Процедура построения скоринговой модели (скоринговой карты)

1. Подготовка данных для построения скоринговой модели
   1.1. Анализ и предварительная обработка
   1.2. Определение зависимой переменной
   1.3. Определение независимых переменных
   1.4. Формирование обучающей и тестовой выборки
   1.5. Определение объема выборки
2. Анализ и корректировка переменных для построения модели
   2.1. Корректировка распределения зависимой переменной
   2.2. Описательный анализ скоринговых переменных (для обнаружения ошибок и заполнения пропусков + оценить силу связи между независимыми переменными)
   2.3. Преобразование количественных переменных и порождение новых признаков
   2.4. Оценка мультиколлинеарности между количественными переменными
   2.5. Категоризация количественных переменных (биннинг)
   2.6. Сегментация выборки
   2.7. Оценка взаимосвязи скоринговых переменных на вероятность дефолта
3. Построение скоринговой карты
   3.1. Выбор модели (логистическая регрессия)
   3.2. Включение независимых переменных в модель
   3.3. Выбор критерии качества модели логистической регрессии
   3.4. Перевод коэффициентов модели в скоринговую карту

Список переменных

Variable	Type	Categories
Loan currency	Nominal	3
Applied amount	Linear	
Monthly payment	Linear	
Tetm of contract	Linear	
Region of the office	Nominal	7
Day of week of scoring	Linear	
Hour of scoring	Linear	
Age	Linear	
Gender	Nominal	2
Marital status	Nominal	4
Education	Ordinal	5
Number of children	Linear	
Industrial sector	Nominal	27
Salary	Linear	
Place of birth	Nominal	94
...	...	...
Car number shown	Nominal	2

Преобразование шкал

- Область деятельности заемщика, номинальная шкала

Nominal	Tourism	Banking	Education
John	1	0	0
Thomas	0	1	0
Sara	0	0	1

- Образование заемщика, ординальная шкала

Ordinal	Primary	Secondary	Higher
John	1	0	0
Thomas	1	1	0
Sara	1	1	1

Группировка признаков: оптимизационная задача

Мы имеем начальную модель, заданную набором индексов $\mathcal{A}$. Добавим полученные в результате группировки признаки и рассмотрим улучшение функционала качества.

$$\begin{array}{rccccccl} \xi = & 1 & 2 & 3 & \dots & c, & c \text{ число категорий}, \xi \in C; \\ & \downarrow & \downarrow & \downarrow & & \downarrow & \\ x_j = & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_c, & |\Gamma| \text{ число групп}, \gamma \in \Gamma. \end{array}$$

Требуется найти функцию

$$h : C \to \Gamma.$$

Задача оптимизации ставится так:

$$(h, |\Gamma|) = \arg \max_{h \in H} S(w)_{\mathcal{A} \cup j}$$

и решается методом полного перебора или генетическим алгоритмом.

Постановка задачи: данные

1. Набор данных: $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$,
$$D = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^i, y^i), \ldots, (\mathbf{x}^m, y^m)\};$$
2. матрица плана $X \in \mathbb{R}^{m \times n}$,
$$X = (\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n);$$
3. целевая переменная $\mathbf{y} \sim \text{Bernoulli}(\boldsymbol{\sigma})$;
$$\mathbf{y} = (y^1, \ldots, y^m)^T,$$
4. модель
$$\mathbf{y} = \boldsymbol{\sigma}(\mathbf{w}) + \varepsilon, \qquad \boldsymbol{\sigma}(\mathbf{w}) = \frac{1}{1 + \exp(-X\mathbf{w})}.$$

Индексы

- объектов $\{1, \ldots, i, \ldots, m\} = \mathcal{I}$, поделены $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$;
- признаков $\{1, \ldots, j, \ldots, n\} = \mathcal{J}$; обозначим $\mathcal{A}$ активное множество признаков.

Постановка задачи: целевая функция

Целевая функция (критерий качества модели) — логарифмическая функция правдоподобия:

$$-\ln P(D|\mathbf{w}) = -\sum_{i\in\mathcal{L}} \Big(y^i \ln \mathbf{w}^T\mathbf{x}^i + (1-y^i)\ln(1-\mathbf{w}^T\mathbf{x}^i) \Big) = S(\mathbf{w}).$$

Требуется найти активное множество признаков $\mathcal{A} \subset \mathcal{J}$ и параметров модели $\mathbf{w}_\mathcal{A}$, доставляющих максимум функции

$$S(\mathbf{w}_{ML})_\mathcal{A} \longrightarrow \min_{\mathcal{A}\subseteq\mathcal{J}, i\in\mathcal{T}},$$

где

$$\mathbf{w}_{ML} = \arg\min_{\mathbf{w}\in\mathcal{W}, i\in\mathcal{L}} S(\mathbf{w}).$$

Индексы

- объектов $\{1,\ldots,i,\ldots,m\} = \mathcal{I}$, разбиты $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$;
- признаков $\{1,\ldots,j,\ldots,n\} = \mathcal{J}$; обозначим $\mathcal{A}$ активное множество признаков.

Структурные параметры и выбор моделей

Полный перебор порожденных обобщенных линейных моделей

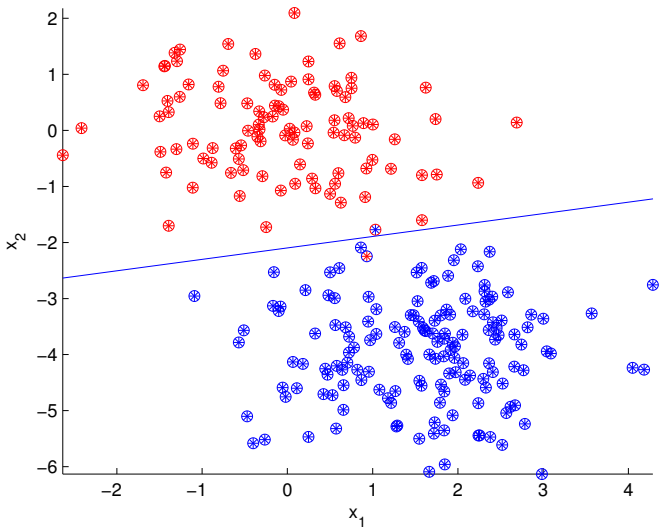
$$\mu(y) = w_0 + \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \ldots + \alpha_R w_R x_R.$$

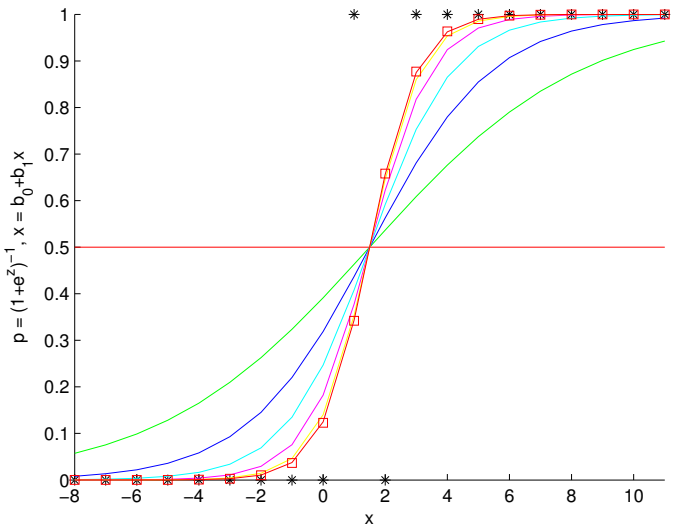
Здесь $\alpha \in \{0, 1\}$ — структурный параметр.

Найти модель, заданную множеством индексов активных признаков $\mathcal{A} \subseteq \mathcal{J}$:

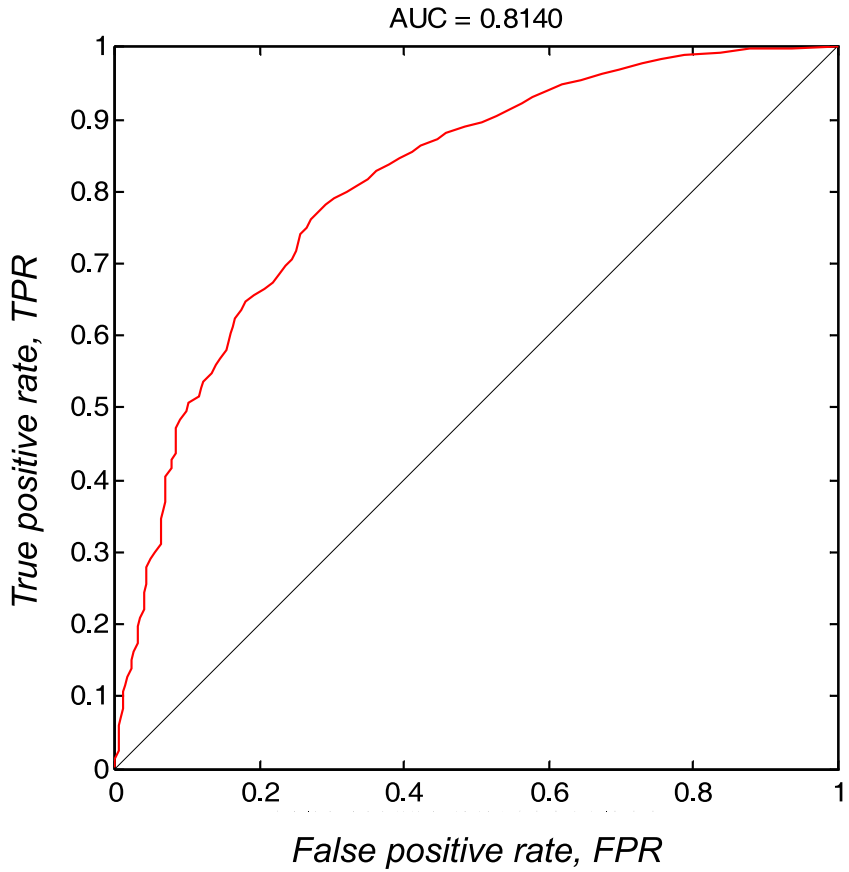
α_1	α_2	$\ldots$	$\alpha_{\lvert\mathcal{J}\rvert}$
1	0	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	$\ldots$	1

Разделяющая плоскость и логистическая кривая





ROC-кривая как дополнительный критерий качества



	P	N
P^*	TP	FP
N^*	FN	TN

$$TPR = TP/P = TP/(TP + FN)$$
$$FPR = FP/N = FP/(FP + TN)$$

Кстати, $2AUC = Gini + 1$

Устойчивость модели во времени

1. Последовательные сегменты времени делят выборку на подвыборки
2. Модель тестируется на подвыборках, результаты представляются в виде пулов
3. Пулы для различных сегментов сравниваются