Метрики и генерация фич

Семинар DMIA 2016

План семинара

- Оптимизация метрик
- Работа с признаками
  - Преобразования целевой переменной
  - Категориальные признаки
  - Вещественные признаки
  - Переменное количество признаков
- Отбор признаков

Оптимизация метрик

Постановка задачи в соревновании - достичь максимального качества

#	Δrank	Team Name ‡ model uploaded * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	Go Polar Bears ‡ *	1.000000	49	Mon, 12 Oct 2015 22:57:38
2	↑1	Alexander Gramolin ‡ *	0.999998	12	Mon, 12 Oct 2015 18:38:07
3	↓1	Josef Slavicek ‡ *	0.999897	25	Mon, 12 Oct 2015 21:49:53
4	–	Michal Wojcik	0.999225	35	Mon, 12 Oct 2015 23:57:46 (-3h)
5	–	rakhlin	0.998338	31	Mon, 12 Oct 2015 23:32:18 (-5.8h)
6	–	Archy ‡	0.997784	47	Mon, 12 Oct 2015 20:31:53 (-7.8h)
7	–	Faron	0.995918	66	Mon, 12 Oct 2015 18:15:46
8	–	Alejandro Mosquera	0.994946	28	Mon, 12 Oct 2015 15:23:51 (-19.7h)

Поиск наилучшего константного решения

$$\frac{1}{n}\sum(y-a)^2 \to \min_a$$

$$\frac{\partial L}{\partial a} = \frac{2}{n}(y-a) = 0$$

$$a = \overline{y}$$

Сведение к известным метрикам

$$RMSLE = \sqrt{\frac{1}{n}\sum(log(1+y) - log(1+a(x)))^2}$$

$$\tilde{y} = log(1+y)$$

$$RMSLE = \sqrt{\frac{1}{n}\sum(\tilde{y} - \tilde{a}(x))^2}$$

Сведение к известным метрикам

$$RMSPE = \sqrt{\frac{1}{n}\sum_{y>0}\left(\frac{a-y}{y}\right)^2}$$

$\tilde{y} = ?$

Сведение к известным метрикам

$$RMSPE = \sqrt{\frac{1}{n}\sum_{y>0}\left(\frac{a-y}{y}\right)^2}$$

$$\tilde{y} = ln(y)$$

$$RMSE = \frac{1}{n}\sum(F(a) - F(y))^2$$

$$\frac{a-y}{y} \approx F(a) - F(y)$$

$$a = y + \delta$$

$$\frac{\delta}{y} \approx F(y+\delta) - F(y) \approx F'\delta$$

$$\frac{1}{y} = \frac{\partial F}{\partial y}$$

$$F(x) = \ln(x)$$

Оптимальное смешивание моделей

Как смешивать если метрика - AUC?

Оптимальное смешивание моделей

Как смешивать если метрика - AUC?

AUC оценивает порядок элементов

Надо смешивать порядки элементов!

preds = a * rankdata(y1) + (1-a) * rankdata(y2)

Генерация и отбор признаков

- Преобразование целевой переменной
- Категориальные признаки
- Вещественные признаки
- Переменное количество признаков

Преобразование целевой переменной

В случае регрессии!

$$\tilde{y} = F(y)$$
$$model.fit(X, \tilde{y})$$
$$preds = model.predict(X_{test})$$
$$submission = F^{-1}(preds)$$

Преобразования:

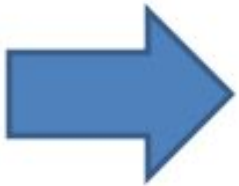
- Логарифм
- Степенное - $\tilde{y} = y^{\alpha}$

Категориальные признаки

- Бинаризация по категориям (one-hot encoding)
- Кодирование другим признаком
- Кодирование целевой переменной
- Hashing trick

Бинаризация по категориям (one-hot encoding)

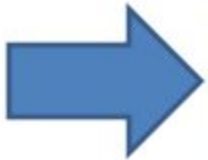
Страна
2
2
1
3



Страна=1	Страна=2	Страна=3
0	1	0
0	1	0
1	0	0
0	0	1

Кодирование другим признаком

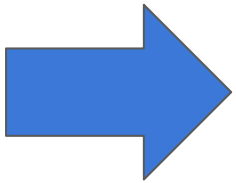
Страна	Доход
2	2,000
2	20,000
1	10,000
3	12,000
2	2,000



Страна	Доход	Средний доход по стране
2	2,000	8,000
2	20,000	8,000
1	10,000	10,000
3	12,000	12,000
2	2,000	8,000

Кодирование целевым признаком

Страна	Невозврат кредита
2	1
2	0
1	1
3	0
2	1



Страна	Невозврат кредита	Доля невозвратов
2	1	0.66
2	0	0.66
1	1	1
3	0	0
2	1	0.66

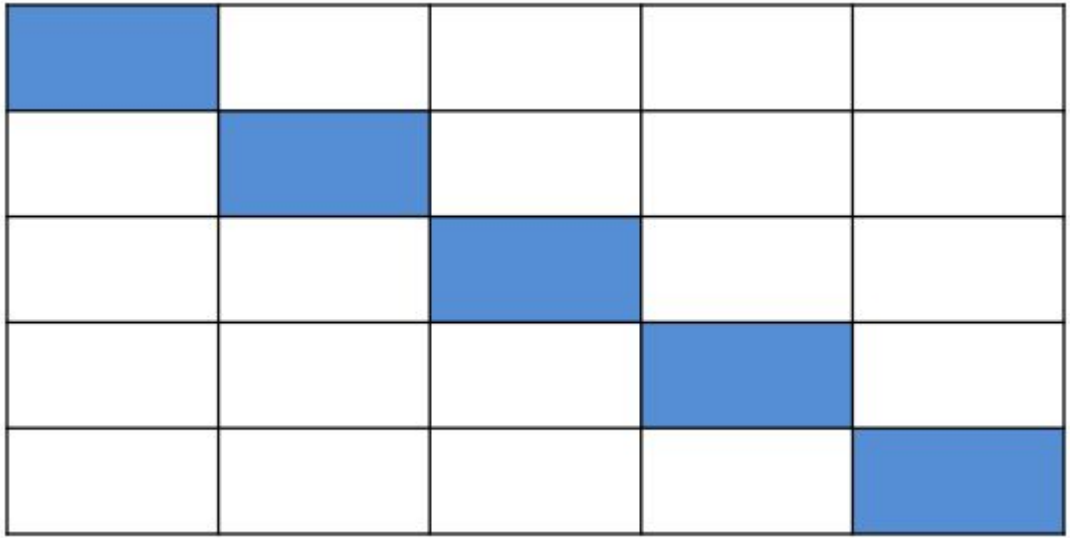
Утечка в данных (Data Leak)

Значение целевой переменной неявно содержится в признаках

Модель обучается использовать эту информацию

Закономерность неверна для теста!

Out-of-Fold кодирование



Сглаживание целевой переменной

Когда категорий очень много!

1. $$\frac{mean_{group} * size_{group} + mean_{global} * C}{size_{group} + C}$$

2. $$mean_{global} + \frac{2}{\pi} \cdot (arctan \ln group_{size}) \cdot (mean_{group} - mean_{global})$$

Hashing trick

features[hash(string) % bin_count] += 1

Страна
2
2
1
3
2

Hashing trick

string = feature_name + ‘_’ + feature_value

string = country_2

Вещественные признаки

Линейные методы и регуляризация

$$L = \frac{1}{n}\sum(y - w \cdot x)^2 + \frac{1}{2}||w||_2^2$$

Нужно приводить признаки к одной шкале!

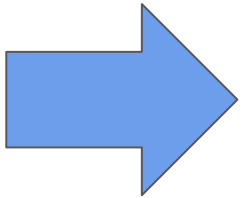
```
from sklearn.preprocessing import scale
X = scale(X)
```

Вещественные признаки

Линейные методы и однородные признаки

Однородности можно добиться превращением признаков в ранки

Год	Зарплата
2014	32,232
2015	25,123
2016	40,576
2017	67,984

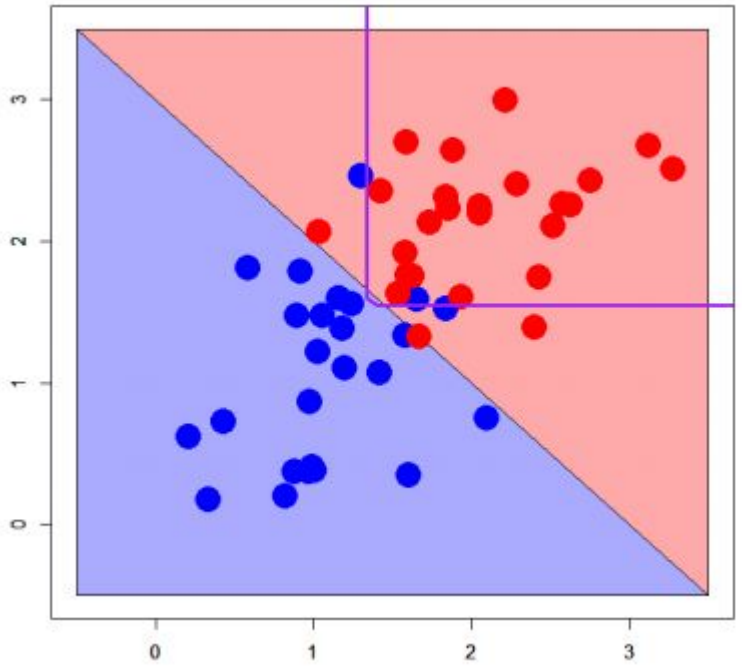


Год	Зарплата
1	2
2	1
3	3
4	4

Вещественные признаки

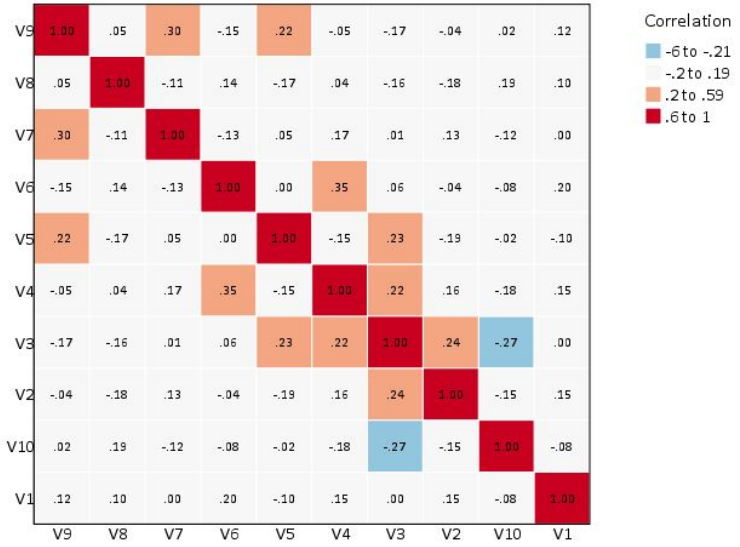
Деревья и линейные комбинации признаков

Деревья плохо моделируют сложение\вычитание (и другие ариф. операции) признаков



Вещественные признаки: как выбрать какие признаки складывать\вычитать?

- По смыслу задачи
- По корреляции между признаками
- Если ничего не помогает - все подряд + отбор признаков



Переменное количество строк/признаков

Покупатель	Магазин	Категория	Количество
CONS1	SHOP1	FOOD	2
CONS1	SHOP2	CLOTHS	1
CONS2	SHOP1	DEVICES	1
CONS3	SHOP1	FOOD	2
CONS3	SHOP1	CLOTHS	3
CONS3	SHOP1	DEVICES	11

Переменное количество признаков

Покупатель	FOOD	DEVICES	CLOTHS	SHOP1	SHOP2
CONS1	2		1	1	1
CONS2		1		1	0
CONS3	2	11	3	1	0

Отбор признаков

- Feature importance из деревьев
  - rf.feature_importance_
- Корреляция\взаимная информация признака и таргета
  - scipy.stats.pearsonr
  - sklearn.metrics.normalized_mutual_info_score
- Последовательное ручное добавление