ПРИМЕНЕНИЕ МЕТОДОВ ПЛАНИРОВАНИЯ И ПРОВЕДЕНИЯ ЭКСПЕРИМЕНТОВ ДЛЯ ОПТИМИЗАЦИИ КАЧЕСТВА МАШИННОГО ОБУЧЕНИЯ ПУТЕМ ВЫБОРА ПАРАМЕТРОВ

М.В. Водолазкая,

О.Л. Моросин , к.т.н.

ФБГОУ ВПО «НИУ «МЭИ», Москва

Работа выполнена при поддержке гранта РФФИ 16-37-00309,
а также гранта Президента РФ МК-2897.2017.9

Санкт-Петербург, 2017

Машинное обучение

- Машинное обучение (*Machine Learning*) — обширное направление исследований в области искусственного интеллекта, изучающее методы построения алгоритмов, способных обучаться.
- Два типа обучения:
  - Дедуктивное обучение.
  - Индуктивное обучение.
- Также существует комбинированное обучение.

Параметры алгоритмов машинного обучения

- Эффективность решения задачи машинного обучения зависит от набора параметров используемого алгоритма машинного обучения.
- Перед исследователем ставится задача анализа влияния параметров на работу алгоритма и подбора оптимального набора параметров для выбранного алгоритма машинного обучения.
- Исследований в области выбора параметров крайне мало, и нет формализованного подхода к выбору набора параметров для получения оптимального решения.
- Как правило, требуется большое количество проб и ошибок, чтобы определить удачное сочетание параметров.
- Разработка метода, позволяющего формализовать подбор наилучших параметров, могла бы сильно упростить эту задачу.

Поиск оптимального набора параметров

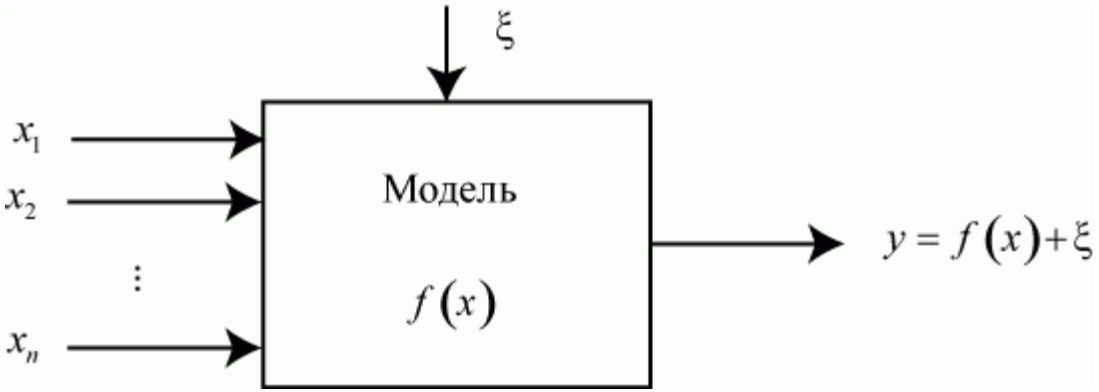
- Набор параметров, при котором достигается удовлетворяющий требованиям исследователя компромисс между точностью обучения и затратам по ресурсам, будем называть ***оптимальным***.
- Если оптимальный набор параметров для решения конкретной задачи существует, то искать его методом полного перебора было бы бессмысленно ввиду большого количества параметров и их комбинаций.
- Для решения подобной задачи уместно применить методы оптимизации и, так как часто обучение модели бывает весьма затратным по ресурсам, методы планирования эксперимента при поиске оптимальных условий.

Планирование эксперимента

- Средством достижения приемлемого компромисса между максимумом информации и минимумом затрат ресурсов является ***план эксперимента***.
- План эксперимента определяет:
  - объем вычислений;
  - порядок проведения вычислений;
  - способы накопления и статистической обработки результатов моделирования.
- Планирование экспериментов имеет следующие цели:
  - сокращение общего времени моделирования при соблюдении требований к точности и достоверности результатов;
  - увеличение информативности каждого наблюдения;
  - создание структурной основы процесса исследования.

Планирование эксперимента

- Математические методы планирования экспериментов основаны на так называемом кибернетическом представлении процесса проведения эксперимента.
- Объектами стратегического планирования являются:
  - выходные переменные (отклики);
  - входные переменные (факторы);
  - уровни факторов.



$x_i, i = \overline{1,n}$ — входные переменные, факторы;

$y = f(x) + \xi$ — выходная переменная (реакция, отклик);

ξ — ошибка, помеха, вызываемая наличием случайных факторов;

$f(x)$— оператор, моделирующий действие реальной системы, определяющий зависимость выходной переменной y от факторов x_i.

Иначе: $f(x)$ — модель процесса, протекающего в системе.

Планирование эксперимента

Планирование эксперимента преследует следующие цели:

1) стремление к минимизации общего числа опытов;

2) одновременное варьирование всеми переменными, определяющими процесс, по специальным правилам — алгоритмам;

3) использование математического аппарата, формализующего многие действия экспериментатора;

4) выбор четкой стратегии, позволяющей принимать обоснованные решения после каждой серии экспериментов.

Планирование эксперимента

Эксперимент, в котором реализуются все возможные сочетания уровней, называется полным факторным экспериментом. Условия эксперимента представляют в виде таблицы – матрицы планирования, где строки соответствуют различным опытам, а столбцы – значениям факторов. Матрица планирования 2^2 для двух факторов показана в таблице:

Номер опыта	Матрица планирования		Выход y
	x_1	x_2	
1	–1	–1	y_1
2	+1	–1	y_2
3	–1	+1	y_3
4	+1	+ 1	y_4

Построение модели

- Для движения к точке оптимума нам нужна линейная модель:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

- Наша цель — найти по результатам эксперимента значения неизвестных коэффициентов модели.
- Эксперимент, содержащий конечное число опытов, позволяет получить выборочные оценки для коэффициентов:

$$b_j = \frac{\sum_{i=1}^{N} x_{ji} y_i}{N}, j = 0, 1, \dots, k$$

- В силу свойства симметричности средние значения $\overline{x_1}$ и $\overline{x_2}$ равны 0. Следовательно, так как уравнение $y = b_0 + b_1 x_1 + b_2 x_2$ верно и для средних значений: $\overline{y} = b_0 + b_1 \overline{x_1} + b_2 \overline{x_2}$, то $\overline{y} = b_0$. Таким образом коэффициент b_0 является средним арифметическим значений параметра оптимизации.

Оценка смешанных взаимодействий

Номер опыта	x_0	x_1	x_2	$x_1 x_2$	y
1	+1	–1	–1	+1	y_1
2	+1	+1	–1	–1	y_2
3	+1	–1	+1	–1	y_3
4	+1	+1	+1	+1	y_4

$$в_i = \frac{\sum_1^N x_{iu}\hat{y}_u}{N};$$

$$в_{ij} = \frac{\sum_1^N x_{iu}x_{ju}\hat{y}_u}{N}; \quad i \neq j$$

$$в_{ijk} = \frac{\sum_1^N x_{iu}x_{ju}x_{ku}\hat{y}_u}{N}; \quad i \neq j \neq k$$

- Теперь модель выглядит следующим образом:

$$y = b_0 x_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2.$$

- Коэффициент b_{12} вычисляется обычным путем.
- Столбцы x_1 и x_2 задают планирование – по ним непосредственно определяются условия опытов, а столбцы x_0 и $x_1 x_2$ служат только для расчета.
- При оптимизации мы стремимся сделать эффекты взаимодействия возможно меньшими.

Проведение эксперимента

Имя атрибута	Что обозначает
survival	Спасение (0 = Нет; 1 = Да)
pclass	Класс пассажира (1 = 1*st*; 2 = 2*nd*; 3 = 3*rd*)
name	Имя
sex	Пол
age	Возраст
sibsp	Количество Братьев(Сестер)/Супругов на борту
parch	Количество Родителей/Детей на борту
ticket	Номер билета
fare	Пассажирский тариф
cabin	Каюта
embarked	Порт посадки (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)

Проведение эксперимента

- Объектом исследования является алгоритм *J48*. *J48* является java-имплементацией алгоритма *C4.5*.
- *C4.5* — алгоритм для построения деревьев решений. Разработанный Джоном Квинланом (англ. *John Ross Quinlan*). *C4.5* является усовершенствованной версией алгоритма *ID3* того же автора.
- Для исследования были выбраны следующие параметры алгоритма:
  - ***confidenceFactor*** — доверительный порог, который используется при обрезке деревьев (малые значения приводит к большему обрезанию);
  - ***minNumObj*** — минимальное количество объектов на листе;
  - ***binarySplits*** — использовать двоичное расщепление на номинальных признаках при построении деревьев.

Проведение эксперимента

Наименование и обозначение факторов	Уровни варьирования			Интервалы варьирования
	-1 Нижний уровень	0 Основной уровень	+1 Верхний уровень	
confidenceFactor	0.1	0.5	0.9	0.4
minNumObj	0	5	10	5
binarySplits	*false*	-	*True*	-

Номер опыта	Матрица планирования								Процент правильно классифицированных объектов, %	Среднеквадратичное отклонение
	x0	*x1*	*x2*	*x3*	*x1x2*	*x1x3*	*x2x3*	*x1x2x3*		
1	+	-	-	-	+	+	+	-	**81.5937**	0.3712
2	+	-	-	+	+	-	-	+	**81.0325**	0.3756
3	+	-	+	-	-	+	-	+	**82.2671**	0.3702
4	+	-	+	+	-	-	+	-	**81.3692**	0.3738
5	+	+	-	-	-	-	+	+	**79.5735**	0.4054
6	+	+	-	+	-	+	-	-	**79.5735**	0.4117
7	+	+	+	-	+	-	-	-	**80.4717**	0.3711
8	+	+	+	+	+	+	+	+	**79.4613**	0.3765

Построение и анализ модели

Проверим адекватность модели:

Однородность дисперсий по критерию Фишера:

$$F_{\text{эксп}} = \frac{s_{max}^2}{\sum_{i=1}^{N} s_i^2} = \frac{0.4117}{0.3819375} = 1.0779 < F_{\text{табл}} = 6.4$$

Следовательно, дисперсии однородны, и их можно усреднять.

Дисперсия воспроизводимости рассчитывается по формуле: $S_{\{y\}}^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n}(y_i - \bar{y})^2}{N(n-1)}$

Где y_i – результат ***i***-того эксперимента, $\bar{y}$ – среднее. Таким образом, $S_{\{y\}}^2 = \frac{\sum_{i=1}^{N} s_i^2}{N}$

где s_i^2 – среднеквадратичное отклонение. $\boldsymbol{S_{\{y\}}^2 = 0.381938}$

Вычислим коэффициенты регрессионной модели:

$$b_j = \frac{\sum_{i=1}^{N} x_{ji} y_i}{N}, j = 0, 1, \dots, k$$

b0	*b1*	*b2*	*b3*	*b12*	*b13*	*b23*	*b123*
80.667809	-0.897809	0.224513	-0.308686	-0.028012	0.056087	-0.168387	-0.084212

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$$

Номер опыта	1	2	3	4	5	6	7	8
$\hat{y}$	81.593689	81.032501	82.267090	81.369202	79.573502	79.573502	80.471695	79.461288
Δy_i	0.925880	0.364693	1.599281	0.701393	1.094307	1.094307	0.196114	1.206520
$\Delta y_i{}^2$	0.857255	0.133001	2.557701	0.491952	1.197508	1.197508	0.038461	1.455691

Дисперсия адекватности: $S^2_{\text{ад}} = \frac{\sum_{i=1}^{N} \Delta y_i}{f}$ Где f – число степеней свободы. $f = 4$. $\boldsymbol{S^2_{\text{ад}} = 1.982269}$

Для проверки используем F-критерий Фишера: $F_{\text{эксп}} = \frac{S^2_{\text{ад}}}{S^2_{\{y\}}} = 5.190035 < F_{\text{табл}} = 6.4$

Таким образом, построенная модель адекватна.

b0	*b1*	*b2*	*b3*	*b12*	*b13*	*b23*	*b123*
80.667809	-0.897809	0.224513	-0.308686	-0.028012	0.056087	-0.168387	-0.084212

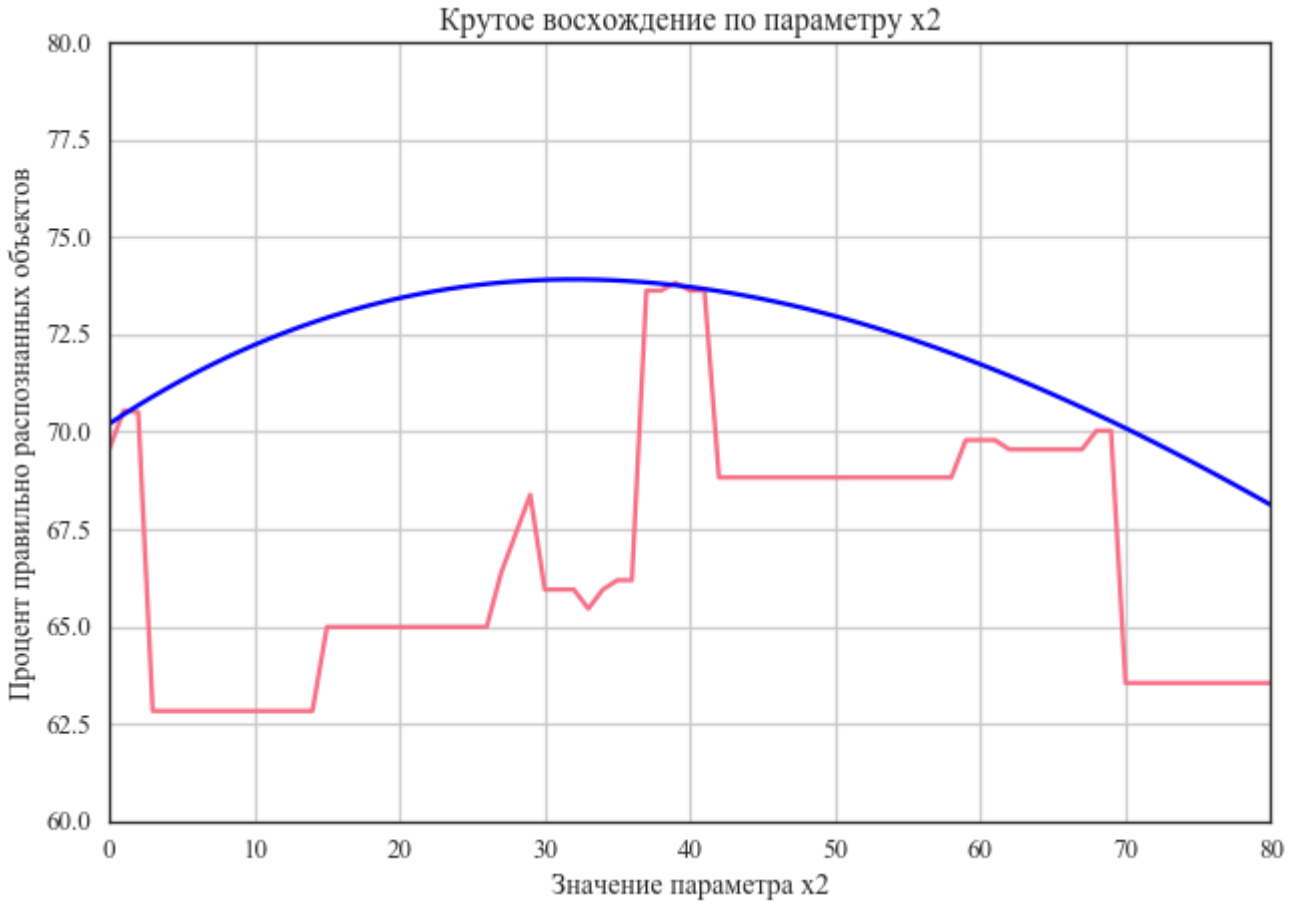
$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$$

С учетом значения дисперсии воспроизводимости $S^2_{\{y\}} = 0.381938$ с доверительной вероятностью $\alpha = 0.95$ находим границы доверительных интервалов для коэффициентов регрессии:

- Дисперсия коэффициентов регрессии: $S^2_{\{b_i\}} = \frac{S^2_{\{y\}}}{N} = 0.04774225$
- Доверительный интервал $\Delta b_j = \pm t \cdot S_{b_i} = 0.606993$
- при $t = 2.778$ – табличное значение критерия Стьюдента при количестве степеней свободы $f = 4$ и $\alpha = 0.05$

Крутое восхождение по параметру *x2*

Значение фактора x2	Процент правильно распознанных объектов, %	СКО
0	69,5444	0,5302
2	**70,5036**	0,5302
3	62,8297	0,5154
8	64,2686	0,5022
10	62,8297	0,5154
18	64,988	0,5
30	65,9472	0,5033
40	**73,6211**	0,4985
43	68,8249	0,4935
50	68,8249	0,4935



Проверка метода на задаче Kaggle

Проверка точности предсказания на платформе Kaggle показала, что используемый метод позволяет существенно улучшить качество предсказания (с 66% до 79%).

Submission	Files	Public Score
Sat, 28 May 2016 12:13:17 j48 -c 0.5 -m 2	predictions2.csv	0.66507

Submission	Files	Public Score
Thu, 16 Jun 2016 13:33:23 j48 -c 0.1 -m 2	predictions.csv	0.78947
Thu, 16 Jun 2016 15:14:51 j48 -c 0.1 -m 40	predictions3.csv	0.77033

Полученные результаты

- В результате выполнения полнофакторного эксперимента построена адекватная регрессионная модель, описывающая поведение параметра оптимизации в зависимости от значений факторов. Параметром оптимизации выбран процент правильно распознанных объектов, а факторами — некоторые параметры алгоритма.
- Показано, что линейная модель, несмотря на её недостаточность для описания подобных процессов, позволяет добиться повышения эффективности работы алгоритмов машинного обучения.
- Показано, что варьирование параметрами алгоритма машинного обучения по методу, описанному в данной работе, позволяет добиться существенного улучшения качества работы алгоритма машинного обучения.

Спасибо за внимание!