Банковский кредитный скоринг: методы автоматического порождения и выбора моделей

Вадим Стрижов

Вычислительный центр РАН

Семинар «Задачи анализа данных в бизнес-аналитике»
17 сентября 2010

Процесс

Заявка-анкета, кредитная история

↓

Скоринговый балл: вероятность невозврата (PD)

Вариант: кредитный контроль

↓

Назначение кредитного лимита

↓

Сообщение о принятии заявки или об отказе

↓

Заключение договора

↓

...

↓

Появление кредитной истории

Виды скоринговых карт

- Выдача кредита (Application scoring)
- Динамика состояния (Behavioral scoring)
- Просроченная задолженность (Collection scoring)

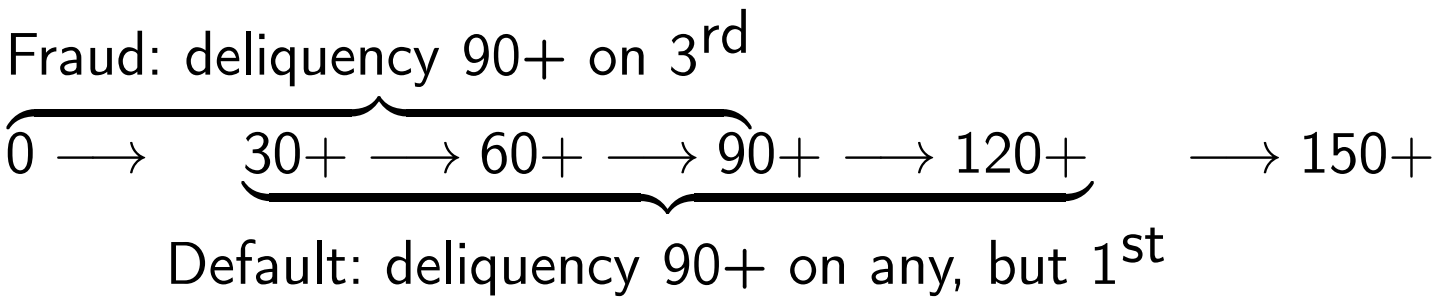
Типы кредитов для физических лиц:

- Потребительский (POS)
- Кредит наличными
- Автокредит
- Ипотечный

Типичное число клиентских записей в базе данных:

- $\sim 10^4$ для «тяжелых» долгосрочных кредитов,
- $\sim 10^6$ для «легких» кредитов,
- $\sim 10^7$ для банковских карт.

Типы невозвратов кредита



- Fraud — мошенничество
- Default — возврат кредита просрочен

Убытки от невозвратов кредита

Примерная просрочка (от недели и выше) по потребительским кредитам на некоторый момент времени

Категория	Количество	Сумма
Все категории товаров	100 000	2 100 М
Бытовая техника	30 000	350 М
Мебель	20 000	300 М
Одежда	15 000	200 М
Телевизоры	10 000	100 М
Мобильные телефоны	15 000	80 М
Фотоаппараты	2 000	20 М

Причины отказа в кредите

Некоторые типичные причины:

- недостаточный скоринговый балл,
- не прошел кредитный контроль,
- в черном списке банка,
- просрочка по данным бюро кредитных историй,
- не гражданин России,
- маленький личный доход,
- клиент моложе (старше) определенного возраста и сумма слишком велика,
- мобильный телефон найден у другого клиента.

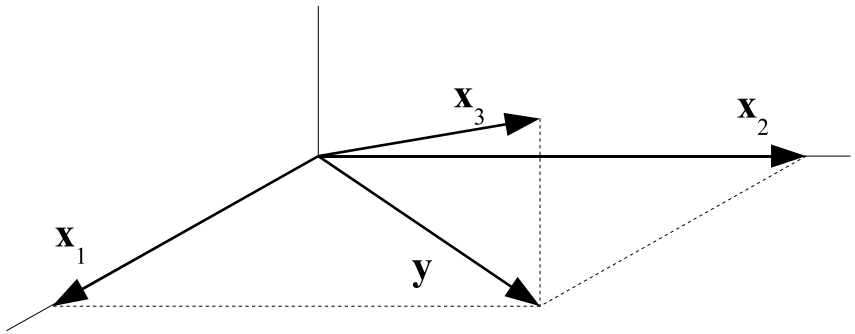
Разработка скоринговых карт

1. Определение целей и характеристик будущей карты
2. Создание матрицы плана, «витрины» (design matrix)
3. Преобразование ординальных и номинальных признаков в бинарные
4. Создание «композитных» признаков
5. Одномерный анализ, отбор наиболее информативых признаков по одному
6. Многомерный анализ, создание регрессионной модели
7. Тестирование модели (на мультиколлинеарность, устойчивость, и т.д.)
8. Определение порога отсечения согласно политике банка
9. Документирование модели и утверждение на заседании правления
10. Программирование, внедрение в информационную систему
11. Передача на эксплуатацию

Различие между классическим и новым подходом

- Отказ от одномерного анализа
- Автоматическое построение композитных признаков

Одномерный или многомерный?



Инструменты

- ORACLE SQL, SQL-Developer
- SAS, SAS-Data Miner
- Ksema-XSEN
- Matlab
- SPSS
- ...
- MS-Excel

Общие сведения о выборке

- Кредиты с просрочкой 90+, дефолты
- Случаи мошенничества (fraud) из выборки исключены
- Всего элементов выборки $\sim 10^4$–10^6
- Доля просрочивших (default rate) $\sim$ 8–16%
- Период наблюдения – не менее 91 дней после заключения контракта
- Число исходных переменных $\sim$ 30–50
- Число пропущенных записей > 0, обычно мало
- Число записей-выбросов > 0, $3\sigma^2$-cutoff

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

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: функционал качества

The quality criterion is the log likelihood function

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

We must find the active set $\mathcal{A} \subset \mathcal{J}$ and the model parameters $\mathbf{w}_\mathcal{A}$, such that

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

where $\mathcal{I} = \mathcal{L} \sqcup \mathcal{T}$.

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.

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

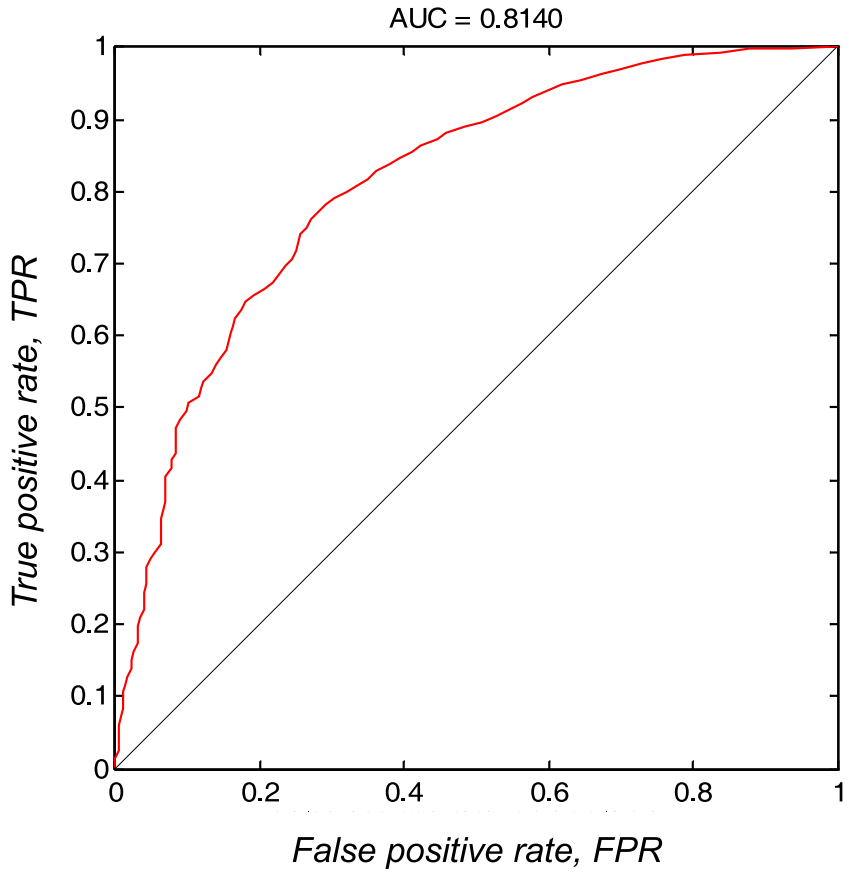
$$h : C \to \Gamma.$$

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.

ROC-кривая как дополнительный критерий качества

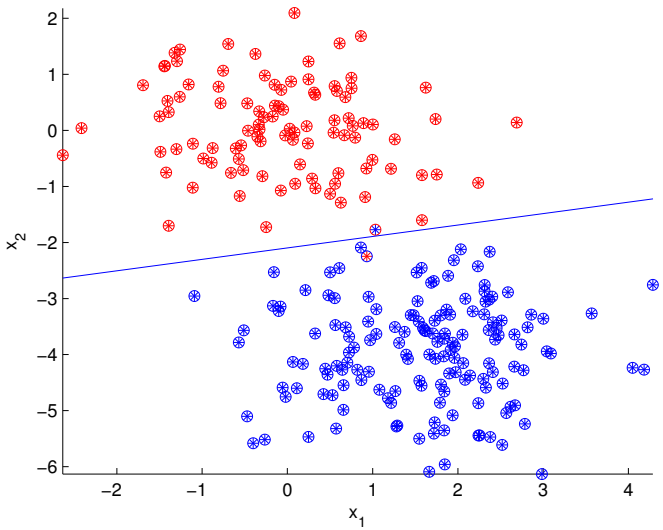


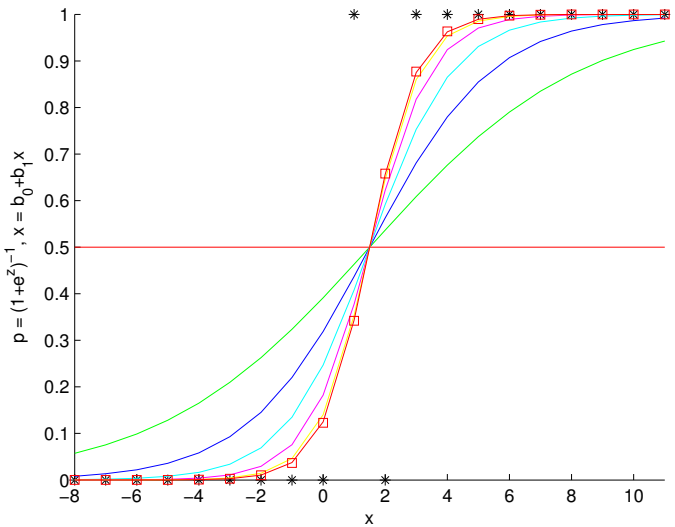
	P	N
P^*	TP	FP
N^*	FN	TN

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

Кстати, $2AUC = Gini + 1$

Разделяющая плоскость и логистическая кривая





Мультикорреляция и VIF

Фактор инфляции дисперсии j-го признака

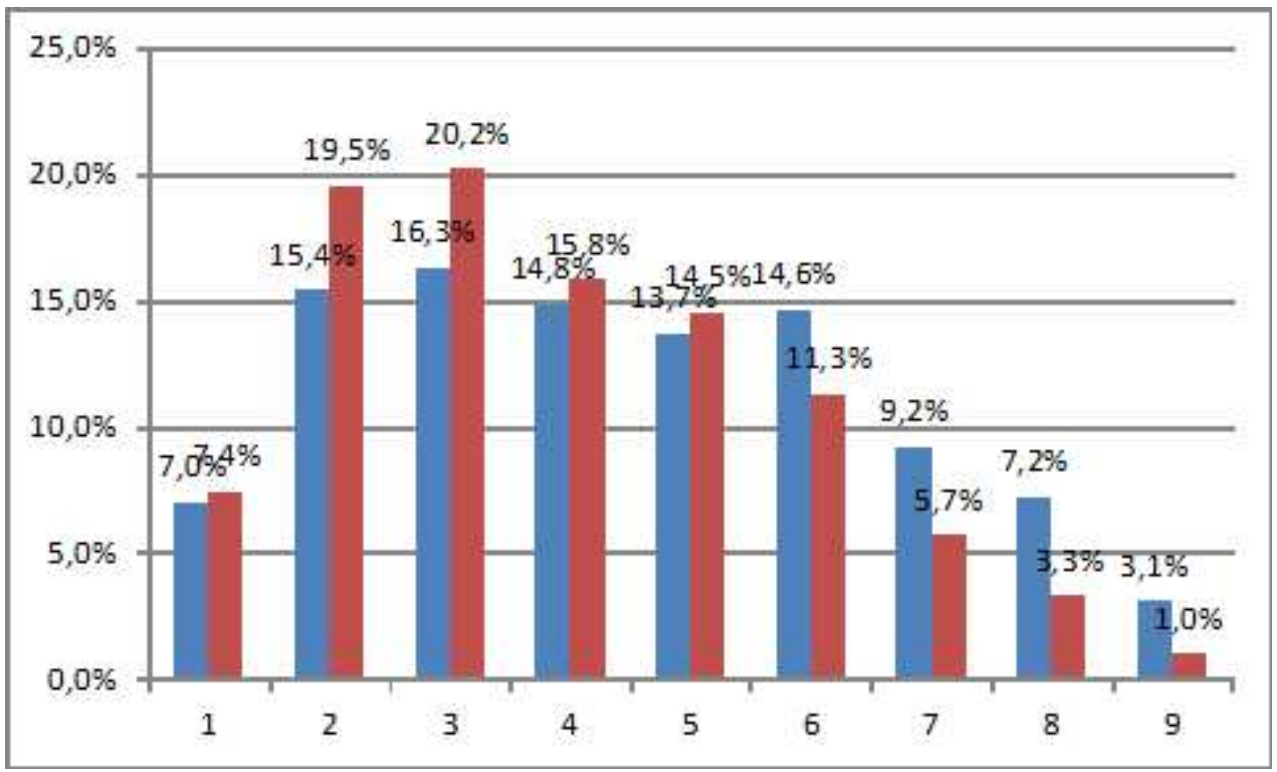
$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

где R_j^2 — коэффициент детерминации,

$$R_j^2 = 1 - \frac{\|\vec{x}_j - \vec{f}(\vec{x}_1, \ldots, \vec{x}_{j-1}, \vec{x}_{j+1}, \ldots, \vec{x}_n)\|^2}{\|\vec{x}_j - \vec{1}\bar{x}_j\|^2}.$$

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



Список порождающих функций

Description	In	N in	Out	N out	Comm	Param
Nominal to binary	nom	1	bin	1–4	-	Yes
Ordinal to binary	ord	1	bin	1–4	-	Yes
Linear to linear segments	lin	1	lin	1–4	-	Yes
Linear segments to binary	lin	1	bin	1–4	-	Yes
Get one column of n-matrix	bin	1–4	bin	1	-	Yes
Conjunction	bin	2–6	bin	1	Yes	-
Dijsunction	bin	2–6	bin	1	Yes	-
Negate binary	bin	1	bin	1	-	-
Logarithm	lin	1	lin	1	-	-
Hyperbolic tangent sigmiod	lin	1	lin	1	-	-
Logistic sigmoid	lin	1	lin	1	-	-
Sum	lin	2–3	lin	1	Yes	-
Difference	lin	2	lin	1	No	-
Multiplication	lin,bin	2–3	lin	1	Yes	-
Division	lin	2	lin	1	No	-
Inverse	lin	1	lin	1	-	-
Polynomial transformation	lin	1	lin	1	-	Yes
Radial basis function	lin	1	lin	1	-	Yes
Monomials: $x\sqrt{x}$, etc.	lin	1	lin	1	-	-

Задача порождения признаков

Даны

- измеряемые признаки $\Xi = \{\xi\}$,
- заданные экспертами порождающие функции $G = \{g(\mathbf{b}, \xi)\}$,
$$g : \xi \mapsto x;$$
- правила порождения: $\mathcal{G} \supset G$, где суперпозиция $g_k \circ g_l \in \mathcal{G}$ построена с учетом ограничений на число типы входных и выходных переменных ;
- правила упрощения суперпозиций: g_u не принадлежит $\mathcal{G}$, если существует правило
$$r : g_u \mapsto g_v \in \mathcal{G}.$$

Результат

набор «композитных» признаков $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n\}$.

Внимание! Число порожденных признаков может превосходить число клиентов!

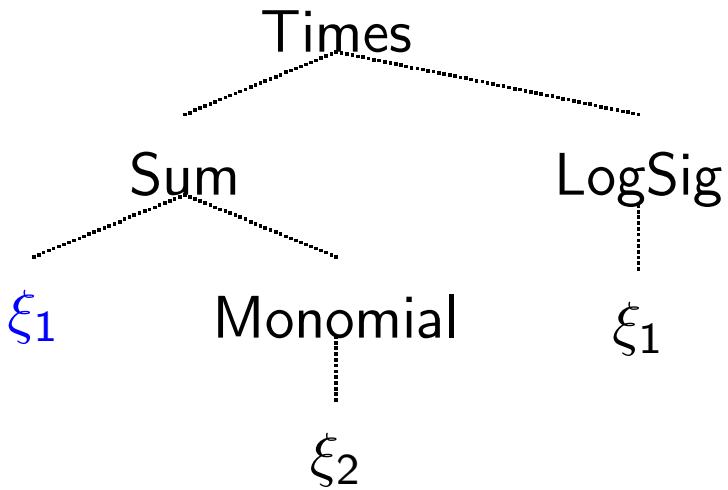
Примеры композитных признаков

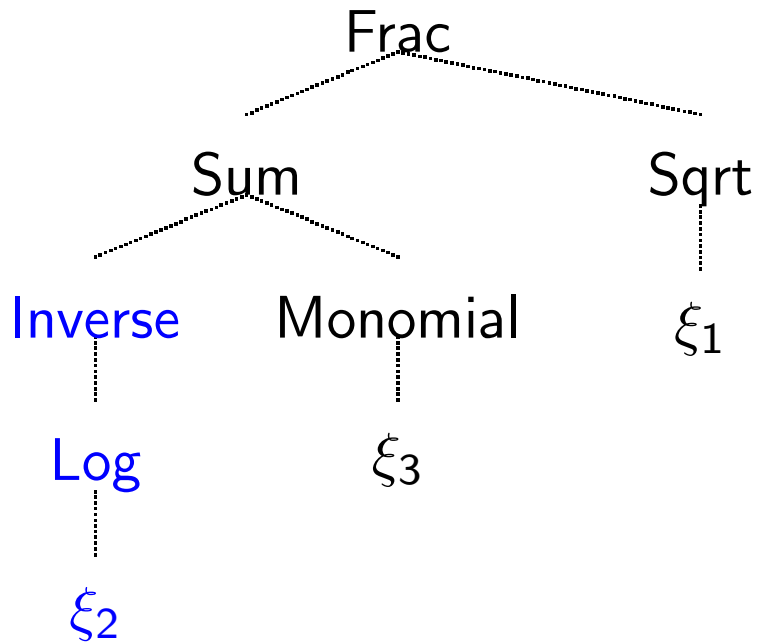
- **Frac**(Period of residence, Undeclared income)
- **Frac**(**Seg**(Period of employment), Term of contract)
- **And**(Income confirmation, Bank account)
- **Times**(**Seg**(Score hour), **Frac**(**Seg**(Period of employment), Salary))

Алгоритм случайного порождения признаков

1. Выбрать случайно узлы двух суперпозиций,
2. обменять соответствующие поддеревья,
3. изменить порождающую функцию на случайном узле.

Любые операции должны учитываю условия допустимости суперпозиций.





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

Связанный Байесовский вывод

$f_1, ..., f_M$ набор конкурирующих моделей,
$P(f_i|D)$ постериорная вероятность, $P(D|f_i)$ — правдоподобие

$$P(f_i|D) = \frac{P(D|f_i)P(f_i)}{\sum_{j=1}^{M} P(D|f_j)P(f_j)}. \quad (1)$$

Модели f_i и f_j сравниваются:

$$\frac{P(f_i|D)}{P(f_j|D)} = \frac{P(D|f_i)P(f_i)}{P(D|f_j)P(f_j)}.$$

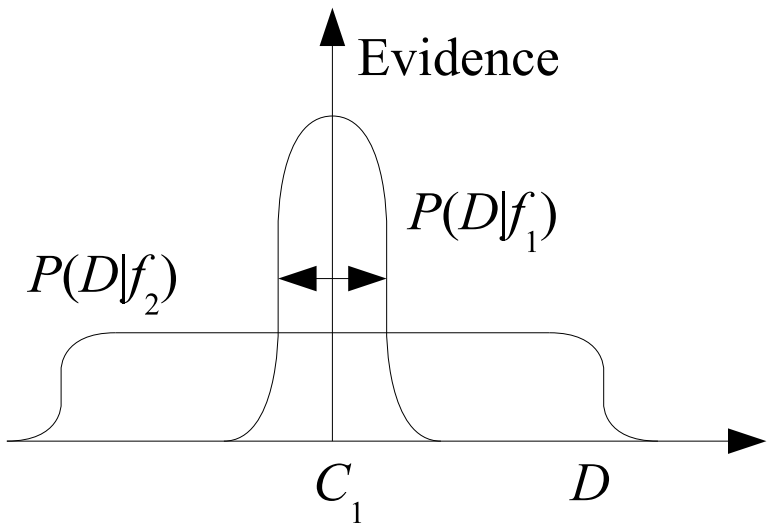
Апостериорное распределение $\mathbf{w}$ при заданных данных D

$$P(\mathbf{w}|D, f_i) = \frac{P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)}{P(D|f_i)}, \quad (2)$$

где правдоподобие модели определено выражением

$$P(D|f_i) = \int P(D|\mathbf{w}, f_i)P(\mathbf{w}|f_i)d\mathbf{w}.$$

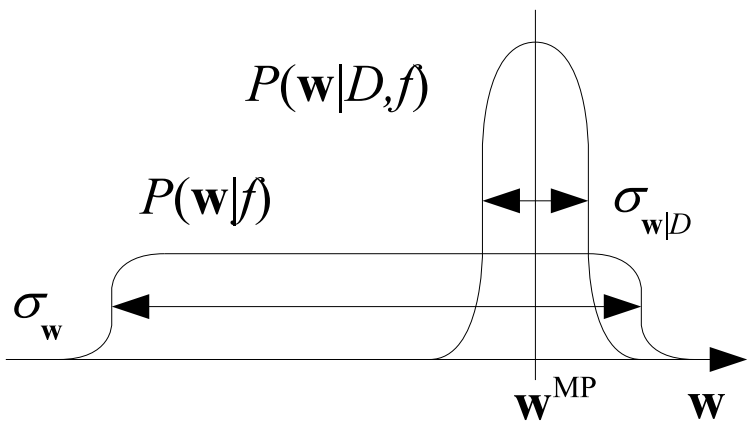
Сравнение моделей

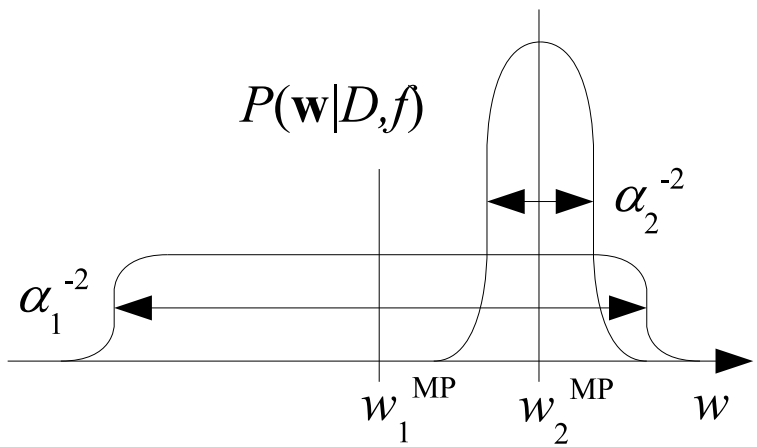


Сравнение признаков

Вектор параметров $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, A)$ — многомерная случайная величина. Рассмотрим ее ковариационную матрицу A, варианты:

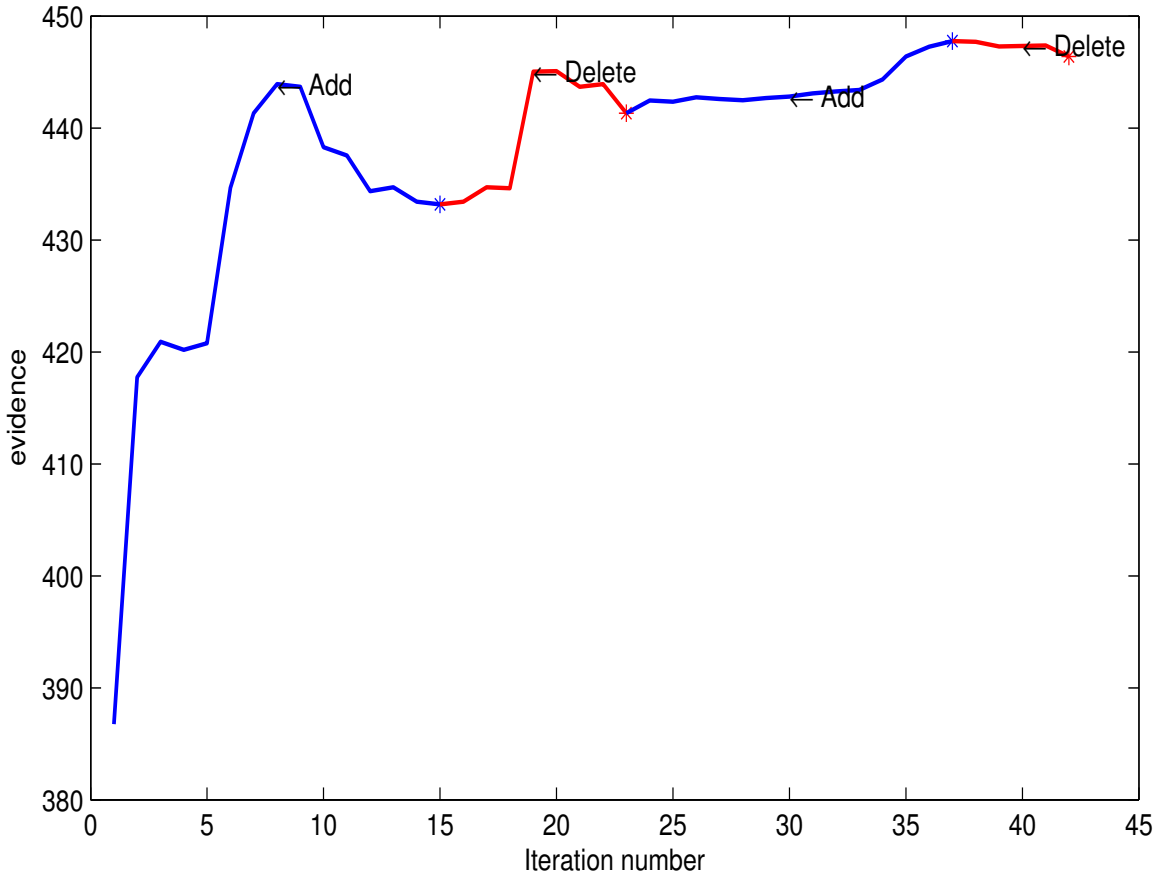
1. $A = \text{diag}(\alpha, \ldots, \alpha)$,
2. $A = \text{diag}(\alpha_1, \ldots, \alpha_W)$,
3. A — недиагональная.



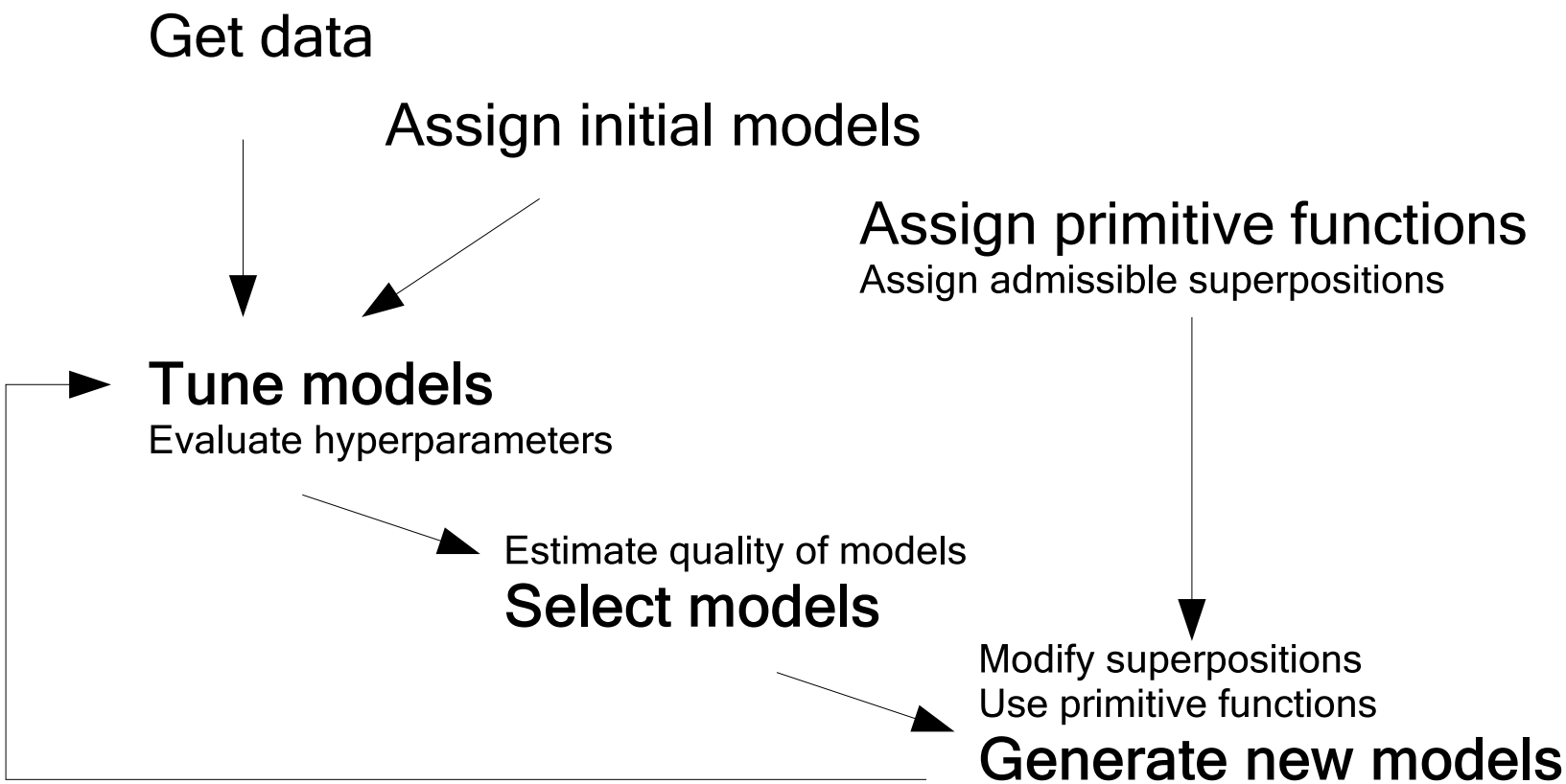


Правдоподобие модели $P(D|f_i) = \int P(D|\mathbf{w}, f_i) P(\mathbf{w}|f_i) d\mathbf{w}.$

Выбор наиболее правдоподобной модели



Процедура построения модели



Литература

Классика скоринга

- N. Siddiqi: Credit Risk Scorecards developing, 2004
- D. Hosmer, S. Lemeshov: Logistic Regression, 2000

Порождение и выбор моделей

- H. Madala: Group Method of Data Handling, 1995
- J. Koza, I. Zelinka: Genetic Programming, 2004
- Y. LeCun: Optimal Brain Surgery, 1985
- C. Bishop, J. Nabney: Model Selection and Coherent Bayesian Inference, 2004
- P. Grunwald: Minimum Description Length Principle, 2009

Полезные ссылки

Исследователи

- Lyn Thomas, School of Management University of Southampton
- David Hand, Imperial College London
- Christophe Mues, School of Management University of Southampton
- Bart Baesens, University of Southampton

Статьи

- Черкашенко В.Н. Этот загадочный Скоринг // Банковское дело, 2006, № 3. С. 42–48.
- Tony Bellotti and Jonathan Crook, Support vector machines for credit scoring and discovery of significant features

Стандарты

- Capital Requirements Directive/Basel 2
- Basel 2 о потребительском кредите: BIPRU 4.6

Открытые данные

Statlog (German Credit Data) Data Set by Dr. Hans Hofmann

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Categorical, Integer
- Associated Tasks: Classification
- Number of Instances: 1000
- Number of Attributes: 20