Чем дальше в случайный лес …

Автор: Павел Сулимов

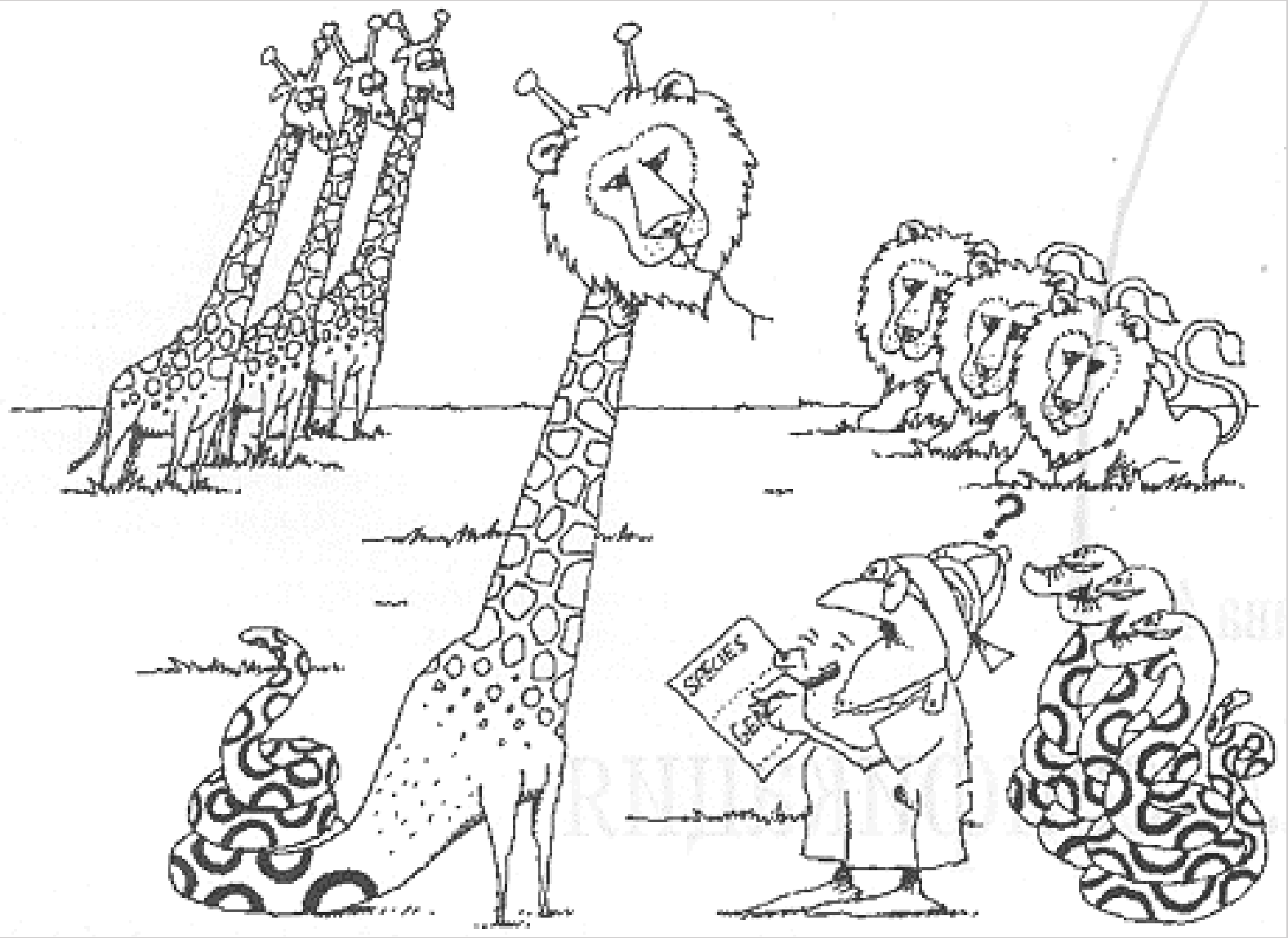
+7 (919) 965 78 56

sulimov@crocode.org

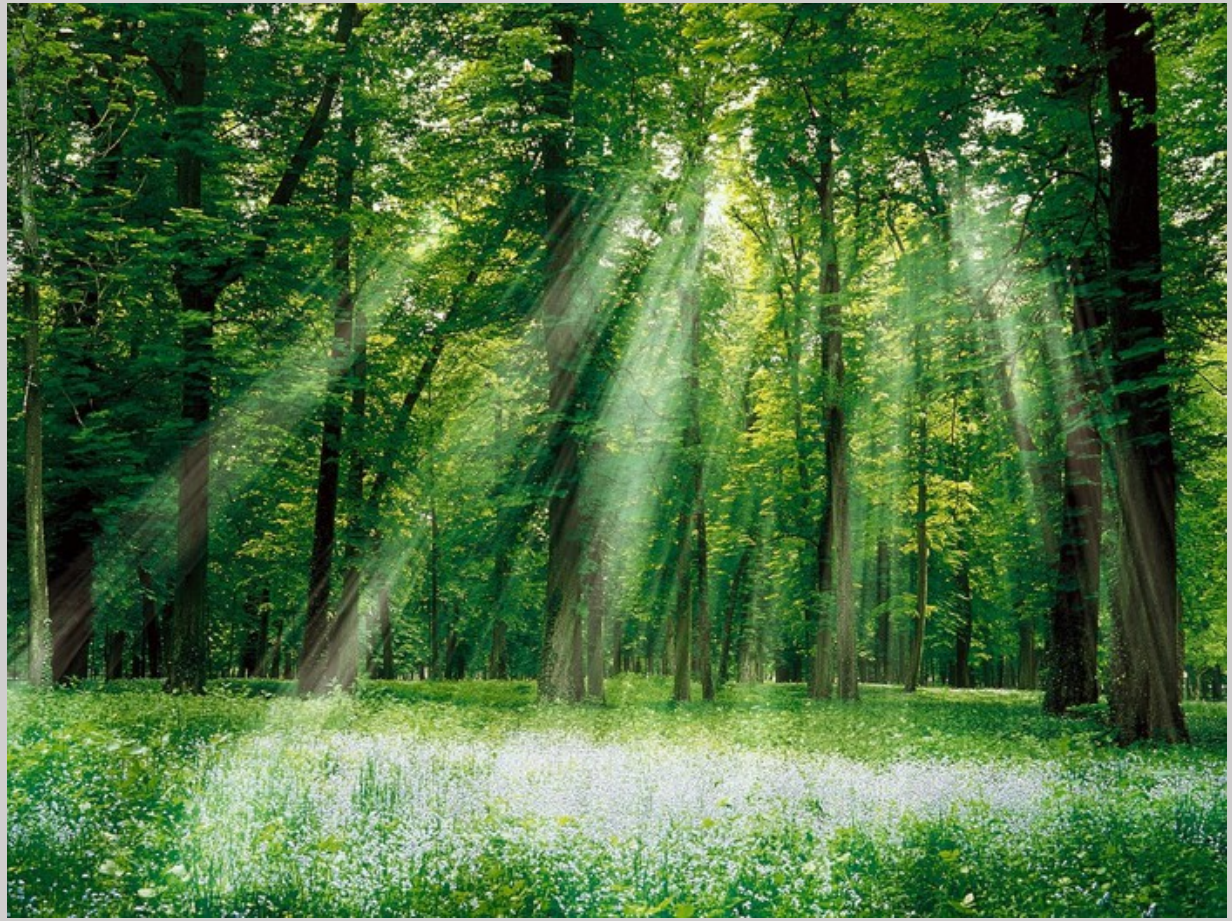
ДИСКЛЕЙМЕР

НЕКОТОРЫЕ ИЗ ПРЕДСТАВЛЕННЫХ ЗАДАЧ МОГУТ ОКАЗАТЬСЯ ДИКИМИ БАЯНАМИ-БАБАЯНАМИ...
НО МНЕ КАК-ТО...

Куда отнести объект (к какому классу)?



Random Forest



Идея Random Forest

- Bagging
- Random subspace method

Алгоритм

Обучающая выборка состоит из N примеров, размерность пространства признаков равна M

1.извлекаем бутстреп-выборку B объема n с возвращением из обучающей выборки (некоторые примеры попадут в неё несколько раз, а примерно $N/3$ примеров не войдут в неё вообще)

Алгоритм

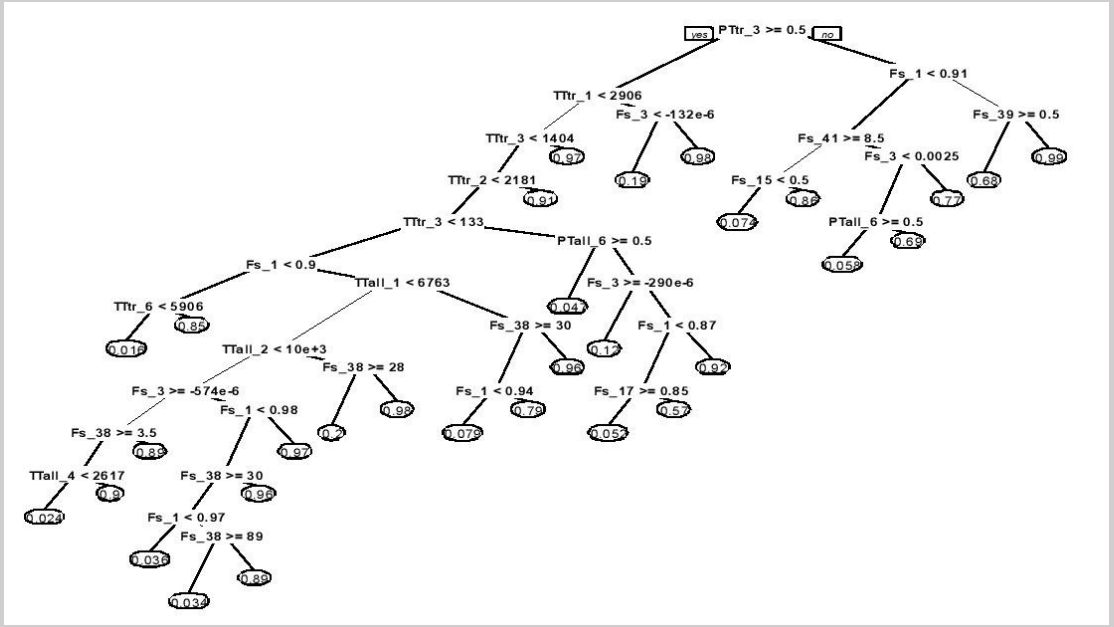
2. Построим решающее дерево, причём в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а лишь из m случайно выбранных

   Выбор наилучшего из этих m признаков обычно осуществляется с использованием критерия Джинни, применяющийся также в алгоритме построения решающих деревьев CART

   Иногда вместо него используется критерий прироста информации

Алгоритм

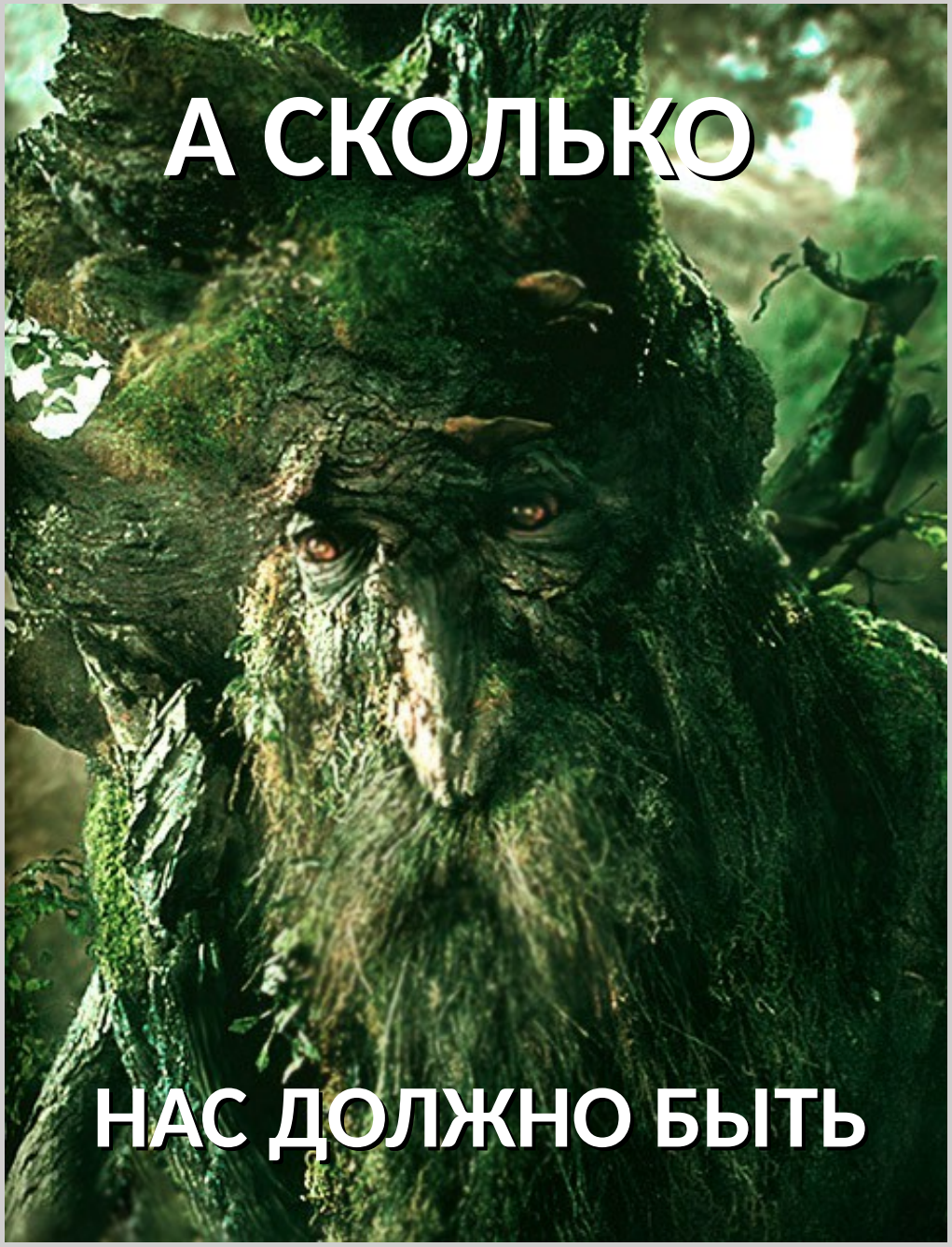
3. Дерево строится до полного исчерпания подвыборки и не подвергается процедуре прунинга



Деревья голосуют

Параметры

- Объем бутстреп-выборки = объем обучающей подвыборки
- Число случайно отбираемых переменных: квадратный корень из m
- Число деревьев: ?????


А СКОЛЬКО
НАС ДОЛЖНО БЫТЬ

Ответ:

Оптимальное число деревьев подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке. В случае её отсутствия, минимизируется оценка ошибки *out-of-bag*: доля примеров обучающей выборки, неправильно классифицируемых комитетом, если не учитывать голоса деревьев на примерах, входящих в их собственную обучающую подвыборку

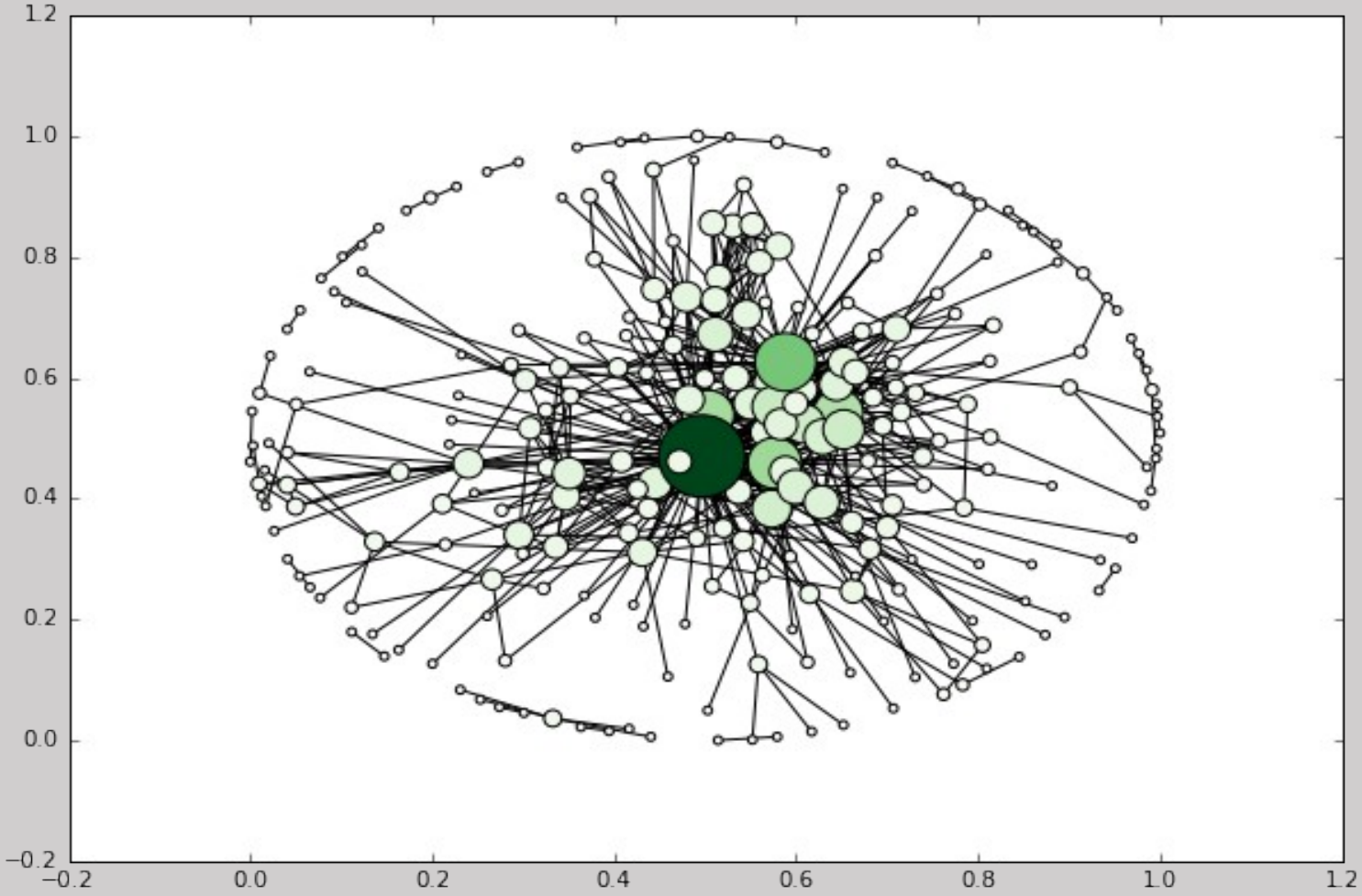
$$\mathrm{OOB} = \sum_{i=1}^{\ell} L\left(y_i, \frac{1}{\sum_{n=1}^{N}[x_i \notin X_n^\ell]} \sum_{n=1}^{N}[x_i \notin X_n^\ell] b_n(x_i)\right)$$

Баяны

Задача постановки диагноза

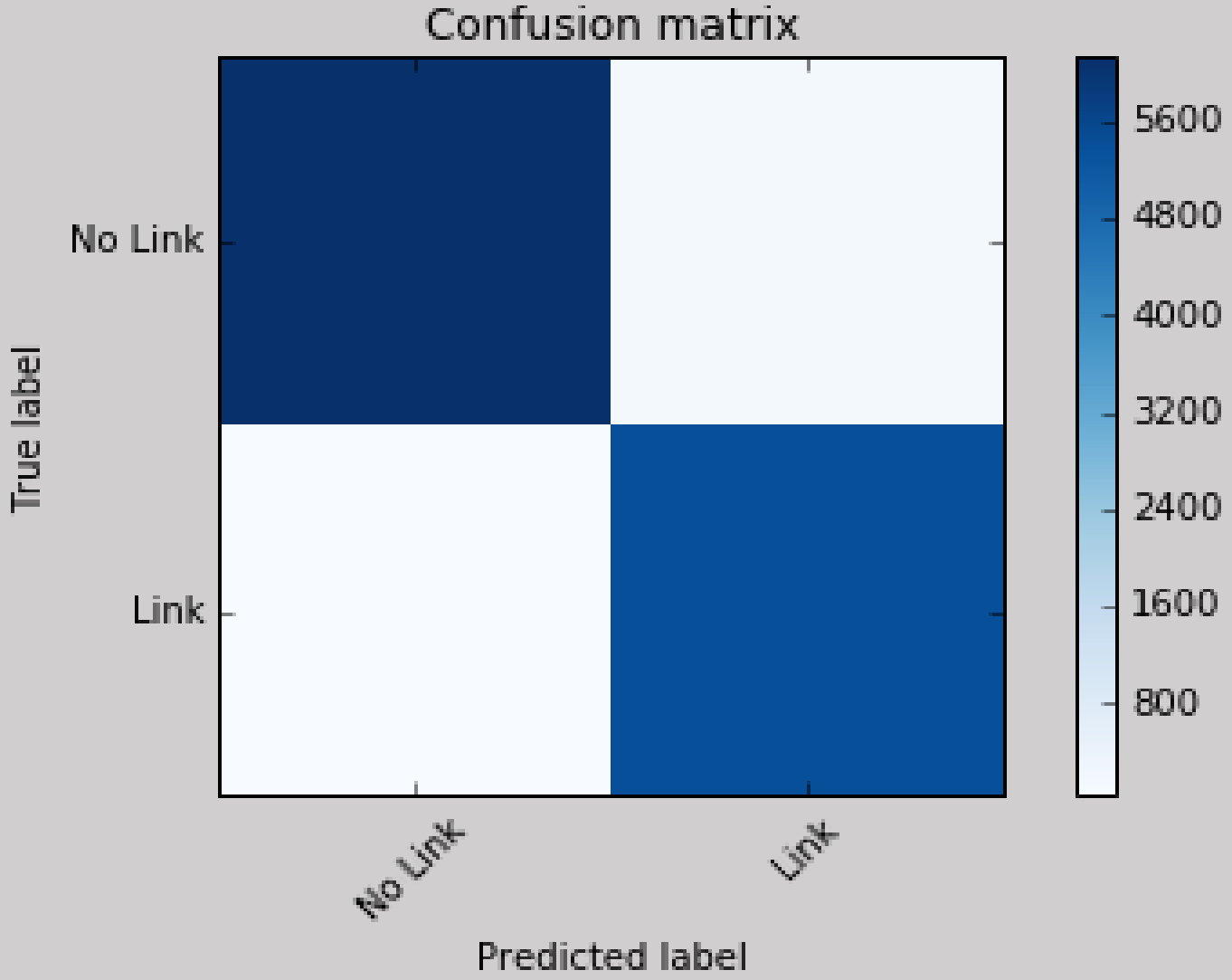
Задача кредитного скоринга

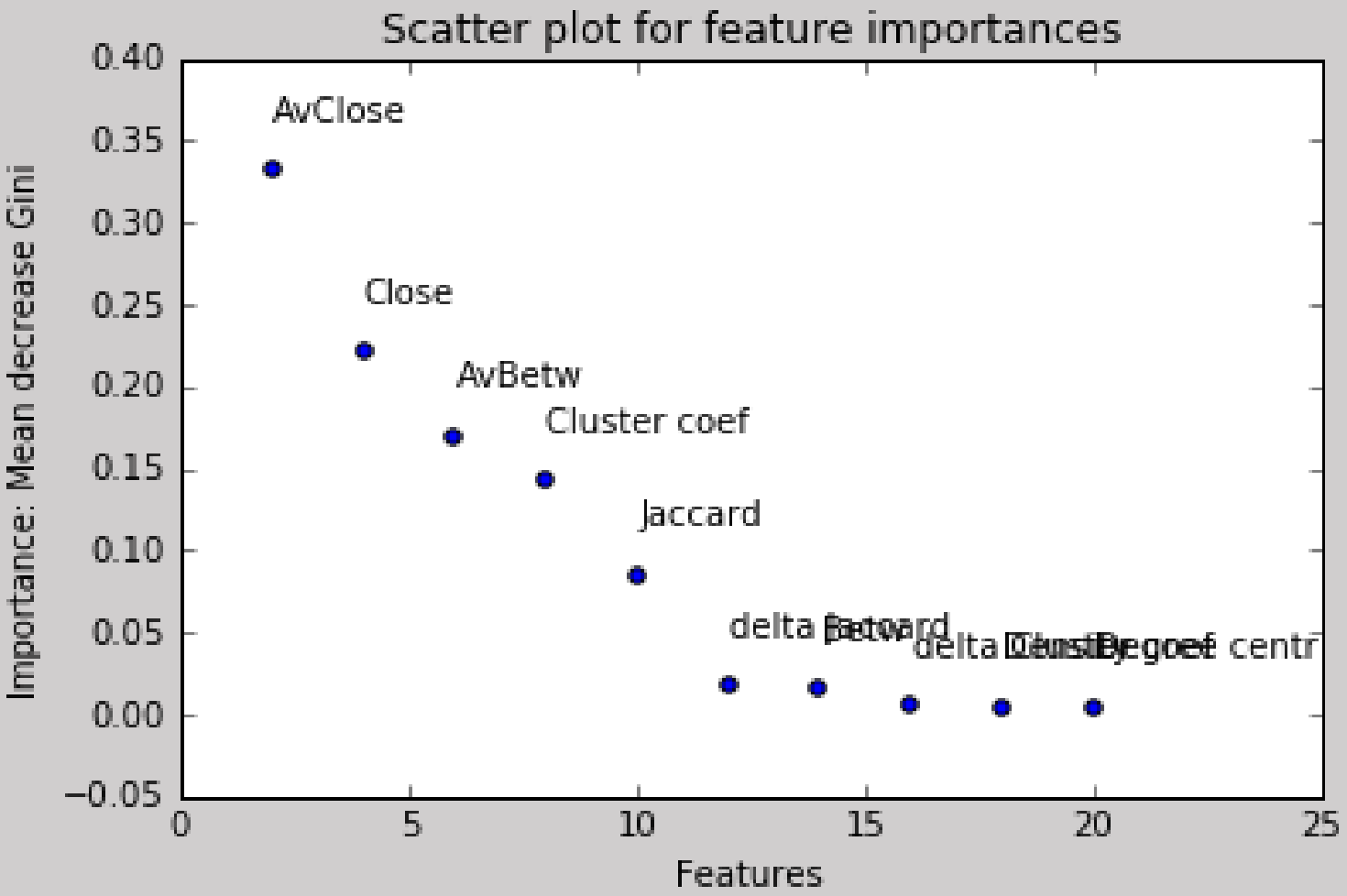
Пример с сетевыми данными (Flickr)

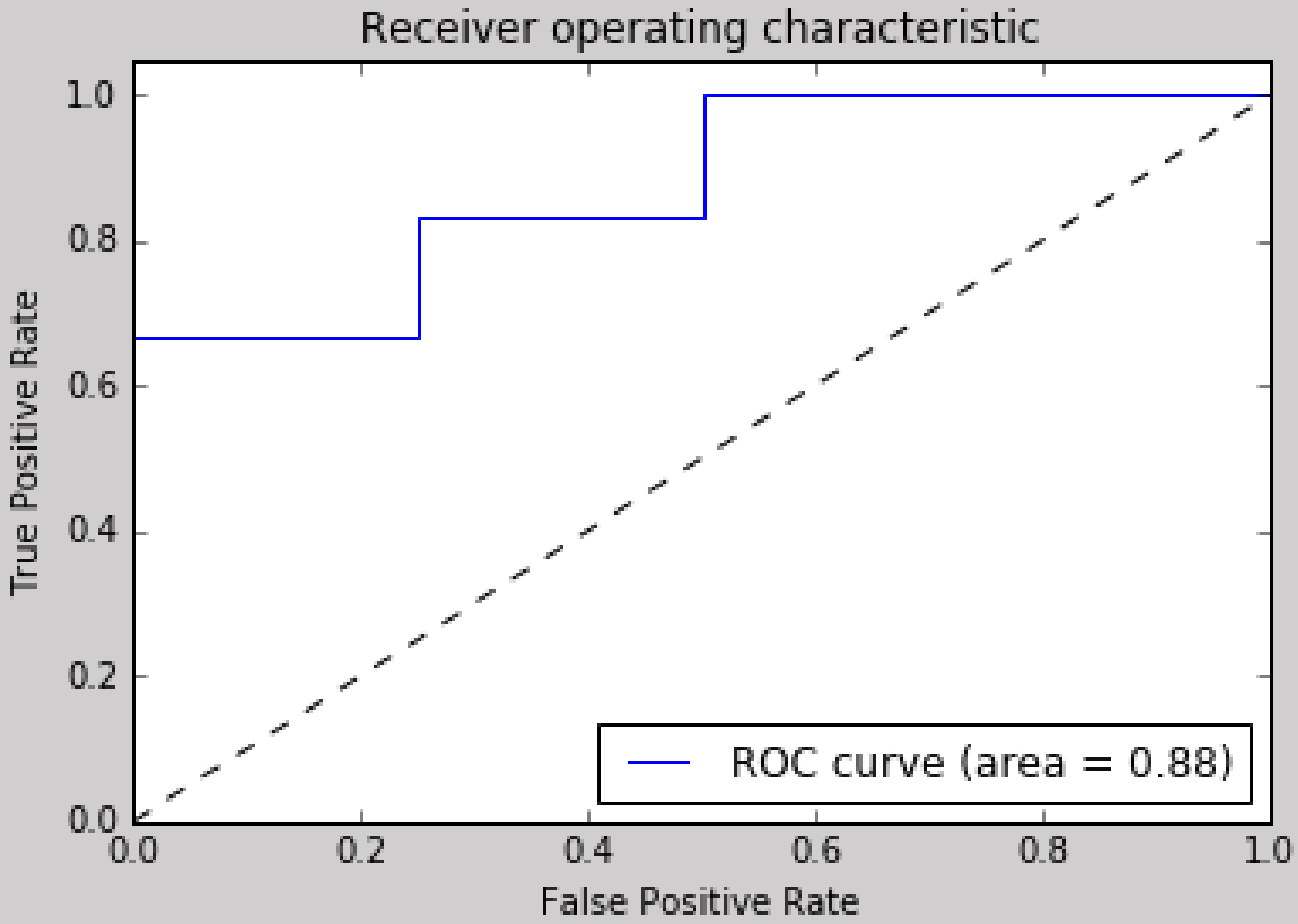


Dataset для модели

	AvClose	AvBetw	Cluster coef	Jaccard	Link
14462	0.21083401955582676	0.02448830043477378	0.04560912182436487	0.15617938997637132	1
21675	0.00022225926543312772	0.0	0.0001666944490748458	0.0	0
18477	0.05797216599970348	0.0	0.0001666944490748458	0.0	0
13170	0.12629614869154945	2.6174432724410186e-07	0.0054010802160432084	1.7833333333333332	1
31537	0.00029632922176538136	7.409877183508647e-08	0.00022224691632403602	0.0	0
8750	0.05096307511125238	0.0	0.00040008001600320064	1.0	0
4634	0.1347643754303316	1.6009604481920797e-08	0.0014002800560112022	0.9333333333333333	0
12778	0.18728900295985224	0.0015675081268454614	0.013602720544108823	0.8284690799396681	1
3705	0.06920029461869692	8.531891251176281e-06	0.000600120024004801	0.3333333333333333	0
27343	0.14751230434319057	0.00021248551321179467	0.0016669444907484582	0.2777777777777778	1


Confusion matrix
No Link
Link
True label
No Link
Link
Predicted label
5600
4800
4000
3200
2400
1600
800


Scatter plot for feature importances
Importance: Mean decrease Gini
Features
AvClose
Close
AvBetw
Cluster coef
Jaccard
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0
5
10
15
20
25


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)

SVM (украдено у Воронцова)

Линейный классификатор:

$$a(x, w) = \text{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^n,\ w_0 \in \mathbb{R}.$$

Пусть выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ линейно разделима:

$$\exists w, w_0 : \quad M_i(w, w_0) = y_i\big(\langle w, x_i\rangle - w_0\big) > 0, \quad i = 1, \ldots, \ell$$

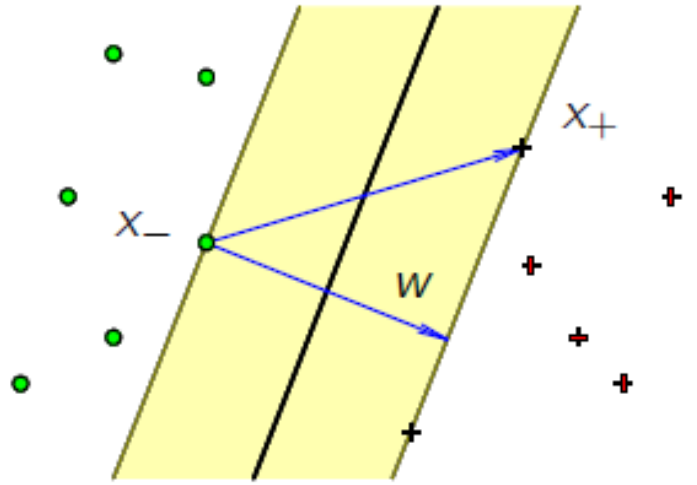
Нормировка: $\min\limits_{i=1,\ldots,\ell} M_i(w, w_0) = 1.$

Разделяющая полоса:

$$\big\{x : \ -1 \leqslant \langle w, x\rangle - w_0 \leqslant 1\big\}.$$

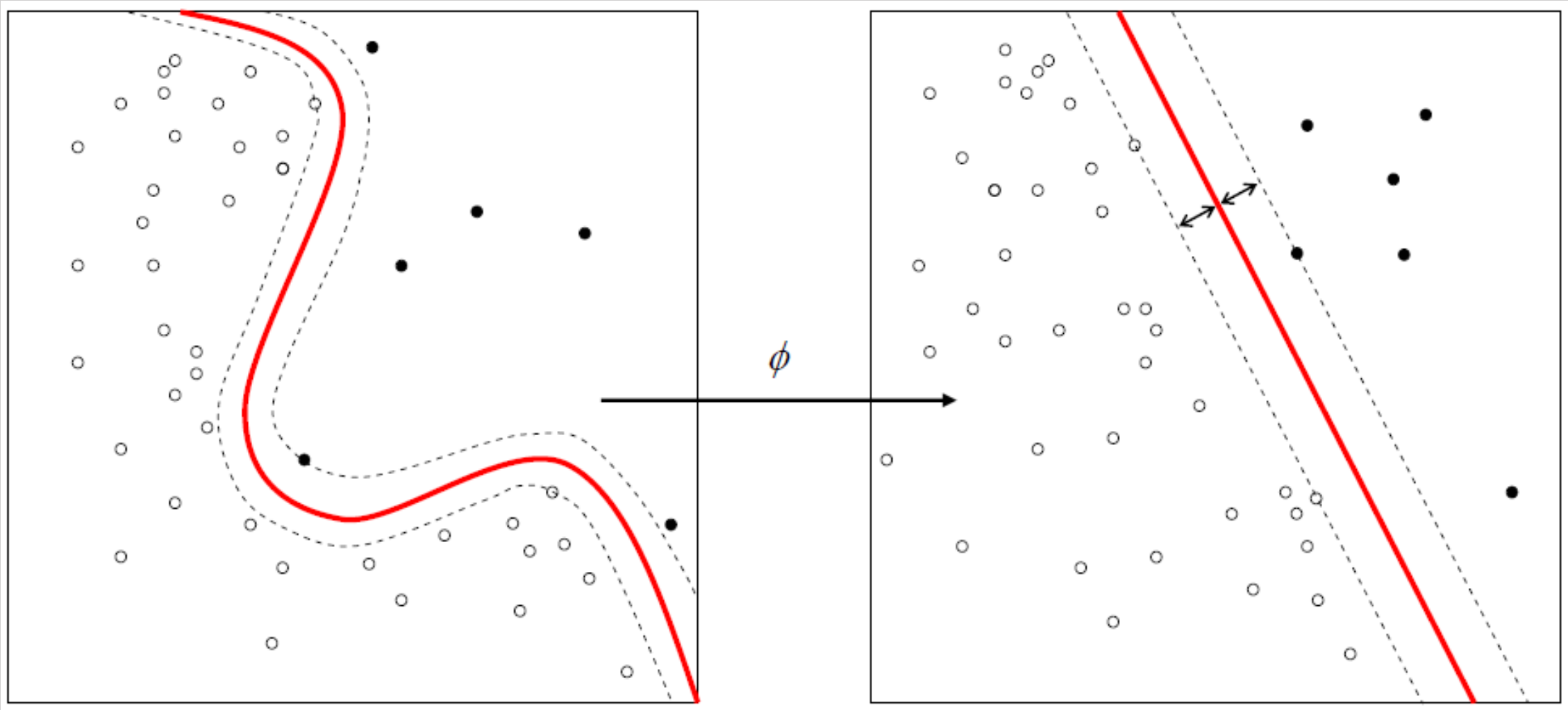
Ширина полосы:

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{2}{\|w\|} \to \max.$$



Примеры ядер (украдено у Воронцова)

1. $K(x,x') = \langle x,x' \rangle^2$
   — квадратичное ядро;
2. $K(x,x') = \langle x,x' \rangle^d$
   — полиномиальное ядро с мономами степени d;
3. $K(x,x') = \big(\langle x,x' \rangle + 1\big)^d$
   — полиномиальное ядро с мономами степени $\leqslant d$;
4. $K(x,x') = \sigma\big(\langle x,x' \rangle\big)$
   — нейросеть с заданной функцией активации $\sigma(z)$
   (не при всех σ является ядром);
5. $K(x,x') = \mathrm{th}\big(k_0 + k_1\langle x,x' \rangle\big), \;\; k_0, k_1 \geqslant 0$
   — нейросеть с сигмоидными функциями активации;
6. $K(x,x') = \exp\big(-\beta\|x - x'\|^2\big)$
   — сеть радиальных базисных функций;


ϕ

Почему же SVM?

- Хорошо работают на разреженных данных
- Такого рода данные возникают, например, при работе с текстами. При работе с текстами формируется столько признаков, сколько всего уникальных слов встречается в текстах, и значение каждого признака равно числу вхождений в документ соответствующего слова.

Что же будем делать мы?

Будем записывать не количество вхождений слова в текст, а TF-IDF.

TF-IDF = TF * IDF,

где TF = отношению числа вхождений слова в документ к общей длине документа, IDF = в скольки документах выборки встречается это слово. Чем больше таких документов, тем меньше IDF.

Таким образом, TF-IDF будет иметь высокое значение для тех слов, которые много раз встречаются в данном документе, и редко встречаются в остальных

Алгоритм работы с SVM

TF-IDF

ИЛИ

Принципы анализа данных

- Делайте предварительную обработку (выбросы, «разреженные» данные и т.д.)
- Используйте кросс-валидацию
- Проверяйте ошибки на обучающих и тестовых выборках
- Используйте знакомые модели
- Будьте осторожны с нейронными сетями

Вместо «Спасибо за внимание»

Приходите на стажировку!

- студенты 3-4 курсов бакалавриата, магистратуры или выпускники
- уверенные знания математической статистики
- стремление получать новые знания и использовать их для решения реальных задач
- желание работать в команде

Анкета стажера:
https://docs.google.com/forms/d/e/1FAIpQLSddh16WHhSrsx-vk60u4PZt6UM9xPGeeFZbvKc3-D-SatuFvw/viewform