Одним из основных аспектов обучения вашей модели машинного обучения является предотвращение переобучения. Модель будет иметь низкую точность, если она будет переобучать. Это происходит потому, что ваша модель слишком усердно пытается уловить шум в наборе обучающих данных. Под шумом мы подразумеваем точки данных, которые на самом деле представляют не истинные свойства ваших данных, а случайный выбор. Изучение таких точек данных делает вашу модель более гибкой с риском переобучения.

Концепция уравновешивания смещения и дисперсии помогает понять явление переобучения.

Одним из способов избежать переобучения является использование перекрестной проверки, которая помогает оценить ошибку по набору тестов и решить, какие параметры лучше всего подходят для вашей модели.

Регуляризация

Это форма регрессии, которая ограничивает/упорядочивает или сужает оценки коэффициентов до нуля. Другими словами, этот метод препятствует изучению более сложной или гибкой модели, чтобы избежать риска переобучения.

Простое соотношение для линейной регрессии выглядит так. Здесь Y представляет изученное отношение, а β представляет оценки коэффициентов для различных переменных или предикторов (X).

$$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$$

Процедура подбора включает функцию потерь, известную как остаточная сумма квадратов или RSS. Коэффициенты выбираются такими, чтобы минимизировать эту функцию потерь.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Теперь это скорректирует коэффициенты на основе ваших тренировочных данных. Если в обучающих данных есть шум, то оценочные коэффициенты не будут хорошо обобщаться на будущие данные. Именно здесь вступает в действие регуляризация, которая сужает или упорядочивает эти изученные оценки до нуля.

Ridge Regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

На изображении выше показана регрессия гребня, где RSS изменяется путем добавления величины сжатия. Теперь коэффициенты оцениваются путем минимизации этой функции. Здесь λ — это параметр настройки, который определяет, насколько мы хотим снизить гибкость нашей модели. Увеличение гибкости модели представлено увеличением ее коэффициентов, и если мы хотим минимизировать указанную выше функцию, то эти коэффициенты должны быть небольшими. Вот как метод регрессии Риджа предотвращает слишком высокий рост коэффициентов. Кроме того, обратите внимание, что мы уменьшаем предполагаемую связь каждой переменной с ответом, за исключением точки пересечения β0. Эта точка пересечения является мерой среднего значения ответа, когда xi1 = xi2 = …= xip = 0.

Когда λ = 0, штрафной член не имеет значения, и оценки, полученные с помощью гребневой регрессии, будут равны методу наименьших квадратов. Однако при λ→∞ влияние штрафа за усадку возрастает, и оценки коэффициента гребневой регрессии будут приближаться к нулю. Как видно, выбор хорошего значения λ имеет решающее значение. Перекрестная проверка пригодится для этой цели. Оценки коэффициентов, полученные этим методом, также известны как норма L2.

Коэффициенты, полученные стандартным методом наименьших квадратов, являются масштабно-эквивариантными, т. е. если мы умножаем каждый вход на c, то соответствующие коэффициенты масштабируются с коэффициентом 1/c. Следовательно, независимо от того, как масштабируется предиктор, умножение предиктора и коэффициента (Xjβj) остается одним и тем же. Однако это не относится к гребневой регрессии, и поэтому нам необходимо стандартизировать предикторы или привести предикторы к одному масштабу перед выполнением гребенчатой регрессии. Формула, используемая для этого, приведена ниже.

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

Lasso

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

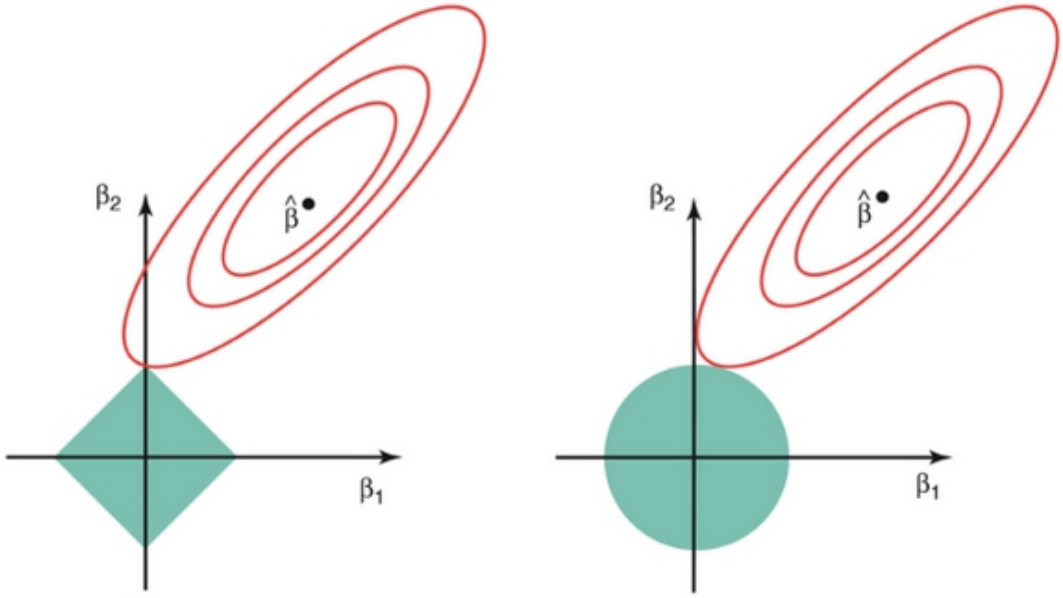
Лассо — еще один вариант, в котором вышеуказанная функция сведена к минимуму. Понятно, что эта вариация отличается от гребневой регрессии только штрафом за высокие коэффициенты. В качестве штрафа он использует |βj|(модуль) вместо квадратов β. В статистике это известно как норма L1.

Давайте взглянем на вышеупомянутые методы с другой точки зрения. Гребневую регрессию можно рассматривать как решение уравнения, в котором сумма квадратов коэффициентов меньше или равна s. А лассо можно рассматривать как уравнение, в котором сумма модулей коэффициентов меньше или равна s. Здесь s — константа, существующая для каждого значения коэффициента усадки λ. Эти уравнения также называются функциями ограничений.

Рассмотрим их 2 параметра в данной задаче. Тогда в соответствии с приведенной выше формулировкой гребневая регрессия выражается как $\beta1^2 + \beta2^2 \leq s$. Это означает, что коэффициенты гребневой регрессии имеют наименьшую RSS (функцию потерь) для всех точек, лежащих в пределах круга, определяемого как $\beta1^2 + \beta2^2 \leq s$.

Точно так же для лассо уравнение принимает вид $|\beta1|+|\beta2|\leq s$. Это означает, что коэффициенты лассо имеют наименьшую RSS (функцию потерь) для всех точек, лежащих в пределах ромба, заданного $|\beta1|+|\beta2|\leq s$.

Изображение ниже описывает эти уравнения.



На изображении выше показаны функции ограничения (зеленые области) для лассо (слева) и регрессии гребня (справа), а также контуры для RSS (красный эллипс). Точки на эллипсе разделяют значение RSS. Для очень большого значения s зеленые области будут содержать центр эллипса, делая оценки коэффициентов обоих методов регрессии равными оценкам методом наименьших квадратов. Но на изображении выше это не так. В этом случае оценки коэффициентов регрессии лассо и гребня задаются первой точкой, в которой эллипс касается области ограничений. Поскольку регрессия гребня имеет круговое ограничение без острых точек, это пересечение обычно не происходит на оси, и поэтому оценки коэффициента регрессии гребня будут исключительно ненулевыми. Однако ограничение лассо имеет углы на каждой из осей, поэтому эллипс часто будет пересекать область ограничения на оси. В этом случае один из коэффициентов будет равен нулю. В более высоких измерениях (где параметры намного больше 2) многие оценки коэффициентов могут одновременно равняться нулю.

Это проливает свет на очевидный недостаток гребневой регрессии, заключающийся в интерпретируемости модели. Это уменьшит коэффициенты для наименее важных предикторов, очень близко к нулю. Но это никогда не сделает их точно равными нулю. Другими словами, окончательная модель будет включать все предикторы. Однако в случае лассо штраф L1 приводит к тому, что некоторые оценки коэффициентов становятся точно равными нулю, когда параметр настройки λ достаточно велик. Следовательно, метод лассо также выполняет выбор переменных и, как говорят, дает разреженные модели.

Что дает регуляризация?

Стандартная модель наименьших квадратов, как правило, имеет некоторую дисперсию, т. Е. Эта модель не будет хорошо обобщаться для набора данных, отличного от ее обучающих данных. Регуляризация значительно снижает дисперсию модели без существенного увеличения ее смещения. Таким образом, параметр настройки λ, используемый в описанных выше методах регуляризации, контролирует влияние смещения и дисперсии. По мере увеличения значения λ уменьшается значение коэффициентов и, таким образом, уменьшается дисперсия. До определенного момента это увеличение λ выгодно, так как оно только уменьшает дисперсию (следовательно, избегая переобучения) без потери каких-либо важных свойств данных. Но после определенного значения модель начинает терять важные свойства, что приводит к смещению в модели и, следовательно, к недообучению. Поэтому следует тщательно выбирать значение λ.