

Машинное обучение: один эксперимент

И. Куралёнок, Н. Поваров

Яндекс

СПб, 2013

Задача на сегодня

Задача: отделить “хороших” студентов от “плохих”

Формально: предсказать средний балл следующей сессии

План работы

- 1 Определиться с тем кто такой студент
- 2 Как из этого определения можно понять хороший он или нет
- 3 Выработать чувство прекрасного
- 4 Решить полученную задачу оптимизации
- 5 Оценить качество полученного решения
- 6 Проанализировать результаты

Векторизация¹

Векторизация: перевод представления о предмете в векторное выражение.

Компоненты полученного в результате векторизации вектора будем называть **факторами**

¹В данном случае не про графику.

Факторы про студента

- Пол
- Средний школьный балл
- Школа номер
- Школа город
- Доля пропущенных лекций
- Оценка по мнению родителей
- Пиво/неделя
- Друзей в ОК/FB/ВК
- Расстояние от дома до универа
- Ряд в аудитории
- Наличие планшета
- Периметр головы

Множество на котором обучаемся L

Пол	СШБ	Город	Тип Шк	Проп.	Род.	Пиво	Друзей D	Ряд	iPad	Голова	Баллы
1	4.5	Красноярск	Гимназия	0.70	5	0	300	2	1	58	19
0	5	Киров	СОШ	0.00	-1	0	70	4.2	2	61	19
1	3	СПб	Лицей	0.09	4	1	300	7	1	57	12
1	4	СПб	Лицей	0.30	3	30	80	5.5	4	56	15
0	5	СПб	Гимназия	0.05	5	0	60	2.5	2	61	17
0	5	Уфа	Лицей	0.14	2	0	45	2	1	60	19
1	3	Иркутск	Гимназия	1.00	4	10	120	2	3	57	16
0	4.8	Рыбинск	СОШ	0.00	4	0	130	2	1	60	16
0	3.5	СПб	Лицей	1.00	-1	0	250	7	4	50	14
1	4.2	СПб	Гимназия	0.14	-1	0.5	800	3	3	54	18
1	4	СПб	Лицей	0.00	4	0	45	2.5	2	55	12
1	5	Сыктывкар	Лицей	0.86	-1	0	100	2	2	59	18
1	5	Норильск	Лицей	0.09	-1	0	208	2.5	2	57	16
0	4.3	Череповец	Лицей	0.05	4	0	200	2	2	54	19
0	5	Нефтеюганск	СОШ	0.05	4.5	0	50	2	1	57	15
1	5	СПб	Лицей	0.45	5	0	25	5.5	2	59	19
1	5	СПб	СОШ	0.00	5	0	65	5	1	61	18
1	4	Ульяновск	Лицей	0.00	5	0	85	2	4	61	20
0	5	Иркутск	Лицей	0.05	5	0	60	2	2	55	19
0	4.5	Ангарск	Лицей	0.00	4	0	46	2	3	52	19
1	4.3	Н.Новгород	Лицей	0.00	-1	0	30	3.5	4	58	20
1	4.3	СПб	СОШ	0.05	-1	0	60	4.5	2	58	20
1	4.5	СПб	Гимназия	0.09	-1	2	181	7.5	2	57	19
1	4	Костомай	Лицей	0.27	5	0	80	2	2	56	16
1	3.3	Ангарск	СОШ	0.21	4.5	2	220	2	4	57	19
1	4.7	Киров	Лицей	0.09	3	0	140	2	3	59	12
0	4.3	Пушкин	СОШ	0.05	-1	0	99	2	4	55	16
1	3.8	СПб	Лицей	0.09	-1	0	100	4.8	1	58	15

Решающая функция

$$h_{w_0}(x) = w_0^T x$$

$$w_0, x \in \mathbb{R}^n$$

- Простая
- Универсальная
- Легко интерпретируемая

Адаптация факторов

Не все факторы подходят для использования в выбранной модели:

- Город – плохой фактор, не из \mathbb{R} , \Rightarrow местные/неместные $\in \{0, 1\}$
- Тип школы \Rightarrow
 - 1 Школа = Лицей/Гимназия $\in \{0, 1\}$
 - 2 Школа = 239 $\in \{0, 1\}$
- Оценка по мнению родителей – есть пропущенные значения, \Rightarrow хуже/неизвестно/лучше чем было в школе $\in \{-1, 0, 1\}$
- Планшеты есть у всех \Rightarrow выкинули

Итоговый вектор факторов

- | | | | |
|----|----------------------------|----|-------------------------------|
| 1 | Пол | 2 | Средний школьный балл |
| 3 | школа = 239 | 4 | Понаехали |
| 5 | Гимназия/Лицей | 6 | Доля пропущенных лекций |
| 7 | Оценка по мнению родителей | 8 | Пиво/неделя |
| 9 | Друзей в ОК/FB/ВК | 10 | Расстояние от дома до универа |
| 11 | Ряд в аудитории | 12 | Периметр головы |

Решающая функция

$$h_{w_0, b_0}(x) = w_0^T x + b_0$$

$$w_0, x \in \mathbb{R}^n, b_0 \in \mathbb{R}$$

- Простая
- Универсальная
- Легко интерпретируемая

Целевая функция

$$(w_0, b_0) = \arg \min_{(w,b)} \sum_{(x_i, y_i) \in L} \|h_{w,b}(x_i) - y_i\|$$

- MSE — хорошее первое приближение в \mathbb{R}
- Интерпретируемое значение \min

Решение и его интерпретация

$$\begin{aligned}(w_0, b_0) &= \arg \min_w \sum_{(x_i, y_i) \in L} \|x_i^T w + b - y_i\| \\ &= \arg \min_w \|X^T w + b - y\|^2 \\ &\begin{cases} w_0 = (XX^T)^{-1} X(y - 1b) \\ b_0 = \frac{1}{n} \sum_{i=1}^n y_i \end{cases}\end{aligned}$$

Вот что получилось:

$$\begin{aligned}w_0 &= -0.24, 1.17, -5.82, 0.32, 0.30, -2.33, 1.77, -0.01, -0.00, 0.13, 0.03, -0.08 \\ b_0 &= 17.04 \\ T &= 83.06 \\ \sqrt{\frac{T(w_0, b_0)}{n}} &= 1.72\end{aligned}$$

Интерпретация результата

Обещали, что результат хорошо интерпретируется.
Влияние факторов:

Школа = 239 (3), вес = -5.82 Вот такая вот выборка :)

Доля пропущенных лекций (6), вес = -2.33 на лекции надо ходить

Оценка по мнению родителей (7), вес = 1.77 честность – это хорошо :)

Интерпретация результата

Обещали, что результат хорошо интерпретируется.
Влияние факторов:

Школа = 239 (3), вес = -5.82 вот такая вот выборка :)

Доля пропущенных лекций (6), вес = -2.33 на лекции надо ходить

Оценка по мнению родителей (7), вес = 1.77 честность – это хорошо :)

Все это фигня!

Нормализация факторов

Хотим простого: 0 мат. ожидание и 1 дисперсию.

$$y_i = \frac{x_i - m_i}{\sigma_i}$$

*Это стандартная техника, но немного (2 слайда)
подумайте насколько она нам подходит.*

$$\begin{aligned}w_0 &= -0.16, 0.74, -1.47, 0.19, 0.16, -0.67, 1.19, -0.05, -0.02, 0.27, 0.11, -0.08 \\b_0 &= 17.04 \\T(w_0, b_0) &= 82.72 \\ \sqrt{\frac{T(w_0, b_0)}{n}} &= 1.72\end{aligned}$$

Интерпретация результата II

Обещали, что результат хорошо интерпретируется.
Влияние факторов:

Школа = 239 (3), вес = -1.47 таки да :)

Оценка по мнению родителей (7), вес = 1.19 честность еще важнее!

Средний школьный балл (2), вес = 0.74 багаж знаний – определяет (догадайтесь какой курс?)

Доля пропущенных лекций (6), вес = -0.67 опустилась вниз

Интерпретация результата II

Обещали, что результат хорошо интерпретируется.
Влияние факторов:

Школа = 239 (3), вес = -1.47 Таки да :)

Оценка по мнению родителей (7), вес = 1.19 честность еще важнее!

Средний школьный балл (2), вес = 0.74 багаж знаний – определяет (догадайтесь какой курс?)

Доля пропущенных лекций (6), вес = -0.67 опустилась вниз

Но и это фигня!

Хорош ли полученный результат

- Отличается ли от КО
- Сколько можно выжать из данных
- Можно ли верить его компонентам
- Воспроизводится ли результат
- Насколько можно доверять предсказанию

По невязке

В нашем случае невязка – целевая функция

$$T(w_0, b_0) = \sum_{(x_i, y_i) \in L} \|h_{w_0, b_0}(x_i) - y_i\|$$

- Просто посчитать
- Рассказывает о работе на тренировочном множестве
- **Ничего не говорит о качестве предсказания**

По невязке на другом множестве I

Можно поделить множество на 2 части, настроить на одной, проверить на другой

$$DS = L \cup T, L \cap T = \emptyset$$

$$(w_0, b_0) = \arg \min_{(w, b)} \sum_{(x_i, y_i) \in L} \|x_i^T w + b - y_i\|$$
$$T_T = \sum_{(x_i, y_i) \in T} \|h_{w_0}(x_i) - y_i\|$$

По невязке на другом множестве II

- Расскажет о качестве предсказания с точностью до деления на L и T
- Использует меньше данных в обучении
- Если исходное множество не показательное, то деление нас не спасет
- Можно посмотреть на T_L и T_T

По стабильности решения

Поделим несколько раз и посмотрим на то как меняется (w_0, b_0) .

- Стабильные компоненты заслуживают веры
- Если все нестабильно — беда-беда

Немного результатов I

$$\begin{aligned} T_L &= 6.82 & T_T &= 1401.87 \\ \sqrt{\frac{T_L(w_0, b_0)}{n}} &= 0.70 & \sqrt{\frac{T_T(w_0, b_0)}{n}} &= 10.01 \\ w_0 &= (3.12, -2.39, 0.13, 1.94, -1.09, 0.53, -7.56, -0.04, 0.95, 2.27, 3.23, 3.80) & b_0 &= 17.07 \end{aligned}$$

$$\begin{aligned} T_L &= 0.00 & T_T &= 2127.48 \\ \sqrt{\frac{T_L(w_0, b_0)}{n}} &= 0.00 & \sqrt{\frac{T_T(w_0, b_0)}{n}} &= 11.19 \\ w_0 &= (-4.49, -6.03, 0.43, -2.16, 0.84, 0.90, -0.91, -0.81, -4.45, 1.98, 1.53, 1.53) & b_0 &= 18.27 \end{aligned}$$

$$\begin{aligned} T_L &= 20.45 & T_T &= 305.87 \\ \sqrt{\frac{T_L(w_0, b_0)}{n}} &= 1.17 & \sqrt{\frac{T_T(w_0, b_0)}{n}} &= 4.85 \\ w_0 &= (0.04, 2.34, -0.33, -0.67, -0.62, -0.65, 2.38, 2.02, -1.34, 0.32, 1.45, -0.34) & b_0 &= 16.87 \end{aligned}$$

Немного результатов II

А если так делать много раз то можно получить оценки:

w_{01}	-0.87 ± 11.45
w_{02}	-0.27 ± 9.28
w_{03}	-1.34 ± 16.78
w_{04}	1.06 ± 17.27
w_{05}	0.46 ± 3.82
w_{06}	0.03 ± 13.75
w_{07}	0.94 ± 6.32
w_{08}	0.08 ± 19.94
w_{09}	-0.94 ± 18.65
w_{010}	0.97 ± 13.10
w_{011}	0.02 ± 6.97
w_{012}	0.57 ± 13.07
b_0	17.04 ± 0.47
<i>Mean error</i>	13.26 ± 46.89

Можно ли сделать лучше?

- Мало данных или много факторов?
 - Все ли факторы одинаково хороши?
 - Может их можно скомбинировать?
 - Стоит ли одинаково верить всем факторам?
- Может быть в данных что-то нечисто?
 - Все ли мы можем объяснить?
 - А набирали данные правильно?
 - Не подсматриваем ли мы в ответ?
 - Все ли важные примеры представлены в данных и репрезентативно ли это представление?

Не все факторы одинаково полезны

- Можем ли мы обойтись без какого-нибудь фактора?
- А если фактор преобразовать, может его станет проще использовать?
- Если есть похожие факторы, наверное это можно учесть.
- Стоит ли рассмотреть комбинации нескольких факторов?
- Что мы делаем, если фактор посчитать нельзя?

Оценка полезности фактора

Данных мало, моделька линейная... *Brute force it!*
Будем по одному выкидывать факторы и смотреть как поведет себя ошибка:

Фактор	Ошибка если выбросить	Изменение ошибки
Пол	11.14	-2.12
Средний школьный балл школа = 239	17.74	4.48
Понаехали	11.74	-1.52
Гимназия/Лицей	11.03	-2.23
Доля пропущенных лекций	13.07	-0.19
Оценка по мнению родителей	12.00	-1.26
Пиво/неделя	13.90	0.64
Друзей в ОК/FB/ВК	7.67	-5.59
Расстояние от дома до универа	10.19	-3.07
Ряд в аудитории	10.58	-2.68
Периметр головы	10.03	-3.23
	18.46	5.2

Немного результатов III

Оценим параметры модели ещё раз:

w_{02}	1.24 ± 0.67
w_{07}	0.90 ± 0.69
w_{012}	-0.02 ± 0.60
b_0	17.03 ± 0.49
<i>Mean error</i> :	2.72 ± 0.42

Оценка полезности студента

Студент	Ошибка	Изменение ошибки
Student 1	2.75	0.03
Student 2	2.69	-0.03
Student 3	2.43	-0.29
Student 4	2.51	-0.21
Student 5	2.87	0.15
Student 6	2.62	-0.10
Student 7	2.31	-0.41
Student 8	2.65	-0.07
Student 9	2.34	-0.38
Student 10	2.42	-0.30
Student 11	2.10	-0.62
Student 12	2.36	-0.36
Student 13	2.18	-0.54
Student 14	2.61	-0.11
Student 15	2.45	-0.27
Student 16	2.54	-0.18
Student 17	2.53	-0.19
Student 18	2.87	0.15
Student 19	2.48	-0.24
Student 20	2.51	-0.21
Student 21	2.83	0.11
Student 22	2.86	0.14
Student 23	2.48	-0.24
Student 24	2.52	-0.20
Student 25	2.52	-0.20
Student 26	2.59	-0.13
Student 27	2.85	0.13
Student 28	2.91	0.19

Итоговая модель

Обещали, что результат хорошо интерпретируется.

$$\begin{aligned} & \text{Оценка за экзамен} \\ & = \\ & 1.24 \times \text{Средний школьный балл} \\ & + \\ & 0.90 \times \text{Мнение родителей} \\ & - \\ & 0.02 \times \text{Периметр головы} \\ & + \\ & 17.03 \end{aligned}$$