Автоматический feature selection на примере линейных моделей

И. Куралёнок

СПб, 2017

Воспоминания о былом

Сегодня будем искать решающую функцию в параметрическом семействе $h(x, \beta)$.

- Будем рассматривать решение задачи оптимизации, как распределение над β
- Можно предположить, что при увеличении количества параметров линейные модели имеют большой разброс решений
- Из курса алгебры (?) мы знаем, что меньше разброс нам не сделать с избранной целевой функцией (теорема Гаусса-Маркова)

$\Rightarrow$ будем менять целевую функцию: вводить условия, или менять саму T.

Хорошие решения регрессии

1. Предположим, что данные y у нас шумные
2. Введем чувство прекрасного $-R(\beta)$

Тогда нашу проблему можно свести к:

$$\arg\min_\beta R(\beta)$$
$$\|h(X,\beta) - y\| < \epsilon$$

Ну или так:

$$\arg\min_\beta \|h(X,\beta) - y\|$$
$$R(\beta) < p$$

Преобразование целевой функции

$$\arg\min_{\beta} \|(X, \beta) - y\| + \lambda R(\beta)$$

Можно найти такой параметр λ, который будет давать решение задачи на предыдущем слайде. В смысле оптимизации проблемы эквивалентны.

С байесовой точки зрения

Строго говоря, мы ввели prior на распределение решений:

$$\arg\max_\beta P(y|X\beta)P(\beta)$$
$$= \arg\max_\beta \sum_i logP(y_i|x_i\beta) + logP(\beta)$$

Если предположить нормальность $P(y_i|x_i\beta)$, то проблема становится очень похожей на то, что мы уже видели:

$$= \arg\min_\beta \|X\beta - y\| - zlogP(\beta)$$

где z — ошметки нормализации.

Виды prior

Какими бывают $-logP(\beta)$:

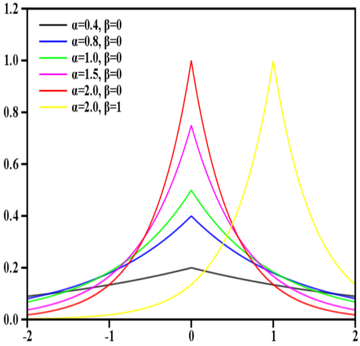
$\|\beta\|_0$ — бритва Оккама, MDL, etc.;

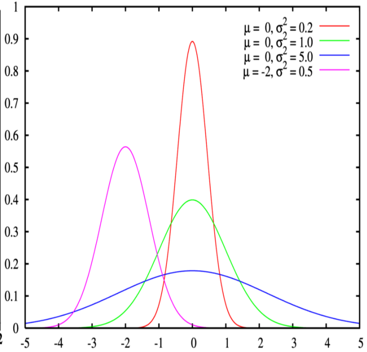
$\|\Gamma\beta\|_2$ — нормальное распределение с $\Sigma = \Gamma^{-1}$ и $\mu = 0$;

$\|\beta\|_1$ — распределение Лапласа;

Как можно видеть все это добро обобщается в l_q.

Сравнение prior'ов в случае l_1 и l_2





Картинки из википедии

Как это относится к ML

Точное решение в несет слишком много информации о обучающей выборке, поэтому точно оптимизировать смысла нету. Таким образом условие:

$$\min R(\beta)$$
$$\|h(X,\beta) - y\| < \epsilon$$

хорошо ложится на ML. При этом возможность выбора R позволяет рассказать наши ожидания от структуры решения.

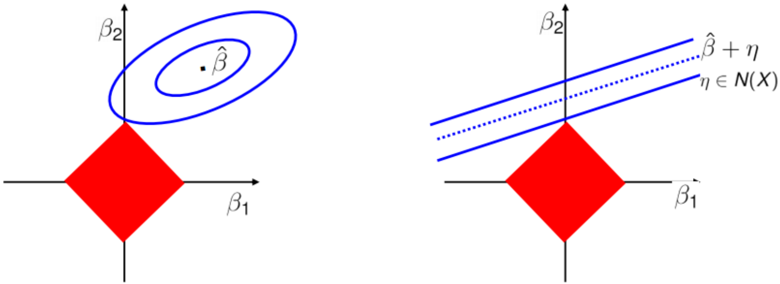
Как это называется

$\|\beta\|_0$ — Best Subset/Akaike information criterion;
$\|\Gamma\beta\|_2$ — регуляризация Тихонова/ridge regression;
$\|\beta\|_1$ — least absolute shrinkage and selection operator (LASSO);

Также рассматривается обощение на l_q.

Геометрия LASSO

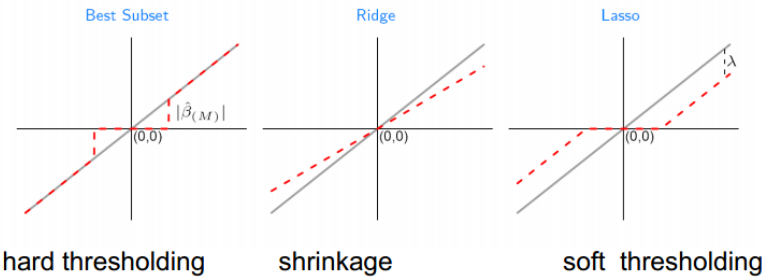
$$\min_\beta \|X\beta - y\|$$
$$\|\beta\|_1 < p$$



Картинки из ICML 2010 Tutorial (Irina Rish, Genady Grabarnik)

Геометрия LASSO II

Рассмотрим зависимость компоненты решения β_0 от решения безусловной системы $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$.



Картинки из The Elements of Statistical Learning (Hastie, Friedman and Tibshirani, 2009)

Пара слов об оптимизации l_0

l_0 — естественное условие на решение: “меньше фишек — легче думать”
Точное решение однако найти очень дорого: проблема *NP*-сложная. Поэтому есть масса работ по тому как найти приближенное решение:

- Если $n < 40$ есть работа “Regressions by Leaps and Bounds” (Furnival and Wilson 1974) + CV для выбора эффективного ограничения
- Взад-назад селекшен (Forward/Backward-Stepwise Selection): выбрасываем или добавляем фичу с наибольшим/наименьшим Z-score
- Приближение с маленьким q
- “Хитрые” переборы многомерного креста

Оптимизация l_1 и l_2

В случае ridge regression все просто:

$$\beta_0 = \left(X^T X + \Gamma^T \Gamma\right)^{-1} X^T y$$

А в случае LASSO все не так просто, так как T негладкая. Поэтому там почти градиентный спускъ. Есть несколько способов:

- Пошаговый LASSO
- LARS
- Покоординатный спуск (посмотреть дома)
- ISTA
- FISTA
- etc.

Пошаговый LASSO

1. Начнем с $\beta = 0$ и зафиксируем шаг w и соответствующие ему шаги по всем ортам $e_i w$
2. Выберем такое направление:

$$\begin{aligned} \hat{i} &= \arg\min_i \|y - X(\beta_t \pm e_i w)\| \\ &= \arg\min_i \|(y - X\beta_t) \pm X e_i w\| \\ &= \arg\min_i \|r_t \pm X e_i w\| \end{aligned}$$

3. $\beta_{t+1} = \beta_t \pm e_{\hat{i}} w$ если $\|y - X\beta_t\| - \|y - X\beta_{t+1}\| > \lambda$

Наблюдение за работой LASSO

Несколько наблюдений:

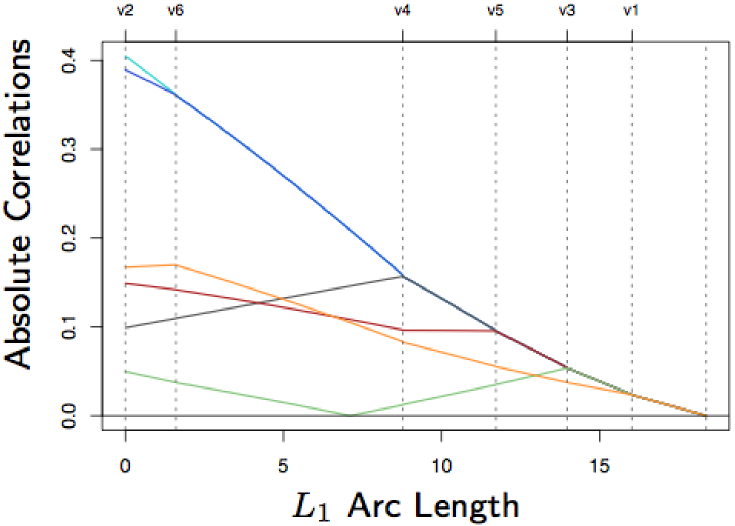
- Мы много раз бродили в одном и том же направлении
- За шаг двигаемся только по одной координате
- Так как шаг w фиксирован приходим не в точное решение, а в приближенное

Least angle regression (LARS)

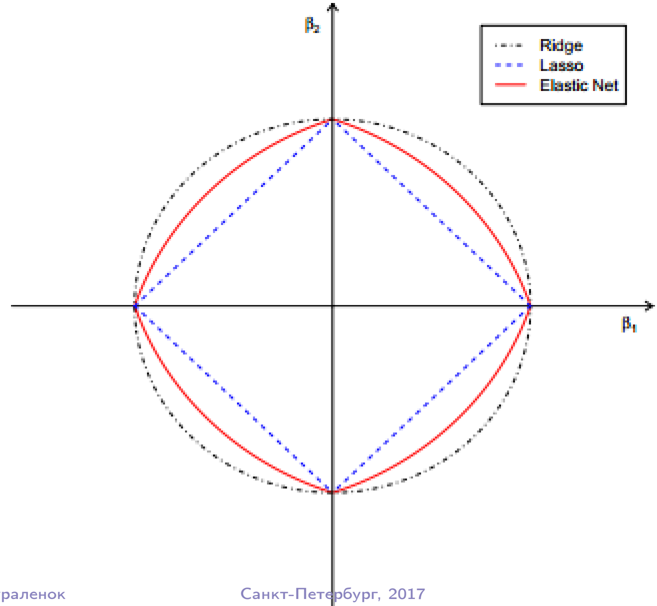
Внимание, 1 дальше — это вектор нужной размерности.

1. Нормируем X так, чтобы $\mu(\mathbf{x_i}) = 0, D(\mathbf{x_i}) = 1$ и y, так чтобы $\mu(\mathbf{y}) = 0$
2. Введем $\beta_1 = 0, r = y$, $A = \emptyset$ — множество всех направлений с максимальной корреляцией с r, s — вектор знаков корреляций.
3. $X_A = \left(s_1 x_{A(1)}, s_2 x_{A(2)}, \dots s_{|A|} x_{A(|A|)}\right)$
4. $a_A = \left(1^T (X_A^T X_A)^{-1} 1\right)^{-\frac{1}{2}}$
5. $u_A = a_A X_A (X_A^T X_A)^{-1} 1_A$ обратите внимание, что $(X_A^T u_A = a_A 1_A)$
6. $c = X^T (y - \beta_t)$
7. $a = X^T u_A$
8. $\gamma = min_j^+ \left(\frac{C - c_j}{a_A - a_j}, \frac{C + c_j}{a_A + a_j}\right)$
9. $\beta_{t+1} = \beta_t + \gamma u_A$

Работа LARS (корреляция)



Немного о касаниях



ElasticNet, $l_1.1$, etc.

Можно пойти 2-мя способами:

1. Добавить компоненту, которая будет отвечать за "круглые бока" (ElasticNet)

$$\arg\min_{\beta} \|X\beta - y\| + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2$$

2. Если жизнь все равно негладкая то гулять так гулять ($1 < q < 2$):

$$\arg\min_{\beta} \|X\beta - y\| + \lambda\|\beta\|_q$$

Grouped LASSO

Иногда мы априорно знаем, что какие-то переменные похожи. Эту информацию можно использовать:

$$\arg\min_{\beta} \|X\beta - y\| + \lambda \sum_{g} \sqrt{\sum_{j \in g} \beta_j^2}$$

Бывают и другие танцы для того, чтобы обеспечить structural sparsity

Как выбирать q?

Есть теоретические работы на эту тему. Не помню кто, не помню как, но доказал, что в ML q зависит от объема выборки. Для малых объемов работает $q = 0$, для бесконечных $q = 2$. Все что посередине должны найти свой любимый q.

Постараюсь к следующему разу привести чуть более точную ссылку

Еще раз о регуляризации

$$\arg\max_\beta P(y|X\beta)P(\beta)$$
$$= \arg\max_\beta \sum_i logP(y_i|x_i\beta) + logP(\beta)$$

Данный фокус гораздо более общий, и подходит не только для линейных решений.