2021

Лекция 6. Регуляризация

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

ПАПУЛИН С.Ю. (papulin.study@yandex.ru)

Дополнительные материалы доступны по ссылке: https://github.com/mlmethods

Содержание

1. Регуляризация

При оценке параметров может произойти ситуация, при которой модель идеально соответствует обучающему множеству. В таком случае мы говорим про переобучение. Хорошая модель должна быть способна адекватно обобщать данные, которые используются для обучения, так, чтобы это обобщение распространялось и на данные, которые модель ранее не видела (например, в качестве таких данных используется тестовое множество). Переобученная модель имеет плохое обобщающее свойство.

Установка значений параметров, при которых модель отлично повторяет обучающие данные и использует для этого параметры с большими значениями, дает менее приемлемый результат, чем при параметрах, которые менее соответствуют обучающим данным, но при этом используют меньшие значения параметров.

Переобучение возникает, когда модель имеет большую дисперсию, что может быть вызвано большой гибкостью выбранной модели или использованием большого количества признаков, сопоставимым с количеством наблюдений. При оценке параметров методом наименьших квадратов (МНК) единственного решения и вовсе не будет, если количество параметров больше количества наблюдений.

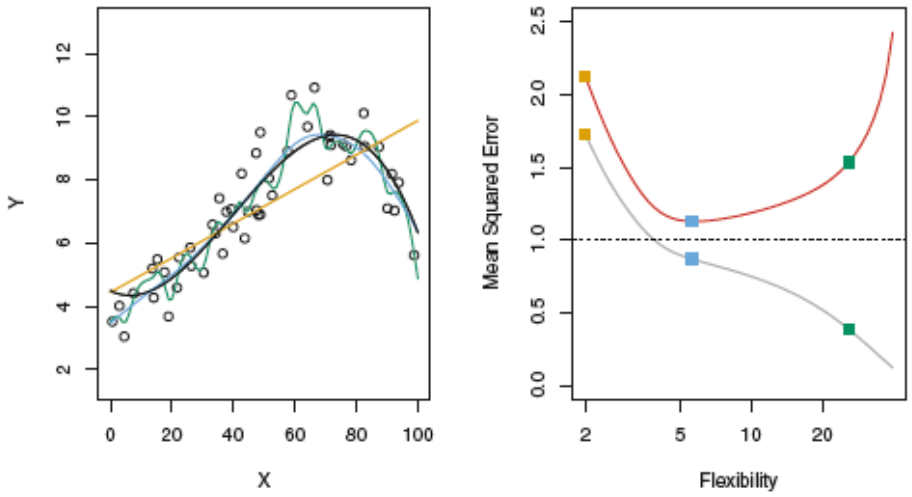


Рисунок – Кривые обучения и тестирования [1]

Чтобы избежать обозначенных проблем, в функцию потерь (целевую функцию) вводится дополнительный компонент – регуляризация, $R(\theta)$.

Существует два основных вида регуляризации: L1 и L2. L2 регуляризация использует L2 норму оценок параметров:

$$R(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{p} \theta_j^2$$

L1 регуляризация использует L1 норму:

$$R(\theta) = \|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|$$

Модели с регуляризацией, как правило, чувствительны к масштабу входных признаков. Поэтому рекомендуется использовать стандартизацию.

2. Регуляризация линейной регрессии

2.1. Ридж-регрессия (Ridge Regression)

Другое название ридж-регрессии – гребневая регрессия или регуляризация Тихонова

В качестве функции потерь $L(\theta)$ при оценке параметров модели линейной регрессии (то есть при обучении) используется MSE или RSS, взаимосвязь между которыми следующая:

$$MSE = \frac{1}{n} RSS,$$

где

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{p} \theta_j x_{ij} \right]^2 .$$

Задача обучения сводится к задаче оценки параметров модели, которую можно записать следующим образом:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} L(\theta).$$

Данную задачу можно решить методом наименьших квадратов или итеративными алгоритмами с использованием частных производных разных порядков или их аппроксимаций (градиентный спуск, стохастический градиентный спуск, методы Ньютона, квазиньютоновские методы и пр.).

В случае с ридж-регрессией функция потерь имеет следующий вид

$$L(\theta) = RSS + \lambda \sum_{j=1}^{p} \theta_j^2 = \underbrace{\sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right]^2}_{RSS} + \underbrace{\lambda \sum_{j=1}^{p} \theta_j^2}_{\text{Штраф}},$$

где $\lambda \geq 0$ – настраиваемый гиперпараметр.

И, соответственно, оценка параметров

$$\hat{\theta}_\lambda^R = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right]^2 + \lambda \sum_{j=1}^{p} \theta_j^2$$

Функция потерь состоит из двух компонент: RSS и штрафа. Первых компонент ридж-регрессии ищет оценки параметров, которые хорошо повторяют обучающие данные, тем самым уменьшая значение RSS. В то же время штраф $\lambda \sum_{j=1}^{p} \theta_j^2$ (shrinkage penalty) имеет небольшое значение, когда $\theta_1, \theta_2, \ldots, \theta_p$ близки к нулю, то есть он имеет эффект сокращения оценок параметров к нулю. Таким образом, параметр λ контролирует относительное влияние каждой составляющей на оценку параметров. Введенный дополнительный компонент штрафа определяет ридж-регрессию, как линейную регрессию с L2 регуляризацией.

При $\lambda = 0$ составляющая, отвечающая за штраф, не имеет никакого эффекта, и оценка параметров ридж-регрессии будет оцениваться исключительно по RSS. Однако при увлечении λ влияние штрафа возрастает, и оценка параметров ридж-регрессия при $\lambda \to \infty$ будет стремиться к нулю.

Ридж-регрессия будет иметь разные оценки параметров $\hat{\theta}_\lambda^R$ для каждого значения λ. Выбор хорошего значения λ критически важен.

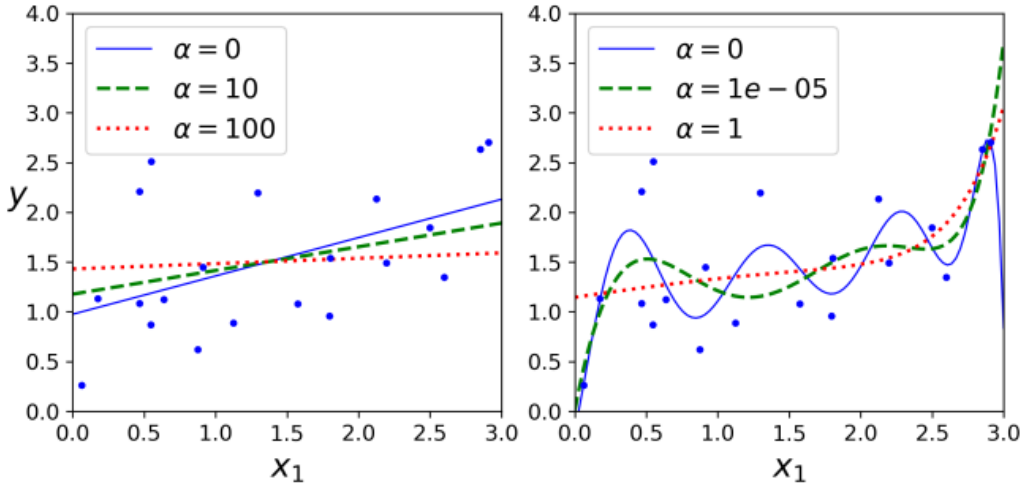


Рисунок x – Пример ридж-регрессии с разными значениям коэффициента регуляризации [4]: а) простая линейная регрессия; б) полиномиальная регрессия со степенью 10 (признаки стандартизованы)

Также следует отметить, что мы хотим сократить влияние каждой оцениваемой переменной на отклик, однако мы не хотим уменьшить смещение θ_0.

Матричная запись функции потерь для ридж-регрессии

Вариант 1. Без учета смещения θ_0

Представим параметр θ_0 как

$$\theta_0 = \bar{y} - \bar{\boldsymbol{x}}^T \boldsymbol{\theta}.$$

Тогда формулу RSS можно записать в следующем виде:

$$\sum_{i=1}^{n} \left[y_i - \theta_0 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta} \right]^2 = \sum_{i=1}^{n} \left[y_i - (\bar{y} - \bar{\boldsymbol{x}}^T \boldsymbol{\theta}) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta} \right]^2 = \sum_{i=1}^{n} [(y_i - \bar{y}) - (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \boldsymbol{\theta}]^2$$

Центрирование (лучше даже стандартизовать):

$$\boldsymbol{x} \leftarrow \boldsymbol{x} - \bar{\boldsymbol{x}}$$

$$y \leftarrow y - \bar{y}$$

Оптимизация без учета θ_0:

$$\hat{\theta}_\lambda^R = \underset{\theta}{\operatorname{argmin}} \left[(\mathbf{y} - X\boldsymbol{\theta})^T(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\right]$$

Оценка параметров:

$$\hat{\theta}_\lambda^R = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$$

и

$$\hat{\theta}_0 = \bar{y} - \bar{\boldsymbol{x}}^T \hat{\boldsymbol{\theta}}_\lambda^R$$

где I – единичная матрица размера p на p; X – матрица признаков размера n на p

Вариант 2. С учетом смещения θ_0

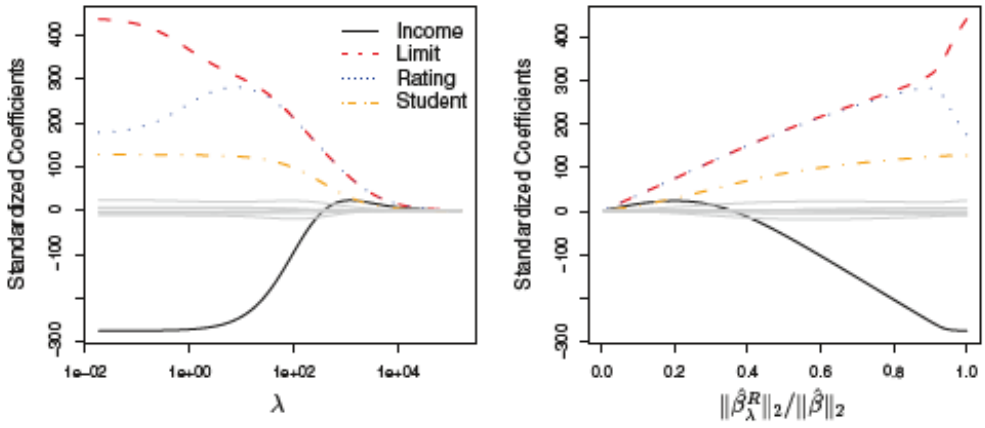
Оценка параметров:

$$\hat{\theta}_\lambda^R = (X^T X + \lambda A)^{-1} X^T \mathbf{y},$$

где A – единичная матрица размера $p+1$ на $p+1$, в которой элемент $a_{0,0} = 0$; X – матрица признаков размера n на $p+1$ (дополнительный первый столбец – это столбец единиц)

Пример

- 10 параметров для 10 признаков
- $\| \hat{\theta}_\lambda^R \| / \| \hat{\theta} \|$, где
  - $\|\theta\|_2 = \sqrt{\sum_{j=1}^{p} \theta_j^2}$,
  - $\hat{\theta}$ – оценка МНК,
  - $\hat{\theta}_\lambda^R$ – оценка с коэффициентом регуляризации λ



- Слева на графике λ имеет небольшое значение. При $\lambda = 0$ значения оценок параметров ридж-регрессии будут равны оценкам обычной линейной регрессии без регуляризации.

- С увеличением λ оценки параметров имеют общую тенденции к уменьшению
- При этом некоторые оценки параметров могут увеличиваться по мере роста λ (например, «income» и «rating»)
- При больших значениях λ оценки параметров стремятся к нулю, что соответствует нулевой модели (null model), в которой нет предикторов
- $\| \hat{\theta}_\lambda^R \|$ всегда уменьшается с ростом λ так же, как и $\| \hat{\theta}_\lambda^R \| / \| \hat{\theta} \|$
- Справа приведен график, который показывает изменения оценок параметров в зависимости от сокращения/уменьшения значения $\| \hat{\theta}_\lambda^R \|$ по отношению к $\| \hat{\theta} \|$, полученных при МНК. Небольшое значение $\| \hat{\theta}_\lambda^R \| / \| \hat{\theta} \|$ говорит о том, что оценки ридж-регрессии близки к нулю

Масштабирование значений признаков

- При МНК неважен масштаб предиктора, так как $\theta_j x_j$ остается постоянным (умножая x_j на константу a, оценка параметра МНК будет θ_j / a)
- При ридж-регрессии напротив оценки параметров могут значительно измениться с умножением предикторов на константу
- Это связанно с тем, что используется дополнительный компонент штрафа $\| \hat{\theta}_\lambda^R \|$ и при умножении на константу предиктора не будет давать простое масштабирование его оценки.
- $\hat{\theta}_{j,\lambda}^R x_j$ будет зависеть не только от λ, но также и от масштаба j-го предиктора
- Правильным вариантом будет стандартизация предикторов перед применением ридж-регрессии

$$\tilde{x}_{ij} = \frac{x_{ij}}{s_j} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

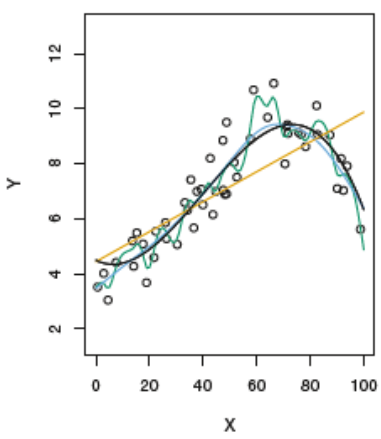
- В итоге все стандартизованные предикторы будут иметь стандартное отклонение равное 1
- Таким образом, конечный результат оценки параметров не будет зависеть от масштаба признаков

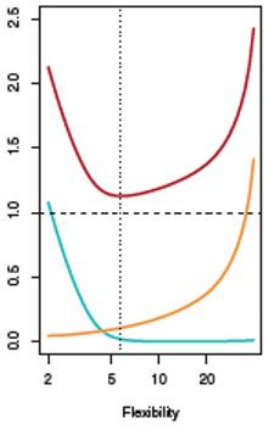
Сравнение с обычной линейной регрессией

- Начальное условие: мы полагаем, что наша модель, полученная посредством МНК, переобучена, например, мы имеем большое количество признаков, а количество наблюдений не сильно его превосходит.
- Преимущество ридж-регрессии над МНК кроется в компромиссе между смещением и дисперсией (bias-variance trade-off)

  Напоминание:

$$\mathrm{E}[(y - \hat{y})^2] = \big(f(x) - h(x)\big)^2 + \mathrm{var}\big[\hat{h}(x)\big] + \mathrm{var}[\varepsilon] = \underbrace{\mathrm{Bias}\big[\hat{h}(x)\big]^2}_{\text{Смещение}} + \underbrace{\mathrm{var}\big[\hat{h}(x)\big]}_{\text{Разброс}} + \underbrace{\sigma^2}_{\text{Шум}}$$

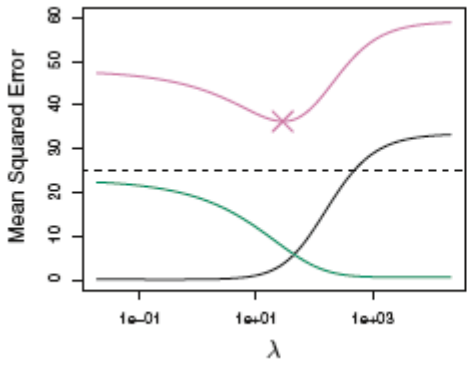


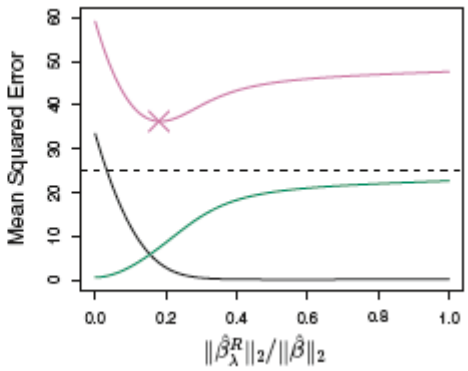


- С увеличением λ гибкость ридж-регрессии уменьшается, что ведет к уменьшению дисперсии, но увеличивается смещение.
- При МНК, что соответствует ридж-регрессии с $\lambda = 0$, дисперсия высокая, но малое смещение
- С увеличением λ уменьшаются оценки параметров, что ведет к существенному уменьшению дисперсии предсказания ценой небольшого увеличения в смещении.

Иллюстрация

- $p = 45$ предикторов and $n = 50$ наблюдений
- Зеленая кривая – дисперсию предсказаний ридж-регрессии как функцию от λ
- Черная кривая – квадрат смещения
- Фиолетовая кривая – тестовая MSE (сумма квадрата смещения, дисперсии предсказания и шума)
- Горизонтальная прерывистая линия – минимально возможная MSE (шум)
- Минимальная MSE достигается приблизительно при $\lambda = 30$
- График справа: при движении слева направо модель становится более гибкой, поэтому смещение уменьшается и увеличивается дисперсия





Применение

- В общем случае в ситуациях, когда взаимосвязь между откликом и предикторами близка к линейной, МНК будет давать небольшое смещение, но может иметь высокую дисперсию.
- Это означает, что небольшое изменение в обучающих данных может привести к большим изменениям в оценках параметров
- В частности, когда количество переменных p приблизительно соответствует количеству наблюдений n, оценки МНК будет очень изменчивы.

- Если $p > n$, то МНК не будет иметь единственного решения, в то время как ридж-регрессия по-прежнему может давать хорошую оценку за счет небольшого увеличения смещения при значительном уменьшении дисперсии.
- Когда много признаков с высокой корреляцией, оценки параметров плохо определимы и показывают высокую дисперсию. В этом случае может быть использована ридж-регрессия
- Таким образом, ридж-регрессия хорошо работает в ситуациях, где МНК имеет высокую дисперсию.

2.2. Лассо регрессия

Лассо регрессия (Least Absolute Shrinkage and Selection Operator – LASSO) – ещё один вид линейной регрессии с регуляризацией.

Проблема

- Ридж-регрессия включает все предикторы в конечную модель. При этом величина оценок параметров будет зависеть от λ
- Это не будет проблемой для точности предсказания, но может затруднить интерпретацию модели в ситуациях, когда количество переменных p велико

Лассо является альтернативой ридж-регрессии, которая позволяет решить обозначенную проблему. Функция потерь лассо-регрессии имеет следующий вид:

$$L(\theta) = RSS + \lambda \sum_{j=1}^{p} |\theta_j| = \underbrace{\sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right]^2}_{RSS} + \underbrace{\lambda \sum_{j=1}^{p} |\theta_j|}_{\text{Штраф}}$$

Таким образом, обучение сводится к оценке параметров, при которых функция потерь будет иметь минимальное значение:

$$\hat{\theta}_\lambda^L = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right]^2 + \lambda \sum_{j=1}^{p} |\theta_j|$$

- Лассо-регрессия соответствует линейной регрессии с L1 регуляризацией
- Подобно ридж-регресии, лассо стремится уменьшить оценки параметров до нуля при увеличении λ
- Однако в случае использования L1 штрафа некоторые оценки могут принять нулевое значение при регулировании параметра λ. Таким образом выполняется выбор/отбор признаков
- Когда часть признаков отпадает из-за нулевых оценок параметров, модель проще интерпретировать
- Так же, как и с ридж-регрессией, важным является правильный выбор λ

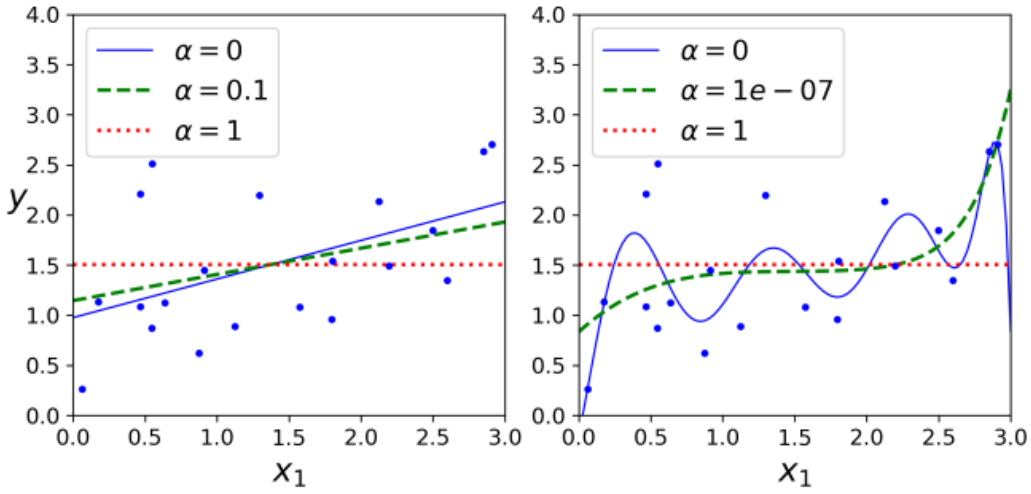
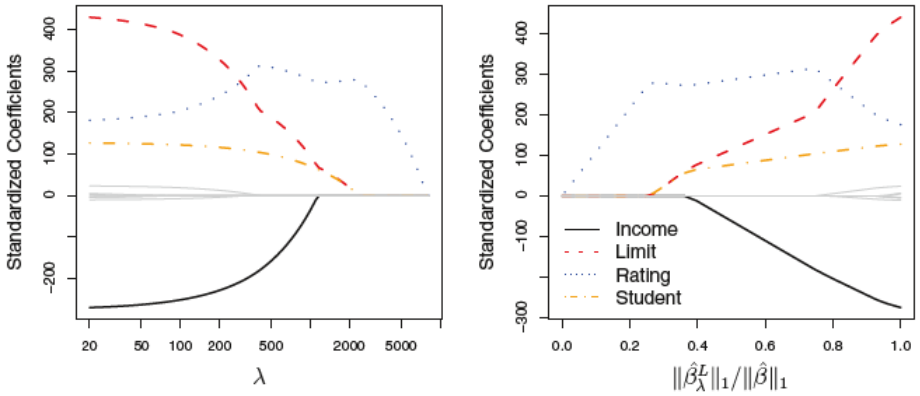


Рисунок x – Пример лассо-регрессии с разными значениям коэффициента регуляризации [4]: а) простая линейная регрессия; б) полиномиальная регрессия со степенью 10 (признаки стандартизованы)

Пример



- Когда $\lambda = 0$, лассо дает МНК оценки
- Когда λ принимает достаточно большое значение, лассо становится нулевой моделью, то есть, когда все оценки равны нулю
- График справа: двигаясь слева направо, мы можем наблюдать, что сначала модель имеет только признак «рейтинг», затем ещё признаки «студент» и «лимит»
- Таким образом, в зависимости от λ модель с лассо может иметь различное количество признаков.

2.3. Формулировка регуляризации через ограничение

Ридж (L2):

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right]^2 \text{ при условии } \underbrace{\sum_{j=1}^{p} \theta_j^2 \leq s}_{\text{Ограничение}}$$

Лассо (L1):

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - \theta_0 - \sum_{j=1}^{p} \theta_j x_{ij} \right]^2 \text{ при условии } \underbrace{\sum_{j=1}^{p} |\theta_j| \leq s}_{\text{Ограничение}}$$

- Для каждого значения λ существует значение s, такое что обозначенные выше выражения будут соответствовать введенным ранее функциям потерь для ридж и лассо регрессий и, соответственно, будут получены одни и те же оценки параметров
- См. метод множителей Лагранжа для задач условной оптимизации

Интерпретация через бюджет

- Ограничение есть бюджет на то, насколько большим может быть значение $\|\theta\|$
- Когда s имеет очень большое значение, бюджет не сильно ограничен, и следовательно оценки параметров могут быть большими
- Если s достаточно большое, то решение МНК может попасть в пределы бюджета. В этом случае мы просто получим МНК оценки.
- И наоборот, если s небольшое, тогда $\|\theta\|$ должно быть небольшим, чтобы не выйти за пределы бюджета

Иллюстрация

- Если s достаточно большое, то ограничивающий регион будет содержать $\hat{\theta}$ и, следовательно, оценки параметров при ридж и лассо регрессиях будут аналогичны МНК
- Как показано на рисунках ниже оценки МНК находятся за пределами окружности и ромба и поэтому оценки ридж и лассо будет отличаться от МНК
- Оценки ридж и лассо будут точки касания RSS и регионов ограничений
- Так как ридж-регрессия имеет окружность в качестве ограничения (то есть без острых углов), пересечение в общем случае не произойдет по осям координат, поэтому оценки будет ненулевыми
- Лассо ограничение имеет углы по каждой оси и поэтому RSS часто имеет пересечение с регионом ограничения вдоль осей координат. Когда это происходит, один из параметров будет равен нулю. В многомерном пространстве большое количество оценок параметров могут быть одновременно равны нулю

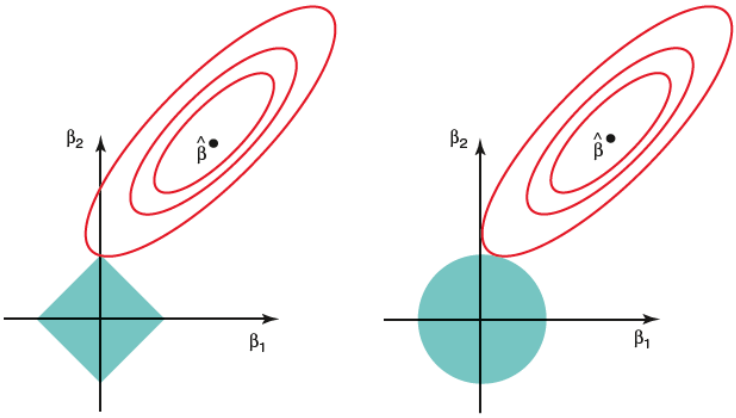


Рисунок – Ограничения для L1 и L2 регуляризации по двум параметрам [1]

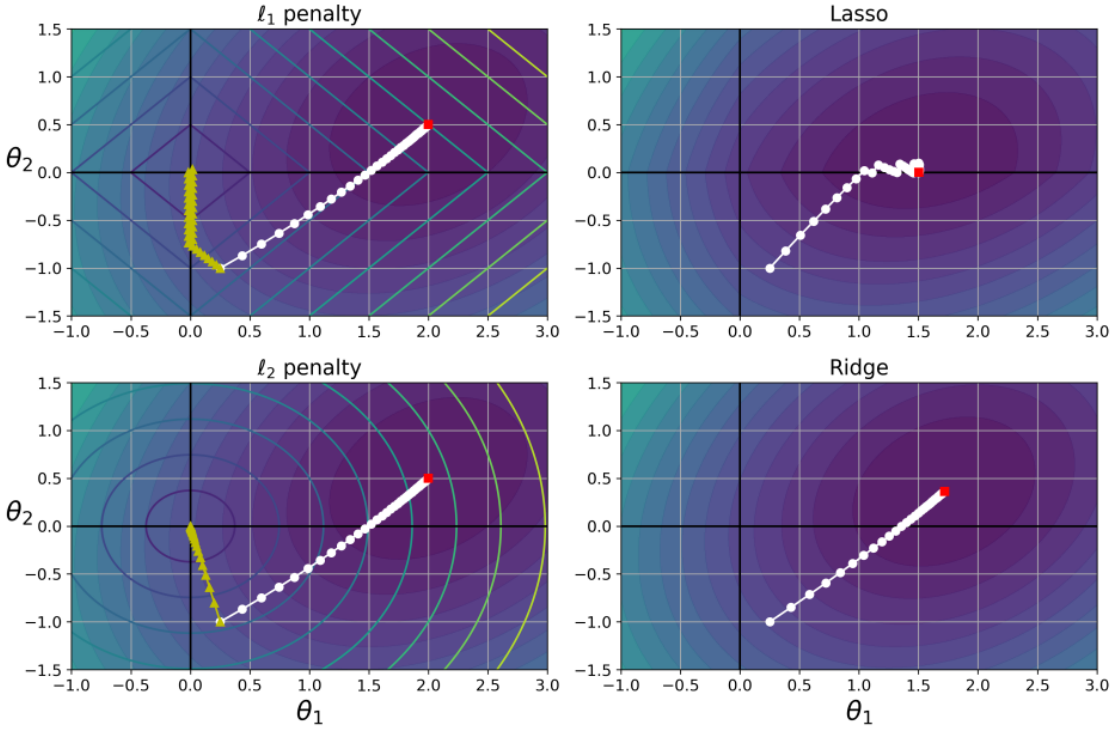
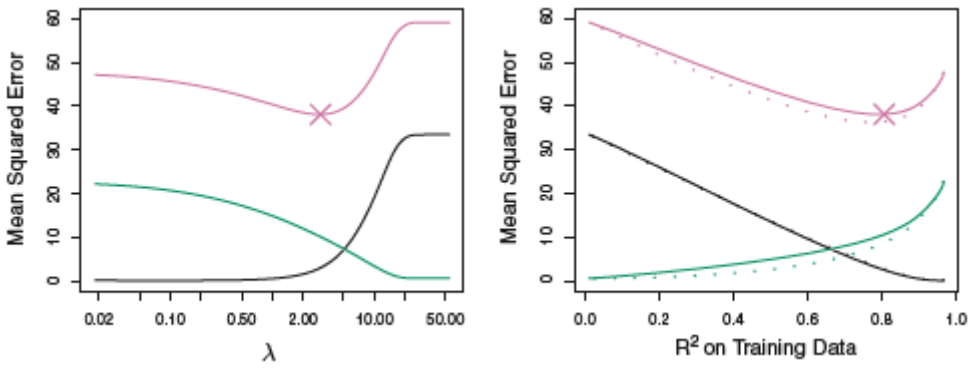


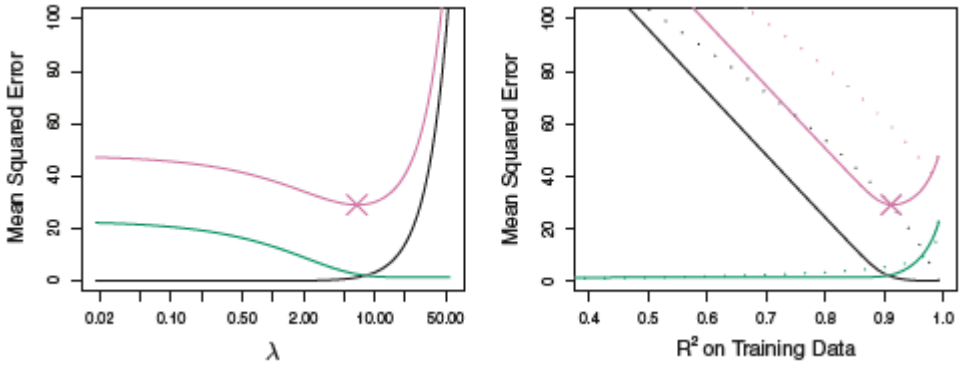
Рисунок – Пример L1 (сверху) и L2 (снизу) регуляризации линейной регрессии с использованием градиентного спуска для оценки параметров [4]: а) коэффициент регуляризации $\lambda = 0$ (белым) и $\lambda \to +\infty$ (желтым); б) коэффициент регуляризации $\lambda = 0.5$

Сравнение L1 и L2

- Важным преимуществом лассо по сравнению с ридж-регрессией является то, что она позволяет получить более простую модель с меньшим количеством предикторов, которую проще интерпретировать
- Пример, когда все действительные параметры $\theta_1, \dots, \theta_{45}$ не равны нулю
    - Дисперсия ридж-регрессии немного ниже, чем дисперсия лассо
    - Квадрат смещения практически совпадают
    - Поэтому минимальная MSE ридж-регрессии немного меньше, чем для лассо



- При использовании лассо мы полагаем, что некоторые параметры в реальности имеют нулевые значения и не оказывают существенного влияния на предсказание
- Пример, когда только два из 45 предикторов являются значимыми
    - Дисперсия и квадрат смещения выше у ридж-регрессии
    - В этом случае лассо имеет тенденцию быть лучше ридж-регрессии



- Ни одна из моделей, ридж и лассо, не имеет преимущества по отношению к другой
- В общем случае можно ожидать, что лассо лучше в ситуации, когда относительно небольшое количество предикторов имеют существенные значения параметров, а остальные предикторы имеют параметры близкие или равные нулю
- Ридж-регрессия будет вести себя лучше, когда отклик есть функция от многих предикторов с приблизительно равными параметрами
- Однако количество предикторов, которые связанны с откликом, никогда заранее неизвестно для реальных данных
- Так же, как и ридж-регрессия, когда оценки МНК имеют высокую дисперсию, с использованием лассо можно уменьшить дисперсию ценой небольшого увеличения смещения, и в результате получить более точные предсказания
- В отличие от ридж-регрессии лассо также выполняет выбор признаков и таким образом получаются модели, которые проще интерпретировать

2.4. Elastic-Net

Elastic-Net – линейная регрессия с L1 и L2 регуляризацией. Такая комбинация позволяет получить небольшое количество ненулевых оценок параметров подобно лассо с сохранением свойств ридж-регрессии. Данная регрессии может быть использована, когда несколько признаков имеют корреляционную зависимость (коррелируют).

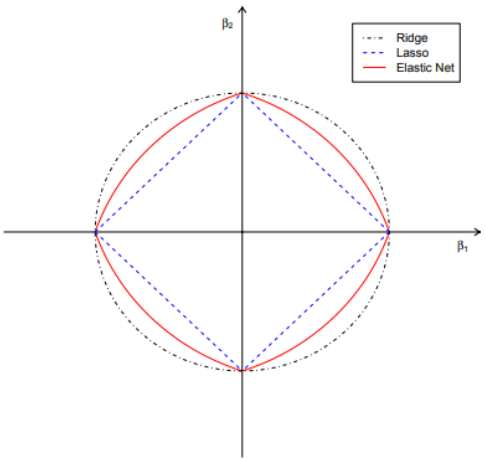
В общем виде оценку параметром можно представить в следующем виде

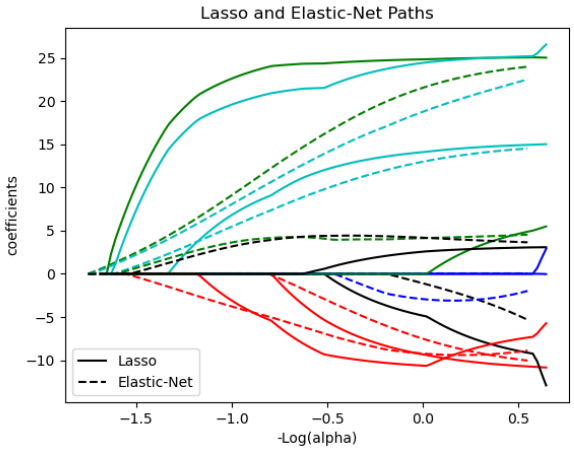
$$\hat{\theta}_{\lambda}^{EN} = \underset{\theta}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n}\left[y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij}\right]^2}_{RSS} + \underbrace{\lambda_1 \sum_{j=1}^{p}|\theta_j|}_{\text{Штраф } L1} + \underbrace{\lambda_2 \sum_{j=1}^{p}\theta_j^2}_{\text{Штраф } L2},$$

где λ_1 – коэффициент L1 регуляризации; λ_2 – коэффициент L2 регуляризации.

Влияние того или иного типа регуляризации можно контролировать через параметр ρ:

$$\hat{\theta}_{\lambda}^{EN} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n}\left[y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij}\right]^2 + \lambda\rho \sum_{j=1}^{p}|\theta_j| + \lambda(1-\rho)\sum_{j=1}^{p}\theta_j^2$$





$\rho = 0.8$

https://scikit-learn.org/stable/modules/linear_model.html#elastic-net

Предпочтительно иметь по крайней мере небольшую регуляризацию, чтобы избежать использования функции потерь исключительно на значении RSS. Поэтому ридж-регрессию можно использовать как базовый вариант. Если есть предположение, что только некоторые признаки имеют существенное значение, то следует перейти на лассо или elastic-net. В целом elastic-net предпочтительнее, чем лассо, так как лассо может везти себя хаотично, если количество признаков больше, чем количество наблюдений или когда некоторые признаки имеет высокую корреляцию [4].

3. Регуляризация логистической регрессии

3.1. Логистическая регрессия

Логистическая функция (сигмоида) имеет вероятностную интерпретацию при использовании в логистической регрессии:

$$p(y_i = 1|x_i, \theta) = h_{\theta,i} = \frac{1}{1 + e^{-\theta^T x}}$$

и

$$p(y_i = 0|x_i, \theta) = 1 - h_{\theta,i}(\theta^T x_i) = \frac{e^{-\theta^T x_i}}{1 + e^{-\theta^T x_i}}$$

В общем виде вероятность принадлежности некоторому классы при заданном наблюдении x_i и параметрах можно записать как

$$p(y_i|x_i, \theta) = p(y_i = 1|x_i, \theta)^{y_i} p(y_i = 0|x_i, \theta)^{1-y_i}$$

Функция правдоподобия:

$$L(y|x, \theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta)$$

Оценка параметров

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(y|x, \theta)$$

Как правило, для упрощения вычислений произведение преобразуют в сумму посредством логарифма следующим образом

$$\log L(y|x, \theta) = \log \prod_{i=1}^{n} p(y_i|x_i, \theta) = \sum_{i=1}^{n} \log p(y_i|x_i, \theta) = \sum_{i=1}^{n} \left[y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right] =$$
$$= \sum_{i=1}^{n} \left[y_i \theta^T x_i - \log\left(1 + e^{\theta^T x_i}\right) \right]$$

В этом случае задача по оценке параметров будет сводиться к следующей

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(y_i|x_i) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[y_i \theta^T x_i - \log\left(1 + e^{\theta^T x_i}\right) \right]$$

Задачу можно записать в виде минимизации с использование кросс-энтропии в качестве функции потерь:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L_{CE}(\theta) = \underset{\theta}{\operatorname{argmin}} -\frac{1}{n} \sum_{i=1}^{n} \log p(y_i|x_i)$$

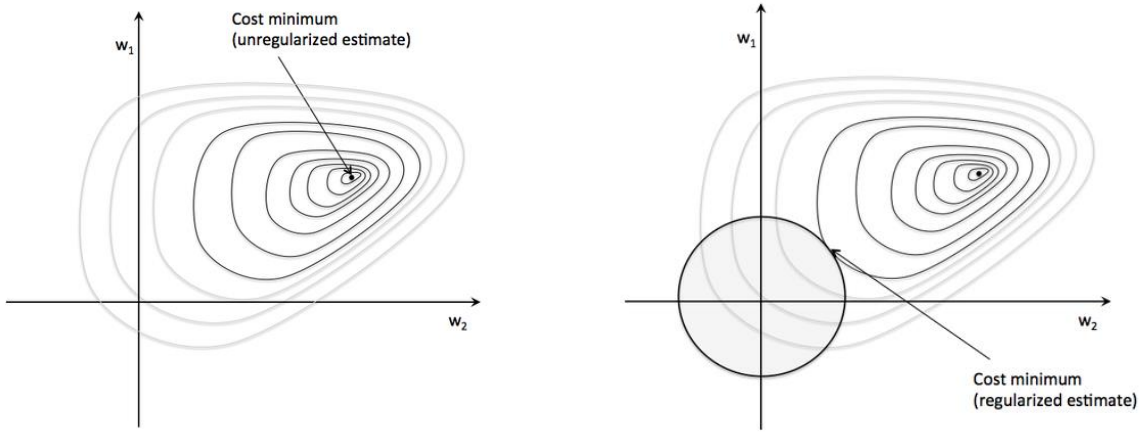
3.2. L2 регуляризация логистической регрессии

При использовании L2 регуляризация для оценки параметров логистической регрессии решается следующая задача

$$\hat{\theta}_{\lambda}^{R} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right] - \lambda \sum_{j=1}^{p} \theta_j^2$$

или через минимизацию

$$\hat{\theta}_{\lambda}^{R} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[-y_i \log h_{\theta,i} - (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right] + \underbrace{\lambda \sum_{j=1}^{p} \theta_j^2}_{\text{Штраф}}$$



https://sebastianraschka.com/faq/docs/probablistic-logistic-regression.html

3.3. L1 регуляризация логистической регрессии

При использовании L1 регуляризация для оценки параметров логистической регрессии решается следующая задача

$$\hat{\theta}_{\lambda}^{L} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right] - \lambda \sum_{j=1}^{p} \left|\theta_j\right|$$

или через минимизацию

$$\hat{\theta}_{\lambda}^{L} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[-y_i \log h_{\theta,i} - (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right] + \underbrace{\lambda \sum_{j=1}^{p} \left|\theta_j\right|}_{\text{Штраф}}$$

- При L2 регуляризации оценки параметров будет уметь небольшие значения
- При L1 часть оценок будут иметь несколько бо́льшие значения, в то время как другие могут иметь нулевые значения
- Таким образом, при L1 получаем более простую модель с меньшим количеством признаков

3.4. Байесовская вероятностная интерпретация

В обоих случаях имеется байесовская вероятностная интерпретация, которая представляется как ограничение на то, какой вид должны иметь априорные вероятности оценок параметров.

В L1 регуляризации оценки параметров распределены в соответствии распределением Лапласа, в то время как для L2 используется нормальное распределение Гаусса с $\mu = 0$.

Рассмотрим байесовскую интерпретацию для L2 регуляризации логистической регрессии. Для этого воспользуемся оценкой апостериорного максимума (Maximum a posteriori estimation – MAP). Используя теорему Байеса, можно записать:

$$p(\theta|x,y) = \frac{p(y|x,\theta)\cdot p(\theta|x)}{p(x)} \propto p(y|x,\theta)\cdot p(\theta|x) = \underbrace{p(y|x,\theta)}_{\substack{\text{Функция}\\ \text{правдоподобия}}} \cdot \underbrace{p(\theta)}_{\substack{\text{Априорная}\\ \text{вероятность}}}$$

Тогда оценка апостериорного максимума примет вид

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|x,y) = \underset{\theta}{\operatorname{argmax}}\, p(y|x,\theta)\cdot p(\theta)$$

или

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|x_i;\theta) \cdot \prod_{i=1}^{p} p(\theta_j)$$

Или то же самое в виде логарифма

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} [\log p(y|x,\theta) + \log p(\theta)]$$

или

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(y_i|x_i;\theta) + \sum_{j=1}^{p} \log p(\theta_j)$$

Априорную вероятность (в данном случае функция плотности) оценки j-го параметра представим в виде нормального закона

$$p(\theta_j) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left[-\frac{(\theta_j - \mu_j)^2}{2{\sigma_j}^2}\right].$$

Если $\mu_j = 0$ и $\sigma_j = \sigma$, то произведение априорных вероятностей примет вид:

$$\prod_{j=1}^{p} p(\theta_j) = \left[2\pi\sigma_\theta^2\right]^{-p/2} \exp\left[-\sum_{j=1}^{p} \frac{\theta_j^2}{2\sigma^2}\right].$$

Получаем

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|x_i;\theta) \cdot \left[2\pi\sigma_\theta^2\right]^{-p/2} \exp\left[-\sum_{i=1}^{n} \frac{\theta_j^2}{2\sigma^2}\right]$$

Возьмем логарифм

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \log p(y_i|x_i;\theta) \underbrace{-\frac{p}{2} \cdot \log(2\pi\sigma^2)}_{\text{константа по } \theta} - \sum_{j=1}^{p} \frac{\theta_j^2}{2\sigma^2} \right]$$
$$= \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \log p(y_i|x_i;\theta) - \sum_{j=1}^{p} \frac{\theta_j^2}{2\sigma^2} \right]$$

Если примем $\lambda = \frac{1}{2\sigma^2}$, то задача оптимизации посредством метода апостериорного максимума будет иметь следующий вид, который соответствует ранее рассмотренной логистической регрессии с регуляризацией L2:

$$\hat{\theta}_{\lambda}^{R} = \hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \log p(y_i|x_i;\theta) - \lambda \sum_{j=1}^{p} \theta_j^2 \right]$$

или в виде минимизации

$$\hat{\theta}_{\lambda}^{R} = \hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmin}} \left[\underbrace{-\sum_{i=1}^{n} \log p(y_i|x_i;\theta)}_{\text{кросс-энтропия}} + \underbrace{\lambda \sum_{j=1}^{p} \theta_j^2}_{\text{штраф}} \right]$$

Вероятностная интерпретация линейной регрессии с L2 регуляризацией

Оценка апостериорного максимума имеет вид

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|x_i;\theta) \cdot \prod_{i=1}^{p} p(\theta_j)$$

Полагаем, что каждое наблюдение подчиняется нормальному закону распределения:

$$\prod_{i=1}^{n} p(y_i|x_i;\theta) = \prod_{i=1}^{n} p(y_i|x_i;\theta,\sigma^2) = [2\pi\sigma^2]^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 \right]$$

Примем, что априорную вероятность (в данном случае функция плотности) оценки j-го параметра также будет соответствовать нормальному закону с $\mu_j = 0$ и $\sigma_j = \sigma_\theta$:

$$\prod_{i=1}^{p} p(\theta_j) = \left[2\pi\sigma_\theta^2\right]^{-p/2} \exp\left[-\frac{1}{2\sigma_\theta^2} \sum_{j=1}^{p} \theta_j^2 \right]$$

Тогда если возьмем логарифм от произведения функции правдоподобия и априорной вероятности, получим

$$\log \prod_{i=1}^{n} p(y_i|x_i;\theta) \cdot \prod_{i=1}^{p} p(\theta_j) = \underbrace{-\frac{n}{2}\log(2\pi\sigma^2) - \frac{p}{2}\log\left(2\pi\sigma_\theta^2\right)}_{\text{константы по } \theta} - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (y_i - \theta^T x_i)^2}_{RSS} - \underbrace{\frac{1}{2\sigma_\theta^2} \sum_{j=1}^{p} \theta_j^2}_{\text{штраф}}$$

Учитывая то, что неизвестны параметры θ, задача оптимизации посредством метода апостериорного максимума примет вид, что соответствует линейной регрессии с L2 регуляризацией:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[-RSS - \lambda \sum_{j=1}^{p} \theta_j^2 \right]$$

или в виде минимизации

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmin}} \left[RSS + \lambda \sum_{j=1}^{p} \theta_j^2 \right]$$

Список литературы

1. Chapter 6. Linear Model Selection and Regularization // An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshir. URL: https://www.statlearning.com/
2. The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani, Jerome Friedman. URL: https://web.stanford.edu/~hastie/ElemStatLearn/
3. Chapter 5. Logistic Regression // Speech and Language Processing. Daniel Jurafsky & James H. Martin URL: https://web.stanford.edu/~jurafsky/slp3/
4. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron