Tokmakova A.A., Strijov V.V. Estimation of linear model hyperparametres for noise or correlated feature selection problem // Intellectual Information Processing. Conference proceedings, 2012 : 156-159.

Оценка гиперпараметров линейных регрессионных моделей при отборе шумовых и мультикоррелирующих признаков*

Токмакова А. А., Стрижов В. В.
aleksandra-tok@yandex.ru
Москва, МФТИ

В работе решается задача отбора признаков при восстановлении линейной регрессии. Принята гипотеза о нормальном распределении вектора зависимой переменной и параметров модели. Для оценки ковариационной матрицы параметров используется аппроксимация Лапласа: логарифм функции ошибки приближается функцией нормального распределения. Исследуется проблема присутствия в выборке шумовых и мультикоррелирующих признаков, так как при их наличии матрица ковариаций параметров модели становится вырожденной. Предлагается алгоритм, производящий отбор информативных признаков. В вычислительном эксперименте приводятся результаты исследования на реальных данных.

Estimation of linear model hyperparametres for noise or correlated feature selection problem*

Tokmakova A. A., Strijov V. V.
MFTI, Moscow, Russia

This paper deals with the problem of feature selection in the linear regression models. To select features the author estimate the covariance matrix of the model parameters. Dependent variable and model parameters are assumed to be normally distributed. The laplace approximation is used for estimation the covariance matrix: the logarithm error function is approximated by the normal distribution function. In the case of noise and correlated features covariance matrix becomes singular. An algorithm for feature selection is proposed.

Ввведение

В работе рассматривается задача отбора признаков при восстановлении линейной регрессии. При дальнейшем развитии статистических методов К.Бишопом было предложено [1] использовать функцию плотности распределения параметров модели для оценки информативности признаков модели. Параметры данной функции стали называть гиперпараметрами. Использование байесовского вывода привело к необходимости оценки совместной функции распределения независимых переменных и зависимой переменной.

Зависимая переменная и параметры модели в этой работе рассматриваются как многомерные случайные величины, распределённые нормально [2, 3]. При наличии в выборке шумовых и мультикоррелирующих признаков получение оценок соответствующих ковариационных матриц затруднено, так как происходит их вырождение. Работа посвящена способу получения невырожденных ковариационных матриц.

При анализе статистических связей между независимыми и зависимыми переменными необходимо решить задачу отбора признаков. Шумовыми признаками называются признаки ортогональные вектору зависимой переменной в пространстве столбцов матрицы плана. Мультикоррелирующими — признаки, которые с некоторым уровнем значимости эквивалентны другим признакам.

Работа выполнена при финансовой поддержке РФФИ, проекты № 10-07-00422, 11-07-13160.

Постановка задачи

В работе рассматривается выборка, состоящая из матрицы плана и вектора зависимой переменной:

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^m = (X, \mathbf{y}),$$

где $\mathbf{x}_i \in \mathbb{R}^n, i = 1, \ldots, m$. Данные аппроксимируются линейной моделью: $\mathbf{y} = \mathbf{f}(\mathbf{w}, X)$, где $\mathbf{f}(\mathbf{w}, X)$ — некоторая параметрическая вектор-функция. Вектор коэффициентов этой функции называется вектором параметров этой модели. Принята гипотеза о нормальном распределении вектора зависимой переменной:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I_m),$$

где $\mathbf{f}$ — вектор-функция, σ^2 — дисперсия распределения, I_m — единичная матрица размерности m. Обозначим $\beta^{-1} = \sigma^2$. Рассматривается набор конкурирующих моделей, определяемых своим набором параметров. Необходимо сформировать множество индексов активных признаков $\mathcal{A} \subseteq \mathcal{J} = \{1, 2, \ldots, n\}$, то есть не шумовых и не мультикоррелирующих, при котором достигается минимум функции ошибки $S(\mathbf{w})$. При этом параметры модели должны также принимать оптимальные значения.

Функция ошибки

Исходя из предположения о нормальном распределении вектора зависимой переменной записы-

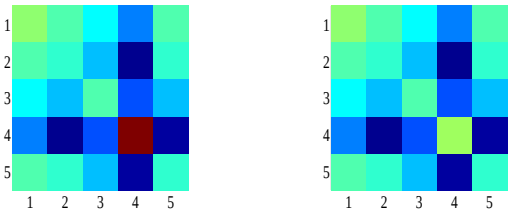

Рис. 1. Удаление шумового признака.

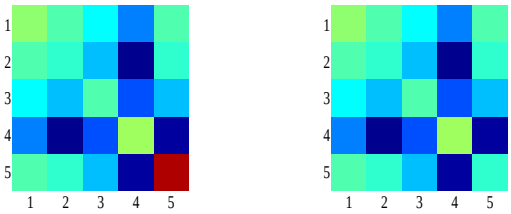

Рис. 2. Удаление мультикоррелирующих признаков.

вается его плотность:

$$p(\mathbf{y} \mid X, \mathbf{w}, \beta, \mathbf{f}) = p(D \mid \mathbf{w}, \beta, \mathbf{f}) = \\ = \frac{1}{(2\pi\beta^{-1})^{\frac{m}{2}}} \exp(-\frac{1}{2}(\mathbf{y} - \mathbf{f})^{\intercal}\beta I(\mathbf{y} - \mathbf{f})).$$

Так как матрица плана не является случайной величиной и используется линейная модель, то вектор параметров также будет распределен нормально, но с другими параметрами:

$$p(\mathbf{w} \mid A, \mathbf{f}) = \\ = \frac{1}{(2\pi)^{\frac{m}{2}}|A^{-1}|} \exp(-\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^{\intercal}A(\mathbf{w} - \mathbf{w}_0)).$$

Используя формулу Байеса и представив апостериорное распределение параметров через функцию правдоподобия данных и априорное распределение параметров, получим:

$$p(\mathbf{w} \mid D, A, \beta, \mathbf{f}) = \frac{p(D \mid \mathbf{w}, \beta, \mathbf{f})p(\mathbf{w} \mid A, \mathbf{f})}{p(D \mid A, \beta, \mathbf{f})},$$

$$\frac{p(D \mid \mathbf{w}, \beta, \mathbf{f})p(\mathbf{w} \mid A, \mathbf{f})}{p(D \mid A, \beta, \mathbf{f})} = \\ = \frac{\exp(-\beta E_D)\exp(-E_{\mathbf{w}})}{Z_D(\beta)Z_{\mathbf{w}}(A)} = \frac{\exp(-(\beta E_D + E_{\mathbf{w}}))}{Z_D(\beta)Z_{\mathbf{w}}(A)}.$$

Записывая функцию ошибки как

$$S = E_{\mathbf{w}} + \beta E_D = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^{\intercal}A(\mathbf{w} - \mathbf{w}_0) + \\ + \frac{1}{2}(\mathbf{y} - \mathbf{f})^{\intercal}\beta I(\mathbf{y} - \mathbf{f}),$$

получим следующее выражение для апостериорного распределения параметров:

$$p(\mathbf{w} \mid D, A, \beta, \mathbf{f}) = \frac{\exp(-S(\mathbf{w}))}{Z_S(A, \beta)},$$

где $Z_S = Z_S(A, \beta)$ — нормирующий коэффициент. Оценка нормировочного коэффициента производится с помощью аппроксимации Лапласа:

$$Z_S = \frac{\exp(-S(\mathbf{w}_0))(2\pi)^{\frac{n}{2}}}{|H|^{\frac{1}{2}}},$$

где $H = -\nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}$ — матрица Гессе функции ошибки.

Оценка ковариационных матриц

Для того чтобы выбрать устойчивую модель с высокой предсказательной способностью, требуется отфильтровать мультикоррелирующие и шумовые признаки. В работе [4] приводится алгоритм, использующий для этого гиперпараметры распределений вектора параметров и вектора зависимой переменной. В данном случае гиперпараметрами являются ковариационные матрицы распределений. Для оценки гиперпараметров используется принцип максимума правдоподобия. Для удобства расчетов рассматривается логарифм правдоподобия:

$$\ln p(D \mid A, \beta, \mathbf{f}) = -\frac{1}{2}\ln|A^{-1}| - \frac{m}{2}\ln 2\pi + \\ + \frac{m}{2}\ln\beta^{-1} - S(\mathbf{w}_0) - \frac{1}{2}\ln|H|.$$

В работе рассматривается случай, когда ковариационная матрица распределения вектора параметров A^{-1} диагональна, а вектор зависимой переменной гомоскедастичен, то есть матрица $B^{-1} = \\ = \beta^{-1}I_m$.

В таких условиях гессиан функции ошибки представим в виде двух слагаемых, одно из которых зависит от исключительно от данных, а другое от параметров:

$$H = -\nabla\nabla S(\mathbf{w}) = -\nabla\nabla(\beta E_D + E_{\mathbf{w}}) = \\ = -\beta\nabla\nabla E_D - \nabla\nabla E_{\mathbf{w}} = H_D + H_{\mathbf{w}},$$

где H_D зависит от β, а $H_{\mathbf{w}}$ зависит от A.

Часть гессиана $H_{\mathbf{w}}$ диагональна по построению:

$$\nabla\nabla E_{w_i} = \nabla\nabla\left(\tfrac{1}{2}\alpha_i(w_i - w_{0i})^2\right) = \alpha_i,$$

где $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^{\intercal}$ — вектор, состоящий из элементов диагонали матрицы A.

Часть гессиана H_D в работе также принимается диагональной, так как возможны две следующих ситуации:

1) если все признаки независимы, то недиагональные элементы H_D равны нулю, так как они являются коэффициентами корреляции величин;
2) если же в выборке присутствуют шумовые или мультикоррелирующие признаки, то наблюдается резкое возрастание диагональных

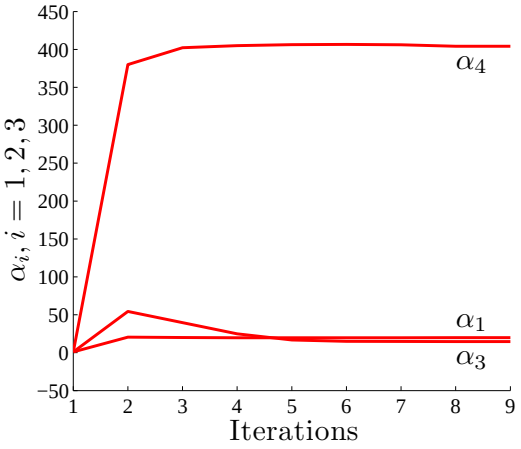

Рис. 3. Элементы, соответствующие информативным признакам.

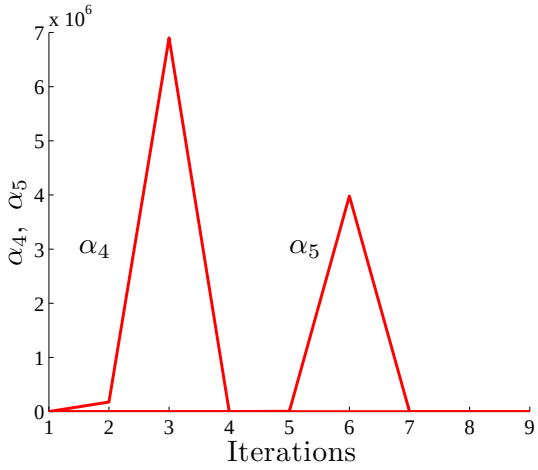

Рис. 4. Элементы, соответствующие неинформативным признакам.

элементов матрицы, при котором недиагональные элементы можно считать пренебрежительно малыми.

Таким образом, представим H_D в следующем виде: $H_D = \mathrm{diag}(h_1, \ldots, h_n)$.

Для нахождения рекуррентных формул, необходимых для нахождения гиперпараметров, используем необходимое условие минимума и приравнивняем к нулю производные логарифма правдоподобия по гиперпараметрам:

$$\alpha_i = \frac{1}{2}\lambda_i\left(\sqrt{1 + \frac{4}{(w_i - w_0)^2\lambda_i}} - 1\right), \quad \lambda_i = \beta h_i;$$

$$\beta = \frac{m - \gamma}{2E_D}, \quad \gamma = \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha_j}.$$

При решении уравнения для гиперпараметра α возникает два корня. Однако в рассматриваемой задаче имеют смысл только положительные корни, так как они являются дисперсиями случайных величин.

При оптимизации гиперпараметров возникает необходимость нахождения минимума функции ошибки $S(\mathbf{w})$. В работе рассматривается общий вид функции ошибки, отличающийся от стандартной суммы квадратов регрессионных остатков:

$$S = E_{\mathbf{w}} + \beta E_D = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^\intercal A(\mathbf{w} - \mathbf{w}_0) + \\ + \frac{1}{2}(\mathbf{y} - \mathbf{f})^\intercal \beta I(\mathbf{y} - \mathbf{f}).$$

Модификация алгоритма Левенберга-Марквардта

Для минимизации используется модифицированный алгоритм Левенберга-Марквардта. Алгоритм Левенберга-Марквардта предназначен для оптимизации параметров регрессионных моделей, заключается в последовательном приближении заданных начальных значений параметров к искомому локальному оптимуму и является обобщением метода сопряженных градиентов и алгоритма Ньютона-Гаусса. В данной работе шаг в пространстве параметров вычисляется по формуле:

$$\Delta\mathbf{w} = \left[\left(A + A^\intercal + X^\intercal(B^\intercal + B)X\right)^{-1}\right]^\intercal \times \\ \times\left(-\mathbf{w}^\intercal(A + A^\intercal) + (\mathbf{y} - X\mathbf{w})^\intercal(B^\intercal + B)X\right)^\intercal.$$

После нахождения рекуррентных формул организовывается итерационный процесс, который продолжается до сходимости как параметров $\mathbf{w}$, так и гиперпараметров $\boldsymbol{\alpha}$ и β, то есть до сходимости функции правдоподобия модели.

Вычислительный эксперимент

При появлении шумовых и мультикоррелирующих признаков происходит вырождение матрицы Гессе функции ошибки из-за возрастания диагональных элементов (большое значение дисперсии свидетельствует о неинформативности признака). Поэтому необходимо принудительно занижать возрастающие диагональные элементы, тем самым производя отбор признаков.

Тестирование алгоритма производится на временном ряде продаж нарезного хлеба в зависимости от времени. Ряд содержит 195 записей. Модель, аппроксимирующая ряд:

$$\mathbf{y} = 0.2256 + 0.1996\boldsymbol{\xi} + 0.0496\sin(10\boldsymbol{\xi}),$$

где $\boldsymbol{\xi} \in \mathbb{R}^n$ — регрессионная выборка. Введем следующие обозначения: $\boldsymbol{\xi}^0$, $\boldsymbol{\xi}^1$ — значение каждого элемента выборки в нулевой и первой степени соответственно, $\sin(10\boldsymbol{\xi})$ — поэлементное применение элементарной функции к вектору $\boldsymbol{\xi}$. На рис. 6 представлена выборка и аппроксимирующая её модель.

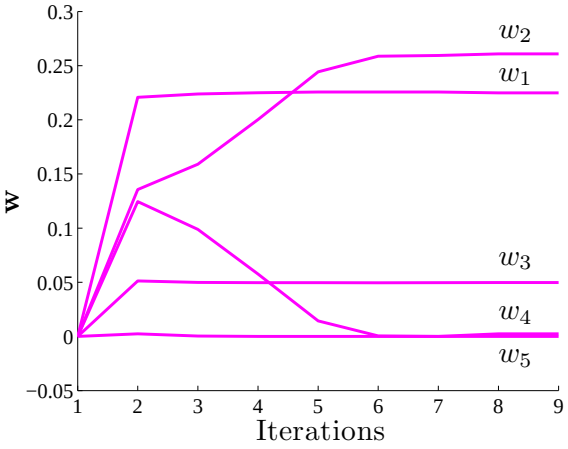

Рис. 7. Изменение параметров модели.

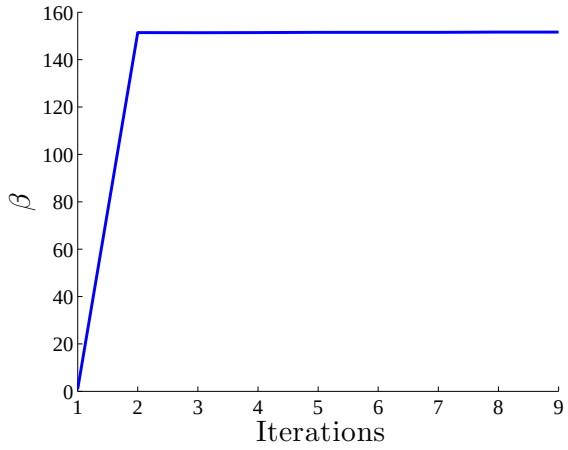

Рис. 5. Изменение скалярного гиперпараметра.

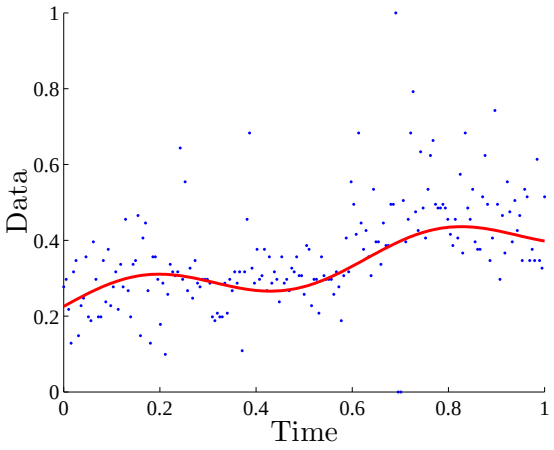

Рис. 6. Данные и аппроксимирующая модель.

Пусть матрица плана X представлена в следующем виде $X = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n]$, где $\boldsymbol{\chi} \in \mathbb{R}^m$. В данном случае она состоит из трёх столбцов: $\boldsymbol{\xi}^0$, $\boldsymbol{\xi}^1$, $\sin(10\boldsymbol{\xi})$. Добавим в матрицу плана шумовой и мультикоррелирующий признаки. Столбцы матрицы плана X: $\boldsymbol{\chi}_1 = \boldsymbol{\xi}^0$, $\boldsymbol{\chi}_2 = \boldsymbol{\xi}^1$, $\boldsymbol{\chi}_3 = \sin(10\boldsymbol{\xi})$, $\boldsymbol{\chi}_4 \sim \mathcal{N}(0,1)$, $\boldsymbol{\chi}_5 = \boldsymbol{\xi}^1 + \mathbf{k}$, где $\mathbf{k} \sim \mathcal{U}[0; 0.05]$.

В данном случае процесс сходится за 9 итераций. На 3-ей и 6-ой итерации гиперпараметр α, соответствующий шумовому и мультикоррелирующему признаку сильно возрастает. С помощью его принудительного занижения происходит отбор информативных признаков. На рис. 1 проиллюстрирована матрица Гессе H на 3-ей и 4-ой итерации. Произошло выявление и удаление шумового признака $\boldsymbol{\chi}_4$. На 6-ой и 7-ой итерациях был удалён мультикоррелирующий признак $\boldsymbol{\chi}_5$. Матрица Гессе представлена на рис. 2. На графиках 3 и 4 представлены диагональные элементы матрицы A. На рис. 5 представлены изменения скалярного гиперпараметры β. На рис. 7 показаны изменения параметров модели $\mathbf{w}$. Параметры модели w_4 и w_5 в конце итерационного процесса равны нулю. Фильтрация неинформативных признаков закончена.

Заключение

В работе решена задача фильтрации неинформативных признаков, а также построен алгоритм отбора, основанный на оценках ковариационных матриц распределений вектора параметров и вектора зависимой переменной. При использовании данного алгоритма не возникает необходимости разделения выборки на обучающую и контрольную части, алгоритм не содержит параметров, нуждающихся в дополнительных внешних оценках.

Литература

[1] *Tipping M. E.* Sparse Bayesian learning and the relevance vector machine // Journal of Machine Learning Research. — 2001. — №. 1. — Pp. 211–244.

[2] *Strijov V. V., Weber G. W.* Nonlinear regression model generation using hyperparameter optimization // Computers and Mathematics with Applications. — 2010. — Vol. 60, №. 4. — Pp. 981–988.

[3] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Вычислительные технологии. — 2007. — №. 1. — С. 93–102.

[4] *Токмакова А. А.* Получение устойчивых оценок гиперпараметров линейных регрессионных моделей // Машинное обучение и анализ данных. — 2017. — Т. 1, № 2. — С. 138–153.