Регуляризация

Определение:

Регуляризация (англ. *regularization*) в статистике, машинном обучении, теории обратных задач — метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Чаще всего эта информация имеет вид штрафа за сложность модели.

Содержание

Мотивация

Как говорилось ранее, регуляризация полезна для борьбы с переобучением. Если вы выбрали сложную модель, и при этом у вас недостаточно данных, то легко можно получить итоговую модель, которая хорошо описывает обучающую выборку, но не обобщается на тестовую.

На примере линейной регрессии

В качестве наглядного примера рассмотрим линейные регрессионные модели. Восстановить зависимость для нескольких точек можно пытаться полиномами разной степени M.

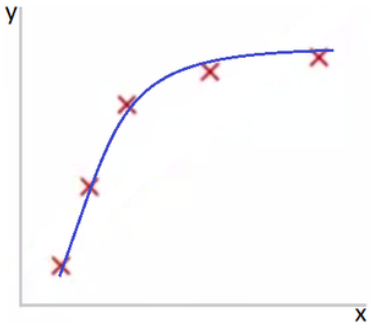

Рис. 1. Норма. M = 2

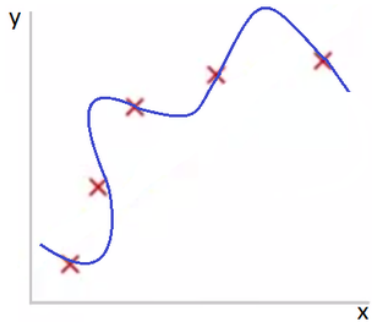

Рис. 2. Переобучение. M = 4

На Рис. 1 представлена зависимость, которая хорошо подходит для описания данных, а на Рис. 2 — модель, слишком сильно заточенная под обучающую выборку.

Однин из способов бороться с негативным эффектом излишнего подстраивания под данные — использование регуляризации, т. е. добавление некоторого штрафа за большие значения коэффициентов у линейной модели. Тем самым запрещаются слишком "резкие" изгибы, и предотвращается переобучение.

На примере логистической регрессии

Необходимость регуляризации можно увидеть и на другом примере — при использовании логистической регресии. Представьте, что ваша обучающая выборка была линейно разделима. В таком случае в процессе оптимизации значения весов модели уйдут в бесконечность, и вместо сигмойды получится "ступенька", представленная на Рис. 3.

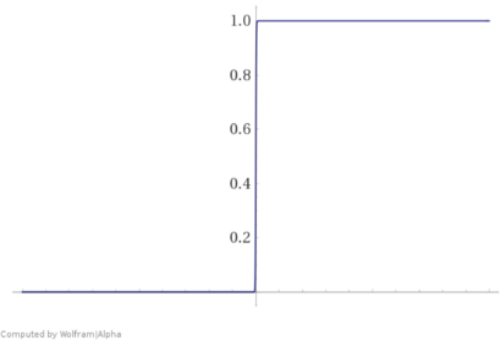

Рис 3. Сигмойда — "ступенька"

Это плохо, ибо произошло затачивание под обучающую выборку. Как и в предыдущем примере, побороться с этим можно путем добавления регуляризатора, не дающего весам принимать слишком большие значения.

Основные виды регуляризации

Переобучение в большинстве случаев проявляется в том, что итоговые модели имеют слишком большие значения параметров. Соответственно, необходимо добавить в целевую функцию штраф за это. Наиболее часто используемые виды регуляризации — L_1 и L_2, а также их линейная комбинация — эластичная сеть.

В представленных ниже формулах для эмпирического риска Q: $\mathcal{L}$ является функцией потерь, а β — вектором параметров $g(x, \beta)$ из модели алгоритма, а λ — неотрицательный гиперпараметр, являющийся коэффициентом регуляризации.

L_2-регуляризация

Определение:

L_2-регуляризация, или регуляризация Тихонова (англ. *ridge regularization* или *Tikhonov regularization*):

$$Q(\beta, X^l) = \sum_{i=1}^{l} \mathcal{L}(y_i, g(x_i, \beta)) + \lambda \sum_{j=1}^{n} {\beta_j}^2.$$

Минимизация регуляризованного соответствующим образом эмпирического риска приводит к выбору такого вектора параметров β, которое не слишком сильно отклоняется от нуля. В линейных классификаторах это позволяет избежать проблем мультиколлинеарности и переобучения.

L_1-регуляризация

Определение:

L_1-регуляризация (англ. *lasso regularization*), или регуляризация через манхэттенское расстояние:

$$Q(\beta, X^l) = \sum_{i=1}^{l} \mathcal{L}(y_i, g(x_i, \beta)) + \lambda \sum_{j=1}^{n} |\beta_j|.$$

Данный вид регуляризации также позволяет ограничить значения вектора β. Однако, к тому же он обладает интересным и полезным на практике свойством — обнуляет значения некоторых параметров, что в случае с линейными моделями приводит к отбору признаков.

Запишем задачу настройки вектора параметров β:

$$Q(\beta) = \sum_{i=1}^{l} \mathcal{L}_i(\beta) + \lambda \sum_{j=1}^{n} |\beta_j|,$$

где $\mathcal{L}_i(\beta) = \mathcal{L}(y_i, g(x_i, \beta))$ — некоторая ограниченная гладкая функция потерь. Сделаем замену переменных, чтобы функционал стал гладким. Каждой переменной β_j поставим в соответствие две новые неотрицательные переменные:

$$\begin{cases} u_j = \dfrac{1}{2}(|\beta_j| + \beta_j) \\ v_j = \dfrac{1}{2}(|\beta_j| - \beta_j) \end{cases}$$

Тогда:

$$\begin{cases} \beta_j = u_j - v_j \\ |\beta_j| = u_j + v_j \end{cases}$$

В новых переменных функционал становится гладким, но добавляются ограничения-неравенства:

$$\begin{cases} Q(u, v) = \sum\limits_{i=1}^{l} \mathcal{L}_i(u - v) + \lambda \sum\limits_{j=1}^{n} (u_j + v_j) \to \min\limits_{u,v} \\ u_j \geq 0, v_j \geq 0, \ j = 1, \ldots, n \end{cases}$$

Для любого j хотя бы одно из ограничений $u_j \geq 0$ и $v_j \geq 0$ обращается в равенство, иначе второе слагаемое в $Q(u, v)$ можно было бы уменьшить, не изменив первое. Если гиперпараметр λ устремить к ∞, в какой-то момент все $2n$ ограничений обратятся в равенство. Постепенное увеличение гиперпараметра λ приводит к

увеличению числа таких j, для которых $u_j = v_j = 0$, откуда следует, что $\beta_j = 0$. Как говорилось ранее, в линейных моделях это означает, что значения j-го признака игнорируются, и его можно исключить из модели.

Эластичная сеть

Определение:

Эластичная сеть (англ. *elastic net regularization*):

$$Q(\beta, X^l) = \sum_{i=1}^{l} \mathcal{L}(y_i, g(x_i, \beta)) + \lambda_1 \sum_{j=1}^{n} |\beta_j| + \lambda_2 \sum_{j=1}^{n} {\beta_j}^2.$$

Приведенная регуляризация использует как L_1, так и L_2 регуляризации, учитывая эффективность обоих методов. Ее полезной особенностью является то, что она создает условия для группового эффекта при высокой корреляции переменных, а не обнуляет некоторые из них, как в случае с L_1-регуляризацией.

Вероятностная интерпретация регуляризации

Эквивалентная вероятностная задача

Перед нами стоит задача — минимизировать эмпирический риск:

$$Q(\beta, X^l) = \sum_{i=1}^{l} \mathcal{L}(y_i, g(x_i, \beta)) \to \min_{\beta}$$

Вероятностная модель данных дает возможность по-другому взглянуть на задачу. Пусть $X \times Y$ — является вероятностным пространством. Тогда вместо $g(x_i, \beta)$ задана совместная плотность распределение объектов и классов $p(x, y|\beta)$.

Для настройки вектора параметров β воспользуемся *принципом максимума правдоподобия*:

$$p(X^l|\beta) = \prod_{i=1}^{l} p(x_i, y_i|\beta) \to \max_{\beta}$$

Удобнее рассматривать логарифм правдоподобия:

$$L(\beta, X^l) = \ln p(X^l|\beta) = \sum_{i=1}^{l} \ln p(x_i, y_i|\beta) \to \max_{\beta}$$

Можно заключить, что задачи в исходном и вероятностном представлении эквивалентны, если положить:

$$-\ln p(x_i, y_i|\beta) = \mathcal{L}(y_i, g(x_i, \beta))$$

Принцип максимума совместного правдоподобия данных и модели

Допустим, что наряду с параметрической моделью плотности распределения $p(x, y|\beta)$ имеется еще и *априорное распределение в пространстве параметров модели* $p(\beta)$. Чтобы ослабить априорные ограничения, вместо фиксированной функции $p(\beta)$ вводится *параметрическое семейство априорных распределений* $p(\beta; \gamma)$, где γ — гиперпараметр.

Принцип максимума правдоподобия теперь будет записываться по-другому, так как не только появление выборки X^l, но и появление модели β также является случайным. Их совместное появление описывается, согласно формуле условной вероятности, плотностью распределения:

$$p(X^l, \beta; \gamma) = p(X^l|\beta)p(\beta; \gamma)$$

Таким образом, приходим к *принципу максимума совместного правдоподобия данных и модели*:

$$L_\gamma(\beta, X^l) = \ln p(X^l, \beta; \gamma) = \sum_{i=1}^{l} \ln p(x_i, y_i|\beta) + \ln p(\beta; \gamma) \to \max_\beta$$

Функционал L_γ распадается на два слагаемых: логарифм правдоподобия и *регуляризатор*, не зависящий от данных. Второе слагаемое ограничивает вектор параметров модели, не позволяя ему быть каким угодно.

В итоге мы получили, что с байесовской точки зрения многие методы регуляризации соответствуют добавлению некоторых априорных распределений на параметры модели. При этом можно определить распределения, которые соответствуют представленным ранее L_1 и L_2 регуляризаторам.

Нормальный регуляризатор

Пусть вектор β имеет *нормальное распределение*[1], все его компоненты независимы и имеют равные дисперсии:

$$\beta \sim N(0, \sigma^2)$$

Логарифмируя, получаем *квадратичный регуляризатор*:

$$\ln p(\beta; \sigma) = \ln\left(\frac{1}{(2\pi\sigma)^{n/2}} \exp\left(-\frac{\|\beta\|^2}{2\sigma} \right) \right) = -\frac{1}{2\sigma}\|\beta\|^2 + const(\beta),$$

где $const(\beta)$ — слагаемое, не зависящее от β, которым можно пренебречь, поскольку оно не влияет на решение оптимизационной задачи. В итоге имеем L_2-регуляризатор.

Лапласовский регуляризатор

Пусть вектор β имеет *распределение Лапласа*[2], все его компоненты независимы и имеют равные дисперсии:

$$\beta \sim Laplace(0, C)$$

Тогда:

$$\ln p(\beta; C) = \ln\left(\frac{1}{(2C)^{n}} \exp\left(-\frac{\|\beta\|_1}{C} \right) \right) = -\frac{1}{C}\|\beta\|_1 + const(\beta), \|\beta\|_1 = \sum_j |\beta_j|$$

Аналогично случаю с нормальным регуляризатором, $const(\beta)$ можно опустить и, таким образом, получаем L_1 -регуляризатор.

Распределение Лапласа имеет более острый пик и более тяжёлые «хвосты», по сравнению с нормальным распределением, как можно видеть на Рис. 4. Дисперсия Лапласовского распределения равна $2C^2$.

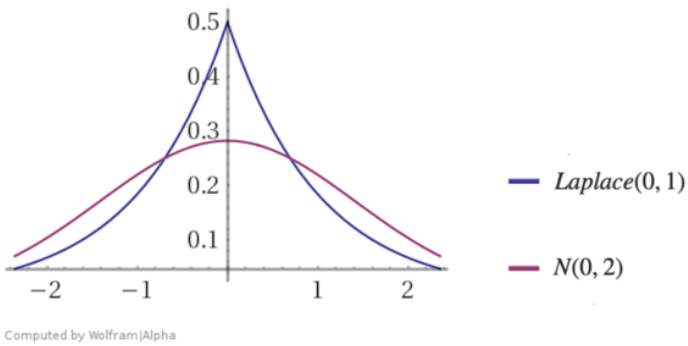

Рис. 4. Сравнение нормального и Лапласовского распределений при одинаковых математических ожиданиях и дисперсиях.

Регуляризация в линейной регрессии

В линейной регрессии моделируется линейная зависимость между зависимой и независимой переменной. Каждому объекту $x \in X^l$ соответствует признаковое описание $(f_1(x), \ldots, f_n(x))$, где $f_j : X \to \mathbb{R}$ — числовые признаки. Модель алгоритмов для линейной регрессии состоит из функций вида:

$$g(x, \beta) = \sum_j^n \beta_j \, f_j(x)$$

В итоге оптимизируемый функционал эмпирического риска выглядит следующим образом:

$$Q(a) = \|F\beta - y\|^2,$$

где $F = (f_j(x_i))_{l \times n}$ — матрица объекты-признаки, $y = (y_i)_{l \times 1}$ — целевой вектор, $\beta = (\beta_j)_{n \times 1}$ — вектор параметров. Приравняв нулю производную $Q(\beta)$ по параметру β, получаем:

$$\beta^* = (F^T F)^{-1} F^T y$$

В итоге, используя сингулярное разложение для представления F и проведя МНК-аппроксимизацию целевого вектора y, имеем выражение для нормы вектора β:

$$\|\beta^*\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j} (v_j^T y)^2$$

К сожалению, могут возникнуть проблемы мультиколлинеарности и переобучения в случае, если ковариационная матрица $\sum = F^T F$ плохо обусловлена. Одним из способов борьбы с этими проблемами, как говорилось ранее, является регуляризация.

В статье о вариациях регрессии представлены модификации линейной регресиии с различными регуляризаторами (L_1 и L_2) и их отличие. Описание в данном разделе будет похожим, однако здесь будет рассмотрен эффект от добавления регуляризаторов немного подробнее.

Гребневая регрессия

В гребневой регрессии к функционалу Q добавляется L_2-регуляризатор.

Итоговый минимизируемый функционал с поправкой:

$$Q_\lambda(\beta) = ||F\beta - y||^2 + \tau||\beta||^2$$

Итоговое выражение для параметра β:

$$\beta_\tau^* = (F^T F + \tau I_n)^{-1} F^T y$$

Таким образом, перед обращением матрицы к ней добавляется "гребень" — диагональная матрица τI_n. При этом все её собственные значения увеличиваются на τ, а собственные векторы не изменяются. В результате матрица становится хорошо обусловленной, оставаясь в то же время «похожей» на исходную.

Оценим эффект, который оказывает добавление гребня. Выразим регуляризованное МНК-решение через сингулярное разложение:

$$\beta_t^* = (UD^2U^T + \tau I_n)^{-1} UDV^T y = U(D^2 + \tau I_n)^{-1} DV^T y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^T y)$$

Теперь найдём регуляризованную МНК-аппроксимацию целевого вектора y:

$$F\beta_\tau^* = VDU^T \beta_\tau^* = V diag\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^T y)$$

Как можно видеть, проекции на собственные векторы сокращаются, умножаясь $\frac{\lambda_j}{\lambda_j + \tau} \in (0, 1)$.

В сравнении с нерегуляризованным случаем, уменьшается и норма вектора β:

$$\|\beta_\tau^*\|^2 = \|D^2(D^2 + \tau I_n)^{-1} D^{-1} V^T y)\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j + \tau} (v_j^T y)^2 < \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^T y)^2 = \|\beta^*\|^2$$

Поэтому данный метод называют также *сжатие* или *сокращение весов*.

Из формул видно, что по мере увеличения параметра τ вектор коэффициентов β_τ^* становится всё более устойчивым и жёстко определённым. Фактически, происходит понижение *эффективной размерности решения* — это второй смысл термина *сжатие*. Роль размерности играет след проекционной матрицы.

В нерегуляризованном случае:

$$n_{effective} = tr\ F(F^T F)^{-1} F^T = tr\ (F^T F)^{-1} F^T F = tr\ I_n = n$$

В случае с гребнем:

$$n_{effective} = tr\ F(F^T F + \tau I_n)^{-1} F^T = tr\ diag\left(\frac{\lambda_j}{\lambda_j + \tau}\right) = \sum_{j=1}^{n} \frac{1}{\lambda_j} < n$$

Лассо регрессия

В лассо регрессии к функционалу Q добавляется L_1-регуляризатор.

Итоговый минимизируемый функционал с поправкой:

$$Q_\tau(\beta) = ||F\beta - y||^2 + \tau||\beta||$$

Запишем систему для этой регрессии в виде минимизации неизменного функционала Q при неравенстве-ограничении:

$$\begin{cases} Q(\beta) = \|F\beta - y\|^2 \to \min\limits_{\beta} \\ \sum\limits_{j=1}^{n} |\beta_j| \leq \chi \end{cases}$$

Так как используется L_1-регуляризатор, коэффициенты β_j постепенно обнуляются с уменьшением χ. Происходит отбор признаков, поэтому параметр χ называют еще *селективностью*. Параметр χ "зажимает" вектор коэффициентов β, отсюда и название метода — лассо (англ. *LASSO, least absolute shrinkage and selection operator*).

Сравнение гребневой и лассо регрессий

Основное различие лассо и гребневой регрессий заключается в том, что первая может приводить к обращению некоторых независимых переменных в ноль (используется L_1-регуляризатор), тогда как вторая уменьшает их до значений, близких к нулю (используется L_2-регуляризатор).

Продублируем наглядный пример из статьи о вариациях регрессии. Рассмотрим для простоты двумерное пространство независимых переменных. В случае лассо регрессии органичение на коэффициенты представляет собой ромб ($|\beta_1| + |\beta_2| \leq t$), в случае гребневой регрессии — круг ($\beta_1^2 + \beta_2^2 \leq t^2$). Необходимо минимизировать функцию ошибки, но при этом соблюсти ограничения на коэффициенты. С геометрической точки зрения задача состоит в том, чтобы найти точку касания линии, отражающей функцию ошибки с фигурой, отражающей ограничения на β. Из Рис. 5 интуитивно понятно, что в случае лассо регрессии эта точка с большой вероятностью будет находиться на углах ромба, то есть лежать на оси, тогда как в случае гребневой регрессии такое происходит очень редко. Если точка пересечения лежит на оси, один из коэффициентов будет равен нулю, а значит, значение соответствующей независимой переменной не будет учитываться.

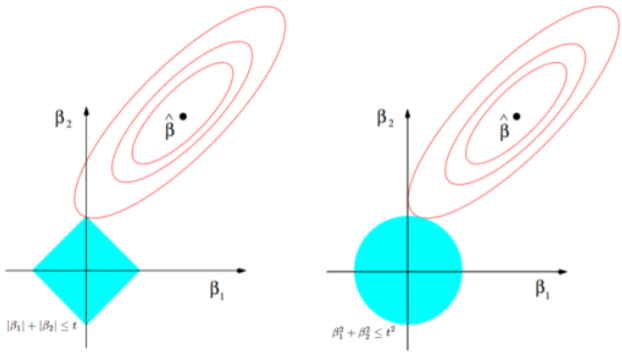

Рис. 5. Сравнение лассо (слева) и гребневой (справа) регрессий, пример для двумерного пространства независимых переменных. Бирюзовые области изображают ограничения на коэффициенты β, эллипсы — некоторые значения функции наименьшей квадратичной ошибки.

Также полезно будет рассмотреть простую модельную задачу. Пусть $l = n$ и матрица объекты-признаки является единичной $F = I$. Тогда МНК-решение дает вектор коэффициентов β:

$$\beta^* = argmin \left(\sum_{i=1}^{l} (\beta_i - y_i)^2 \right)$$
$$\beta_j^* = y_j$$

В случае с гребневой регрессией:

$$\beta_j^* = \frac{y_j}{1+\lambda}$$

В случае с лассо регрессией:

$$\beta_j^* = \begin{cases} y_j - \lambda/2, y_j > \lambda/2 \\ y_j + \lambda/2, y_j < -\lambda/2 \\ 0, |y_j| \leq \lambda/2 \end{cases}$$

В итоге на Рис. 6 на графиках с зависимостями β_j^* от y_j можно увидеть описанные ранее особенности данных регуляризованных линейных регрессий.

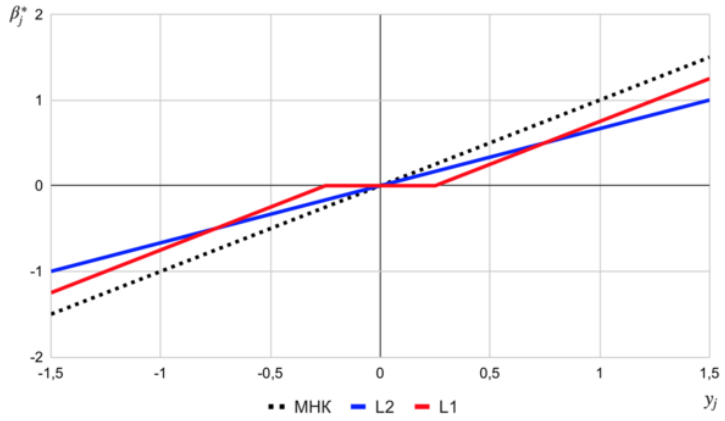

Рис. 6. Сравнение лассо и гребневой регрессий, пример с простой модельной задачи.

Регуляризация в алгоритмах

Градиентный спуск

Алгоритм градиентного спуска используют для нахождения аппроксимирующей зависимости, определяя вектор весов $w \in \mathbb{R}^n$, при котором достигается минимум эмпирического риска:

$$Q(w, X^l) = \sum_{i=1}^{l} \mathcal{L}(y_i, \langle w, x_i \rangle) \to \min_w$$

В этом методе выбирается некоторое начальное приближение для вектора весов w, затем запускается итерационный процесс, на каждом шаге которого вектор w изменяется в направлении наиболее быстрого убывания функционала Q — противоположно вектору градиента $Q'(w) = (\frac{\partial Q(w)}{\partial w_j})_{j=1}^n$:

$$w := w - \eta Q'(w),$$

где $\eta > 0$ — величина шага в направлении антиградиента.

Регуляризация — одна из эвристик улучшения градиентных методов обучения. Основным способом уменьшить переобучение является квадратичная регуляризация, называемая также *сокращением весов*. Чтобы ограничить рост абсолютных значений весов, к минимизируемому функционалу $Q(w)$ добавляется штрафное слагаемое:

$$Q_\tau(w) = Q(w) + \frac{\tau}{2}\|w\|^2$$

Это приводит к появлению аддитивной поправки в градиенте:

$$Q'_\tau(w) = Q'(w) + \tau w$$

В результате правило обновления весов принимает вид:

$$w := w(1 - \eta\tau) - \eta Q'(w)$$

Таким образом, вся модификация сводится к появлению неотрицательного множителя $(1 - \eta\tau)$, приводящего к постоянному уменьшению весов.

Регуляризация предовтращает паралич, повышает устойчивость весов в случае мультиколлинеарности, повышает обобщающую способность алгоритма и снижает риск переобучения. Однако есть и недостатки — параметр τ необходимо выбирать с помощью кросс-валидации, что связано с большими вычислительными затратами.

Метод опорных векторов

Метод опорных векторов (SVM) используется для задач классификации и регрессии. В нем строится гиперплоскость, разделяющая объекты выборки оптимальным образом.

К сожалению, зачастую выборка является линейно неразделимой. В таком случае приходится "ослаблять ограничения", позволяя некоторым объектам попадать на территорию другого класса. Для каждого объекта от отступа отнимается некоторая положительная величина ξ_i, но требуется, чтобы введенные поправки были минимальны. В итоге постановка задачи *SVM с мягким отступом* (англ. *soft-margin SVM*) выглядит следующим образом: $\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{l}\xi_i \to \min\limits_{w,b,\xi} \\ M_i(w,b) \geq 1 - \xi_i, \quad i = 1,\ldots,l \\ \xi_i \geq 0, \quad i = 1,\ldots,l \end{cases}$

Как показано в соответствующем данному методу разделе, эквивалентной задачей безусловной минимизации является: $Q(w,b) = \dfrac{1}{2C}\|w\|^2 + \sum\limits_{i=1}^{l}(1 - M_i(w,b))_+ \to \min\limits_{w,b}$

В силу неравенства $[M_i < 0] \leq (1 - M_i)_+$, функционал $Q(w,b)$ можно рассматривать как верхнюю оценку эмпирического риска, к которому добавлен регуляризатор $\dfrac{1}{2C}\|w\|^2$.

С введением регуляризатора устраняется проблема мультиколлинеарности, повышается устойчивость алгоритма, улучшается его обобщающая способность.

В результате получаем, что принцип оптимальной разделяющей гиперплоскости или максимизации ширины разделяющей полосы в случае неразделимой выборки тесно связан с L_2-регуляризацией, которая возникает естественным образом из постановки задачи.

Также существуют разновидности SVM с другими регуляризаторами.

- Метод релевантных векторов (англ. *RVM, Relevance vector Machine*):

$$\frac{1}{2}\sum_{i=1}^{l}\left(\ln w_i + \frac{\lambda_i^2}{w_i}\right)$$

- Метод опорных векторов с лассо (англ. *LASSO SVM*):

$$\mu\sum_{i=1}^{n}|w_i|$$

- Метод опорных признаков (англ. *Support feature machine*):

$$\sum_{i=1}^{n} R_\mu(w_i), \begin{cases} 2\mu|w_i|, |w_i| < \mu \\ \mu^2 + w_i^2, |w_i| \geq \mu \end{cases}$$

Другие использования регуляризации

Логистическая регрессия

Как было показано в мотивационном примере, для логистической регрессии может быть полезно использовать регуляризацию.

Для настройки вектора коэффициентов β по обучающей выборке X^l максимизируют логарифм правдоподобия:

$$L(\beta, X^l) = log_2 \prod_{i=1}^{l} p(x_i, y_i) \to \max_\beta$$

$$L(\beta, X^l) = \sum_{i=1}^{l} log_2 \sigma(\langle \beta, x_i \rangle y_i) + const(\beta) \to \max_\beta$$

L_2-регуляризация:

$$L(\beta, X^l) = \sum_{i=1}^{l} log_2 \sigma(\langle \beta, x_i \rangle y_i) - \lambda \|\beta\|^2 + const(\beta) \to \max_\beta$$

L_1-регуляризация:

$$L(\beta, X^l) = \sum_{i=1}^{l} log_2 \sigma(\langle \beta, x_i \rangle y_i) - \lambda \|\beta\|_1 + const(\beta) \to \max_\beta$$

Аналогично можно использовать и другие регуляризаторы.

Нейронные сети

Регуляризация также используется и в нейронных сетях для борьбы со слишком большими весами сети и переобучением. Однако, в этом случае зануление коэффициентов при использовании L_1-регуляризатора не несет в себе смысл "отбора признаков", как в случае с линейными моделями. К сожалению, регуляризация не снижает число параметров и не упрощает структуру сети.

Для нейронной сети помимо добавления штрафного слагаемого к эмпирическому риску активно используют и другой метод борьбы с переобучением — *прореживание сети* (англ. *dropout*), в ходе которого упрощают сеть, руководствуясь правилом — если функция ошибки не изменяется, то сеть можно упрощать и дальше. Подробнее об этом можно почитать в статье, рассказывающей о практике реализации нейронных сетей.

См. также

- Переобучение
- Модель алгоритма и её выбор
- Байесовская классификация
- Вариации регрессии
- Линейная регрессия
- Логистическая регрессия
- Стохастический градиентный спуск
- Метод опорных векторов (SVM)
- Нейронные сети, перцептрон

- Практики реализации нейронных сетей

Примечания

1. Нормальное распределение (https://ru.wikipedia.org/wiki/Нормальное_распределение)
2. Распределение Лапласа (https://ru.wikipedia.org/wiki/Распределение_Лапласа)

Источники информации

- Воронцов К.В. — Математические методы обучения по прецедентам (http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf)
- Википедия — Регуляризация (математика) (https://ru.wikipedia.org/wiki/Регуляризация_(математика))
- coursea.org — Регуляризация (https://www.coursera.org/lecture/supervised-learning/rieghuliarizatsiia-sR94Q)
- machinelearning.ru — L1-регуляризация линейной регрессии (http://www.machinelearning.ru/wiki/images/7/7e/VetrovSem11_LARS.pdf)
- medium.com — 5 видов регрессии и их свойства (https://medium.com/nuances-of-programming/5-видов-регрессии-и-их-свойства-f1bb867aebcb)
- Wikipedia — Elastic net regularization (https://en.wikipedia.org/wiki/Elastic_net_regularization)
- Keng B. — A Probabilistic Interpretation of Regularization (http://bjlkeng.github.io/posts/probabilistic-interpretation-of-regularization/)
