборки $20\%, 30\%$

Отложенная выборка

- Однако найденные коэффициенты и качество модели будут зависеть от разбиения. Например, у нас есть важный бинарный признак. В обучающую выборку попали объекты только с одним значением этого признака, в этом случае коэффициент при этом признаке не будет отражать его настоящую важность, ведь признак ведёт себя как константный. И на тестовой выборке качество может быть небольшим.
- Можно проделать разбиение n раз и посмотреть на среднее значение функционала ошибки на получившихся n тестовых выборках.
- Однако нет гарантии, что все объекты хотя бы один раз попадут в обучающую выборку. А нам бы хотелось этого: чем больше вариантов объектов мы предъявим алгоритму, тем лучше будет его обобщающая способность.
- Идея: используем кросс-валидацию

Кросс-валидация

- Фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную.
- Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке.
- Оценивается его ошибка на объектах контрольной подвыборки.
- Оценка CV - средняя по всем разбиениям величина ошибки на контрольных подвыборках.
- Если выборка независима, то средняя ошибка скользящего контроля даёт несмещённую оценку вероятности ошибки.



Стратегии кросс-валидация

1. Оценка строится по всем $N = C_L^k$ разбиениям.
2. Контроль по отдельным объектам (leave-one-out CV).
3. Контроль по q блокам (q-fold CV).
   Выборка случайным образом разбивается на q непересекающихся блоков одинаковой (или почти одинаковой) длины $k_1, \ldots, k_q$:

$$X^l = X_1^{k_1} \cup \cdots \cup X_q^{k_q}, \; k_1 + \cdots + k_q = l.$$

   Каждый блок по очереди становится контрольной подвыборкой, при этом обучение производится по остальным q-1 блокам. Критерий определяется как средняя ошибка на контрольной подвыборке:

$$CV(\mu, X^l) = \frac{1}{q} \sum_{n=1}^{q} Q\big(\mu(X^l \setminus X_n^{k_n}), X_n^{k_n}\big).$$

4. Контроль по $r \times q$ блокам ($r \times q$-fold CV).
   Контроль по q блокам (q-fold CV) повторяется r раз. Каждый раз выборка случайным образом разбивается на q непересекающихся блоков.

Причины плохого качества модели

- Обнаружили плохое качество модели. Как понять, в чём причина?
- Регрессионная модель: $y = f(x) + \varepsilon$, $f(x)$-истинная зависимость. Мы находим алгоритм $\hat{a}(x)$, максимально приближающий истинную зависимость. Математическое ожидание отклонения ответа алгоритма от истинного ответа можно представить в виде (докажем далее в курсе)

$$E(\hat{a}(x) - y)^2 = (E\hat{a}(x) - f(x))^2 + D\hat{a}(x) + \sigma^2,$$

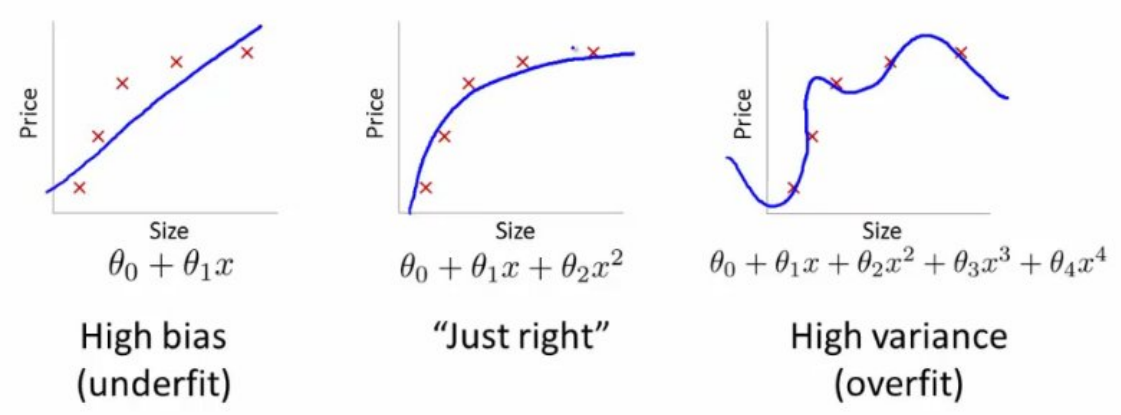
где $E(\hat{a}(x) - f(x))$ -смещение среднего ответа нашего алгоритма от ответа истинной зависимости,

$D\hat{a}(x)$ -дисперсия ответов нашего алгоритма,

σ^2 - шум (неустранимая ошибка).

Причины плохого качества модели

- Рассмотрим пример. Допустим, хотим аппроксимировать целевую зависимость полиномом. Подбираем степень полинома

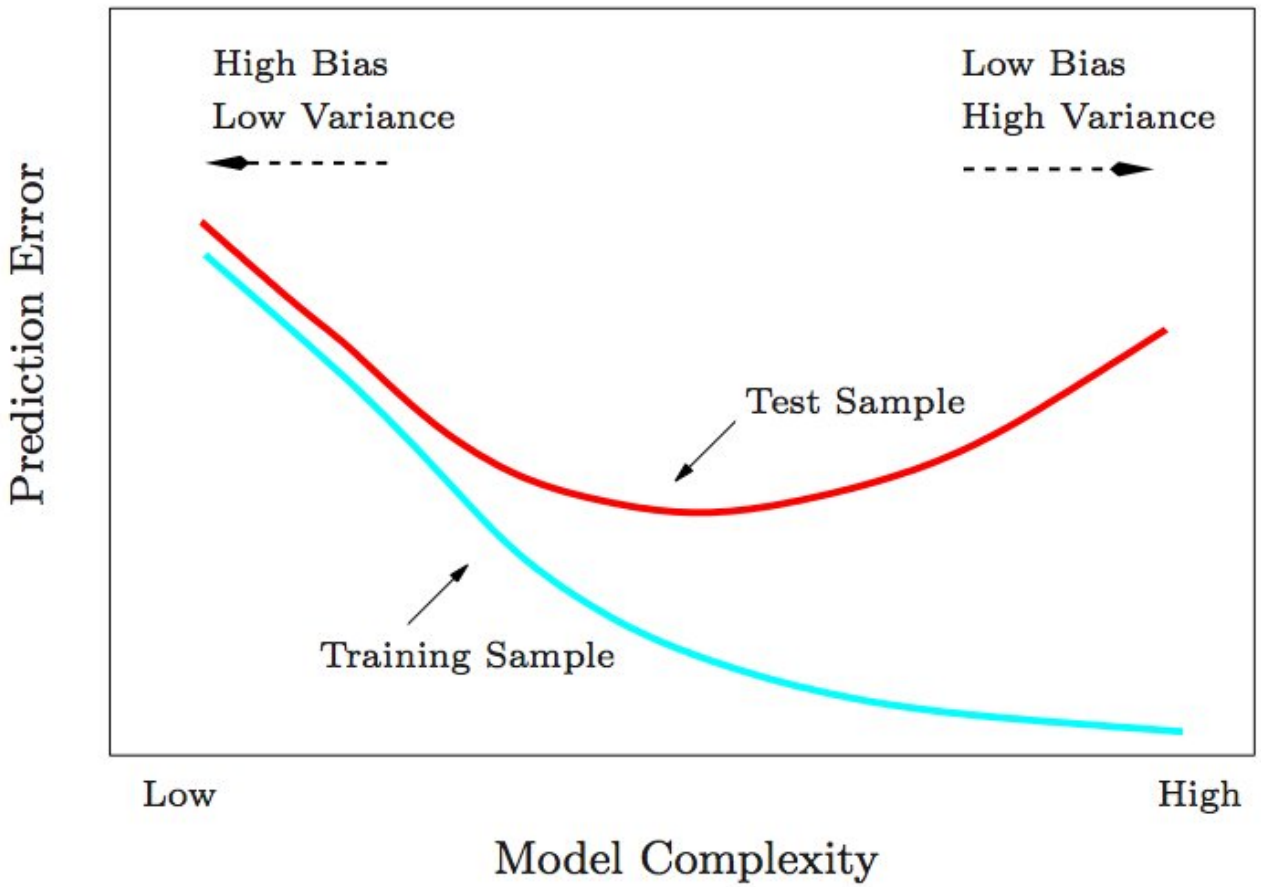


- Модель может быть слишком простой и не учитывать закономерности в данных. Тогда её смещение велико, речь идёт о недообучении

- Либо модель слишком сложная и подстраивается под обучающую выборку. Тогда велика дисперсия модели - чувствительность алгорима к изменениям в обучающей выборке. Имеем дело с переобучением.

Причины плохого качества модели

- Недообученная модель показывает плохое качество и на обучающей, и на тестовой выборке
- Переобученная модель показывает хорошее качество на обучающей выборке, но плохое на новых данных



Причины переобучения

- Мало данных
- Слишком сложная модель
- Мультиколлинеарность

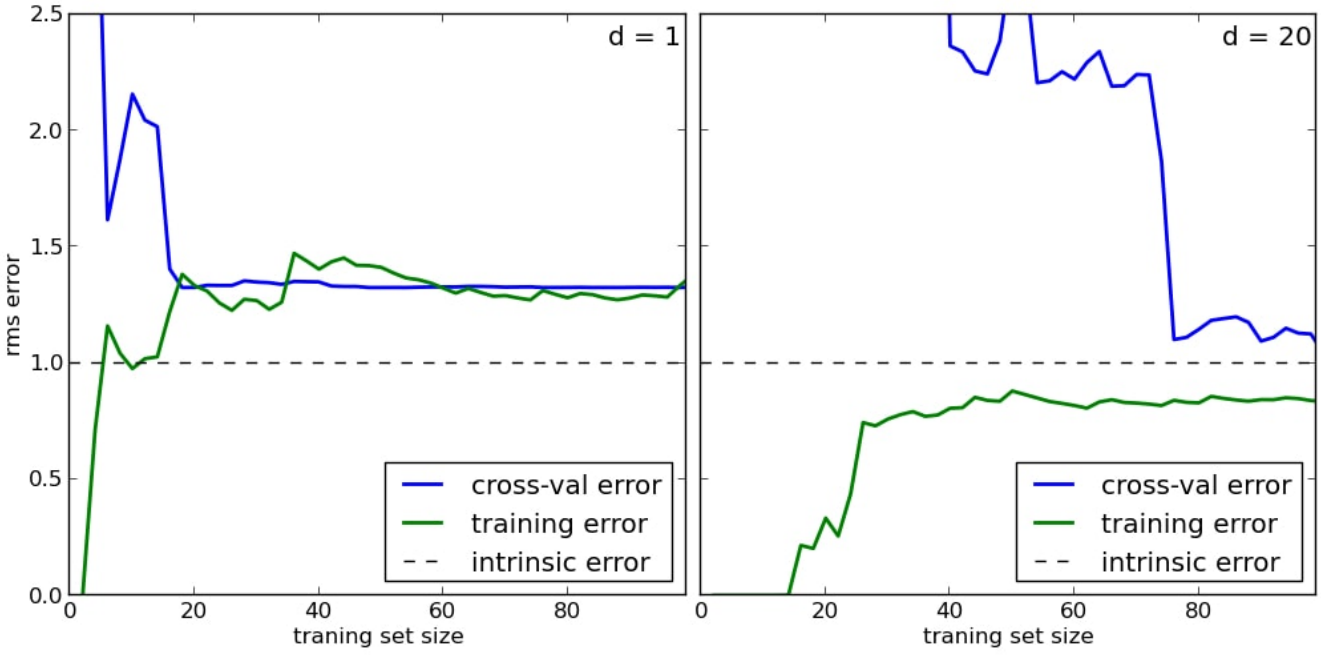
О важности размера выборки

- Рассмотрим снова аппроксимацию истинной зависимости полиномом. Пусть у нас есть 7 точек обучающей выборки, и степень полинома равна 6. В этом случае мы можем подобрать коэффициенты полинома так, чтобы ошибка на обучающей выборке будет равна 0. Однако модель будет переобученной и даст плохое качество на новых данных.
- Если же в обучающей выборке 100 точек, но полином степени 6 может быть вполне подходящей моделью.

- Т.е. добавление данных может помочь справиться с переобучением.

- Но как понять, поможет ли добавление объектов в нашем случае?

Поможет ли добавление новых данных?



Learning curves

- Слева learning curve для полинома степени 1. Это слабая модель, которая и на обучающей выборке даёт плохое качество.
- Справа полином степени 20. Если точек меньше 21, то ошибка на обучающейи выборке равна 0. Но тогда обучающей выборки ошибка на тестовой выборке велика. С ростом количества данных ошибка на обучающей выборке растёт, т.к. модели становится всё сложнее подстроиться под большое число точек. Однако растёт обобщающая способность: ошибка на тестовой выборке падает. И наша модель вполне может считаться оптимальной.

Борьба с мультиколлинеарностью

Как бороться с переобучением, если причина - мультиколлинеарность?

- Отбор признаков
- Преобразование признаков
- Регуляризация

Ridge-регрессия

- Признаком мультиколлинеарности являются большие значения весов
- Будем искать решение, ограничивая при этом величину весов, тем самым не позволим модели слишком подстраиваться под обучающую выборку

$$\begin{cases} Q(a) = || F\alpha - y ||^2 \to \min_\alpha \\ \sum_{j=1}^n \alpha_j^2 \leqslant C \end{cases}$$

NB На α_0 ограничение не накладываем

- Перепишем в эквивалентном виде (ищем точку минимума функции Лагранжа):

$$Q_\tau = \sum_{i=1}^l (y_i - \alpha_0 - \sum_{j=1}^n \alpha_j f_j(x_i))^2 + \tau \sum_{j=1}^n \alpha_j^2 \to \min_\alpha$$

где $\tau \in (0, 1)$ - коэффициент регуляризации.

- Перейдём к модели без смещения: заменим признаки центрированными $f_j^*(x_i) = f_j(x_i) - \bar{f}_j$, аналогично поступим и с вектором ответов $y_i^* = y_i - \bar{y}$:

$$Q_\tau = \sum_{i=1}^l (y_i^* - \sum_{j=1}^n \alpha_j f_j^*(x_i))^2 + \tau \sum_{j=1}^n \alpha_j^2 \to \min_\alpha$$

Ridge-регрессия

- МНК-решение задачи Ridge с $l \times n$-матрицей признаков F:

$$\hat{\alpha}_\tau = (F^T F + \tau I_k)^{-1} F^T y$$

- Запишем решение,используя SVD, и увидим, что веса действительно уменьшились

$$\alpha^* = F^+ y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^T, y);$$

Lasso-регрессия

- Аналогично можно ввести ограничение на слишком большие абсолютные значения коэффициентов:

$$Q_\tau = \sum_{i=1}^{l}(y_i - \alpha_0 - \sum_{j=1}^{n}\alpha_j f_j(x_i))^2 + \tau\sum_{j=1}^{n}|\alpha_j| \to \min_\alpha$$

  где τ — параметр регуляризации.
- Лассо осуществляет отбор информативных признаков.
  Введём:

$$\alpha_j^+ = 1/2(\mid \alpha_j \mid + \alpha_j) \geqslant 0, \quad \alpha_j^- = 1/2(\mid \alpha_j \mid - \alpha_j) \geqslant 0,$$

$$\alpha_j = \alpha_j^+ + \alpha_j^-$$

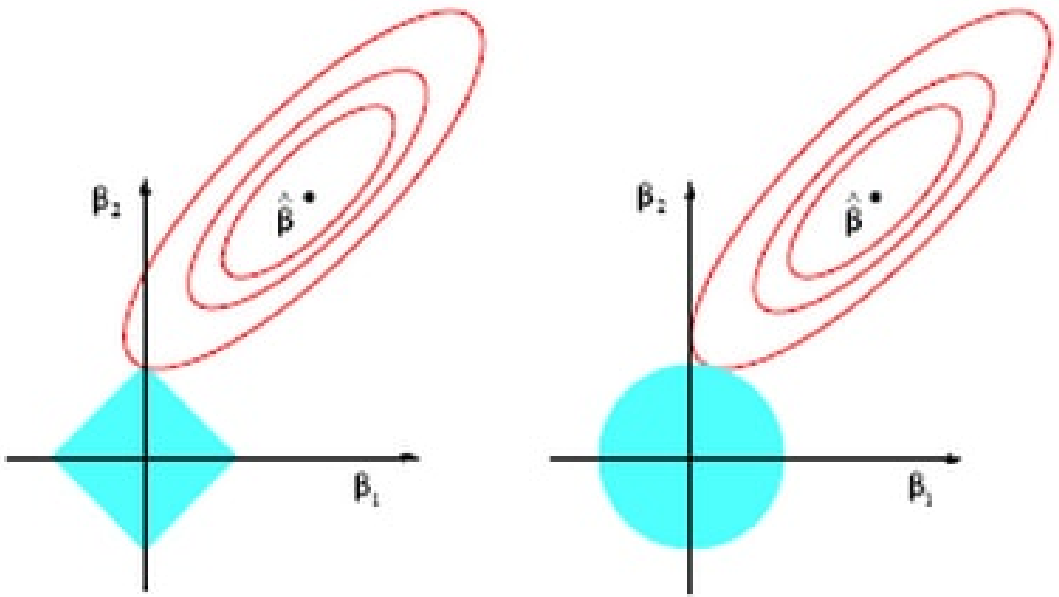
  .
- Тогда минимизируемый функционал останется квадратичным по новым переменным, но станет гладким, а ограничение примет линейный вид:

$$\alpha_j^+ + \alpha_j^- \leqslant \tau$$

- Чем больше τ, тем больше весов обратится в 0

Графическая интерпретация регуляризации

Посмотрим на графическую интерпретацию Lasso и Ridge.
Мы находим точку первого соприкосновения линий уровня и круга в случае Ridge, линий уровня и ромба в случае Lasso



Обоснование использования регуляризации

- Введём обозначения

$$MSE(\hat{\alpha}) = E(\hat{\alpha} - \alpha)^2$$ среднеквадратическая ошибка оценки

$$Var(\hat{\alpha}) = E(\hat{\alpha} - E\hat{\alpha})^2$$ – дисперсия оценки вектора параметров

$Bias = E\hat{\alpha} - \alpha$ – смещение среднего оценки относительно истинного значения

- $MSE(\hat{\alpha}) = E(\hat{\alpha} - \alpha)^2 = E(\hat{\alpha} - E\hat{\alpha})^2 + (E\hat{\alpha} - \alpha)^2 = Var(\hat{\alpha}) + Bias^2(\hat{\alpha})$

- Теорема Гаусса-Маркова утверждает, что МНК-оценки имеют наименьшую дисперсию среди всех несмещенных оценок (bias МНК-оценок =0)

- Однако могут существовать смещенные оценки, разброс имеющие меньшее среднеквадратическое отклонение. Наличие смещение компенсируется небольшим разбросом, и $MSE(\hat{\alpha})$ получается меньше. Именно так и происходит при регуляризации

Параметр регуляризации

- Как подобрать параметр регуляризации?
- Используем кросс-валидацию: выберем тот параметр, для которого средняя ошибка будет наименьшей

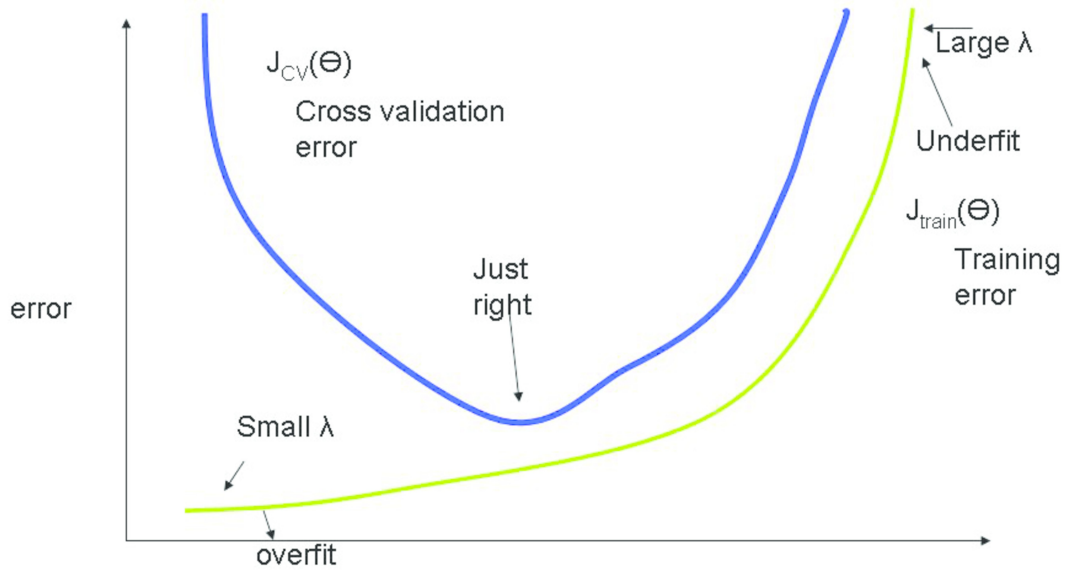


График зависимости значения функционала ошибки от параметра регуляризации λ

Для чего ещё используют кросс-валидацию

- Выбор гиперпараметра

  Гиперпараметр - параметр, значение которого нельзя получить из обущающей выборки. Примеры: степень многочлена, приближающего функцию; параметр регуляризации τ.

- Выбор лучшего алгоритма

  NB Если для выбора лучшего алгоритма использовать отложенную выбоку, результат будет необъективным: результат сравнения будет зависеть от разбиения. Модель, оказавшаяся лучшей на одной отложенной выборке, может сработать хуже остальных при другом разбиении.

  Можно также разделить выборку на две части, выбрать лучший на одной части с помощью CV, затем ещё раз проверить выбранный алгоритм на отложенной выборке.

Работа с категориальными признаками

Для использования категориальных признаков введём бинарные фиктивные переменные (dummy variables). Пусть признак принимает m различных значений, тогда для его кодирования необходимо m фиктивных переменных, из которых только одна принимает значение 1 на объекте.

Родной язык	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_2$
Русский	0	0	1
Английский	0	1	0
Французский	1	0	0
Английский	0	1	0

Стандартизация признаков

- Линии уровня функционала ошибки представляют собой эллипсы. Чем больше разнится масштаб признаков, тем дольше может сходиться градиентный спуск.
- Регуляризация штрафует модели с большими весами, а в случае разных масштабов признаков веса как раз могут получить большие значения, хотя это не будет говорить о переобучении.
- Для избежания перечисленных проблем выполняют предварительную стандартизацию признаков
  - Нормализация

$$f_j(x_i) := (f_j(x_i) - \bar{f_j})/\sigma_j, \quad j = 1, ..., n, \quad i = 1, ..., l,$$

    где $\bar{f_j} = \frac{1}{l}\sum_{i=1}^{l} f_j(x_i)$ — выборочное среднее, $\sigma_j^2 = \frac{1}{l}\sum_{i=1}^{l}(f_{ij} - \bar{f_j})^2$ — выборочная дисперсия j-го признака.
  - Отображение на $[0, 1]$

$$f_j(x_i) := \frac{f_j(x_i) - m_j}{f_j(x_i) - M_j},$$

    где m_j и M_j -соответственно минимальное и максимальное значение признака j

Спрямляющее пространство

- Пусть искомая зависимость y от x носит более сложный, чем линейный, характер. Введём новые признаки и произведём замену переменных. Т.о. перейдём к пространству, где y будет хорошо описываться линейной зависимостью.

- Например $y = \alpha_0 + \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_1^2(x)$ перепишем в виде $y = \alpha_0 + \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x)$