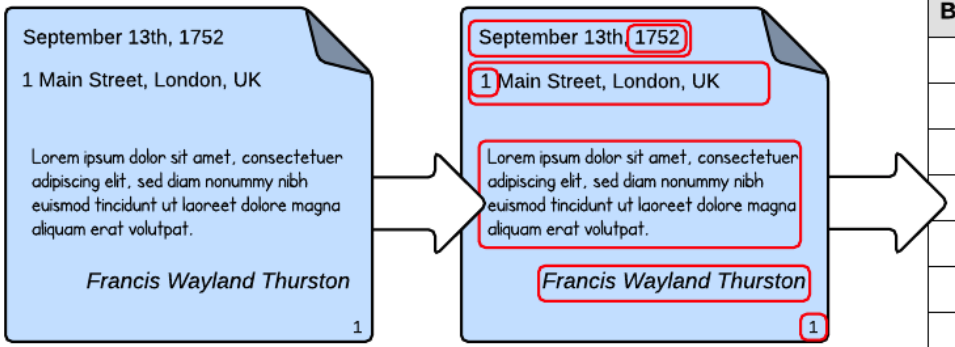
Tradeshift Text Classification

Гущин Александр

Text Classification

Выделение фич из документа



Box nr.	Box text
1	September 13th, 1752
2	1752
3	1
4	1 Main Street, London, UK
5	Lorem ipsum dolor sit amet, consectetuer adipiscing ...
6	Francis Wayland Thurston
7	1

Выборки

Типы переменных:

1. хэши
2. boolean
3. числовые

Необходимо предсказать много классов (какая между ними зависимость?):

train.csv

id	x1	x2	x3	x4	x5	x6
1	m268i97y	0	NO	105.4	14	
2	j0gheu6	1	YES	25.631	12	
3	26fmap6u	1	NO	12.0	12	m268i97y
4	of64nasl	0	NO	140.12	14	m268i97y
5	13e5dbzp	0	NO	150.92	40	of64nasl
6	8n4t73wy	0	YES	135.01	14	13e5dbzp
7	26fmap6u	1	YES	10.53	10	8n4t73wy

trainLabels.csv

id	y1	y2	y3	y4
1	1	0	0	0
2	0	1	0	1
3	0	0	0	1
4	0	1	0	0
5	0	0	1	1
6	0	0	1	0
7	0	0	0	1

Logloss

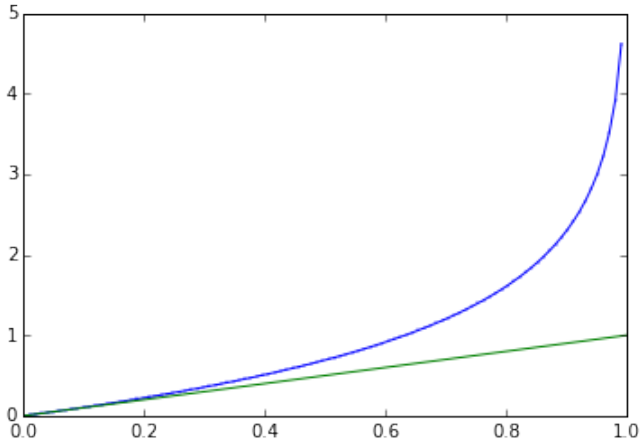
$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

Выгодней сделать много незначительно отличающихся от истины предсказаний, чем мало, отличающихся значительно

Пример:

усредняя два сабмишена с ошибками .0074 и .0069 после подбора коэфициентов линейной комбинации получаем .0056

Вывод: Нужно смешивать разные решения.



Мета-уровни

Необходимо использовать хеш-переменные. Один из подходов:

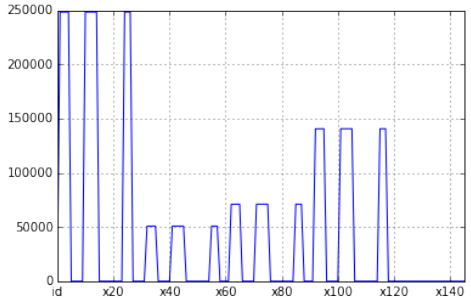
1. Разобъём обучающую выборку (X,Y) на две части - base и meta

2. Первый уровень - учимся на base (X,Y), предсказываем Y2 для meta и test
   a. Хеш-переменные - в спарс-матрицы
      i. LinearSVC
      ii. libFM
   b. Числовые переменные
      i. randomForest
      ii. XGBoost

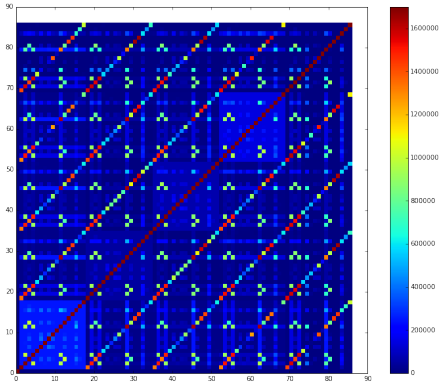
3. Второй уровень - учимся на meta (Y2 x 4,Y), для test предсказываем Y по Y2
   a. randomForest
   b. XGBoost

Фичи

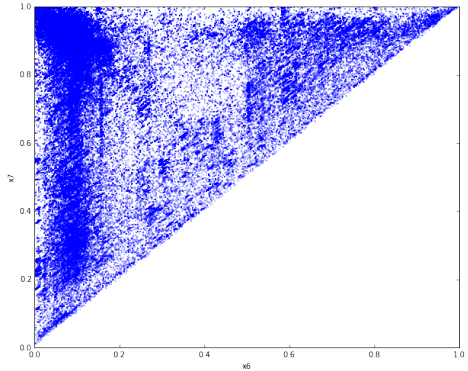
Переменные явно сгруппированы!



Как можно создать новые переменные? Например, вычисляя разницу между одинаковыми переменными в разных группах



Или можно посмотреть картинки (feature1, feature2) на фичах в одной группе

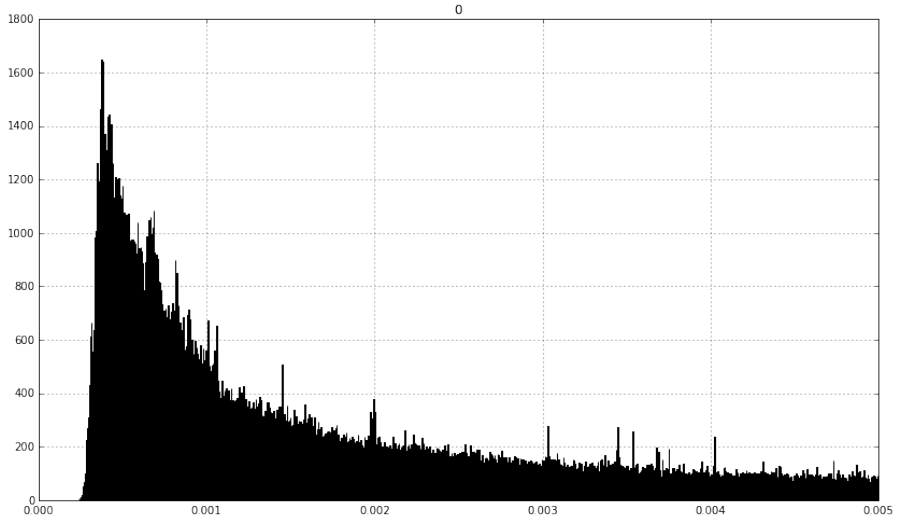


Особенности задачи (1)

Всего 144 уникальных строк y1:y33

Можно “подгонять”, находить ближайшие ответы и округлять.

Для более чем 50% ответов ближайшие к ним находятся на расстоянии менее 0.005 по эвклидовой метрике.



Особенности задачи (2)

Некоторые Y приносят большую ошибку

y33 - около 25% всей ошибки

y6, y9, y12 - около 15% всей ошибки

Подход:

1. решать отдельную задачу для каждой из этих переменных (y33)

2. решать отдельную задачу для группы переменных (y6, y9, y12) - можно заметить, что значения этих переменных часто совпадают