Data Mining in Action

Лекция 4
Отбор и генерация признаков

Гущин Александр

Отбор признаков

1. Статистические методы
2. С помощью регуляризации L1
3. Жадный отбор (Add-del)
4. С помощью деревьев (Boruta)

Регуляризация

Могут привести к переобучению

1. В линейных моделях, нейронных сетях: слишком большие веса

2. В решающих деревьях: слишком глубокие разбиения

Пример

x_1	x_2	y
2	-1	3
4	-2	6
0	0	0

$$y_1 = -3x_2$$

$$y_2 = \frac{3}{2}x_1$$

$$y = y_1(1+b) - y_2 \cdot b$$

$$L_1: \left|-3(1+b)\right| + \left|-\frac{3}{2}b\right| \to \min$$

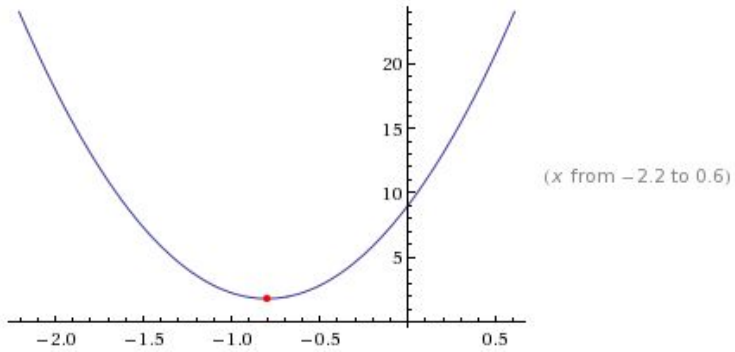
$$L_2: \left(-3(1+b)\right)^2 + \left(-\frac{3}{2}b\right)^2 \to \min$$

Пример

Global minimum:

$$\min\left\{\left(\frac{3\,b}{2}\right)^2 + (3+3\,b)^2\right\} = \frac{9}{5} \text{ at } b = -\frac{4}{5}$$

Plot:

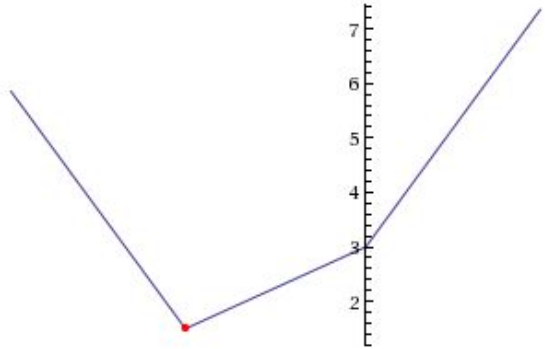


(x from -2.2 to 0.6)

L2: y = x1 * 6/5 - 3/5 * x2

$$\min\left\{\left|\frac{3\,b}{2}\right| + |3+3\,b|\right\} = \frac{3}{2} \text{ at } b = -1$$

Plot:

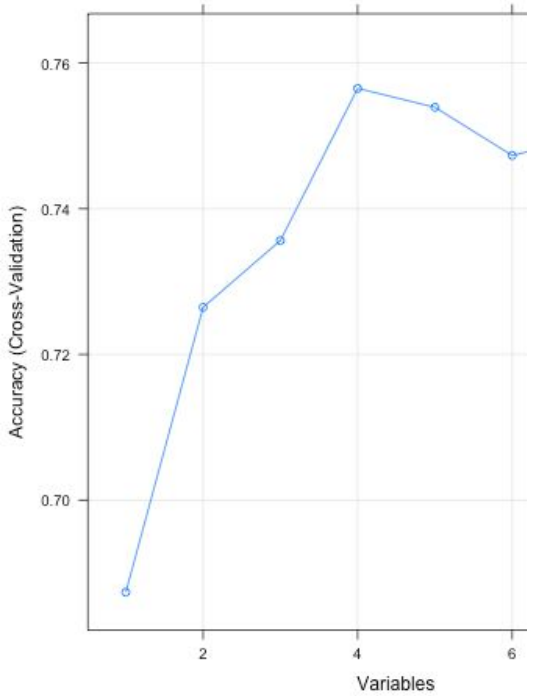


(x from -2 to 1)

L1: y = - 3/2 * x1

Жадный отбор признаков

Чередование добавления и удаления признаков

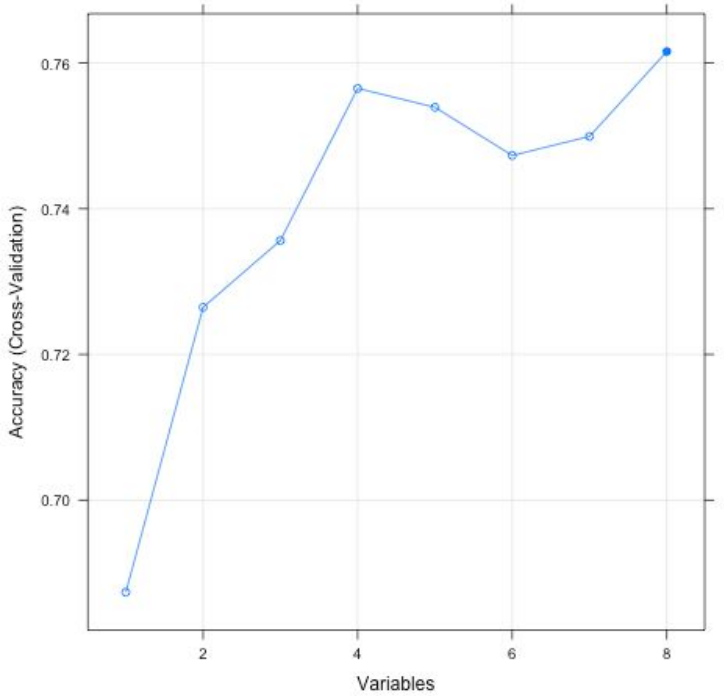


Жадный отбор признаков

Чередование добавления и удаления признаков

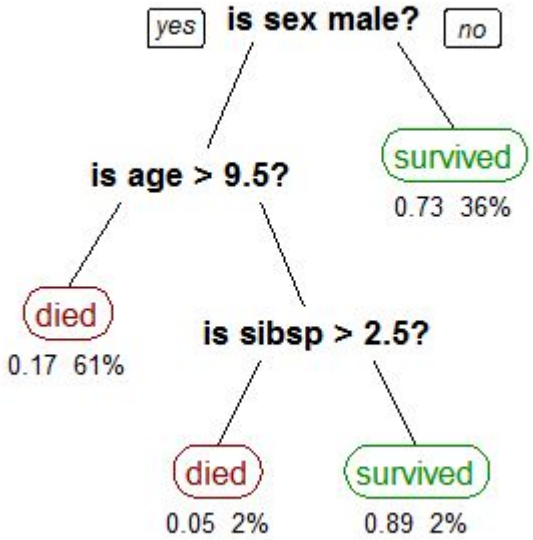
Этап добавления: добавляем лучшие признаки

Этап удаления: удаляем худшие признаки



Деревья

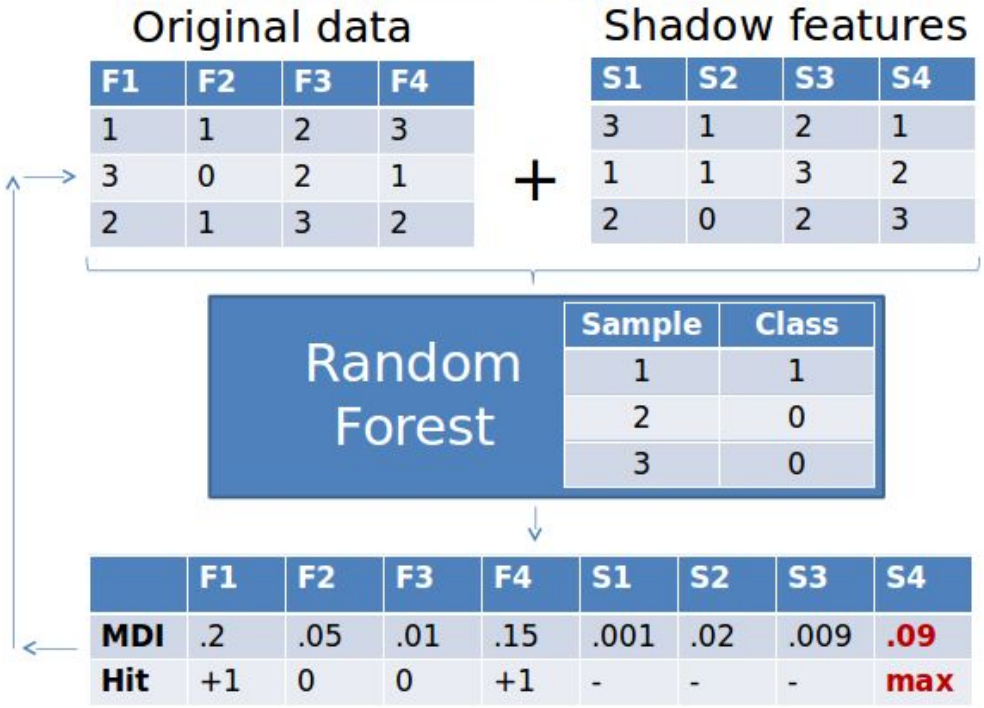
clf.feature_importances_



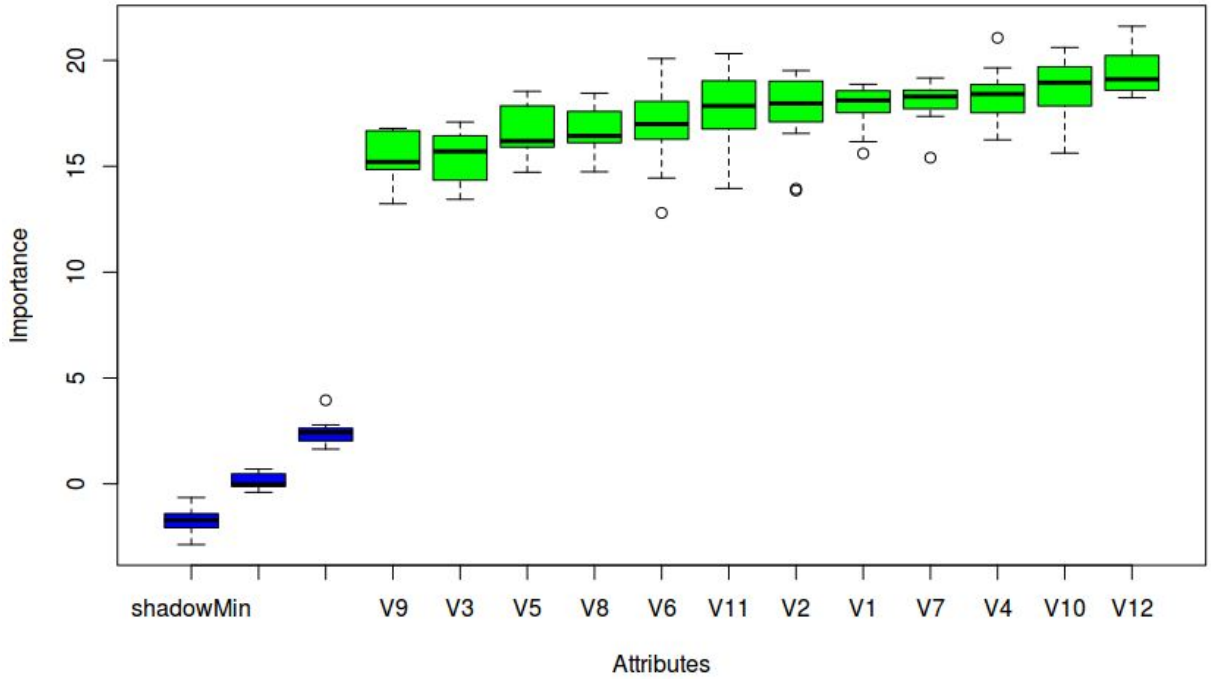
sklearn.tree.DecisionTreeClassifier

sklearn.ensemble.RandomForestClassifier

Boruta

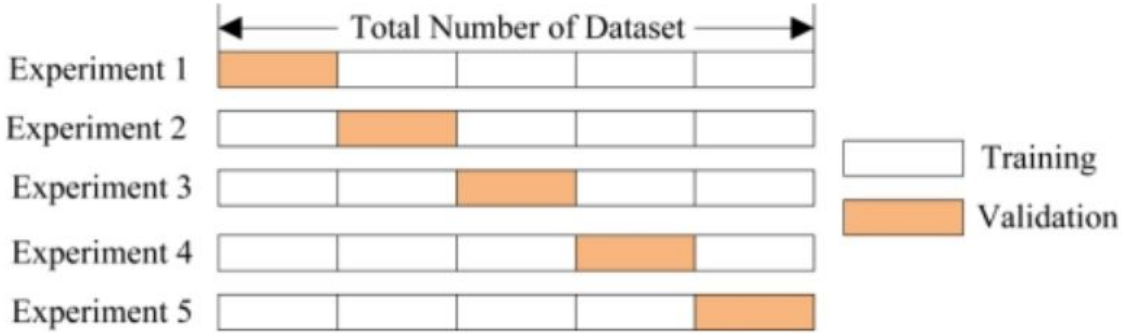

Original data
F1 F2 F3 F4
1 1 2 3
3 0 2 1
2 1 3 2
+
Shadow features
S1 S2 S3 S4
3 1 2 1
1 1 3 2
2 0 2 3
Random Forest
Sample Class
1 1
2 0
3 0
F1 F2 F3 F4 S1 S2 S3 S4
MDI .2 .05 .01 .15 .001 .02 .009 .09
Hit +1 0 0 +1 - - - max

Boruta


Importance
0
5
10
15
20
shadowMin
V9
V3
V5
V8
V6
V11
V2
V1
V7
V4
V10
V12
Attributes

Кросс-валидация

K-Fold cross validation:



На картинке k = 5, обычно такое k и используют. Другие частые варианты – 3 и 10.

Виды признаков

Какие бывают признаки:

1. Числовые
2. Порядковые
3. Категориальные
4. Даты и время
5. Координаты

Даты и время

1. Количество прошедших секунд
   например, с 00:00:00 UTC, 1 January 1970
2. Использование периодичности
   a. номер дня в году, в месяце, в неделе
   b. час, минута, секунда
3. Время до/после важных событий
   Например, количество дней, оставшихся до ближайшего праздника

Координаты

1. Повороты системы координат на 45 градусов, 22.5 градусов, etc
2. Добавление расстояний до:
   a. Других объектов из выборки
   b. Центров кластеров
   c. Инфраструктурных зданий - магазинов, школ, больниц

Категориальные признаки (строки)

Из колонок “name”, “ticket”, “cabin” можно сгенерировать новые признаки

	A	B	C	D	E	F	G	H	I	J	K
1	survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
2	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
5	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
12	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
13	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S

Категориальные признаки

Бинаризация

feature
a
b
c
b

feature == a	feature == b	feature == c
1		
	1	
		1
	1	

pd.get_dummies, sklearn.feature_extraction.DictVectorizer

Категориальные признаки

Hashing trick (n_features = 2)

feature
a
b
c
b

feature == a or feature == c	feature == b
1	
	1
1	
	1

sklearn.feature_extraction.FeatureHasher

Какие ещё признаки можно сгенерировать

1. Агреггации по нескольким переменным
2. Закодировать среднее значение в категориальной переменной
3. Метапризнаки

Категориальные признаки (аггрегация)

Задача определения кредитоспособности клиента

Обучающие данные

Customer_id	target
1	1
2	0
3	1

Данные о транзакциях:

Customer_id	datetime	amount
1	2016-09-01	4000
1	2016-09-02	7000
2	2016-09-01	2500

Категориальные признаки (аггрегация)

Обучающие данные

Customer_id	target
1	1
2	0
3	1

Данные о транзакциях:

Customer_id	datetime	amount
1	2016-09-01	4000
1	2016-09-02	7000
2	2016-09-01	2500

transactions.groupby(‘customer_id’).amount.sum()

Категориальные признаки

Кодирование средним значением целевой переменной

KFold или Leave-one-out:

Split	User ID	Y	mean(Y)	random	Exp_UID
Training	A1	0	.667	1.05	0.70035
Training	A1	1	.333	.97	0.32301
Training	A1	1	.333	.98	0.32634
Training	A1	0	.667	1.02	0.68034
Test	A1	-	.5	1	.5
Test	A1	-	.5	1	.5
Training	A2	0			

Категориальные признаки

Кодирование средним значением целевой переменной последовательно:

Split	User ID	Y	mean(Y)
Training	A1	0	NaN
Training	A1	1	0
Training	A1	1	0.5
Training	A1	0	0.66
Test	A1	-	.5
Test	A1	-	.5
Training	A2	0	

Предыдущих записей нет

np.mean([0])

np.mean([0, 1])

np.mean([0, 1, 1])

np.mean([0, 1, 1, 0])

np.mean([0, 1, 1, 0])

Метапризнаки

Использование ответов других алгоритмов

	xgb_prediction	knn_prediction	svm_prediction	target
train	0.192	0.293	0.122	0
train	0.789	0.890	0.670	1
test	0.542	0.310	0.173	?

Осторожно с переобучением: используйте KFold, LOO

Генерация признаков

Для решения задачи нужно использовать разные типы данных

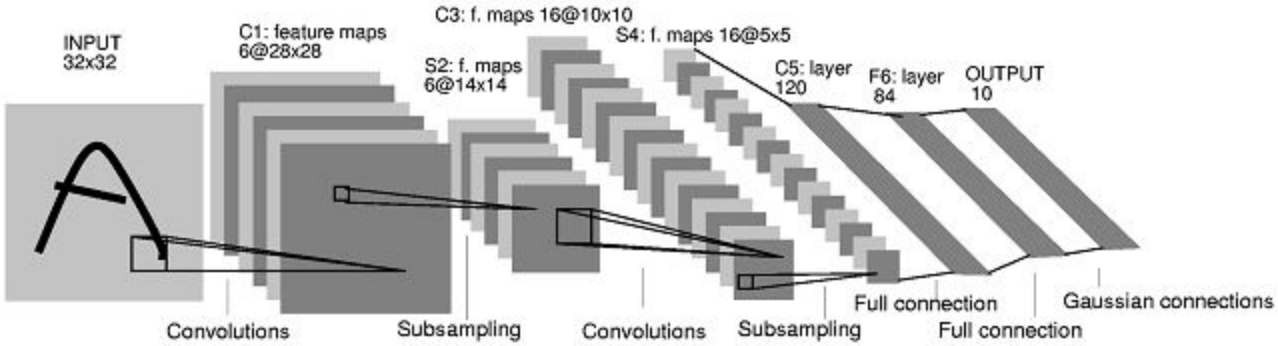
Пример: задача рекомендации музыки

1. Музыкальные треки
2. Тексты песен
3. Плейлисты

Проблема: нужно преобразовать к одному формату - матрице “объекты-признаки”

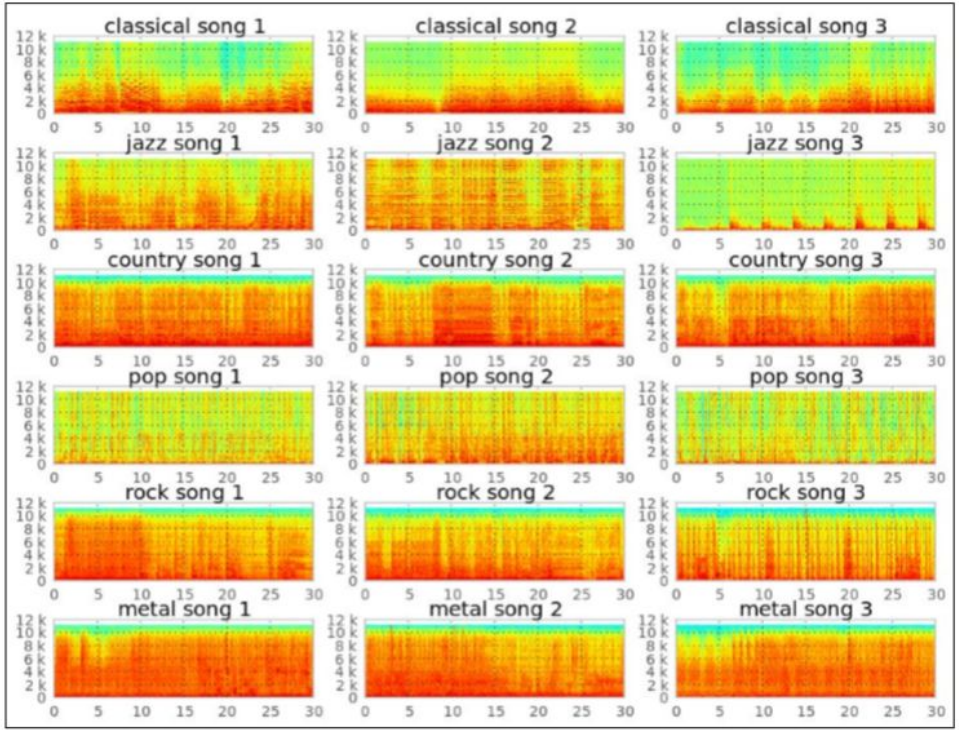
Картинка -> вектор

1. Descriptors/Embeddings



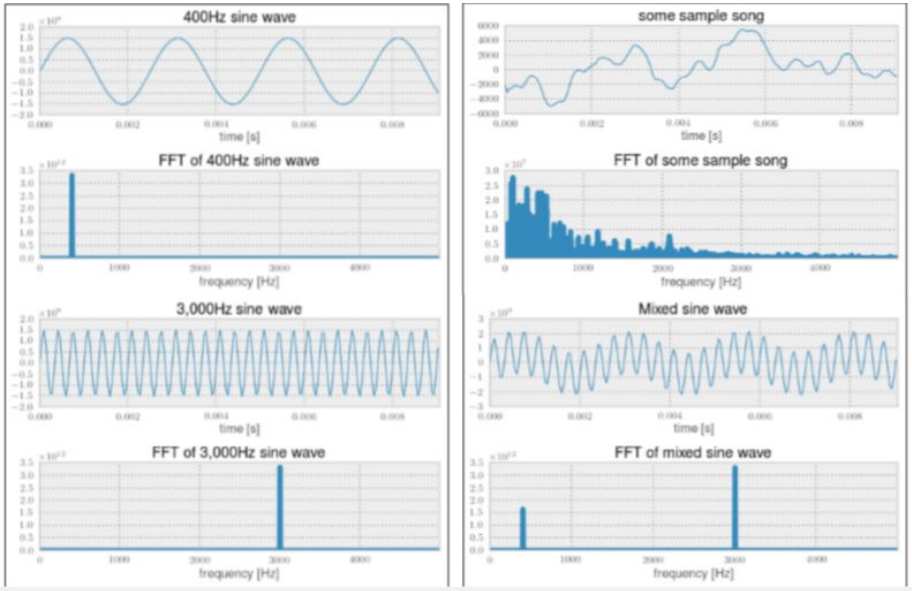
Можно использовать значения слоя FC6 как сжатое представление изображения

Звук: спектрограммы



Звук: спектрограммы -> вектора

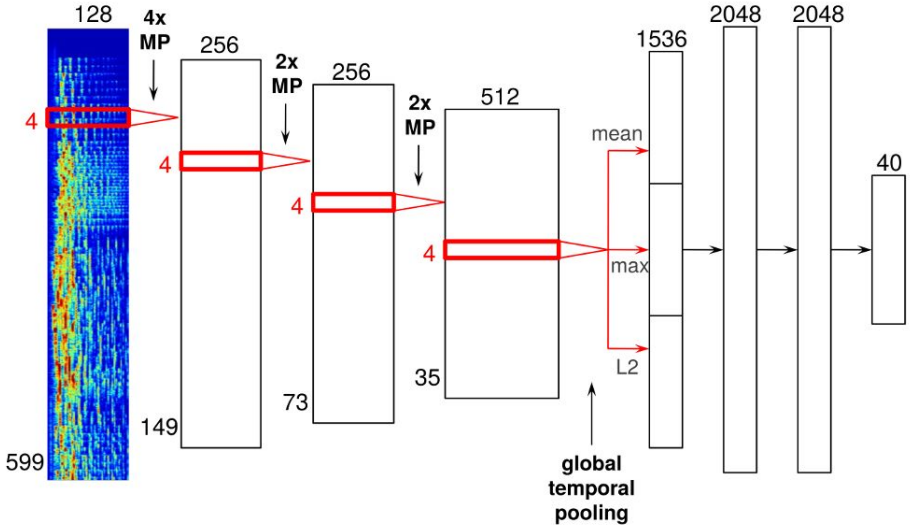
1. MFCC - преобразование Фурье логарифма спектра



Звук: спектрограммы -> вектора

1. MFCC - преобразование Фурье логарифма спектра

2. Embeddings с помощью нейронных сетей:
   a. Как вектора:
      полносвязные нейросети
   b. Как картинки:
      сверточные нейросети

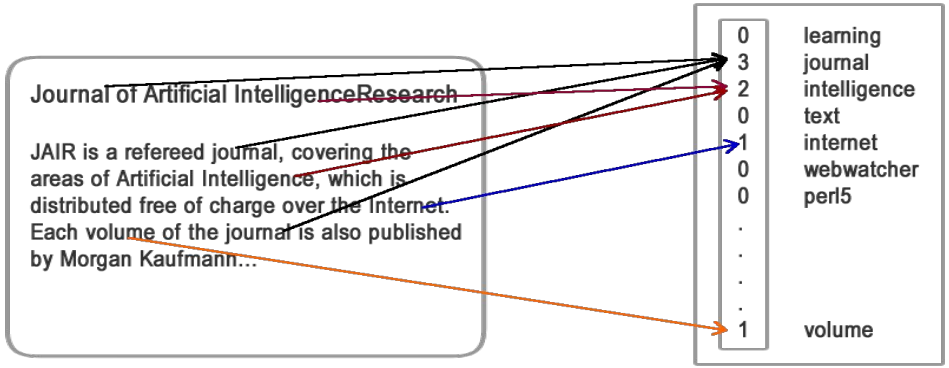


Текст

1. Bag of words:

Очередность слов потеряна

Смысл каждого значения вектора известен



2. Embeddings с помощью нейросетей (~word2vec)

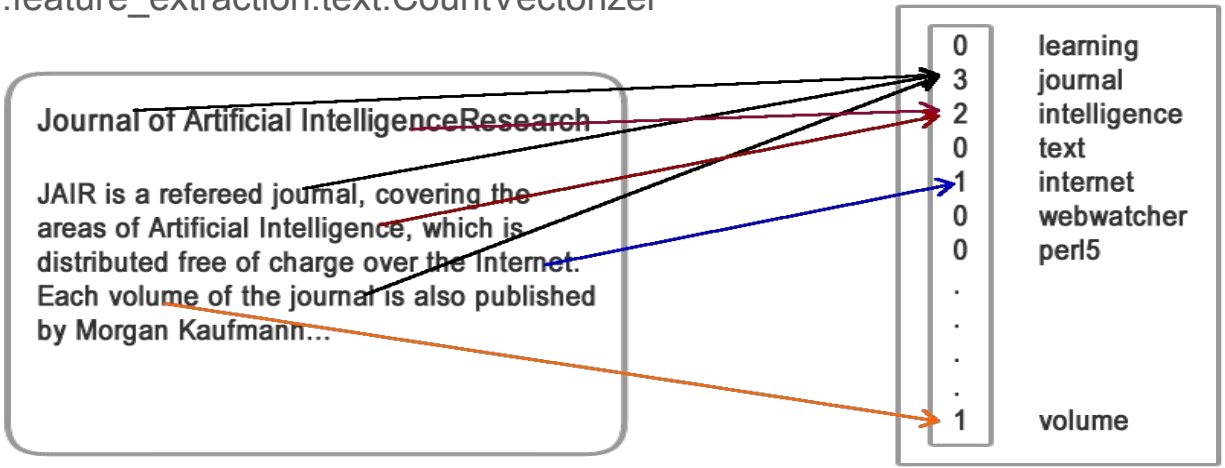
Очередность слов использована

Смысл каждого значения вектора может быть выяснен (скорее всего)

Текст -> числовые признаки

Аналогично бинаризации для категориальных переменных

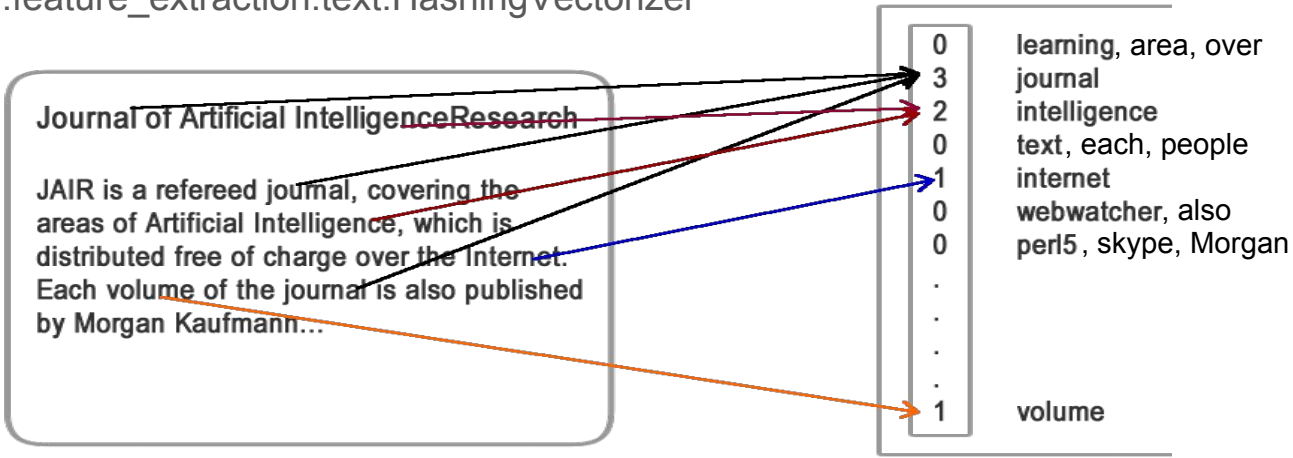
Sklearn.feature_extraction.text.CountVectorizer



Текст -> числовые признаки

Аналогично “Hashing trick” для категориальных переменных

Sklearn.feature_extraction.text.HashingVectorizer



Текст -> числовые признаки

Sklearn.feature_extraction.text.TfidfVectorizer (TfidfTransormer)

Term Frequency

$$\mathrm{tf}_{\mathrm{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

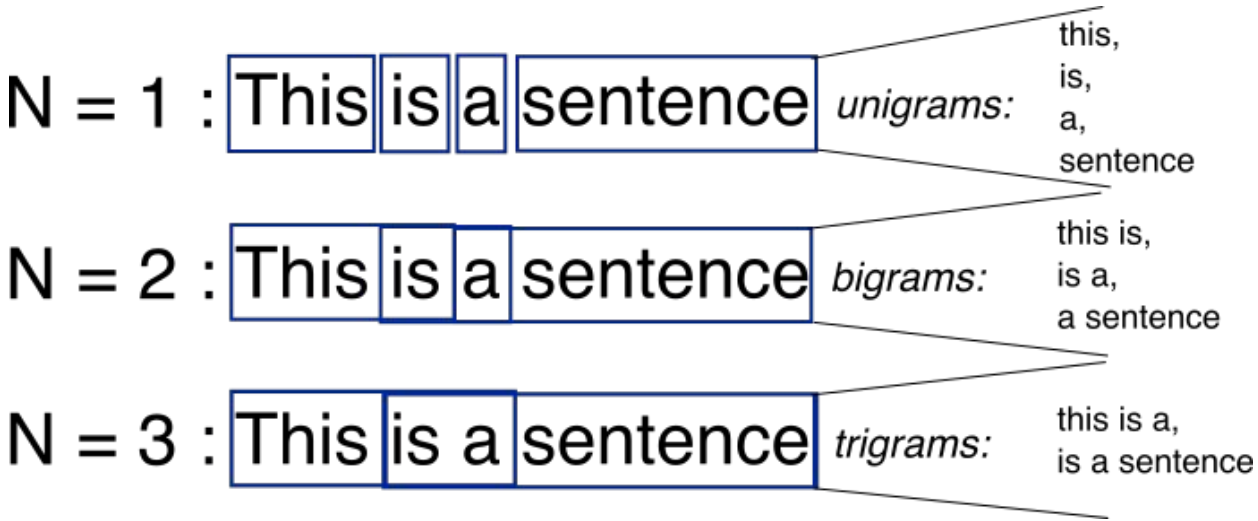
Inverse Document Frequency

$$\mathrm{idf}_{\mathrm{i}} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

Текст -> числовые признаки

N-grams

Sklearn.feature_extraction.text.CountVectorizer(ngram_range=(minN, maxN))



Ссылки

Boruta:
http://danielhomola.com/2015/05/08/borutapy-an-all-relevant-feature-selection-method/

Recommending music on Spotify with deep learning:
http://benanne.github.io/2014/08/05/spotify-cnns.html

Encoding of categorical features:
https://www.kaggle.com/c/caterpillar-tube-pricing/forums/t/15748/strategies-to-encode-categorical-variables-with-many-categories

Tips for data science competitions:
http://www.slideshare.net/OwenZhang2/tips-for-data-science-competitions