Применение машины релевантных объектов в задачах восстановления числовых зависимостей

Разин Н. А.[1]***, Черноусова Е. О.***[2]***, Красоткина О. В.***[3]***, Моттль В. В.***[4]
nrmanutd@gmail.com[1], elchernousova@inbox.ru[2], o.v.krasotkina@yandex.ru[3], vmottl@yandex.ru[4]
МФТИ[1,2]; ТулГу[3,4]; ВЦ РАН[3,4]

В работе рассматривается задача беспризнакового распознавания образов в предположении, что объекты попарно сравниваются при помощи произвольной действительной функции. Такой подход является гораздо более общим, чем традиционный метод потенциальных функций (кернелов), требующий положительной полуопределенности матрицы функции сравнения объектов. Последнее требование в большинстве случаев является чрезмерным, причем обучение еще более осложняется, если существует несколько различных способов сравнительного представления объектов. В таких случаях экспериментатор вынужден решать задачу исключения как избыточных базисных объектов для сравнительного представления объектов обучающей совокупности, так и способов сравнения. В терминах общего пространства попарного сравнительного представления объектов предлагаемая постановка становится математически аналогичной классической задаче отбора признаков. Получившийся выпуклый критерий обучения аналогичен методу релевантных векторов Типпинга, но является существенно более общим, поскольку содержит структурный параметр, контролирующий селективность отбора.

Ключевые слова: *задача восстановления числовых зависимостей, дважды регуляризованный svm, машина релевантных векторов, алгоритмы регуляризации.*

Elastic-Net Relevance Vector Machines for selective multimodal regression estimation

Razin N. A.[1]***, Chernousova E. O.***[2]***, Krasotkina O. V.***[3]***, Mottl V. V.***[4]
MIPT[1,2]; Tula State University[3,4]; Computing Centre of the Russian Academy of Sciences[3,4]

The the problem of regression estimation is addressed under the assumption that pair-wise comparison of objects is arbitrarily scored by real numbers. Such a linear embedding is much more general than the traditional kernel-based approach, which demands positive semidefiniteness of the matrix of object comparisons. This demand is frequently prohibitive and is further complicated if there exists a large number of comparison functions, i. e., multiple modalities of object representation. In these cases, the experimenter typically also has the problem of eliminating redundant modalities and objects. In the context of the general pair-wise comparison space, this problem becomes mathematically analogous to that of wrapper-based feature selection. The resulting convex training criterion based on the principle of Elastic Net regression estimation is analogous to Tipping's Regression Relevance Vector Machine, but essentially generalizes it via the presence of a structural parameter controlling the selectivity level.

Keywords: *regression, elastic net, relevance vector machine, regularization path.*

Введение

Одна из современных задач информатики — это восстановление зависимостей по эмпирическим данным. Пусть задано множество объектов реального мира $\omega \in \Omega$, каждому из которых поставлено в соответствие значение ненаблюдаемой переменной $y \in \mathbb{Y}$. Предположим, что наблюдателю известны значения функции $y(\omega) : \Omega \to \mathbb{Y}$ только на объектах обучающей совокупности.

$$\Omega^* \Rightarrow \{\omega_j, y(\omega_j), j = 1, ..., N\} \subset \Omega. \tag{1}$$

Необходимо продолжить эту функцию на все множество объектов $\hat{y}(\omega) : \Omega \to \mathbb{Y}$ таким образом, чтобы было возможно оценить скрытую характеристику объектов $\omega \in \Omega \setminus \Omega^*$ [1]. В частности, если скрытая переменная принимает значение из множества действительных чисел $\hat{y}(\omega) : \Omega \to \mathbb{Y} = \mathbb{R}$, то такая задача известна как задача восстановления числовой зависимости (задача восстановления регрессии).

В простейшем случае предположим, что каждый объект представлен в компьютере вектором действительных чисел $\mathbf{x}(\omega) = \big(x_i(\omega), i \in \mathbb{I}\big) \in \mathbb{R}^n, \mathbb{I} = \{1, ..., n\}$, а числовая зависимость рассматривается как линейное решающее правило:

$$\hat{y}(\omega) = \hat{y}\big(\mathbf{x}(\omega)\big) = \sum_{i \in \mathbb{I}} a_i x_i(\omega) + b, \quad a_i, b \in \mathbb{R}. \tag{2}$$

То, что решающее правило (2) линейное, не является ограничивающим фактором, так как всегда можно выбрать такой способ описания объектов и задания их признаков $x_i(\omega)$, чтобы они погрузились ровно в такое линейное пространство, которое необходимо.

В качестве альтернативы к признаковому и кернельному подходам Дьюин и его соавторы предложили в [2] беспризнаковый подход, в котором объекты представлены произвольной функцией попарного сравнения $S(\omega', \omega'') : \Omega \times \Omega \to \mathbb{R}$. Идея заключалась в использовании значения этой функции между конкретным объектом $\omega \in \Omega$ и всеми остальными объектами обучающей совокупности $\{\omega_j, j = 1, ..., N\}$ как вектора вторичных признаков $\mathbf{x}(\omega) = \big(x_j(\omega) = S(\omega_j, \omega), j = 1, ..., N\big)$ и применении стандартных методов восстановления числовой зависимости (2), основанных на признаковом описании объекта в $\mathbb{R}^N$ с $\mathbb{J} = \{j = 1, ..., N\}$ вместо $\mathbb{I} = \{i = 1, ..., n\}$.

Позже Бишоп и Типпинг предложили «машину» (метод) релевантных векторов (Relevance Vector Machine – RVM) [3], где основной идеей был отбор только небольшого количества наиболее информативных релевантных объектов из обучающей совокупности. Авторы называли их релевантными векторами, поскольку $S(\omega', \omega'')$ рассматривалась как кернел, погружающий объекты в линейное пространство. В исходном методе RVM селективность релевантных векторов намеренно была принята очень высокой, чтобы придать им аналогию с опорными векторами в модели SVM. К тому же, метод RVM основан на байесовском принципе отбора, поэтому приводит к невыпуклой задаче обучения.

В данной работе рассматривается более общая ситуация, типичная для приложений, когда заданы несколько модальностей попарного представления объектов $S_k(\omega', \omega'')$ в виде различных функций парного сравнения $k \in \mathbb{K} = \{1, ..., m\}$, то количество вторичных признаков каждого объекта $n = mN$ превышает объем обучающей совокупности N:

$$\mathbf{x}(\omega) = \big(x_i(\omega) = x_{kj}(\omega) = S_k(\omega_j, \omega),\ k \in \mathbb{K},\ j \in \mathbb{J},\ i \in \mathbb{I} = \mathbb{K} \times \mathbb{J}\big). \tag{3}$$

Таким образом, чтобы избежать эффекта переобучения, необходимо выполнить отбор наиболее информативных вторичных признаков $(kj) \in \hat{\mathbb{I}} \subset \mathbb{I}$ и исключить остальные $(kj) \notin \hat{\mathbb{I}}$.

Мы рассмотрим класс задач обучения, которые, являясь выпуклыми в отличие от классической машины RVM [3], позволяют отбирать не только релевантные объекты $\hat{\mathbb{J}} \subset \mathbb{J}$, но и релевантные модальности их попарного представления $\hat{\mathbb{K}} \subset \mathbb{K}$, а также контролировать степень отбора с помощью специального параметра селективности $\mu \geqslant 0$.

Пусть дана обучающая совокупность $\{(\tilde{\mathbf{x}}_j, \tilde{y}_j), j = 1, ..., N\}$. Задача восстановления регрессии $\tilde{y}_j \in \mathbb{R}$ с избыточным количеством признаков $\mathbb{I}$ (вторичных признаков в случае задачи восстановления регрессии через RVM $\mathbb{I} = \mathbb{K} \times \mathbb{J}$) может быть записана как задача восстановления гребневой регрессии, т.е. взвешенной суммы квадратов ошибок отклонения $\sum_{j=1}^{N} \left[\tilde{y}_j - \left(\sum_{i \in \mathbb{I}} \tilde{a}_i \tilde{x}_{ij} + \tilde{b}\right)\right]^2 \to \min(\tilde{a}_i, i \in \mathbb{I}, \tilde{b})$ и штрафа на сумму квадратов коэффициентов $\beta \sum_{i \in \mathbb{I}} \tilde{a}_i^2 \to \min(\tilde{a}_i, i \in \mathbb{I})$, где β — коэффициент, взвешивающий эти два слагаемых между собой (trade-off coefficient). Очевидно, что подобный подход имеет смысл, только если исходная обучающая совокупность $\{(\tilde{\mathbf{x}}_j, \tilde{y}_j), j = 1, ..., N\}$ центрирована и нормирована:

$$\{(\mathbf{x}_j, y_j), j = 1, ..., N\},\ \mathbf{x}_j = \left(x_{ij}, i \in I\right) \in \mathbb{R}^n,\ y_j \in \mathbb{R}, \tag{4}$$

$$\frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j = \mathbf{0},\ \frac{1}{N} \sum_{j=1}^{N} y_j = 0,\ \frac{1}{N} \sum_{j=1}^{N} x_{ij}^2 = 1,\ i \in \mathbb{I}. \tag{5}$$

В противном случае штраф на сумму квадратов β_i должен быть уникальным для каждого признака $i \in \mathbb{I}$, чтобы скомпенсировать разницу в их масштабах. Если $\{(\tilde{\mathbf{x}}_j, \tilde{y}_j), j = 1, ..., N\}$ исходная обучающая совокупность, то нормализация может быть выполнена следующим образом:

$$x_{ij} = \frac{\tilde{x}_{ij} - \bar{x}_i}{\tilde{c}_i},\ y_j = \tilde{y}_j - \bar{y},\ \bar{x}_i = \frac{1}{N} \sum_{j=1}^{N} \tilde{x}_{ij},\ \bar{y} = \frac{1}{N} \sum_{j=1}^{N} \tilde{y}_j,\ \tilde{c}_i = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (\tilde{x}_{ij} - \bar{x}_i)^2}. \tag{6}$$

Обратно, когда модель регрессии $\hat{y}(\mathbf{x}) = \hat{\mathbf{a}}^T \mathbf{x}$ построена по нормализованной обучающей совокупности, она должна быть преобразована таким образом, чтобы корректно работать с произвольным новым объектом $\omega \in \Omega$, подаваемым на вход и описанным оригинальными признаками: $\tilde{\mathbf{x}}(\omega) = \left(\tilde{x}_i(\omega), i \in \mathbb{I}\right)$:

$$\hat{\tilde{y}}(\tilde{\mathbf{x}}) = \sum_{i=1}^{m} \tilde{\hat{a}}_i \tilde{x}_i + \tilde{b},\ \tilde{\hat{a}}_i = \frac{\hat{a}_i}{\tilde{c}_i},\ \tilde{b} = \bar{y} - \sum_{i=1}^{m} \frac{\hat{a}_i}{\tilde{c}_i} \bar{x}_i. \tag{7}$$

Таким образом, предполагая, что обучающая совокупность нормирована и центрирована (4)-(5), задача восстановления гребневой регрессии записывается в следующей форме:

$$F_{RR}(\mathbf{a} | \beta, \hat{\mathbb{I}}) = \beta \sum_{i \in \hat{\mathbb{I}}} a_i^2 + \sum_{j=1}^{N} \left(y_j - \sum_{i \in \hat{\mathbb{I}}} a_i x_{ij}\right)^2 \to \min(a_i, i \in \hat{\mathbb{I}}). \tag{8}$$

Критерий гребневой регрессии содержит два структурных параметра — коэффициент β и подмножество активных признаков $\hat{\mathbb{I}}$. Выбор первого из них не является большой проблемой, так как достаточно взять малое значение коэффициента $\beta > 0$ с той лишь целью, чтобы гарантировать строгую выпуклость критерия в случае коллинеарности векторов признаков из $\hat{\mathbb{I}}$. Но что касается выбора $\hat{\mathbb{I}}$, то это очень сложная задача.

Если задан конкретный набор признаков, то, воспользовавшись следствием из формулы Шермана-Вудбери-Моррисона [4], можно применить эффективный алгоритм [5] для вычисления ошибки leave-one-out критерия обучения гребневой регрессии (8). Тем не менее это невозможно сделать для всех 2^n наборов признаков $\hat{\mathbb{I}} \subseteq \mathbb{I}$. В нашем случае, в

соответствии с (3), должно быть найдено лучшее подмножество признаков $i = (kj) \in \mathbb{I} = \mathbb{K} \times \mathbb{J}$, количество которых $n = mN$ — произведение числа модальностей $m = |\mathbb{K}|$ и числа объектов в обучающей совокупности $N = |\mathbb{J}|$.

В данной работе разработана машина релевантных объектов (Relevance Object Machine) с возможностью отбора модальностей для задачи восстановления регрессии. В основу предлагаемого метода мы положили критерий обучения Elastic Net, представленный Zou и Hastie в [6] как результат комбинирования регуляризующего слагаемого гребневой регрессии $\beta \sum_{i\in\mathbb{I}} a_i^2 \to \min$ и Лассо Тибширани: $\mu \sum_{i\in\mathbb{I}} |a_i| \to \min$ [7]:

$$F_{EN}(\mathbf{a}|\beta,\mu) = \beta \sum_{i\in\mathbb{I}} a_i^2 + \mu \sum_{i\in\mathbb{I}} |a_i| + \sum_{j=1}^{N} \Big(y_j - \sum_{i\in\mathbb{I}} a_i x_{ij} \Big)^2 \to \min(a_i, i \in \mathbb{I}). \tag{9}$$

В отличие от гребневой регрессии (8) суммирование здесь выполняется по всем возможным признакам из $i \in \mathbb{I}$, т.е. вторичным признакам $(kj) \in \mathbb{I} = \mathbb{K} \times \mathbb{J}$ (3), и, следовательно, искомый набор признаков $\hat{\mathbb{I}} \subset \mathbb{I}$ не будет найден при помощи стандартного Elastic Net.

Наличие регуляризующего слагаемого Lasso наделяет данный критерий возможностью строго обнулять избыточные признаки и тем самым находить набор информативных признаков $\hat{\mathbb{I}}_{\beta,\mu} = \big\{ i : \hat{a}_{i,\beta,\mu} \neq 0 \big\} \subset \mathbb{I}$, не выполняя полного перебора. Степень отбора признаков регулируется параметром μ.

Когда коэффициент гребневой регрессии положителен, т.е. $\beta > 0$, критерий обучения Elastic Net (9) является строго выпуклым, однако более не является квадратичным, в отличие от гребневой регрессии (8). Тем не менее минимизировать его легко.

Одним из наиболее удобных преимуществ Elastic Net является простота, с которой можно реализовать алгоритм регуляризации [6]. Такой алгоритм позволяет в один проход найти последовательность наборов признаков, охватывающих все возможные значения параметра селективности, от тех, при которых все признаки исключены из решающего правила, и до тех, при которых все включены. Тем не менее выбор наиболее подходящего уровня селективности остается задачей внешних механизмов, например, кросс-валидации.

В данной работе предлагается встроить алгоритм регуляризации в алгоритм машины релевантных векторов регрессии для поиска последовательности наборов вторичных признаков $\hat{\mathbb{I}}_l = \hat{\mathbb{K}}_l \times \hat{\mathbb{J}}_l = \big\{ i = (kj), k \in \hat{\mathbb{K}}_l, j \in \hat{\mathbb{J}}_l \big\}$, $l = |\hat{\mathbb{J}}_l \times \hat{\mathbb{K}}_l|$, начиная от единственного признака $l = 1$, состоящего из одного объекта и одной модальности, $x_1(\omega) = x_{k_1 j_1}(\omega) = S_{k_1}(\omega_{j_1}, \omega)$, $\hat{\mathbb{K}}_1 = \{k_1\}$, $\hat{\mathbb{J}}_1 = \{j_1\}$, и заканчивая полным Декартовым произведением $\big\{ x_i(\omega) = x_{kj}(\omega) = S_k(\omega_j, \omega), (kj) \in \mathbb{I} = \mathbb{K} \times \mathbb{J} \big\}$, $l = n = mN = |\mathbb{K} \times \mathbb{J}|$.

Как только найдены все подмножества-кандидаты вторичных признаков $\hat{\mathbb{I}}_l \subset \mathbb{I}$, $l = 1, 2, 3, ..., n$, остается лишь выбрать лучшее из них с точки зрения обобщающей способности. Больше не нужен регуляризационный член Лассо $\mu \sum_{i\in\mathbb{I}} |a_i| \to \min$, и мы возвращаемся к обычной модели гребневой регрессии (8) с целью использования ее квадратичной формы для того, чтобы вычислить ее ошибку leave-one-out, применив эффективный беспереборный алгоритм, описанный в [5].

Статья начинается разделом 1, в котором математически тщательно описывается главное свойство критерия Elastic Net в точке минимума (9) — разбивать множество признаков на два подмножества: активные и неактивные. Активные, в свою очередь, разбиваются на положительные и отрицательные коэффициенты регрессии Elastic Net. Этот факт становится особенно очевидным в сравнении со стандартными прямыми попытками обоснования в [6], если выписать двойственную задачу для задачи оптимизации Elastic Net.

В разделе 2 предложен общий итеративный алгоритм для решения двойственной задачи при фиксированных параметрах регуляризации и поиска подмножества активных коэффициентов регрессии.

Далее, в разделе 3, излагается алгоритм регуляризации, а именно, алгоритм, решающий задачу Elastic Net в один проход для всех возможных значений параметра селективности.

Раздел 4 посвящен быстрому способу вычисления ошибки leave-one-out, представленному в [5], но адаптированному к двойственной задаче исходного критерия обучения.

Наконец, в разделе 5, приводятся результаты экспериментов на тестовых и реальных данных.

В статье пропущены элементарные доказательства некоторых математических предложений.

Двойственная задача для критерия Elastic Net и разбиение множества признаков

Обобщение критерия обучения Elastic Net (9) превращает его в многомодальную машину релевантных объектов, которая отличается от исходной более сложным смыслом вторичных признаков $i = (k, l)$ и их количеством $|\mathbb{I}| = mN$. Для того чтобы явно увидеть механизм отбора признаков, удобно записать критерий Elastic Net в его наглядной форме — выпуклой оптимизационной задачи с условиями-равенствами:

$$\begin{cases} \beta \sum\limits_{i\in\mathbb{I}} a_i^2 + \mu \sum\limits_{i\in\mathbb{I}} |a_i| + \sum\limits_{j=1}^{N} \delta_j^2 \to \min\big(a_i \in \mathbb{R}, i \in \mathbb{I}; \delta_j \in \mathbb{R}, j = 1, ..., N\big), \\ \delta_j = y_j - \sum\limits_{i\in\mathbb{I}} a_i x_{ij},\ j = 1, ..., N. \end{cases} \tag{10}$$

Утверждение 1. *Решение выпуклой задачи Elastic Net с ограничениями* (10) *выражается через решение двойственной выпуклой задачи без ограничений, т.е. через коэффициенты* $\boldsymbol{\delta} = (\delta_1, ..., \delta_N) \in \mathbb{R}^N$, *которые играют роль множителей Лагранжа:*

$$\begin{gathered} W(\boldsymbol{\delta}|\beta, \mu) = \frac{1}{\beta} \sum_{i\in\mathbb{I}} \Big\{ \min\Big[\frac{\mu}{2} + \bar{\mathbf{x}}_i^T\boldsymbol{\delta},\ 0,\ \frac{\mu}{2} - \bar{\mathbf{x}}_i^T\boldsymbol{\delta}\Big]^2 \Big\} + (\boldsymbol{\delta} - \mathbf{y})^T(\boldsymbol{\delta} - \mathbf{y}) = \\ \frac{1}{\beta} \sum_{i\in\mathbb{I}} \Big\{ \min\Big[\frac{\mu}{2} + \sum_{j=1}^{N} \delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N} \delta_j x_{ij}\Big]^2 \Big\} + \textstyle\sum_{j=1}^{N} (\delta_j - y_j)^2 \to \min, \end{gathered} \tag{11}$$

где $\bar{\mathbf{x}}_i = (x_{i,1}, ..., x_{i,N}) \in \mathbb{R}^N$ *— векторы соответствующих вторичных признаков в обучающей выборке* $i \in \mathbb{I} = \mathbb{K} \times \mathbb{J}$ (3).

Решение $\hat{\boldsymbol{\delta}}_{\beta\mu} = (\hat{\delta}_{1,\beta\mu}, ..., \hat{\delta}_{N,\beta\mu})$ *двойственной задачи* (11) *в свою очередь определяет разбиение множества признаков* $\hat{P}_{\beta\mu} = (\hat{\mathbb{I}}^-_{\beta\mu}, \hat{\mathbb{I}}^0_{\beta\mu}, \hat{\mathbb{I}}^+_{\beta\mu})$ *на активные* $\hat{\mathbb{I}}_{\beta\mu} = \hat{\mathbb{I}}^-_{\beta\mu} \bigcup \hat{\mathbb{I}}^+_{\beta\mu}$ *и неактивные* $\hat{\mathbb{I}}^0_{\beta\mu} = \big\{i \in \mathbb{I} : \hat{a}_{i,\beta\mu} = 0\big\}$:

$$\begin{gathered} \hat{\mathbb{I}}^-_{\beta\mu} = \Big\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} < -\frac{\mu}{2}, \hat{a}_{i,\beta\mu} < 0\Big\},\ \hat{\mathbb{I}}^+_{\beta\mu} = \Big\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} > \frac{\mu}{2}, \hat{a}_{i,\beta\mu} > 0\Big\}, \\ \hat{\mathbb{I}}^0_{\beta\mu} = \Big\{i \in \mathbb{I} : -\frac{\mu}{2} \leqslant \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} \leqslant \frac{\mu}{2}, \hat{a}_{i,\beta\mu} = 0\Big\}. \end{gathered} \tag{12}$$

Полученные активные коэффициенты регрессии являются линейными функциями от остатков, найденных из (11)

$$\hat{a}_{i,\beta\mu} = \frac{1}{\beta}\Big(\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} + \frac{\mu}{2}\Big) < 0 \text{ если } i \in \hat{\mathbb{I}}^-_{\beta\mu}, \quad \hat{a}_{i,\beta\mu} = \frac{1}{\beta}\Big(\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} - \frac{\mu}{2}\Big) > 0 \text{ если } i \in \hat{\mathbb{I}}^+_{\beta\mu}, \qquad (13)$$

тогда как оставшиеся коэффициенты равны нулю $\hat{a}_{i,\beta\mu} = 0$ *если* $i \in \hat{\mathbb{I}}^0_{\beta\mu}$.

Таким образом, решение задачи Elastic Net (9)-(10) полностью определяется решением двойственной задачи (11), чья целевая функций строго выпукла, как сумма двух функций, одна из которых выпуклая функция, а другая квадратичная. Следовательно, решение $\hat{\boldsymbol{\delta}}_{\beta\mu} = (\hat{\delta}_{1,\beta\mu}, ..., \hat{\delta}_{N,\beta\mu})$ всегда единственно.

Наглядно видно из (12) что, если коэффициент гребневой регрессии $\beta > 0$ зафиксирован, то штраф на сумму модулей $\mu > 0$ действует как параметр селективности, контролирующий количество признаков, попавших в выбранное множество активных признаков, или, что эквивалентно, регулирует количество отличных от нуля коэффициентов регрессии $\hat{\mathbf{a}}_{\beta\mu} = \big(\hat{a}_{i,\beta\mu}, i \in \mathbb{I}\big)$.

Когда $\mu = 0$, задача Elastic Net становится стандартной задачей гребневой регрессии. Коэффициенты гребневой регрессии (8) вычисляются через оптимальное решение двойственной задачи (11) в соответствии с (13):

$$W\big(\boldsymbol{\delta}|\beta, \mu = 0\big) = \frac{1}{\beta}\boldsymbol{\delta}^T\Big(\sum_{i\in\mathbb{I}}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\Big)\boldsymbol{\delta} + (\boldsymbol{\delta} - \mathbf{y})^T(\boldsymbol{\delta} - \mathbf{y}) \to \min(\boldsymbol{\delta}), \qquad (14)$$

$$\hat{\boldsymbol{\delta}}_{\beta 0} = \beta\Big(\sum_{i\in\mathbb{I}}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}\mathbf{y}, \;\; \hat{a}_{i,\beta 0} = \frac{1}{\beta}\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0}, \;\; i \in \mathbb{I}, \;\; \hat{\mathbb{I}}^0_{\beta 0} = \varnothing. \qquad (15)$$

Небольшое увеличение параметра селективности μ не изменит исходно пустое множество неактивных признаков, в то время как вектор остатков $\hat{\boldsymbol{\delta}}_{\beta 0}$ удовлетворяет неравенствам $\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0} \leqslant -(\mu/2)$ или $\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0} \geqslant (\mu/2)$ для всех $i \in \mathbb{I}$.

Напротив, когда μ слишком велико, все признаки становятся неактивными. В самом деле, второе слагаемое в целевой функции двойственной задачи (11) является квадратичной функцией $(\boldsymbol{\delta}-\mathbf{y})^T(\boldsymbol{\delta}-\mathbf{y})$. Если ее точка минимума $\boldsymbol{\delta} = \mathbf{y}$ удовлетворяет неравенствам $-(\mu/2) < \bar{\mathbf{x}}_i^T\mathbf{y} < (\mu/2)$ для всех $i \in \mathbb{I}$, то первое слагаемое становится равным нулю, и вектор $\mathbf{y}$ является решением двойственной задачи.

Таким образом, в соответствии с (12), для того чтобы перебрать все значения селективности, начиная с тех, при которых активных признаков нет вообще, $\big(\hat{\mathbb{I}}^0_{\beta\mu} = \varnothing, \; \hat{\mathbb{I}}_{\beta\mu} = \mathbb{I}\big)$ и заканчивая теми значениями, когда все признаки являются активными, $\big(\hat{\mathbb{I}}^0_{\beta\mu} = \mathbb{I}, \; \hat{\mathbb{I}}_{\beta\mu} = \varnothing\big)$, достаточно менять селективность в интервале

$$2\min_{i\in\mathbb{I}}|\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0}| = \mu_{\min} \leqslant \mu \leqslant \mu_{\max} = 2\max_{i\in\mathbb{I}}|\bar{\mathbf{x}}_i^T\mathbf{y}|. \qquad (16)$$

Итерационный алгоритм решения задачи Elastic Net для фиксированных параметров регуляризации

Пусть параметры регуляризации $\beta > 0$ и $\mu > 0$ в критерии обучения (9)-(10) зафиксированы. Для того чтобы описать итерационный алгоритм решения выпуклой двойственной задачи оптимизации (11), мы введем понятие произвольного разбиения P набора признаков $i \in \mathbb{I}$ на три подмножества:

$$P = \big\{\mathbb{I}^-_P, \mathbb{I}^0_P, \mathbb{I}^+_P\big\}, \quad \mathbb{I}^-_P \bigcup \mathbb{I}^0_P \bigcup \mathbb{I}^+_P = \mathbb{I}. \qquad (17)$$

Пусть, далее, $P_{\boldsymbol{\delta}} = \left\{\mathbb{I}^-_{\boldsymbol{\delta}}, \mathbb{I}^0_{\boldsymbol{\delta}}, \mathbb{I}^+_{\boldsymbol{\delta}}\right\}$ обозначает разбиение, связанное с конкретным вектором остатков $\boldsymbol{\delta} = (\delta_1, ..., \delta_N) \in \mathbb{R}^N$:

$$\mathbb{I}^-_{\boldsymbol{\delta}} = \left\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\boldsymbol{\delta} < -\frac{\mu}{2}\right\},\ \mathbb{I}^0_{\boldsymbol{\delta}} = \left\{i \in \mathbb{I} : \frac{\mu}{2} \leqslant \bar{\mathbf{x}}_i^T\boldsymbol{\delta} \leqslant \frac{\mu}{2}\right\},\ \mathbb{I}^+_{\boldsymbol{\delta}} = \left\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\boldsymbol{\delta} > \frac{\mu}{2}\right\}. \tag{18}$$

В частности, разбиение в точке минимума критерия Elastic Net (12) может быть переписано в этих терминах как:

$$\hat{P}_{\beta\mu} = P_{\hat{\boldsymbol{\delta}}_{\beta\mu}},\ \hat{\mathbb{I}}^-_{\beta\mu} = \mathbb{I}^-_{\hat{\boldsymbol{\delta}}_{\beta\mu}},\ \hat{\mathbb{I}}^0_{\beta\mu} = \mathbb{I}^0_{\hat{\boldsymbol{\delta}}_{\beta\mu}},\ \hat{\mathbb{I}}^+_{\beta\mu} = \mathbb{I}^+_{\hat{\boldsymbol{\delta}}_{\beta\mu}}.$$

Наконец, мы зафиксируем разбиение P в двойственной задаче оптимизации (11) и рассмотрим квадратичную задачу условной оптимизации и ее очевидное решение:

$$\begin{aligned} &W(\boldsymbol{\delta}|P) = \frac{1}{\beta}\left\{\sum_{i\in\mathbb{I}^-_P}\left(\bar{\mathbf{x}}_i^T\boldsymbol{\delta} + \frac{\mu}{2}\right)^2 + \sum_{i\in\mathbb{I}^-_P}\left(\bar{\mathbf{x}}_i^T\boldsymbol{\delta} - \frac{\mu}{2}\right)^2\right\} + (\boldsymbol{\delta} - \mathbf{y})^T(\boldsymbol{\delta} - \mathbf{y}) \to \min(\boldsymbol{\delta}),\\ &\hat{\boldsymbol{\delta}}_P = \Big(\sum_{i\in\mathbb{I}_P}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}\Big[\mathbf{y} + \frac{1}{2}\Big(\sum_{i\in\mathbb{I}^+_P}\bar{\mathbf{x}}_i - \sum_{i\in\mathbb{I}^-_P}\bar{\mathbf{x}}_i\Big)\mu\Big], \quad \mathbb{I}_P = \mathbb{I}^-_P \bigcup \mathbb{I}^+_P. \end{aligned} \tag{19}$$

Матрица $\sum\limits_{i\in\mathbb{I}_P}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}$, через которую выражается решение $\hat{\boldsymbol{\delta}}_P$ в (19), всегда обратима.

Утверждение 2. *Пусть P обозначает произвольное разбиение множества признаков и $\hat{\boldsymbol{\delta}}_P$ является решением соответствующей задачи условной оптимизации* (19). *Пусть, кроме этого, $P_{\hat{\boldsymbol{\delta}}_P}$ обозначает разбиение, полученное из $\hat{\boldsymbol{\delta}}_P$* (18). *Совпадение этих двух разбиений $P_{\hat{\boldsymbol{\delta}}_P} = P$ — необходимое и достаточное условие того, чтобы $\hat{\boldsymbol{\delta}}_P$ являлось решением двойственной задачи* (11).

Утверждение 2 описывает итерационный алгоритм решения двойственной задачи (11) и, следовательно, задачи Elastic Net (9)-(10), согласно утверждению 1.

Алгоритм. *Пусть $\hat{\boldsymbol{\delta}}_k$ является некоторым приближением искомого решения двойственной задачи и $P_k = P_{\hat{\boldsymbol{\delta}}_k}$* (18) *является разбиением множества признаков, соответствующим данному решению. Решение условной задачи оптимизации* (19) *с $P = P_k$ кладется равным следующему приближению $\hat{\boldsymbol{\delta}}_{k+1}$. Алгоритм останавливается, когда разбиения $P_{k+1} = P_{\hat{\boldsymbol{\delta}}_{k+1}}$ и P_k совпадают.*

Опыт показывает, что целесообразно начать с $\hat{\boldsymbol{\delta}}_0 = \mathbf{y}$, т. е. с тривиального разбиения $P = \left\{\mathbb{I}^-_P = \varnothing, \mathbb{I}^0_P = \mathbb{I}, \mathbb{I}^+_P = \varnothing\right\}$, при котором все признаки неактивны. В этом случае алгоритм, описанный выше, постепенно расширяет множество активных признаков $\hat{\mathbb{I}}_k = \hat{\mathbb{I}}^-_k \bigcup \hat{\mathbb{I}}^+_k$, но при этом расширение множества практически никогда не является жадным. С таким исходным приближением алгоритм сходится обычно за 15–30 итераций.

Количество итераций часто сокращается до 1–2 только если исходный вектор $\hat{\boldsymbol{\delta}}_0$ близок к решению. Алгоритм регуляризации использует это очень удачное свойство.

Алгоритм регуляризации

Идея алгоритма регуляризации впервые была сформулирована для критерия Лассо ($\beta = 0$) в [7] и затем расширена на случай Elastic Net ($\beta > 0$) в [6]. Мотивацией послужило предположение, что поиск зависимости вектора коэффициентов регрессии от параметра селективности $\hat{\mathbf{a}}_{\beta\mu} = \hat{\mathbf{a}}_\beta(\mu)$ в один проход сразу для всех значений (16) должен быть

вычислительно гораздо эффективнее, чем независимое вычисление искомого вектора коэффициентов регрессии для нескольких отдельных значений селективности.

Пусть коэффициент $\beta > 0$ взят достаточно малым, чтобы гарантировать строгую выпуклость целевой функции Elastic Net (9). В основе теоретического обоснования алгоритма регуляризации для всех значений параметра селективности $\mu > 0$ лежит математический факт, что зависимость $\hat{\mathbf{a}}_\beta(\mu)$, определенная обучающим критерием Elastic Net, — непрерывная, кусочно-линейная функция, имеющая конечное количество точек бифуркации. Будем рассматривать эти точки как уменьшающуюся последовательность $\big(\mu_{max} = \mu_0 > \mu_1 >, ..., > \mu_M = 0\big)$. Можно доказать, что количество нетривиальных точек M удовлетворяет неравенству $n \leqslant M \leqslant 2^n$, где $n = |\mathbb{I}|$ — количество признаков.

Это значит, что в терминах теоретического подхода, освещенного в разделах 1 и 2, для каждой обучающей совокупности (4) убывающая последвательность точек бифуркации $\mu_{max} \to \mu_k \to 0$ (16) задает соответствующую дискретную последовательность разбиений множества признаков P_k, начиная с разбиения, в котором все признаки отсутствуют $\hat{\mathbb{I}}_0 = \hat{\mathbb{I}}_0^- \bigcup \hat{\mathbb{I}}_0^+ = \varnothing$, и заканчивая разбиением, включающим в себя все признаки $\hat{\mathbb{I}}_M = \mathbb{I}$. Между двумя соседними точками бифуркации разбиение остается неизменным $\mu_{k-1} > \mu > \mu_k$, но следует отметить, что, вообще говоря, размер множества активных признаков $|\hat{\mathbb{I}}_k|$ не является монотонно возрастающей функцией от значения параметра селективности μ.

Теоретически всего точек бифуркации может быть 2^n. Однако такое число чрезмерно велико для применения предлагаемого алгоритма машины релевантных объектов, базирующегося на принципе отбора признаков. В нашем случае количество вторичных признаков $n = |\mathbb{I}| = mN$ превышает размер обучающей совокупности $N = |\mathbb{J}|$ в такое же количество раз, сколько задано функций парного сравнения объектов $m = |\mathbb{K}|$. Поэтому мы будем использовать «урезанный» алгоритм регуляризации.

Пусть $\mu_{\max}$ обозначает максимальное значение параметра селективности, вычисленное по обучающей совокупности (4) в соответствии с (16), и M — количество значений селективности, которое мы хотим попробовать. Предлагаемая последовательность $(\mu_0 > \mu_1 >, ..., > \mu_M)$ вычисляется как

$$\mu_k = \Big(1 - \frac{k}{M}\Big)\mu_{\max}, \quad k = 0, 1, ..., M, \quad \text{при этом} \quad \mu_0 = \mu_{\max}, \; \mu_M = 0. \tag{20}$$

Окончательным результатом алгоритма регуляризации является последовательность подмножеств активных признаков $\hat{\mathbb{I}}_k$, $k = 0, 1, ..., M$. Мощность этих подмножеств $|\hat{\mathbb{I}}_k|$ в общем случае возрастает от 0 до полного количества признаков $n = mN$, но ее зависимость от параметра селективности необязательно монотонная.

Нет необходимости вычислять коэффициенты регрессии Elastic Net, потому что мы используем эту модель только лишь для отбора признаков. Вместо этого мы вычисляем коэффициенты гребневой регрессии (13) для каждого из отобранных подмножеств активных признаков в соответствии с (15):

$$\hat{\boldsymbol{\delta}}_{\hat{\mathbb{I}}_k} = \big(\hat{\delta}_{1,\hat{\mathbb{I}}_k}, ..., \hat{\delta}_{N,\hat{\mathbb{I}}_k}\big) = \Big(\sum_{i \in \hat{\mathbb{I}}_k} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T + \beta \mathbf{I}\Big)^{-1} \mathbf{y}, \quad \hat{a}_{i,\hat{\mathbb{I}}_k} = \begin{cases} (1/\beta)\bar{\mathbf{x}}_i^T \hat{\boldsymbol{\delta}}_{\hat{\mathbb{I}}_k}, & i \in \hat{\mathbb{I}}_k, \\ 0, & i \notin \hat{\mathbb{I}}_k. \end{cases} \tag{21}$$

Бесперебoрная процедура верификации leave-one-out для подмножеств активных признаков

На каждом k-ом шаге алгоритма регуляризации полученное среднее значение суммы квадратов остатков гребневой регрессии (8) для заданного подмножества признаков $\hat{\mathbb{I}}_k \subseteq \mathbb{I}$ дается выражениями:

$$Q_{\hat{\mathbb{I}}_k} = \frac{1}{N}\sum_{j=1}^{N}(\hat{\delta}_{j,\hat{\mathbb{I}}_k})^2, \tag{22}$$

где вектор остатков $\hat{\boldsymbol{\delta}}_{\hat{\mathbb{I}}_k} = \big(\hat{\delta}_{1,\hat{\mathbb{I}}_k}, ..., \hat{\delta}_{N,\hat{\mathbb{I}}_k}\big)$ вычисляется согласно (21). Отметим, что остатки являются непосредственным решением двойственной задачи (11), а вычисление коэффициентов регрессии $\hat{a}_{i,\hat{\mathbb{I}}_k}$ не является обязательным для того, чтобы посчитать $Q_{\hat{\mathbb{I}}_k}$.

Ошибка leave-one-out

$$Q^{LOO}_{\hat{\mathbb{I}}_k} = \frac{1}{N}\sum_{j=1}^{N}(\hat{\delta}^{(j)}_{j,\hat{\mathbb{I}}_k})^2, \tag{23}$$

является средним арифметическим квадратов элементов вектора ошибок, каждый из которых вычисляется по всей обучающей совокупности, исключая соответствующий j-й элемент, в отличие от (21):

$$\begin{aligned}
&\hat{\boldsymbol{\delta}}^{(j)}_{\hat{\mathbb{I}}_k} = \big(\hat{\delta}^{(j)}_{1,\hat{\mathbb{I}}_k}, ..., \hat{\delta}^{(j)}_{N,\hat{\mathbb{I}}_k}\big) = \Big(\sum_{i\in\hat{\mathbb{I}}_k}\tilde{\tilde{\mathbf{x}}}^{(j)}_i(\tilde{\tilde{\mathbf{x}}}^{(j)}_i)^T + \beta\tilde{\mathbf{I}}\Big)^{-1}\tilde{\mathbf{y}}^{(j)},\\
&\tilde{\tilde{\mathbf{x}}}^{(j)}_i = \big(x_{il}, l\neq j\big)\in\mathbb{R}^{N-1},\ \tilde{\mathbf{y}}^{(j)} = \big(y_l, l\neq j\big)\in\mathbb{R}^{N-1},\ \ \tilde{\mathbf{I}}\big[(N-1)\times(N-1)\big].
\end{aligned} \tag{24}$$

Как показано в [5], квадратичная форма критерия гребневой регрессии (8) позволяет избежать множественных обращений матриц при вычислении ошибки leave-one-out. В терминах напрямую посчитанных остатков в (23)-(24) при решении двойственной задачи (11), ошибка leave-one-out определяется выражением

$$Q^{LOO}_{\hat{\mathbb{I}}_k} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{\delta}_{j,\hat{\mathbb{I}}_k}}{1-q_{j,\hat{\mathbb{I}}_k}}\right)^2,\ \ q_{j,\hat{\mathbb{I}}_k} = \left[\Big(\sum_{i\in\hat{\mathbb{I}}_k}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\Big)\Big(\sum_{i\in\hat{\mathbb{I}}_k}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}\right]_{jj}, \tag{25}$$

где $[...]_{jj}$ означает j-й диагональный элемент квадратной матрицы.

Можно заметить, что обратная матрица $\Big(\sum_{i\in\hat{\mathbb{I}}_k}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}$ вычисляется лишь однажды на каждом шаге алгоритма регуляризации при оценке остатков гребневой регрессии по всей обучающей совокупности (21).

Результаты экспериментов

Эксперименты на искусственных данных

Продемонстрируем использование предложенной машины релевантных объектов с отбором модальностей на двухмерных искусственных данных. Данные устроены таким образом, что по ним явно можно понять истинное множество релевантных объектов и релевантных модальностей. Набор данных состоит из объектов $\omega\in\Omega$, представленных двумя наблюдаемыми признаками $\mathbf{z}(\omega) = \big(z_1(\omega), z_2(\omega)\big)\in\mathbb{R}^2$ и одним скрытым признаком $y(\omega)\in\mathbb{R}$.

Генератор случайных данных основан на предзаданной функции $\mathbb{R}^2 \to \mathbb{R}$, определенной следующим образом:

$$\begin{array}{l} y^*(\mathbf{z}) = K(\mathbf{z}_1^*, \mathbf{z}) + K(\mathbf{z}_2^*, \mathbf{z}) - K(\mathbf{z}_3^*, \mathbf{z}) - K(\mathbf{z}_4^*, \mathbf{z}), \\ K(\mathbf{z}^*, \mathbf{z}) = \exp\{-5.5||\mathbf{z}^* - \mathbf{z}||^2\}, \\ \mathbf{z}_1^* = (-0.26,\, 0.69),\ \mathbf{z}_2^* = (0.47,\, 0.76),\ \mathbf{z}_3^* = (-0.7,\, 0.32),\ \mathbf{z}_4^* = (0.25,\, 0.4). \end{array} \tag{26}$$

Каждый случайный объект представлен парой $(\mathbf{z}, y)$, где $\mathbf{z}$ — случайный вектор, равномерно распределенный в прямоугольнике $\big(-1 \leqslant z_1 \leqslant 1,\ -0.2 \leqslant z_2 \leqslant 1.2\big)$ и y — нормально распределенная случайная величина с условным математическим ожиданием $E(y|\mathbf{z}) = y^*(\mathbf{z})$ и дисперсией $Var(y) = 0.1$.

Набор данных содержит $|\Omega| = 1250$ объектов, которые случайно разбиты на $N = 150$ объектов для обучения и $N_{test} = 1100$ объектов для контроля. На рис. 1 показана заданная скрытая функция $y^*(\mathbf{z})$, которую нужно оценить, и обучающая совокупность $\big\{(\mathbf{z}_j, y_j), j = = 1, ..., N\big\}$. Контурные линии заданной скрытой функции хорошо демонстрируют два «холма» в верхней части координатной плоскости и две «впадины» в нижней.

Далее мы применили к обучающей совокупности предложенную в работе машину релевантных объектов с отбором модальностей для фиксированного значения параметра $\beta = 0.1$ и для $m = 4$ функций парного сравнения:

$$\begin{array}{l} S_1(\omega', \omega'') = S_1(\mathbf{z}', \mathbf{z}'') = \exp\big\{-5.5\big[(z_1' - z_1'')^2 + (z_2' - z_2'')^2\big]\big\}, \\ S_2(\omega', \omega'') = S_2(\mathbf{z}', \mathbf{z}'') = |z_1' - z_1''|, \\ S_3(\omega', \omega'') = S_3(\mathbf{z}', \mathbf{z}'') = |z_2' - z_2''|, \\ S_4(\omega', \omega'') = S_4(\mathbf{z}', \mathbf{z}'') = |(z_2' - z_2'') - (z_1' - z_1'')|. \end{array} \tag{27}$$

Особенность скрытой заданной функции (26) указывает на то, что алгоритм должен отобрать только первую функцию $S_1(\mathbf{z}', \mathbf{z}'')$ как единственную адекватную функцию искомой зависимости, а в качестве релевантных объектов отобрать объекты обучающей совокупности, расположенные преимущественно в окрестности «холмов» и «впадин».

Таким образом, мы имеем полный набор $\mathbb{I} = \big\{i = (kj)\big\}$ из $n = mN = 400$ вторичных признаков (3), сформированный из $m = 4$ модальностей и $N = 100$ объектов обучения. Для подобной экспериментальной структуры максимальное значение параметра селективности (16) равняется $\mu_{max} \approx 11$.

Применение алгоритма регуляризации привело к тому, что лучшее значение селективности равно $\mu = 1.21$, давая минимальное среднее значение ошибки leave-one-out $Q_{\hat{\mathbb{I}}}^{LOO} = 0.089$, вычисленное согласно (25). Средний квадрат ошибки на контроле оказался немного выше: $Q_{\hat{\mathbb{I}}}^{test} = 0.125$. Обе оценки хорошо согласуются со значением дисперсии $Var(y) = 0.1$, для которой генерировались искусственные данные.

Соответствующее подмножество релевантных вторичных признаков $\hat{\mathbb{I}} = \big\{(kj)\big\}$ содержит 18 элементов. Среди них 17 элементов были сформированы разными релевантными объектами $\hat{\mathbb{J}} = \{j\} \subset \mathbb{J}$, связанными, как и ожидалось, с одной и той же функцией парного сравнения $S_1(\mathbf{z}', \mathbf{z}'')$. И только один релевантный объект предпочел другую функцию парного сравнения $S_4(\mathbf{z}', \mathbf{z}'')$ (28). Релевантные объекты расположились в окрестностях «холмов» и «впадин» (рис. 1), как и ожидалось.

Эксперименты на реальных данных

Эффективность машины релевантных объектов мы проверили на задаче восстановления числовой зависимости, которая заключается в моделировании намагниченности

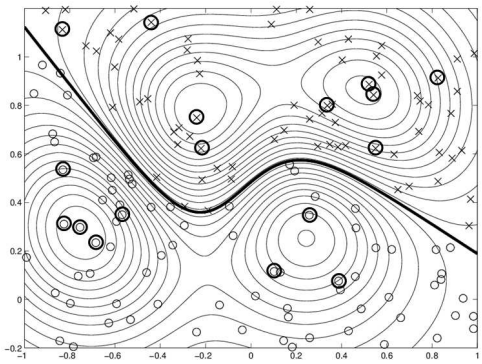

Рис. 1: Контурные линии функции $y^*(\mathbf{z})$ и обучающая совокупность $\{(\mathbf{z}_j, y_j), j = 1, ..., N\}$, $N = 150$. Жирная линия обозначает множество точек, для которых $y^*(\mathbf{z}) = 0$, маркеры $\times$ и $\odot$ обозначают объекты обучающей совокупности $\mathbf{z}_j$ и, соответственно, положительные и отрицательные значения наблюдаемой целевой переменной y_j. Релевантные объекты подсвечены

в сверхпроводниках [8]. Данные для этой задачи предоставлены Национальным технологическим институтом стандартов (National Institute of Standards and Technology) `http://www.nist.gov/index.html`.

Набор данных состоит из объектов $\omega \in \Omega$, представленных одним наблюдаемым признаком $\mathbf{z}(\omega) = z_1(\omega) \in \mathbb{R}$ и одним скрытым признаком $y(\omega) \in \mathbb{R}$.

Скрытый признак — это измеренная намагниченность сверхпроводника, наблюдаемый признак — это логарифм от времени, прошедшего с момента старта эксперимента до момента его завершения. Истинная зависимость между скрытым и наблюдаемым признаком описывается $y = -2000(50 + z_1)^{-10/9}$.

Набор данных содержит $|\Omega| = 154$ объекта, которые случайно разбиты на $N = 75$ объектов для обучения и $N_{test} = 79$ объекта для контроля. На рис. 2 показана заданная скрытая функция $y^*(\mathbf{z})$, которую нужно оценить, и обучающая совокупность $\{(\mathbf{z}_j, y_j), j = = 1, ..., N\}$.

Далее мы применили к обучающей совокупности предложенную в работе машину релевантных объектов с отбором модальностей для фиксированного значения параметра $\beta = 15$ и для $m = 4$ функций парного сравнения:

$$\begin{aligned} S_1(\omega', \omega'') &= S_1(\mathbf{z}', \mathbf{z}'') = (1 + |z_1' - z_1''|)^{-10/9}, \\ S_2(\omega', \omega'') &= S_2(\mathbf{z}', \mathbf{z}'') = \exp(-1.5(z_1' - z'')^2), \\ S_3(\omega', \omega'') &= \exp(-1.5 * |z_1' - z_1''|)/(1 + (z_1' + z_1'')^2), \\ S_4(\omega', \omega'') &= S_4(\mathbf{z}', \mathbf{z}'') = \exp(-1.5|z_1' - z_1''|) \end{aligned} \tag{28}$$

Особенность скрытой заданной функции указывает на то, что алгоритм должен отобрать только первую функцию $S_1(\mathbf{z}', \mathbf{z}'')$ как единственную адекватную функцию искомой зависимости, а в качестве релевантных объектов отобрать небольшое количество объектов обучающей совокупности.

Таким образом, мы имеем полный набор $\mathbb{I} = \{i = (kj)\}$ из $n = mN = 300$ вторичных признаков (3), сформированный из $m = 4$ модальностей и $N = 75$ объектов обучения. Для подобной экспериментальной структуры максимальное значение параметра селективности (16) равняется $\mu_{max} \approx 94.5$.

Применение алгоритма регуляризации привело к тому, что лучшее значение селективности равно $\mu = 186.4$, давая минимальное среднее значение ошибки leave-one-out

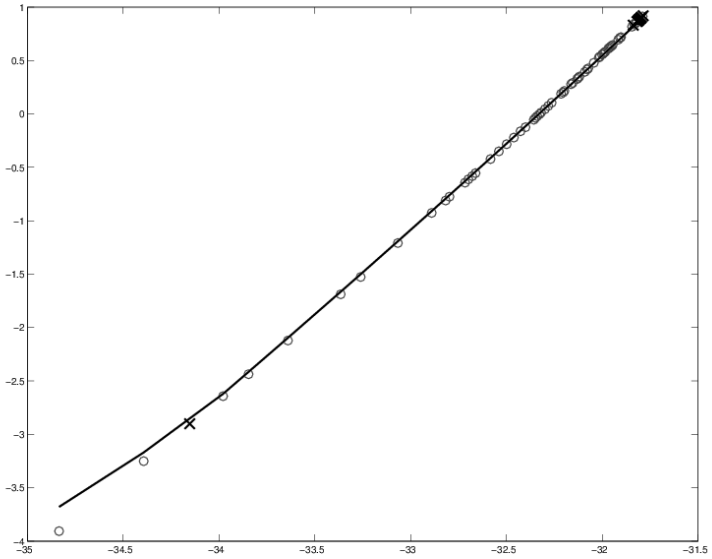

Рис. 2: Обучающая совокупность $N = 75$ и восстановленная по ней зависимость $y^*(\mathbf{z})$. Маркеры $\odot$ и $\times$ обозначают объекты обучающей совокупности $\mathbf{z}_j$ и, соответственно, релевантные объекты

$Q^{LOO}_{\hat{\mathbb{I}}} = 0.000695$, вычисленное согласно (25). Средний квадрат ошибки на контроле оказался немного выше: $Q^{test}_{\hat{\mathbb{I}}} = 0.000805$. Обе оценки хорошо согласуются со значением квадрата отклонения истинной зависимости от наблюдаемых значений скрытой переменной $Var(y) = 0.00052$.

Соответствующее подмножество релевантных вторичных признаков $\hat{\mathbb{I}} = \{(kj)\}$ содержит 11 элементов. Все эти элементы были сформированы разными релевантными объектами $\hat{\mathbb{J}} = \{j\} \subset \mathbb{J}$, связанными с одной и той же функцией парного сравнения $S_1(\mathbf{z}', \mathbf{z}'')$, как и ожидалось.

Выводы

В данной работе предложена методология решения задач восстановления регрессии на основе множества произвольных действительнозначных функций парного сравнения, не обязанных удовлетворять трудновыполнимым на практике условиям положительной полуопределенности. Предложенная разработка является обобщением относительного дискриминантного анализа (Relational Discriminant Analysis) Роберта Дьюина, с одной стороны, и машины релевантных векторов типпинга — с другой.

Критерий обучения отличается от озвученных ранее тем, что он явно включает в себя структурный параметр, контролирующий степень отбора признаков, описывающих объекты, каждый из которых связан с соответствующим объектом в обучающей совокупности и с конкретной функцией парного сравнения. Кроме того, многомодальная машина релевантных объектов с селективностью является выпуклой, следовательно, гарантированно имеет единственное решение.

Явное наличие параметра селективности делает критерий удобным для задач, в которых присутствуют несколько функций парного сравнения, т.е. биоинформатику и многомодальную биометрию.

Литература

[1] *Vapnik V.* Estimation of dependencies based on empirical data. Springer, 1982.

[2] *Duin R., Pekalska E., de Ridder D.* Relational discriminant analysis // *Pattern Recognition Lett.* 1999. Vol. 20. Pp. 1175–1181.

[3] *Bishop C., Tipping M.* Variational Relevance Vector Machines // *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2000. Pp. 46–53.

[4] *Bartlett M. S.* An inverse matrix adjustment arising in discriminant analysis // *Ann. Math. Stat.* 1951. Vol. 22, no. 1. Pp. 107–111.

[5] *Cawley G. C., Talbot N. L. C.* Fast exact leave-one-out cross-validation of sparse least-squares support vector machines // *Neural Networks*. 2004. Vol. 17. Pp. 1467–1475.

[6] *Zou H., Hastie T.* Regularization and variable selection via the elastic net // *J. R. Stat. Soc.* 2005. Vol. 67. Pp. 301–320.

[7] *Tibshirani R.* Regression shrinkage and selection via the lasso // *J. R. Stat. Soc.* 1996. Vol. 58, no. 1. Pp. 267–288.

[8] *Bennett L., Swartzendruber L., Brown H.* Superconductivity magnetization modeling. NIST, 1994