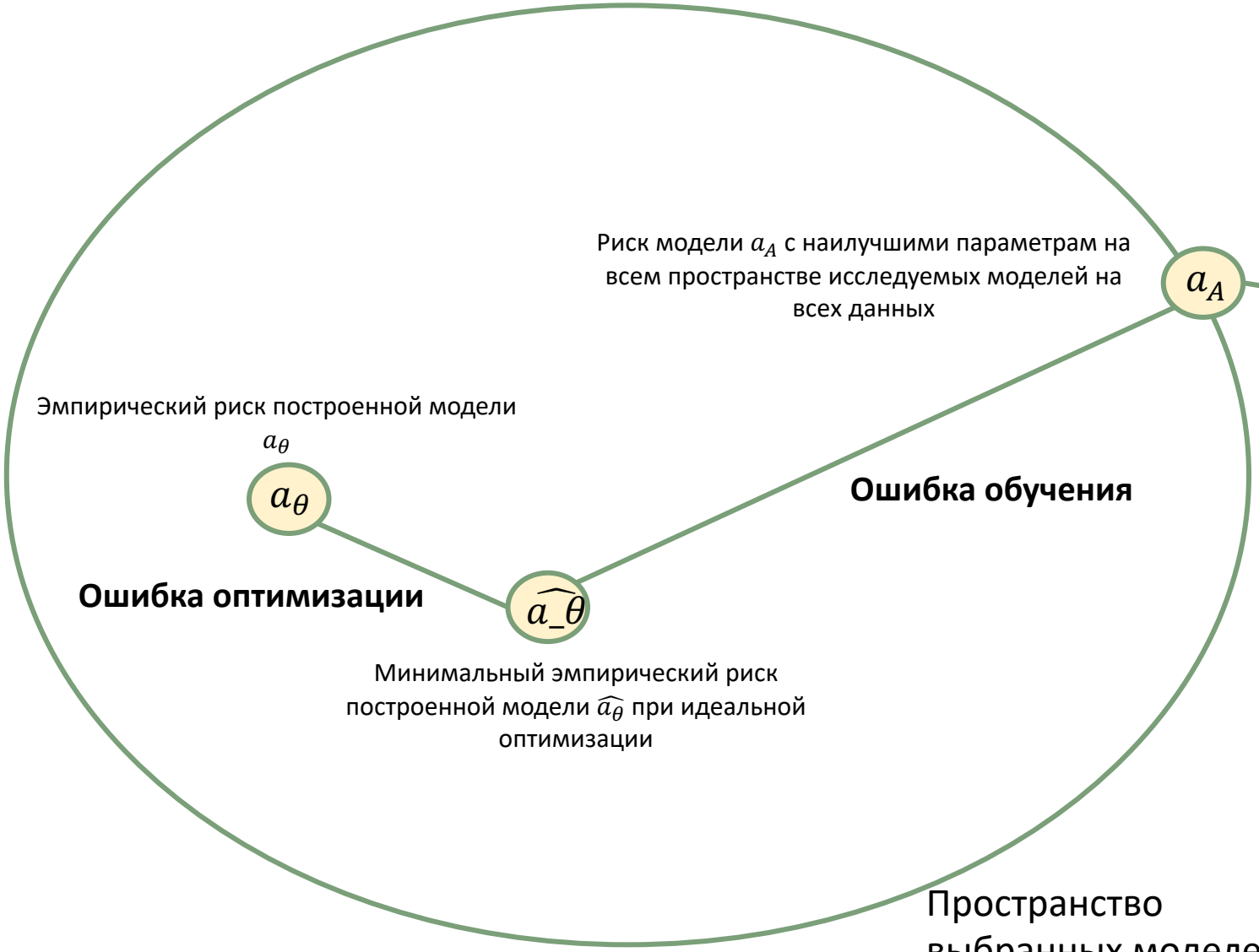
Переобучение в машинном обучении. Регуляризация.

Сбертех, МФТИ

Поиск оптимальной модели в задаче ML



- Чем более сложные модели мы выбираем для нашей выборки, тем меньше ошибка оценки
- Чем более сложные модели мы выбираем для нашей выборки, тем больше ошибка обучения

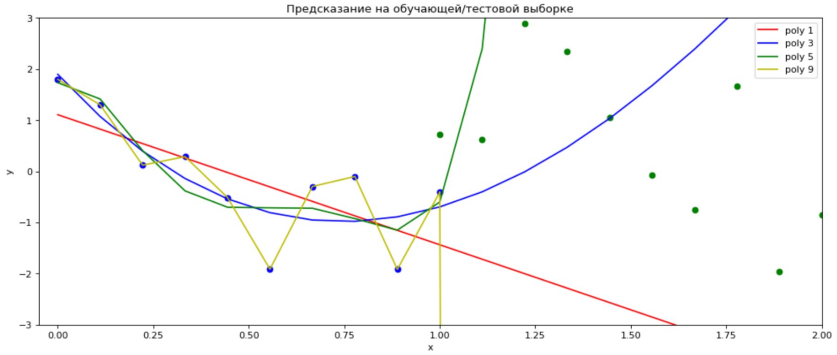
Как найти модель с наилучшими параметрами, сбалансировав между ошибкой обучения и ошибкой оценки?

Переобучение в линейной регрессии

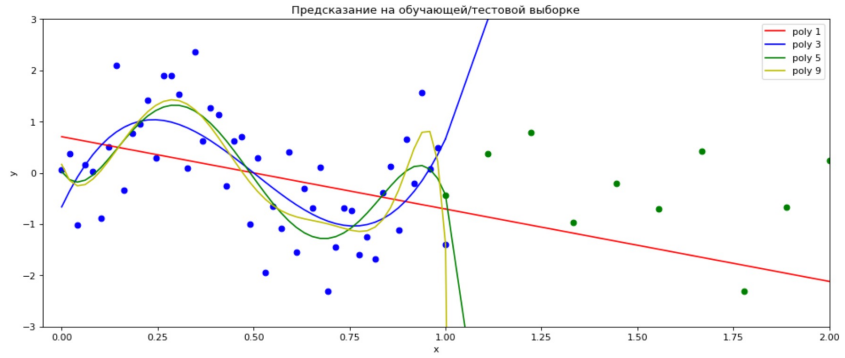
Пусть есть линейная регрессия модель линейной регрессии $y = X\theta$

$$X = \begin{bmatrix} \varphi_1(x_{11}) & \cdots & \varphi_n(x_{1n}) \\ \vdots & \ddots & \vdots \\ \varphi_1(x_{1n}) & \cdots & \varphi_n(x_{nn}) \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$$

1. Как у нас выражается переобучение в линейной регрессии? - **В больших весах.**
2. **Почему появляются большие веса?** Модель подбирает коэффициенты чтобы минимизировать среднеквадратичную ошибку на выборке. Это достигается изменением весов при базисных функциях. Чтобы ошибка была минимальна, модель подбирает большие веса при базовых функциях.
3. **Из-за чего так происходит?**
    1. Мало данных для выбранной функции, а функция задающая зависимость значительно проще моделируемой функции – функция полностью запоминает выборку
    2. Мультиколлинеарность
4. **Почему это плохо?**
    1. Подобранные веса соответствуют необходимому весу базовых функций только на одной выборке. Модель не научилась обобщать свои знания насчет генеральной выборке, она запомнила только тренировочную выборку.
    2. Малейшие изменения в аргументе ведут к большим изменениям в функции



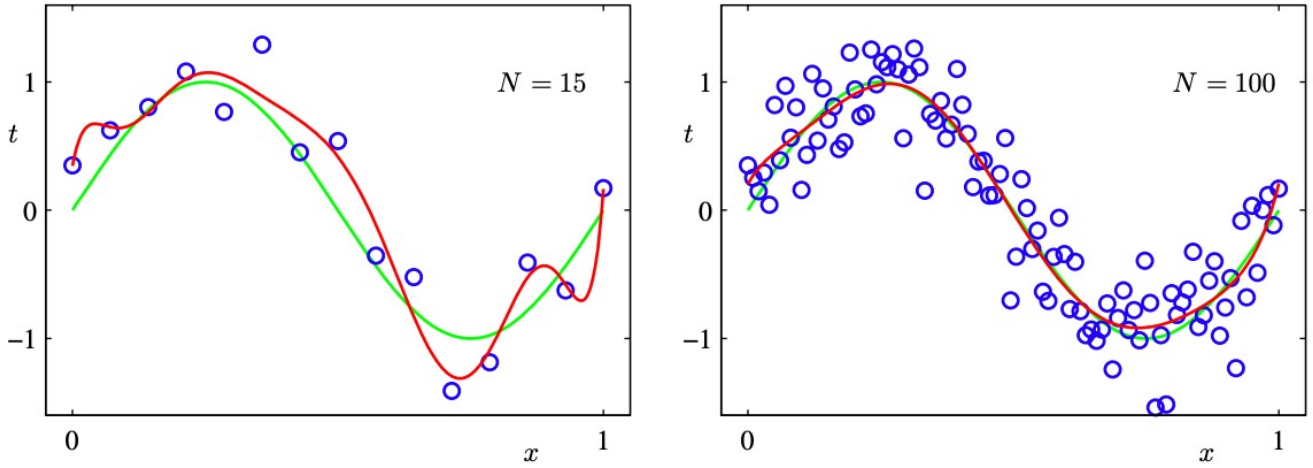
Моделирование функции $y = \sin(2\pi x) + \epsilon$ полиномами разных степеней



Графики полиномов при увеличенном кол-ве объектов

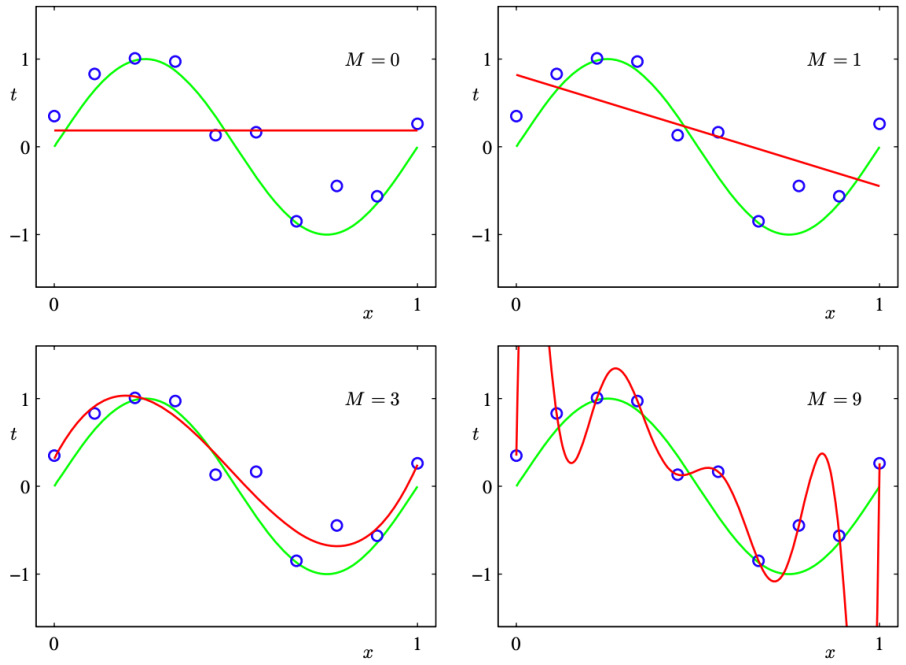
Переобучение при небольшом размере обучающей выборки

Обучение полинома степени 9 на 15 примерах и 100 примерах обучающей выборки. Чем больше данных, тем лучше мы справляемся с переобучением.



Чем больше степеней свободы у полинома, тем ниже дисперсия на обучающей выборке, тем ниже способность к обобщению на новые данные

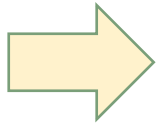
Коэффициенты при полиномах разной степени



	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Линейно зависимые признаки

1. Пускай решаем задачу предсказания стоимости аренды проживания в городах и для этого собираем ответы респондентов для обучения модели предсказания. Среди всех, в анкете есть вопросы:
   1. Средние затраты на ЖКХ в месяц – f_1
   2. Средние затраты на аренду в месяц – f_2
   3. Сколько денег вы откладываете– f_3



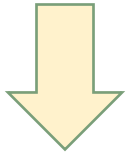
$$f3 \sim f1 + f2 \rightarrow \sum \theta_i f_i \approx 0$$

2. Значит возможно добавить такие константы $\gamma_1, \gamma_2, \gamma_3$ к нашим весам $\theta_1, \theta_2, \theta_3$, что значение решающей функции не изменится:

$$(\theta + \gamma, x)$$

Следовательно существует множество весов θ, которые являются решением нашей задачи:

$$f(x_1, x_2) = 1f_1 + 2f_2 - 3f_3 \approx 0$$
$$f(x_1, x_2) = 45f_1 + 15f_2 - 60f_3 \approx 0$$
$$f(x_1, x_2) = 336x_1 + 228f_2 - 564f_3 \approx 0$$

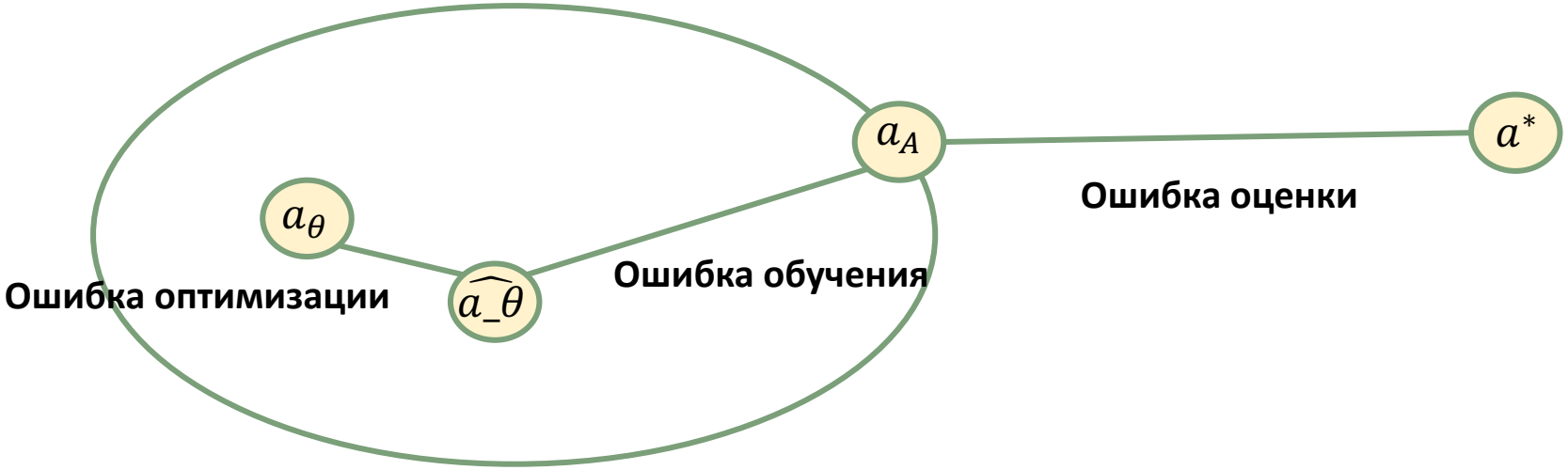


Большие веса решающей функции ведут к переобучению.

Регуляризация – общее понимание

$$\boldsymbol{E(a,y) = variance(a) + \sigma^2 + bias^2(a,y)}$$

- Регуляризация – способ, уменьшающий качество модели на обучающей выборке, для будущего роста на тестовой выборке.
- Регуляризация – способ, увеличивающий bias модели, но снижающий ее variance.
- Регуляризация – возможность найти модель a_θ, близкую к a_A, при этом не переобучится.



Регуляризация

Избежать переобучения(больших весов θ) можно с помощью добавления нового слагаемого в функцию ошибки $\lambda\left|\left|\theta\right|\right|_N$

$$L_\theta(x, y) = L_\theta(X, y) + \frac{\lambda}{2}\theta^2 \rightarrow min$$

Градиент функции ошибки:

$$\nabla_\theta L_\theta(x_i, y_i) = \nabla_\theta L_\theta(x_i, y_i) + \lambda\theta$$

Значит в градиентном спуске

$$\theta_{i+1} = \theta_i - \mu(\nabla L_\theta + \lambda\theta_i)$$

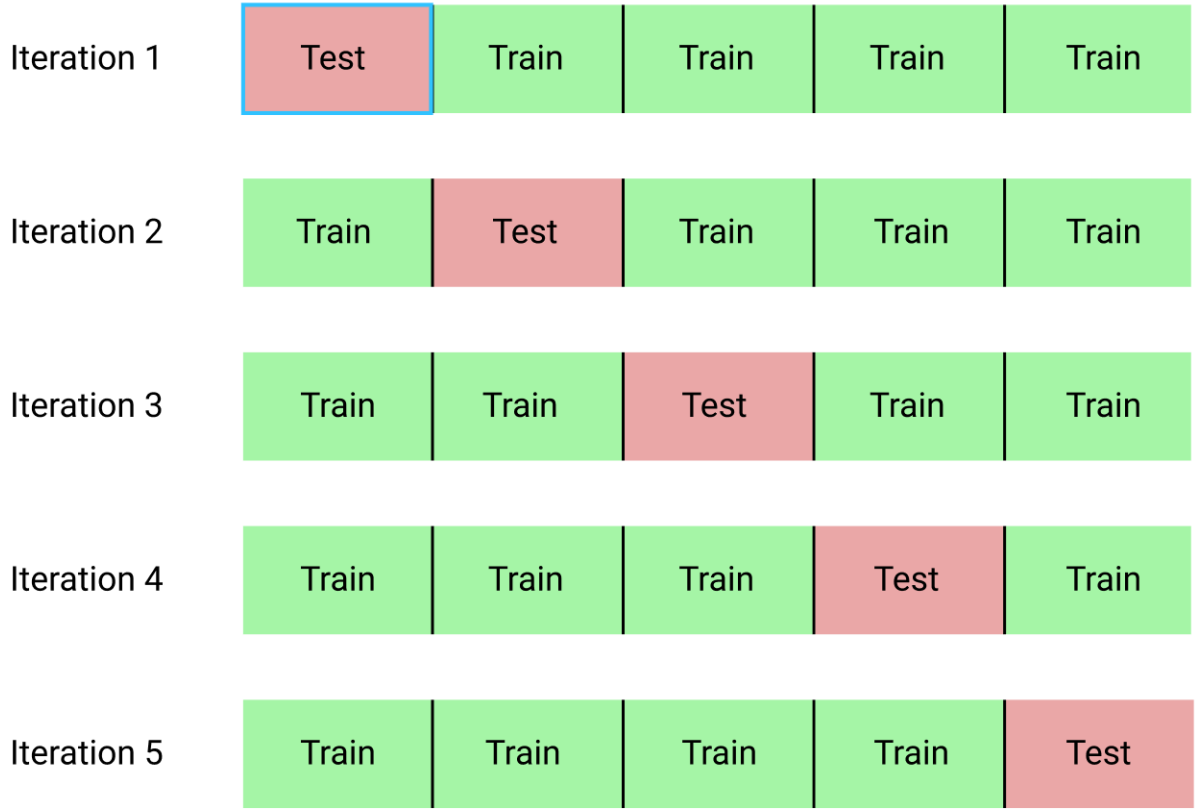
$$\theta_{i+1} = \theta_i(1 - \mu\lambda) - \mu\nabla L_\theta$$

Уменьшение текущего веса на каждой итерации

Чем больше коэффициент регуляризации, тем больше мы уменьшаем веса. В градиентном спуске эта процедура называется уменьшение веса **(weight decay)**

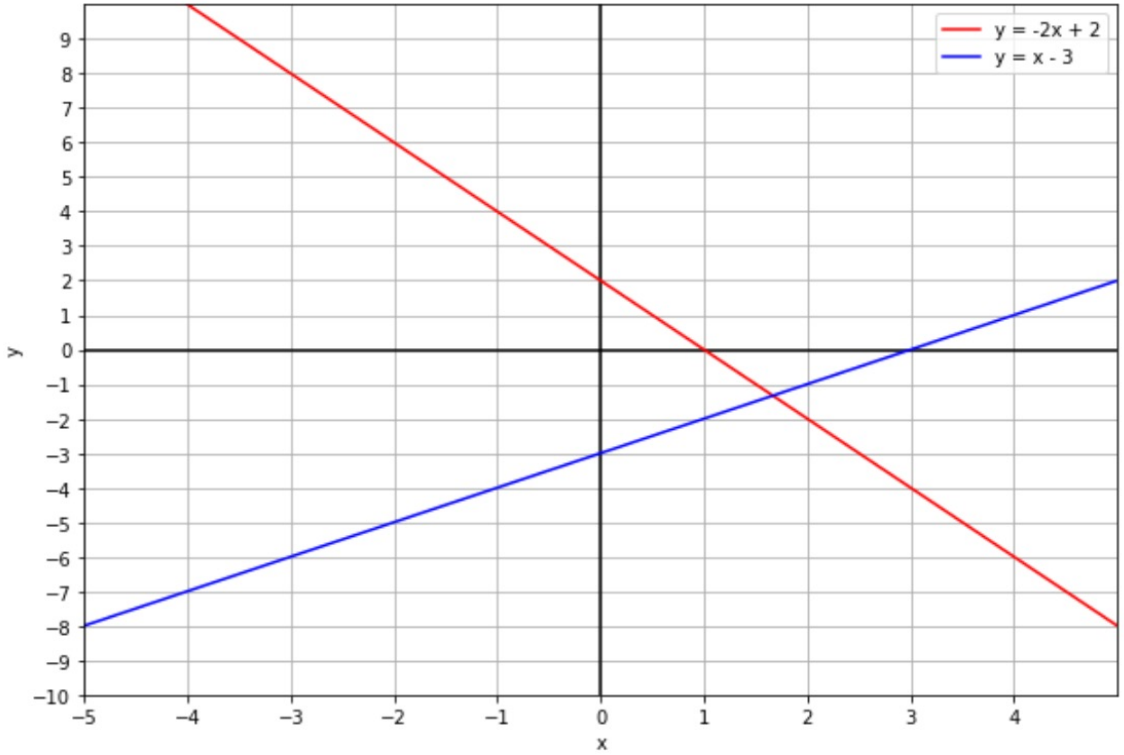
Подбор параметра регуляризации

- Коэфициент λ – гиперпараметр, его нельзя подобрать на обучающей выборке
- Подбор гиперпараметров – на отложенной выборке или кросс-валидации.

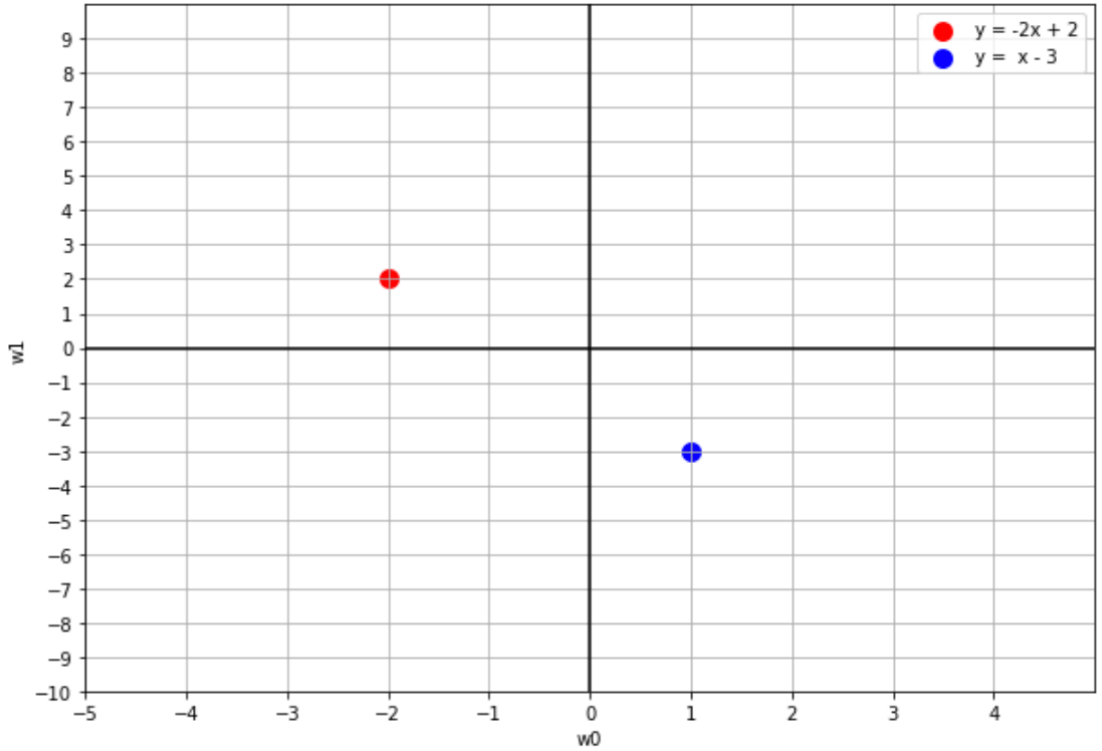


Линейная регрессия и пространство весов

Рассмотрим линейную регрессию в двумя параметрами $y = w_1 * x + w_0$ и отобразим параметры линейной регрессии в пространстве весов



Пространство моделей

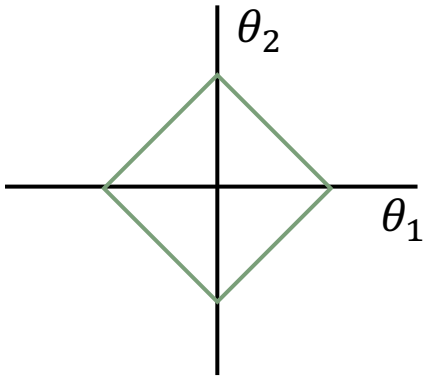


Пространство весов

L1 & L2

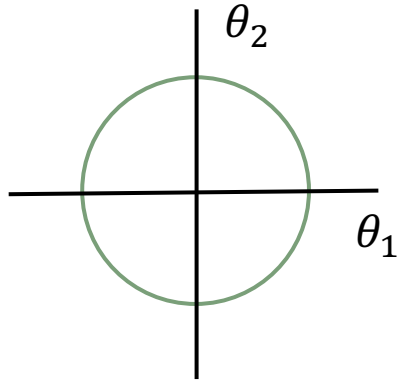
- Ограничимся 2-мерным линейным пространством
- Рассмотрим пространство линейных функций $\mathrm{F} = \{f(x) = \theta_1 x_1 + \theta_2 x_2\}$, которое отражается в $\{\theta_1, \theta_2\} \in R^2$
- Регуляризация – добавление нового взвешенного слагаемого $\lambda||\theta||_q^q$, где λ – масштаб эффекта регуляризации, q – норма весов($q \geq 1$)

$$\mathrm{L}(\theta) = \frac{1}{n}||X\theta - y||^2 + \lambda||\theta||_n - \text{эллипс}$$



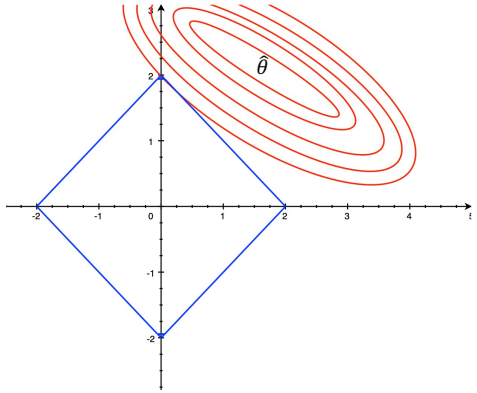
L1 регуляризация(Lasso)

$|\theta_1| + |\theta_2| \leq r$

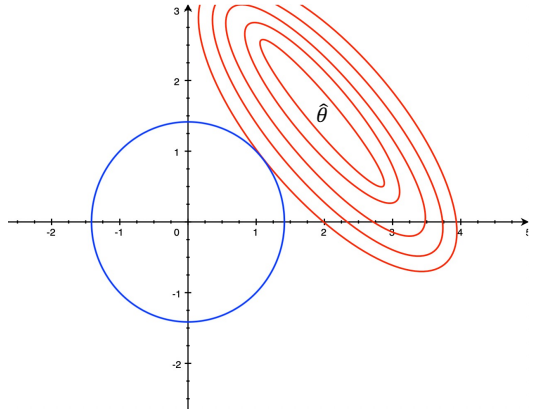


L2 регуляризация(Ridge)

$\theta_1^2 + \theta_2^2 \leq r^2$



L1(Lasso)



L2(Ridge)

Оптимизация с ограничениями при регуляризации

Рассмотрим задачу оптимизации с ограничениями:

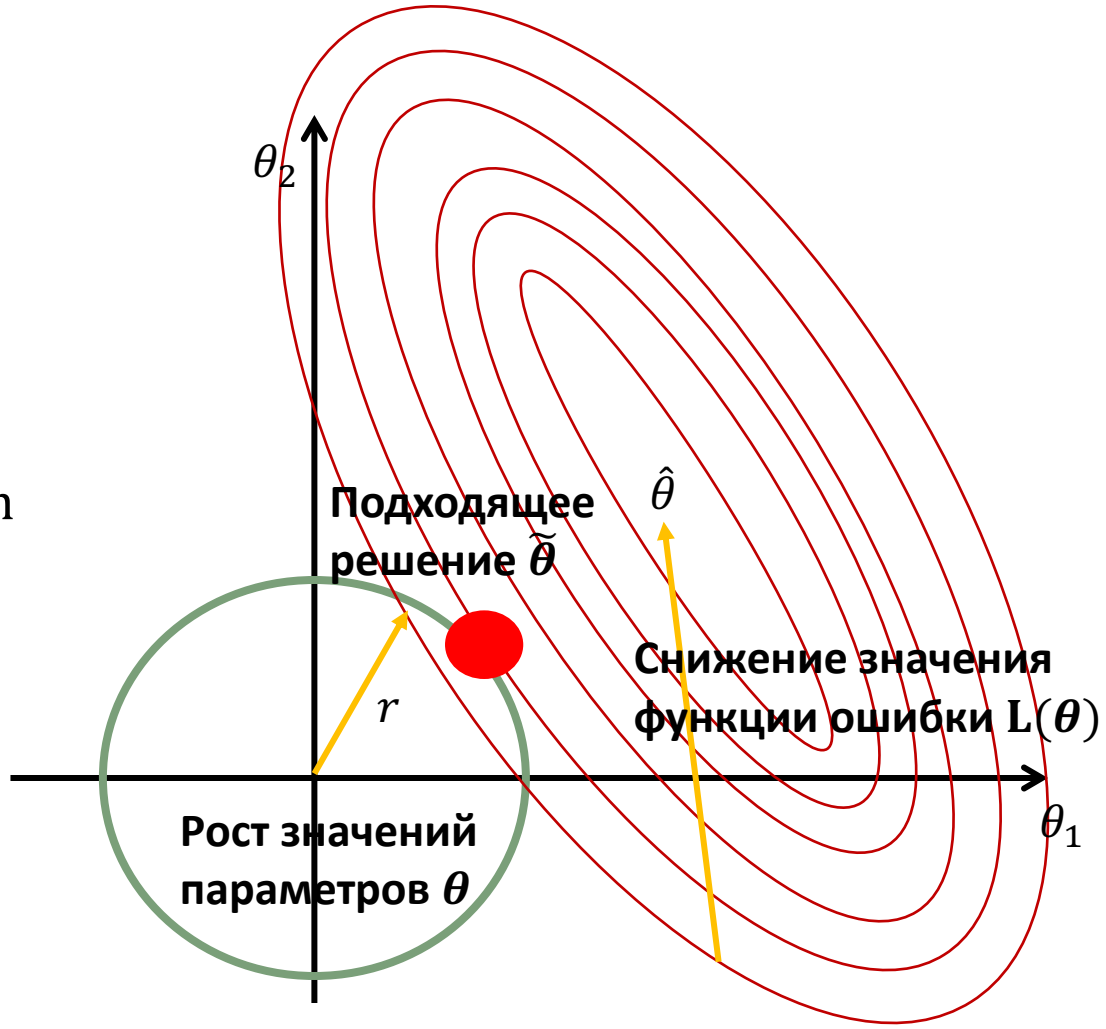
$$L(\theta) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i^1 - \theta_2 x_i^2\right) \to \min \qquad \text{s.t}\ \theta_1^2 + \theta_2^2 \le r^2$$

Для решения этой оптимизационной задачи с ограничениями воспользуемся методом множителей Лагранжа:

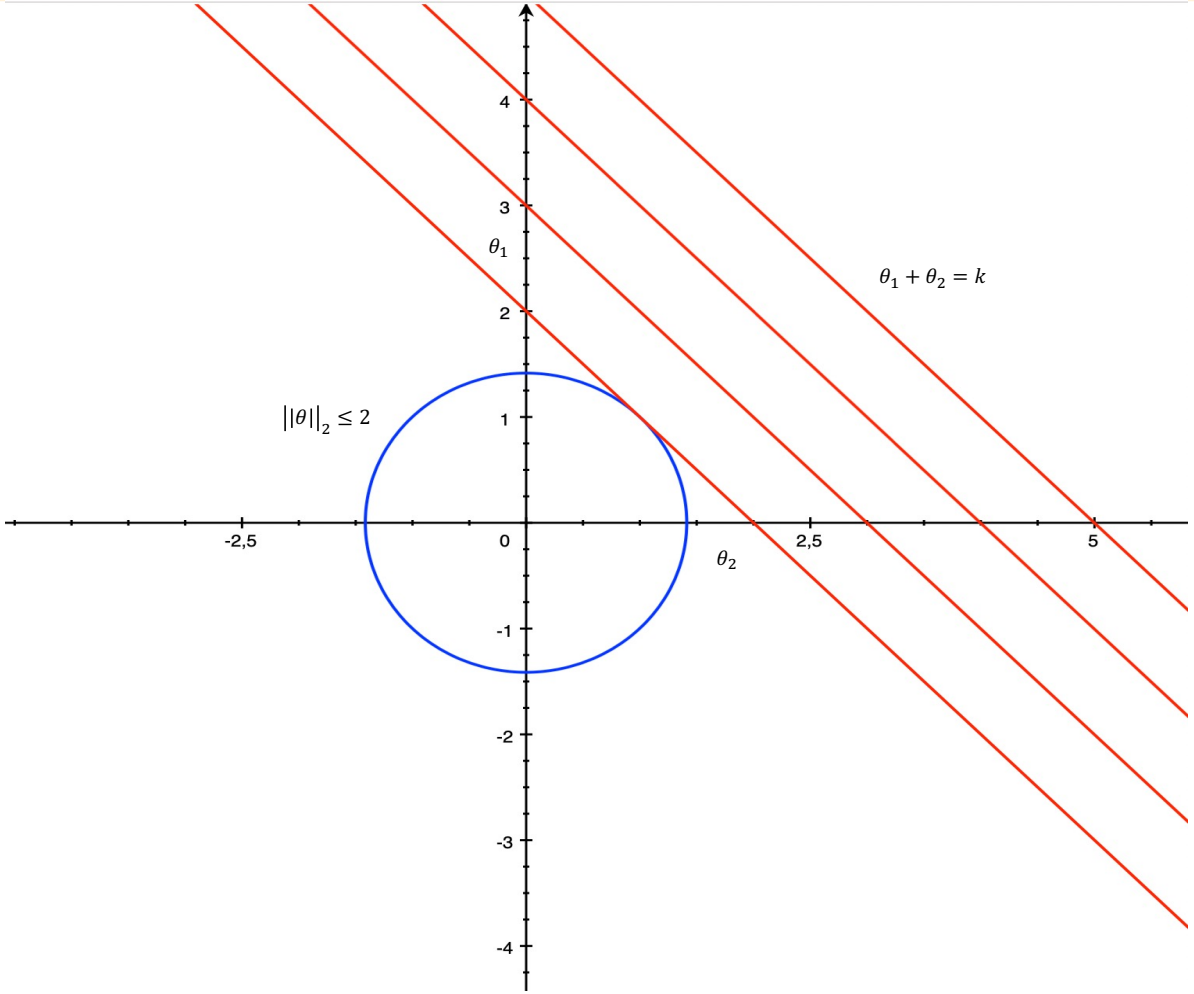
$$L(\theta, \lambda) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i^1 - \theta_2 x_i^2\right) + \lambda(\theta_1^2 + \theta_2^2 - r^2) \to \min_{\theta,\lambda}$$

Так как λ – это гиперпараметр (просто число) и r^2 - тоже, мы можем исключить их из задачи оптимизации. Получаем следующее выражение:

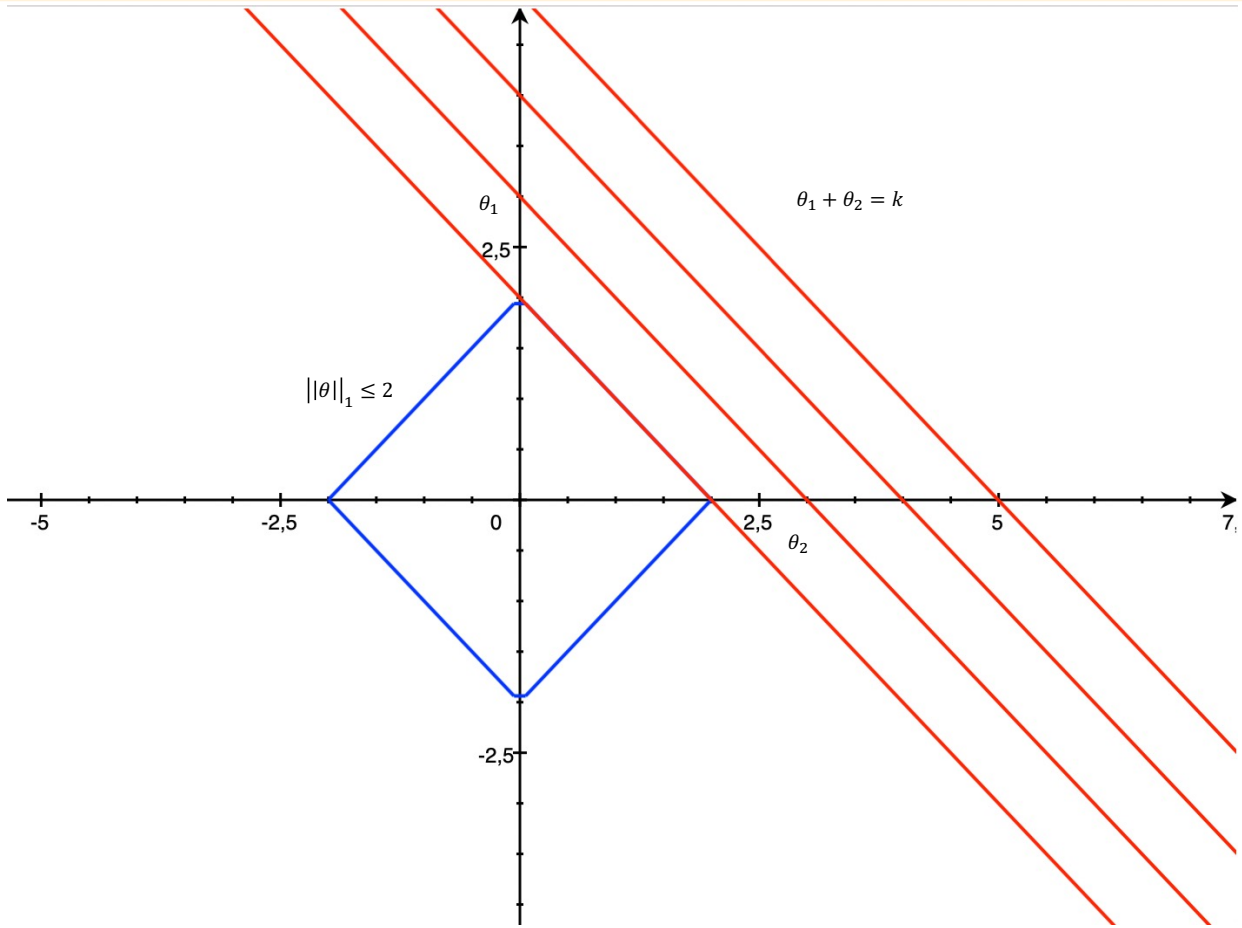
$$L(\theta) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i^1 - \theta_2 x_{x_i^2}\right) + \lambda(\theta_1^2 + \theta_2^2) \to \min_{\theta}$$



Линейно зависимые признаки

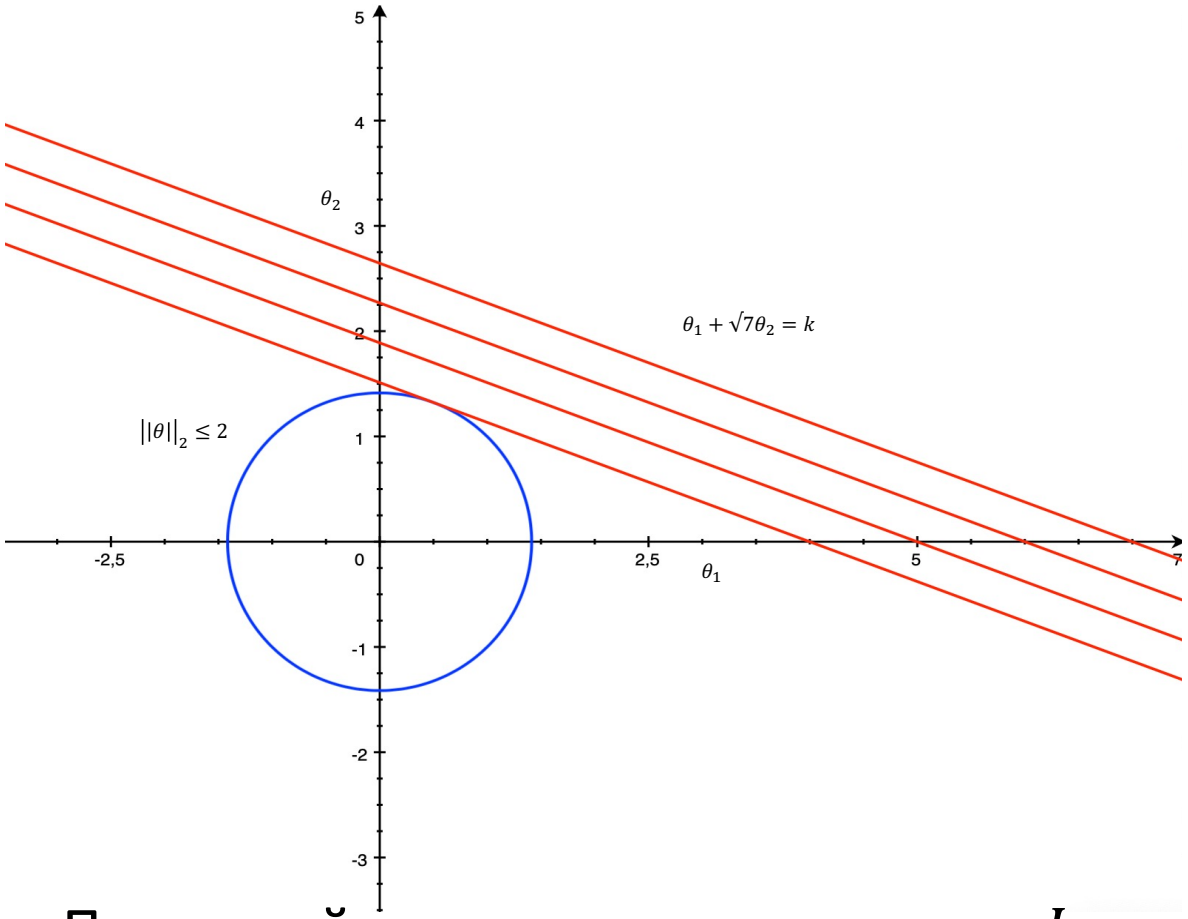


При равных признаках в L2 – вес распределяется одинаково

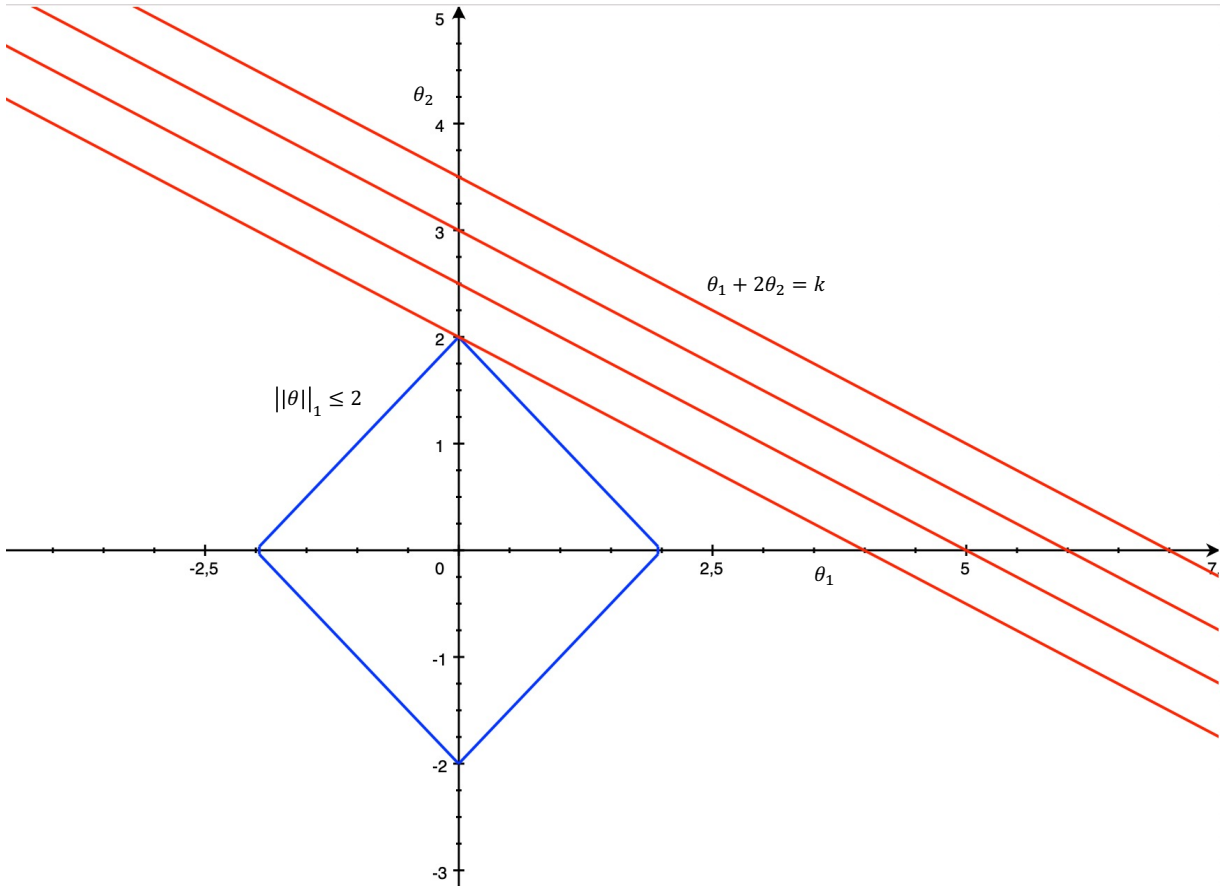


При равных признаках в L1 – вес признаков распределяется произвольно (либо в θ_1 либо в θ_2)

Линейно зависимые признаки



При линейно зависимых признаках L_2 регрессия распределяет веса соответственно масштабу признаков



При линейно зависимых признаках L_1 регрессия выбирает признаки с большим масштабом – другим зануляет веса

Ключевые формулы теории вероятности

Условное распределение случайной величины(с.в.) x при условии влияния с. в. y:

$\boldsymbol{p(x|y)}$ - Правдоподобие (likelihood)

$$\boldsymbol{p(x|y)} = \frac{p(x,y)}{p(y)} \qquad Conditional = \frac{Joint}{Marginal}$$

При независимости с. в. x и y:

$$p(x|y) = \frac{p(x) * \cancel{p(y)}}{\cancel{p(y)}} = p(x)$$

Формула произведения вероятностей (Product rule):

$$p(x,y) = p(x|y) * p(y) \Rightarrow p(x_1, \dots, x_n) = p(x_n|x_1, \dots, x_{n-1}) * p(x_{n-1}|x_1, \dots, x_{n-2}) * \dots * p(x_2|x_1) * p(x_1)$$

Любое совестное распределение вероятностей можно декомпозировать

За счет декомпозиции:

$$p(y|x) * p(x) = p(x|y) * p(y) \Rightarrow \boldsymbol{p(y|x)} = \frac{p(x|y) * p(y)}{p(x)}$$

Формула Байеса

Формула полной вероятности – вероятность принять значение x, при условии принятия значений y:

$$p(x) = \int p(x|y)p(y)dy = \int p(x,y)dy$$

Метод максимального правдоподобия

Пускай есть выборка X из неизвестного распределения, нам нужно оценить параметры θ этого распределе

$$X = (x_1, x_2, \dots, x_n) \sim p_\theta(x) = p(x|\theta)$$

Оценка максимального правдоподобия $L(\theta|X)$:

- $L(\theta|X)$ - дифференцируема по θ
- $L(\theta|X)$ - унимодальная функция

$$L(\theta|X) = \prod_{i=1} p(x_i|\theta) \to \max_\theta \qquad \log L\,(\theta|X) = \sum_{i=1} \log p(x_i|\theta) \to \max_\theta$$

Для поиска θ^*_{ML} возьмем градиент логарифма правдоподобия и приравняем к 0:

$$\frac{\partial}{\partial\theta} logL(\theta|X) = 0$$

θ^*_{ML} - эффективная, состоятельнае, ассимптотически нормальная оценка θ

Связь метода максимального правдоподобия и минимизации эмпирического риска

Метод максимального правдоподобия

$$\log L(\theta) = \sum_{i=1}^{n} \log p(y_i|x_i,\theta) \to \max_{\theta}$$

Аппроксимация эмпирического риска

$$\widehat{R_n}(\theta) = \sum_{i=1}^{n} L(\mathrm{a}_\theta(x_i), y_i) \to \min_{\theta}$$

$$-\log L(\theta) = L(\mathrm{a}_\theta(x_i), y_i)$$

Модель $p(y_i|x_i,\theta)$ ⇄ Модель $\mathrm{a}_\theta(x_i)$ и функция ошибки L

Расмотрим линейную модель $a_\theta(x_i) = < x, \theta >$, заданную уравнением

$$y_i = < x_i, \theta > + \epsilon, \text{где } \epsilon \in N(0, \sigma_i^2)$$

В таком случае $y_i \sim N(< x_i, \theta >, \sigma_i^2)$, где $< x_i, \theta >$ - мат. Ожидание нормального распределения. Запишем оценку на y_i , как

$$p(y_i|x_i,\theta) \sim N(y_i|(< x_i, \theta >, \sigma_i^2)$$

ММП и МНК совпадают, а веса объектов пропорциональны дисперсии шума. **Чем более шумный признак – тем меньше у него вес**

Запишем функцию правдоподобия нормального распределения и прологарифмируем с минусом

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{(2\pi)}} \exp(-\frac{1}{2\sigma_i^2}(y_i - < x_i, \theta >)^2)$$

$$-\log L(\theta) = \log \frac{1}{\sigma_i\sqrt{2\pi}} + \frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(< x_i, \theta > - y_i)^2 \to \min_{\theta}$$

Вероятностный смысл регуляризации

Рассмотрим совместное распределение весов θ и данных X. Вектор весов θ теперь также задан распределением. Тогда формула правдоподобия выражается как:

$$p(X, \theta) = p(X|\theta) * p(\theta)$$

Принцип максимума апостериорной вероятности (Maximum posteriory probability, MAP):

$$L(\theta) = \log p(X, \theta) = \sum_{i=1}^{n} \log p(y_i|x_i, \theta) + \log p(\theta) \to \max_{\theta}$$

Оценка ММП обоснована при $n >> d$, в ином случае – мы не можем пользоваться ММП для нахождения θ^*

Пускай веса θ распределены независимо, $E\theta = 0, D\theta = C$

Распределение Гаусса $p(\theta) = \frac{1}{\sqrt{(2\pi C)}} \exp(-\frac{||\theta||^2}{2C})$ $-\log p(\theta) = \sout{const} + \frac{1}{2C}||\theta||^2$ **L2 регуляризация**

Распределение Лапласа $p(\theta) = \frac{1}{2C} \exp(-\frac{||\theta||}{C})$ $-\log p(\theta) = \sout{const} + \frac{1}{C}||\theta||$ **L1 регуляризация**

$\frac{1}{C}$ - коэффициент регуляризации

Другая интерпретация регуляризации

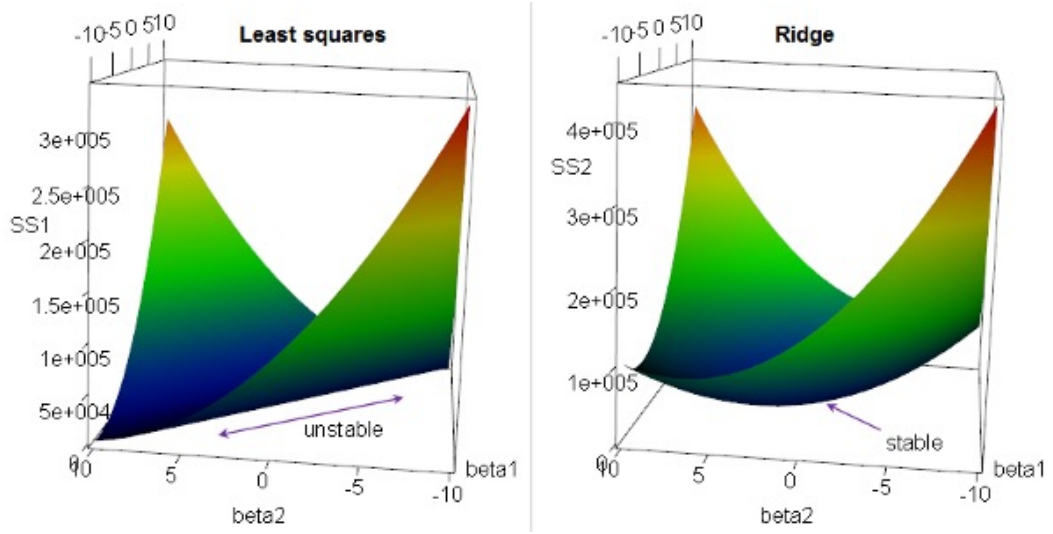
Ridge

$$L_\theta(x,y) = \frac{1}{n}\left|\left|X\theta\ - y\right|\right|^2 + \lambda\theta^T\theta$$

$$\theta^* = \left(\mathrm{X^T X} + \lambda I\right)^{-1}\mathrm{X^T y}$$

$\mathrm{X^T X}$ – может быть не обратимой при линейной зависимости признаков

$\mathrm{X^T X} + \lambda I$ – обратимая матрица, которая гарантирует решение MSE



В случае MSE без регуляризации может быть множество оптимальных параметров. L_2 добавляет возможность уменьшить пространство выбора.

Lasso

$$L_\theta(x,y) = \frac{1}{n}\left|\left|X\theta\ - y\right|\right|^2 + \lambda|\theta|$$

$$\theta^{k+1} = S\left(\theta^k\ - \mu\nabla L_\theta(x,y)\right)$$

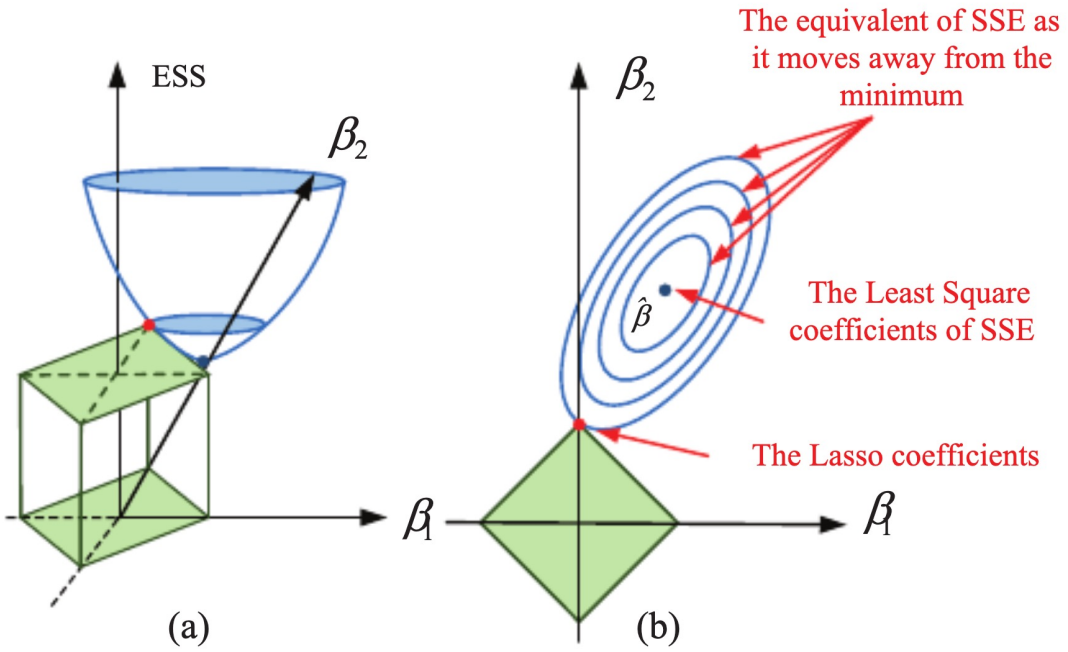
S – покомпонентная функция по вектору θ

$$S(\theta_i) = \begin{cases} \theta_i\ - \mu\lambda\ if\ \theta_i > \mu\lambda \\ 0\ if\ |\theta_i| < \mu\lambda \\ \theta_i + \mu\lambda\ if\ \theta_i < -\mu\lambda \end{cases}$$

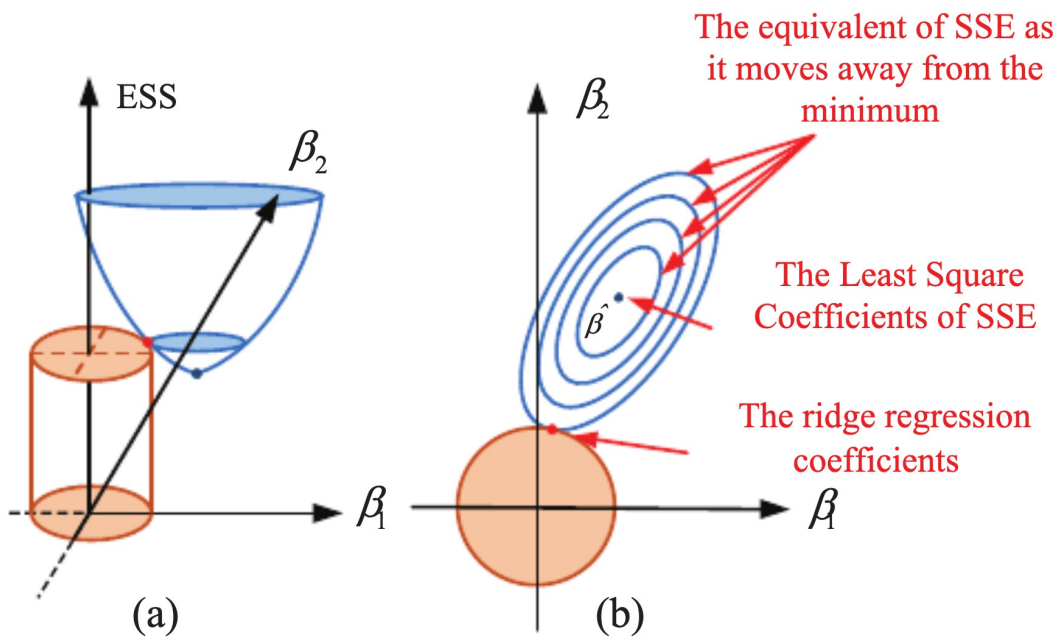
Если на данном шаге значение некоторого веса не очень большое, то на следующем шаге этот вес будет обнулён, причём чем больше коэффициент регуляризации, тем больше весов будут обнуляться

Проксимальный метод

Lasso vs Ridge



Lasso выбор признаков



Ridge выбор признаков

https://www.researchgate.net/publication/334159821_Quality_Monitoring_and_Root_Cause_Diagnosis_for_Industrial_Processes_Based_on_Lasso-SAE-CCA#pf5