Линейные классификаторы с приложением в медицине

Крютченко Ольга Игоревна, гр. 422

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.
Рецензент: к. ф.-м. н., Ананьевская П.В.

Санкт-Петербург
2018 г.

Введение: постановка задачи

Рассматривается проблема классификации многомерных данных с пропусками на примере логистической регрессии.

Задачи:

1. применение проективного метода классификации для адаптации логистической регрессии на случай неполных данных,
2. проверка обобщающей способности предложенного адаптированного алгоритма,
3. редукция размерности множества классификаторов с помощью анализа главных компонент,
4. прогноз рубцевания полости в лёгких у больных туберкулёзом.

Используемые методы: логистическая регрессия

Выборка $(\mathbf{x}_i, y_i)_{i=1}^n$, где

- $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ – вектор значений независимых переменных,
- $y_i \in \{0, 1\}$ – значение зависимой переменной.

Логистическая регрессия – статистическая модель, предсказывающая вероятность принадлежности к классу 1 по значениям предикторов:

$$\pi(\mathbf{x}) = P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\theta_0 - \theta_1 x_1 - \ldots - \theta_m x_m)},$$

где $(\theta_0, \ldots, \theta_m)$ – **параметры** (подбираются с помощью принципа максимума правдоподобия по обучающей выборке).

Отклик:

$$\hat{y}(\mathbf{x}) = \begin{cases} 1 & \text{при } \pi(\mathbf{x}) \geqslant C \\ 0 & \text{при } \pi(\mathbf{x}) < C \end{cases}, \text{где } C \text{ – } \textbf{порог отсечения}.$$

Используемые методы: скользящий контроль

Скользящий контроль — процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам.
В работе использовался контроль по отдельным объектам.

1. Выборка $D = (\mathbf{x}_i, y_i)_{i=1}^n$ разбивается n различными способами на тестовую $D_h^m = (\mathbf{x}_h, y_h)$ и обучающую $D_h^l = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n; i \neq h\}$.
2. n раз подбираются параметры θ по выборке D_h^l и находится значение $\hat{y}_h$.
3. Оценка скользящего контроля (cross-validation)
$$\mathrm{CV} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq \hat{y}_i\}}.$$

Используемые методы: проективный метод классификации

Проективный метод классификации (ПМК) [Алексеева, Горлова, Бондаренко, 2017] с дискриминантной функцией в качестве классификатора был адаптирован для логистической регрессии.

1. На подмножествах $\mathbf{x}_\tau = (x_{\tau\eta_1}, \ldots, x_{\tau\eta_k})$ строятся частные классификаторы $\pi(\mathbf{x}_\tau)$.

   $\hat{y}(\mathbf{x}_\tau) = \mathbb{1}_{\{\pi(\mathbf{x}_\tau) \geqslant C\}}$ — отклик частного классификатора.

2. Для построения интегрального классификатора отбираются классификаторы с заданными свойствами.

ПМК: способы вычисления итогового отклика

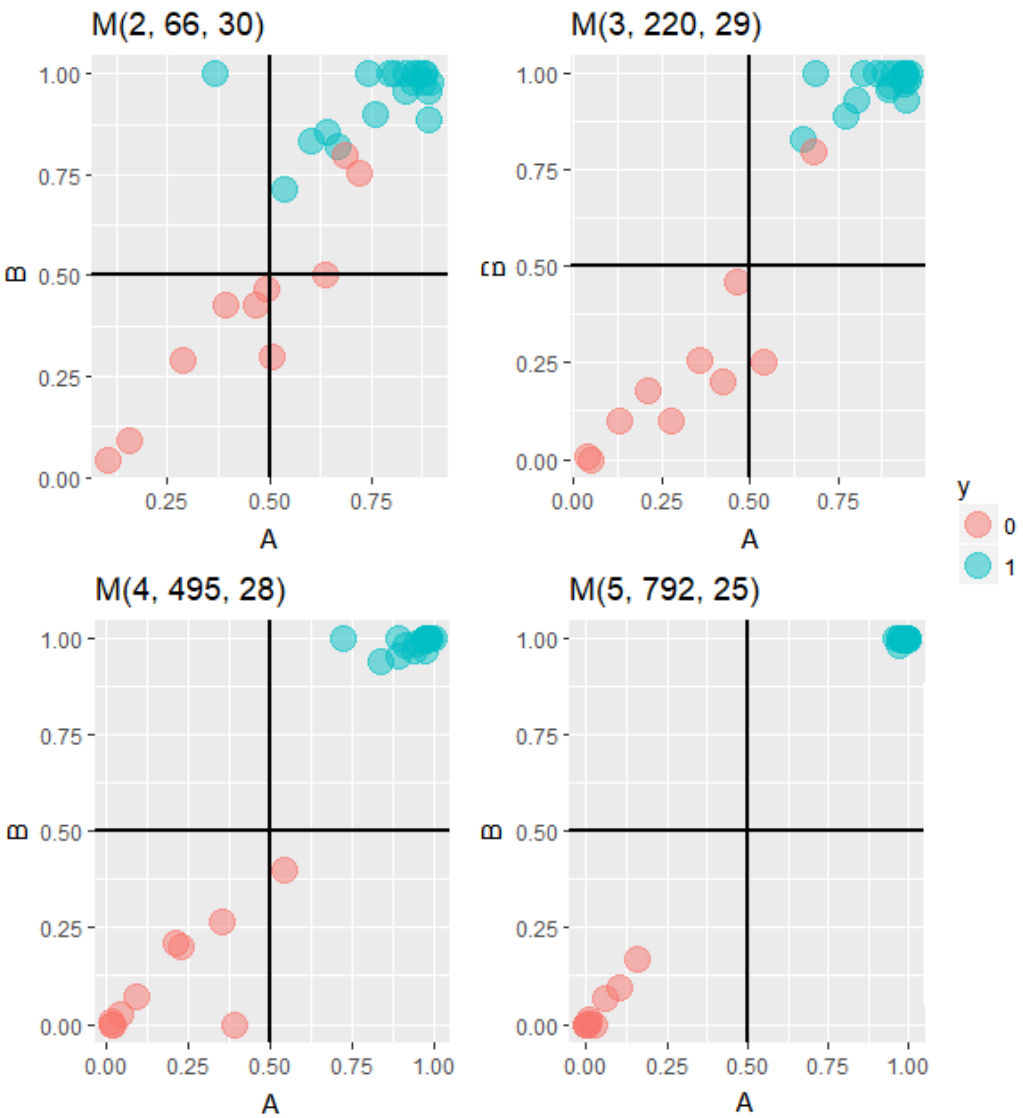
1. По значениям S логистических функций

$$\hat{y}_*(\mathbf{x}_\tau) = \mathbb{1}_{\{\bar{\pi}(\mathbf{x}_\tau) \geqslant 0.5\}}, \text{ где } \bar{\pi}(\mathbf{x}_\tau) = \frac{1}{S}\sum_{s=1}^{S} \pi(\mathbf{x}_{\tau^s}).$$

2. По откликам S частных классификаторов

$$\hat{y}^*(\mathbf{x}_\tau) = \mathbb{1}_{\{\bar{y}(\mathbf{x}_\tau) \geqslant 0.5\}}, \text{ где } \bar{y}(\mathbf{x}_\tau) = \frac{1}{S}\sum_{s=1}^{S} \hat{y}(\mathbf{x}_{\tau^s}).$$

Результаты: сравнительный анализ моделей



$M(k, S_k, n_k)$ – модель:
k – число предикторов в классификаторе,
S_k – число классификаторов,
n_k – число охваченных индивидов.
A — величина усреднённой логистической функции,
B — величина усреднённого отклика.

Классы лучше разделены при вычислении итогового отклика по усреднённому отклику.

ПМК: Отбор логистических функций

Задача: найти наименьшее по мощности множество номеров логистических функций (S_{opt}), максимизирующее долю верной классификации.

- S — множество номеров логистических функций, $s \in S$
- $n(s)$ — количество индивидов, вошедших в s-ый классификатор
- $P(s)$ — доля верной классификации s-го классификатора
- $S(\delta) = \{s : P(s) \geq \delta, n(s) \geq n_0\}, \delta \in [\delta_1, \delta_2]$
- $P(S(\delta))$ — доля верной классификации по $S(\delta)$
- $P_{\max} = \max\limits_{\delta \in [\delta_1, \delta_2]} P(S(\delta))$
- $\delta_{\text{opt}} = \arg\min\limits_{\delta \in [\delta_1, \delta_2]} \{|S(\delta)| : P(S(\delta)) = P_{\max}\}$

$$S_{\text{opt}} = S(\delta_{\text{opt}})$$

ПМК: скользящий контроль

Задача: Адаптировать метод контроля по отдельным объектам для проективного метода классификации.

1. Разбиваем выборку $D = (\mathbf{x}_i, y_i)_{i=1}^n$ на обучающую $D_h^l = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n; i \neq h\}$ и тестовую $D_h^m = (\mathbf{x}_h, y_h)$.
2. С помощью отбора получаем набор логистических функций с номерами из S_{opt}, минимизирующих ошибку CV на обучающей выборке D_h^l.
3. $\bar{y}(\mathbf{x}_i) = \frac{1}{|S_{\text{opt}}|} \sum\limits_{s \in S_{\text{opt}}} \hat{y}(\mathbf{x}_{is})$,

   $\hat{y}^*(\mathbf{x}_i) = \mathbb{1}_{\{\bar{y}(\mathbf{x}_i) \geqslant 0.5\}}$ — итоговый отклик.

Повторяя эти шаги n раз, получаем вектор $\{\hat{y}^*(\mathbf{x}_i) | i \in \mathcal{I}\}$.
Вычисляем оценку скользящего контроля (cross-validation)

$$\text{CV} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq \hat{y}^*(\mathbf{x}_i)\}}.$$

Результаты: отбор логистических функций

Чувствительность (sensitivity): $\mathrm{SNS} = \sum_{i=1}^{n} \mathbb{1}_{\{\hat{y}_i + y_i = 0\}} / \sum_{i=1}^{n} \mathbb{1}_{\{y_i = 0\}}$.

Специфичность (specificity): $\mathrm{SPC} = \sum_{i=1}^{n} \mathbb{1}_{\{\hat{y}_i y_i = 1\}} / \sum_{i=1}^{n} \mathbb{1}_{\{y_i = 1\}}$.

Решая задачу прогноза рубцевания полости по показателям металлопротеиназ (12 шт.), был произведён отбор логистических функций и оценена обобщающая способность этого алгоритма.

	SNS	SPC	$P(S_{\mathrm{opt}})$	$n(S_{\mathrm{opt}})$
без кросс-валидации	0.89	0.80	0.83	30
кросс-валидация	0	0.88	0.56	25

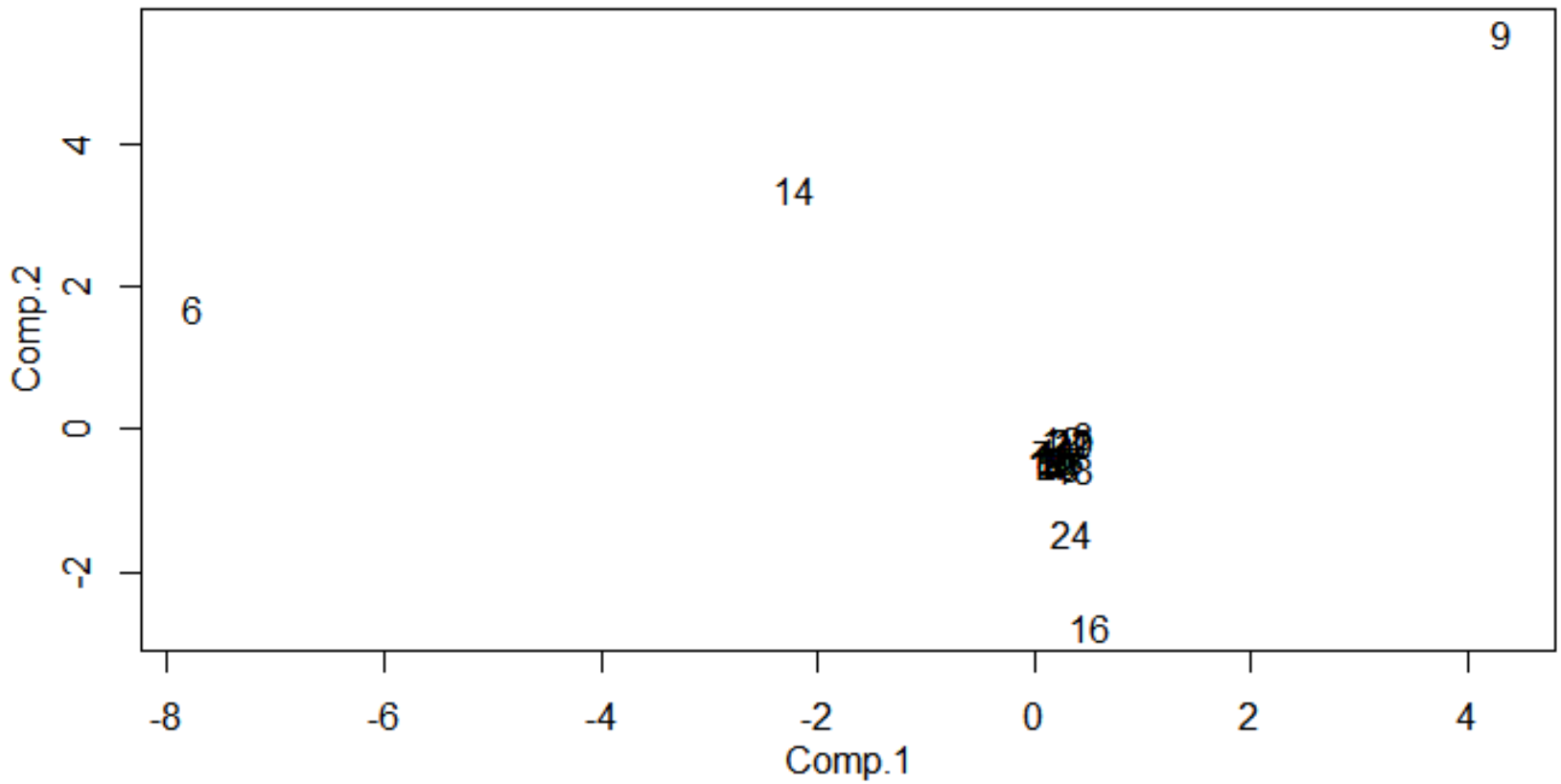
Сделан вывод о том, что имеет место переобучение.

В рассмотрение включены дополнительные показатели (23 шт.).

	SNS	SPC	$P(S_{\mathrm{opt}})$	$n(S_{\mathrm{opt}})$
без кросс-валидации	0.82	1	0.93	29
кросс-валидация	0.22	1	0.72	25

Результаты: анализ главных компонент

Функции с номерами $6, 9, 14, 16, 24$ вносят наибольший вклад.
Редукции размерности: в классификации используются только эти пять функций и одна из оставшихся.
Несколько наборов функций дают следующий результат:
$P(S_{\text{opt}}) = 1$, при $n(S_{\text{opt}}) = 28$.



Результаты: параметры логистических функций

- Приведён один из наборов по 6 функций, который хорошо классифицирует данные и легко поддаётся интерпретации.
- В таблице приведены параметры этих логистических функций, рекомендованные эксперементаторам для интерпретации.

	уров. бактеривыдел.	распространённость	резистентность	деструкция	объём полости	$Timp_{1,1}$	$Timp_{1,2}$	$MMP_{1,1}$	$MMP_{1,2}$	$MMP_{9,1}$
5						-0.002	-0.005	0.089		
6	554.3	85.25	-3.517		-168.8					
9	46.78		4.009	-6.765		0.038				
14	-29.45				-103.8		-0.511		100.3	
16	5.103				-3.050			2.005		0.003
24					-2.021		-0.008	0.779		0.002

Заключение

- Решена задача адаптации логистической регрессии на случай неполных данных для прогноза рубцевания полости в лёгких у больных туберкулёзом.
- Сформулирован и реализован алгоритм, позволяющий использовать проективный метод с логистической регрессией в качестве классификатора.
- На языке программирования R осуществлён отбор классификаторов по двум множествам признаков. После вычисления обобщающих способностей выбран наиболее эффективный.
- В результате редукции размерности множества классификаторов с помощью анализа главных компонент были отобраны классификаторы, позволяющие легко интерпретировать результаты.