Линейные модели: уменьшаем variance

И. Куралёнок, Н. Поваров

Яндекс

СПб, 2013

Переобозначение параметра решающей функции!

λ — мало. С сегодняшнего дня вместо него β

$$F(x, \lambda) \to F(x, \beta)$$

Слайды предыдущих лекций переделаем.

Воспоминания о былом

- Из второй лекции мы знаем, что линейные модели имеют большой разброс решений
- Из 7-й лекции 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

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

Тогда нашу проблему можно свести к:

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

Ну или так:

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

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:

$$\begin{aligned} &\arg\max_\beta P(y|X\beta)P(\beta) \\ &= \arg\max_\beta \sum_i logP(y_i|x_i\beta) + logP(\beta) \end{aligned}$$

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 нормальности.

Виды prior

Какими бывают $-logP(\beta)$:

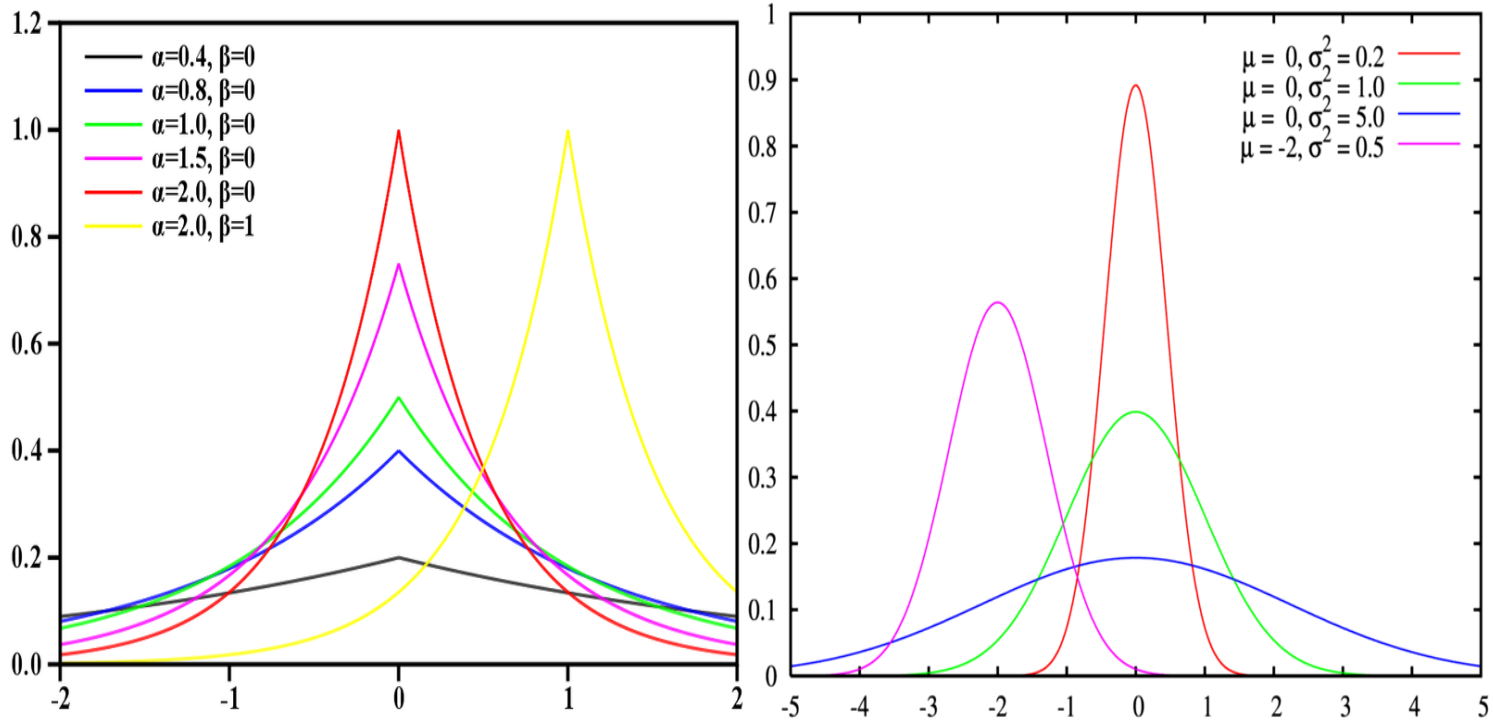
$\|\beta\|_0$ — бритва Оккама, MDL, etc.;

$\|\Gamma\beta\|_2$ — 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 L несет слишком много информации о обучающей выборке, поэтому точно оптимизировать смысла нету. Таким образом условие:

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

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

Как это называется

$\|\beta\|_0$ — Best Subset;

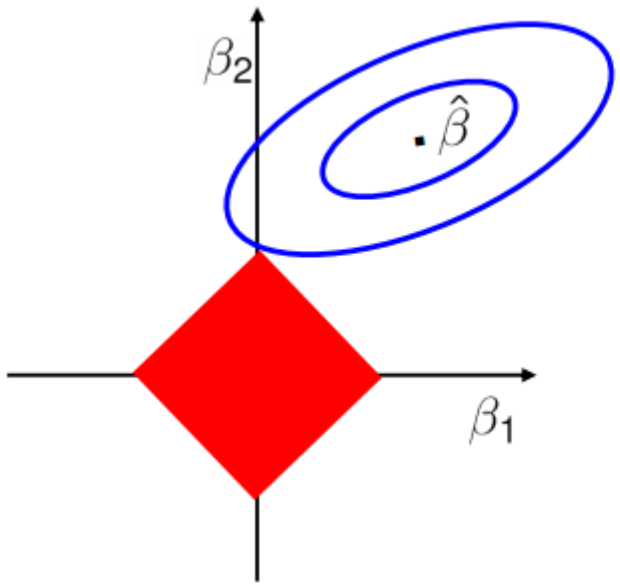
$\|\Gamma\beta\|_2$ — регуляризация Тихонова/ridge regression;

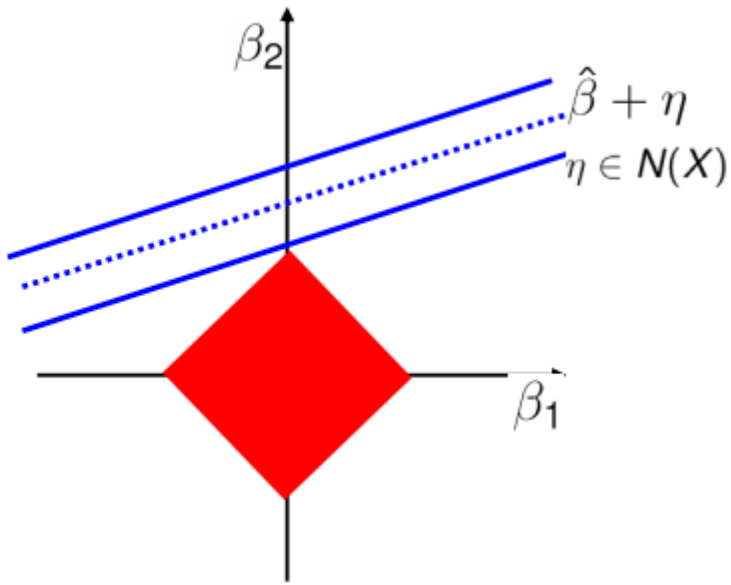
$\|\beta\|_1$ — least absolute shrinkage and selection operator (LASSO);

Также рассматривается обощение на l_q.

Геометрия LASSO

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

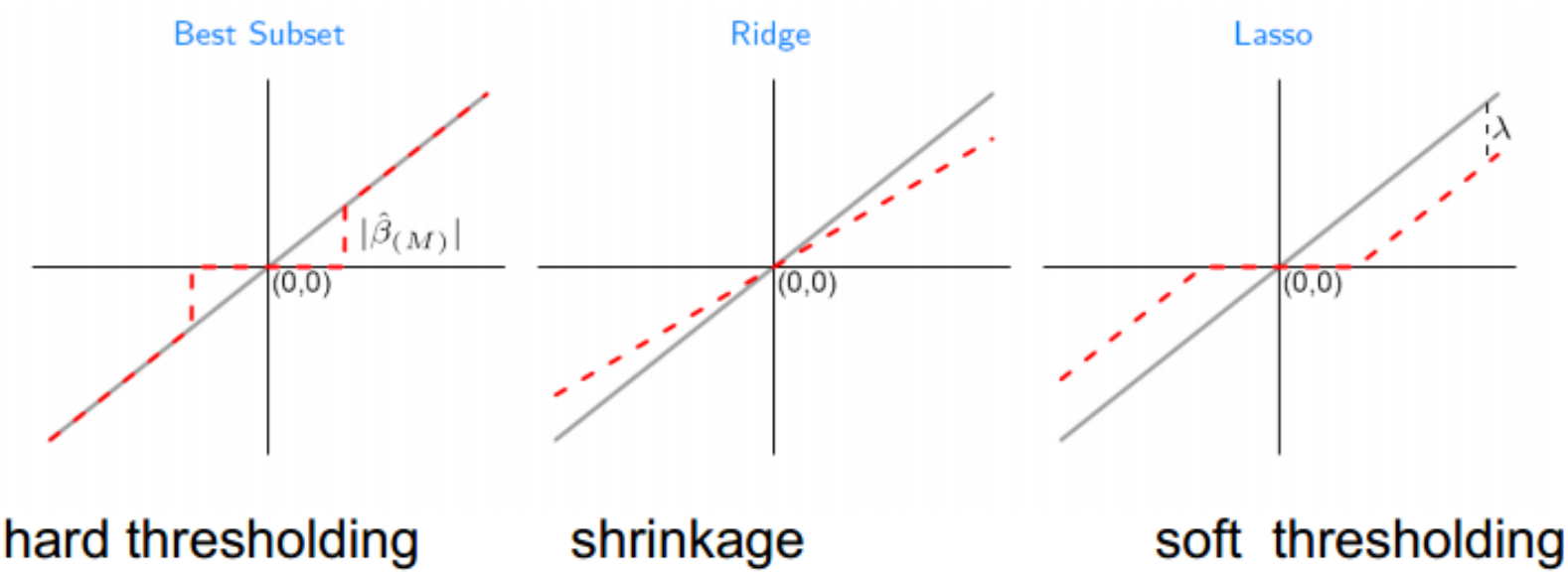




Картинки из ICML 2010 Tutorial (Irina Rish, Genady Grabarnik)

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:

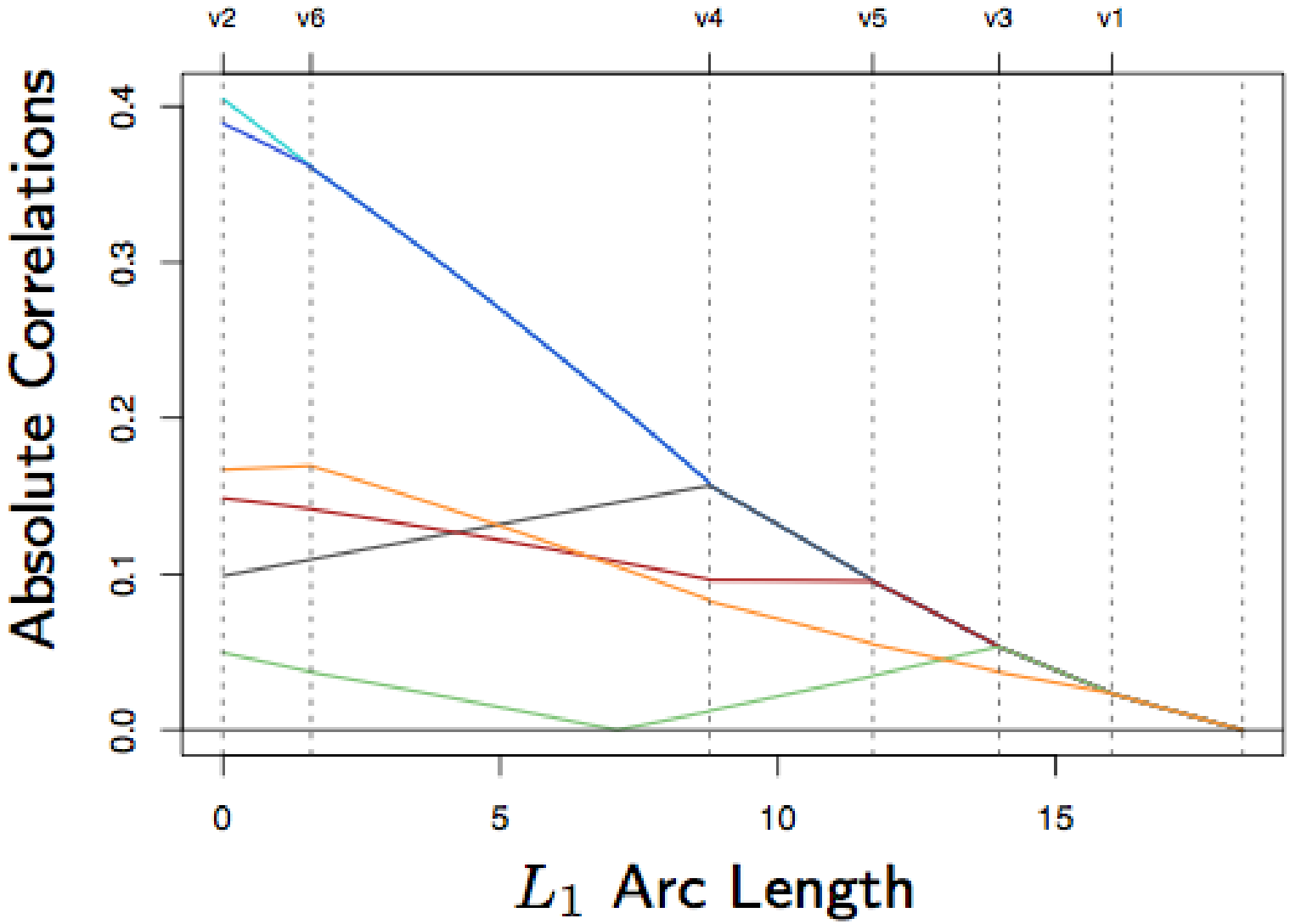
- Мы много раз бродили в одном и том же направлении
- За шаг двигаемся только по одной координате
- Так как шаг w фиксирован приходим не в точное решение, а в приближенное

Least angle regression (LARS)

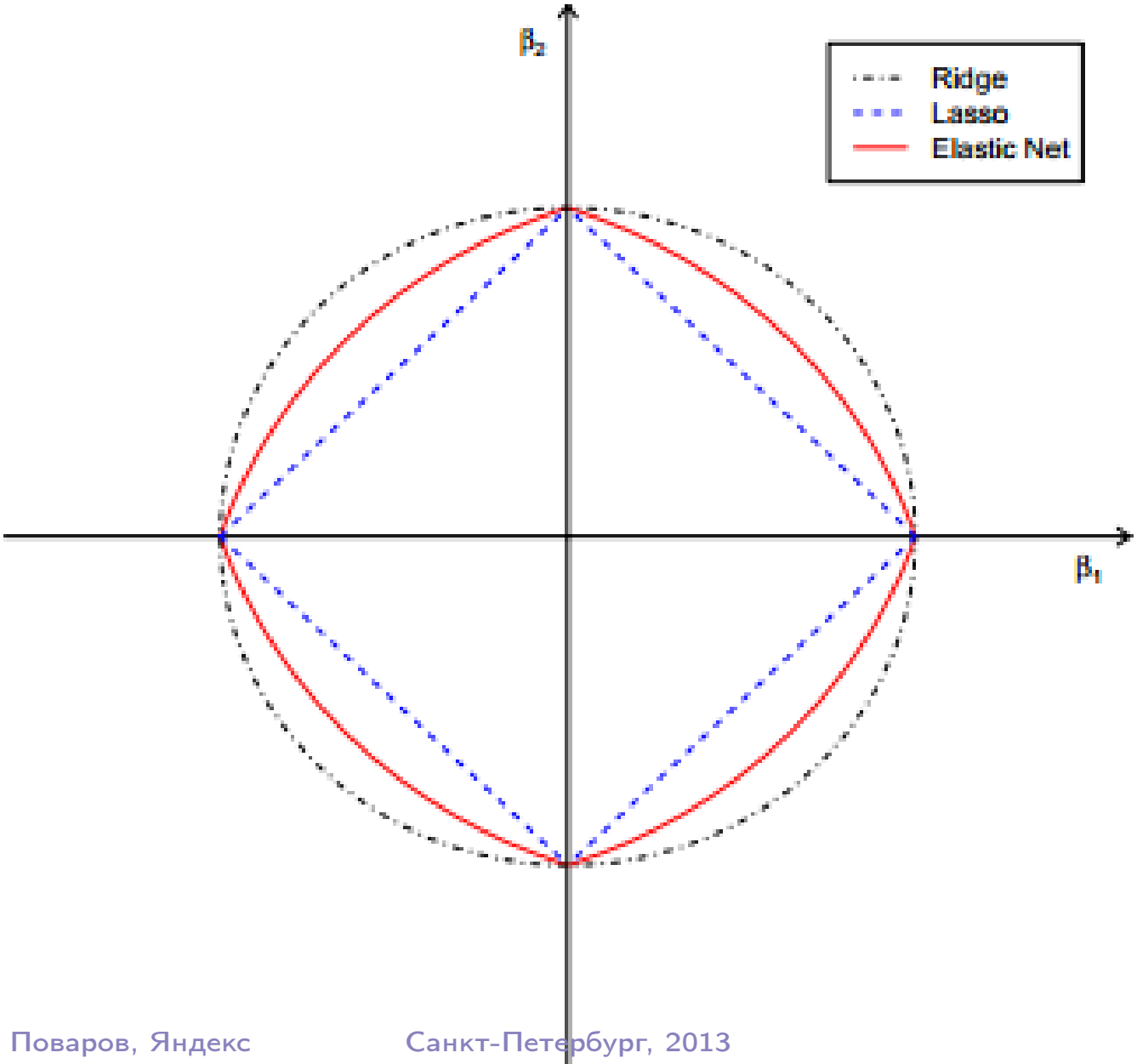
Внимание, 1 дальше — это вектор нужной размерности.

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 $(X^T u_A = a_A 1)$
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. $\beta_{t+1} = \beta + \gamma u_A$

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.

Результаты ДЗ второй недели

1. deb8e1
2. 1371ef
3. 5fd1e0
4. 85cd01
5. 16d288
6. aa0155
7. f54b78
8. a3807d
9. 099f77
10. fe14e8
11. 611bf4
12. 95afb4
13. 11e31c
14. fbcd3c
15. 6fc4ef
16. a434b2
17. a8cb51
18. f6c031
19. 234ec9
20. 3c287f
21. 1bf2d7
22. 15059b
23. f737ec
24. e5bda2
25. dfba4d
26. 91bd7c

Домашнее задание

- Датасет в домашнем задании прошлой недели и этой - одинаковый.
- Задача - мультиклассификация.
- Для ДЗ прошлой лекции можно имплементировать методы из прошлой лекции.
- Для ДЗ текущей лекции можно имплементировать методы из текущей лекции.
- Рейтинг будет строиться исходя из качества полученной мультиклассификации для каждой недели отдельно.
- В svn точно всё лежит.