О схеме применения методов ML

Евгений Борисов

схема применения методов ML

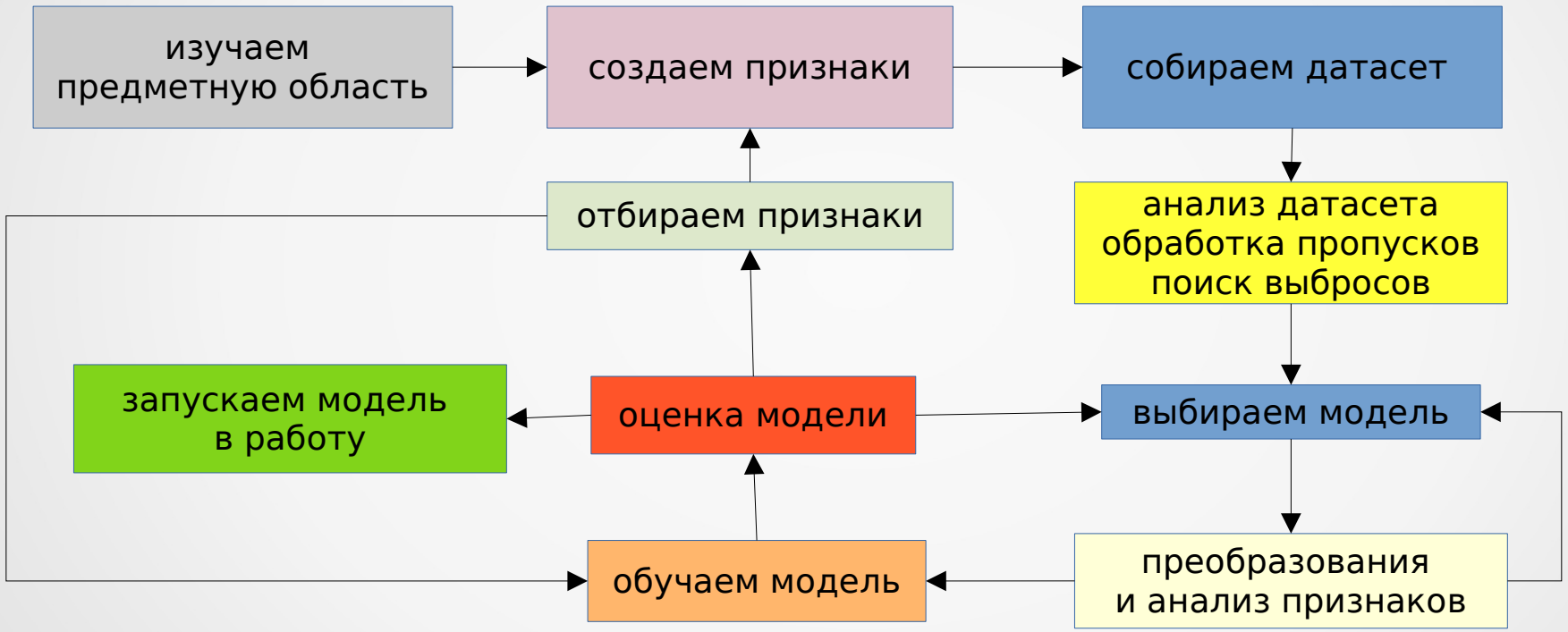


схема применения методов ML

создаем признаки *(feature extraction / feature engineering)*

отображение данных, специфических для предметной области,
в точки пространства признаков

схема применения методов ML

создаем признаки *(feature extraction / feature engineering)*

отображение данных, специфических для предметной области,
в точки пространства признаков

Типы признаков

- бинарные (да/нет)
- категориальные
- количественные ($\mathbb{R}$)
- порядковые

схема применения методов ML

создаем признаки *(feature extraction / feature engineering)*

отображение данных, специфических для предметной области,
в точки пространства признаков

Типы признаков

- бинарные (да/нет)
- категориальные
- количественные ($\mathbb{R}$)
- порядковые

примеры признаков

для текстов
- TF-IDF
- Word2Vec

для изображений:
- Haar-like features,
- HOG (Histogram of Oriented Gradients)

собираем признаки формируем учебный датасет

схема применения методов ML

анализ датасета

обработка пропусков

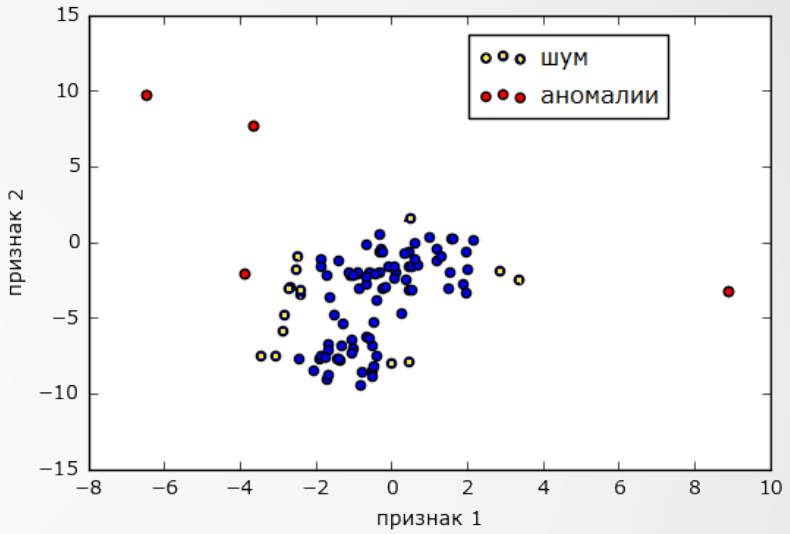
- удалить объект из выборки
- заполнить средним (медианой) вещественных переменных
- заполнить наиболее частым значением для категориальных
- заменить пропуск на редкое (мало вероятное) значение
- заменить на соседнее значение для упорядоченных данных

схема применения методов ML

анализ датасета

обработка пропусков

- удалить объект из выборки
- заполнить средним (медианой) вещественных переменных
- заполнить наиболее частым значением для категориальных
- заменить пропуск на редкое (мало вероятное) значение
- заменить на соседнее значение для упорядоченных данных



поиск выбросов / Outlier Detection

выброс или аномалия это то, что не вписывается в общие правила

Статистические тесты - отсечение по персентелю 0.95

Метрические методы - у выброса мало соседей

Итерационные методы - последовательное удаление выпуклых оболочек.

Модельные тесты - строим модель данных, точки, которые сильно отклоняются от модели - аномалии

Специальные модели ML - IsolationForest, выбросы попадают в листья на небольшой глубине дерева

схема применения методов ML

выбор модели

- тип задачи (классификация, регрессия, кластеризация...)

- особенности датасета (линейная разделимость и т.п.)

профили компактности

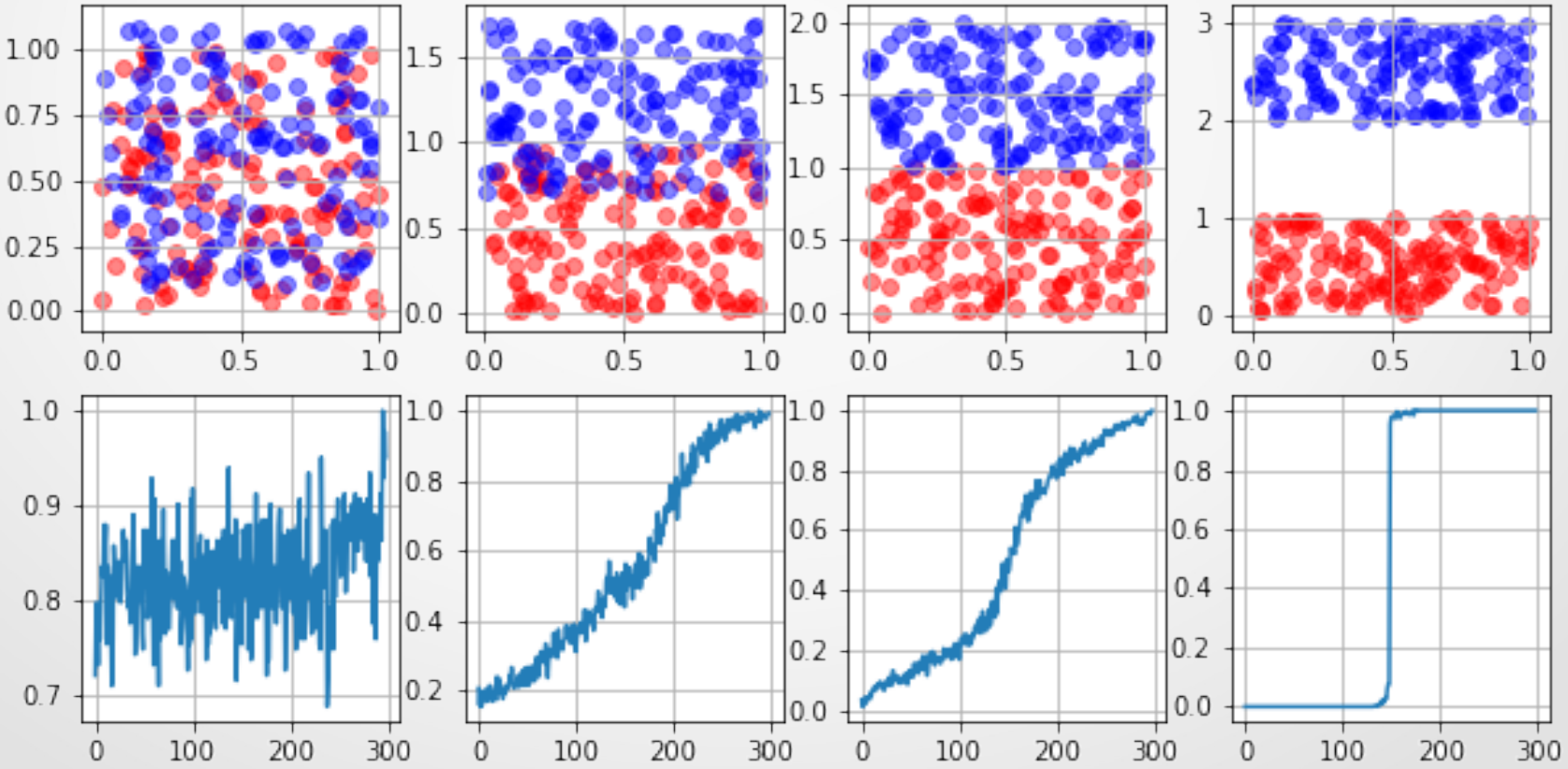


схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

преобразования признаков (feature transformation)

- *масштабирование в отрезок*
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]

$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

преобразования признаков (feature transformation)

- масштабирование в отрезок
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]

$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

- *стандартизация* ; приведение к μ=0 и σ=1 ;
  улучшает ситуацию с выбросами;
  можно применять с метрическими методами ;

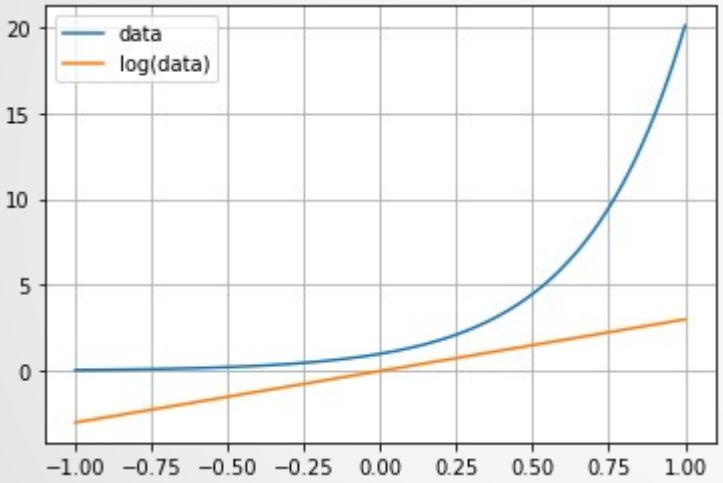
$$x := \frac{x - \mu}{\sigma}$$

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели



преобразования признаков (feature transformation)

- масштабирование в отрезок
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]
$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

- стандартизация приведение к μ=0 и σ=1 ;
  улучшает ситуацию с выбросами;
  можно применять с метрическими методами ;
$$x := \frac{x - \mu}{\sigma}$$

- *логарифмирование*
  помогает сделать значения более равномерными
$$x := \log(x)$$

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

преобразования признаков (feature transformation)

- масштабирование в отрезок
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]

$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

- стандартизация приведение к μ=0 и σ=1 ;
  улучшает ситуацию с выбросами;
  можно применять с метрическими методами ;

$$x := \frac{x - \mu}{\sigma}$$

- логарифмирование
  помогает сделать значения более равномерными

$$x := \log(x)$$

- *метод пространственных знаков (spatial sign)*
проецирует значения на поверхность многомерной сферы,
данные становятся равноудаленными от центра этой сферы;
применяется после стандартизации всех признаков

$$x_j := \frac{x_j}{\sum_k x_k^2}$$

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

значения признака

	16	37	18	81	9	40	29
x	0.53	0.73	0.08	0.89	0.92	0.38	0.23

преобразования признаков (feature transformation)

- масштабирование в отрезок
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]

$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

- стандартизация приведение к μ=0 и σ=1 ;
  улучшает ситуацию с выбросами;
  можно применять с метрическими методами ;

$$x := \frac{x - \mu}{\sigma}$$

- логарифмирование
  помогает сделать значения более равномерными

$$x := \log(x)$$

- метод пространственных знаков (spatial sign)
проецирует значения на поверхность многомерной сферы,
данные становятся равноудаленными от центра этой сферы;
применяется после стандартизации всех признаков

$$x_j := \frac{x_j}{\sum_k x_k^2}$$

- *категоризация по шкале и бинаризация*
  уход от избыточной детализации;
  помогает улучшить результаты некоторых типов моделей

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

значения признака

	16	37	18	81	9	40	29
x	0.53	0.73	0.08	0.89	0.92	0.38	0.23

шкала персентилей

	min	10%	25%	50%	75%	95%	max
x	0.00	0.08	0.21	0.43	0.75	0.94	0.97

преобразования признаков (feature transformation)

- масштабирование в отрезок
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]
$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

- стандартизация приведение к μ=0 и σ=1 ;
  улучшает ситуацию с выбросами;
  можно применять с метрическими методами ;
$$x := \frac{x - \mu}{\sigma}$$

- логарифмирование
  помогает сделать значения более равномерными
$$x := \log(x)$$

- метод пространственных знаков (spatial sign)
проецирует значения на поверхность многомерной сферы,
данные становятся равноудаленными от центра этой сферы;
применяется после стандартизации всех признаков
$$x_j := \frac{x_j}{\sum_k x_k^2}$$

- *категоризация по шкале и бинаризация*
  уход от избыточной детализации;
  помогает улучшить результаты некоторых типов моделей

схема применения методов ML

анализ признаков
- оценка зависимости (корреляции)

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков не позволяет однозначно оценить параметры модели

значения признака

	16	37	18	81	9	40	29
x	0.53	0.73	0.08	0.89	0.92	0.38	0.23

шкала персентилей

	min	10%	25%	50%	75%	95%	max
x	0.00	0.08	0.21	0.43	0.75	0.94	0.97

категоризация по шкале

	x	cat	bin
16	0.53	3	[0, 0, 0, 1, 0, 0]
37	0.73	3	[0, 0, 0, 1, 0, 0]
18	0.08	1	[0, 1, 0, 0, 0, 0]
81	0.89	4	[0, 0, 0, 0, 1, 0]
9	0.92	4	[0, 0, 0, 0, 1, 0]
40	0.38	2	[0, 0, 1, 0, 0, 0]
29	0.23	2	[0, 0, 1, 0, 0, 0]

преобразования признаков (feature transformation)

- масштабирование в отрезок
  полезен для визуализации,
  легко перенести признаки на отрезок [0, 255]
$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

- стандартизация приведение к μ=0 и σ=1 ;
  улучшает ситуацию с выбросами;
  можно применять с метрическими методами ;
$$x := \frac{x - \mu}{\sigma}$$

- логарифмирование
  помогает сделать значения более равномерными
$$x := \log(x)$$

- метод пространственных знаков (spatial sign)
проецирует значения на поверхность многомерной сферы,
данные становятся равноудаленными от центра этой сферы;
применяется после стандартизации всех признаков
$$x_j := \frac{x_j}{\sum_k x_k^2}$$

- *категоризация по шкале и бинаризация*
  уход от избыточной детализации;
  помогает улучшить результаты некоторых типов моделей

схема применения методов ML

обучение модели

задача оптимизации: минимизация функции потери в пространстве параметров модели

схема применения методов ML

обучение модели

задача оптимизации: минимизация функции потери в пространстве параметров модели

оценка модели (применяем кроссвалидацию)

- кластеризация: отношение средних внутрикластерного и межкластерного расстояний

- классификация: погрешность, точность, полнота, ROC AUC

- регрессия: среднеквадратичное отклонение

схема применения методов ML

обучение модели

задача оптимизации: минимизация функции потери в пространстве параметров модели

оценка модели (применяем кроссвалидацию)

- кластеризация: отношение средних внутрикластерного и межкластерного расстояний

- классификация: погрешность, точность, полнота, ROC AUC

- регрессия: среднеквадратичное отклонение

результаты оценки модели

- успешное завершение

- замена модели

- коррекция (отбор) признаков

схема применения методов ML

методы отбора признаков

цель: минимизация ошибки модели на контроле

на каждой итерации переобучаем и оцениваем модель

- полный перебор подмножеств признаков

- добавление признаков по одному с минимизацией ошибки (жадный)

- поочередное добавление/удаление

схема применения методов ML

Конкурс BigData от Beeline

https://special.habrahabr.ru/beeline/

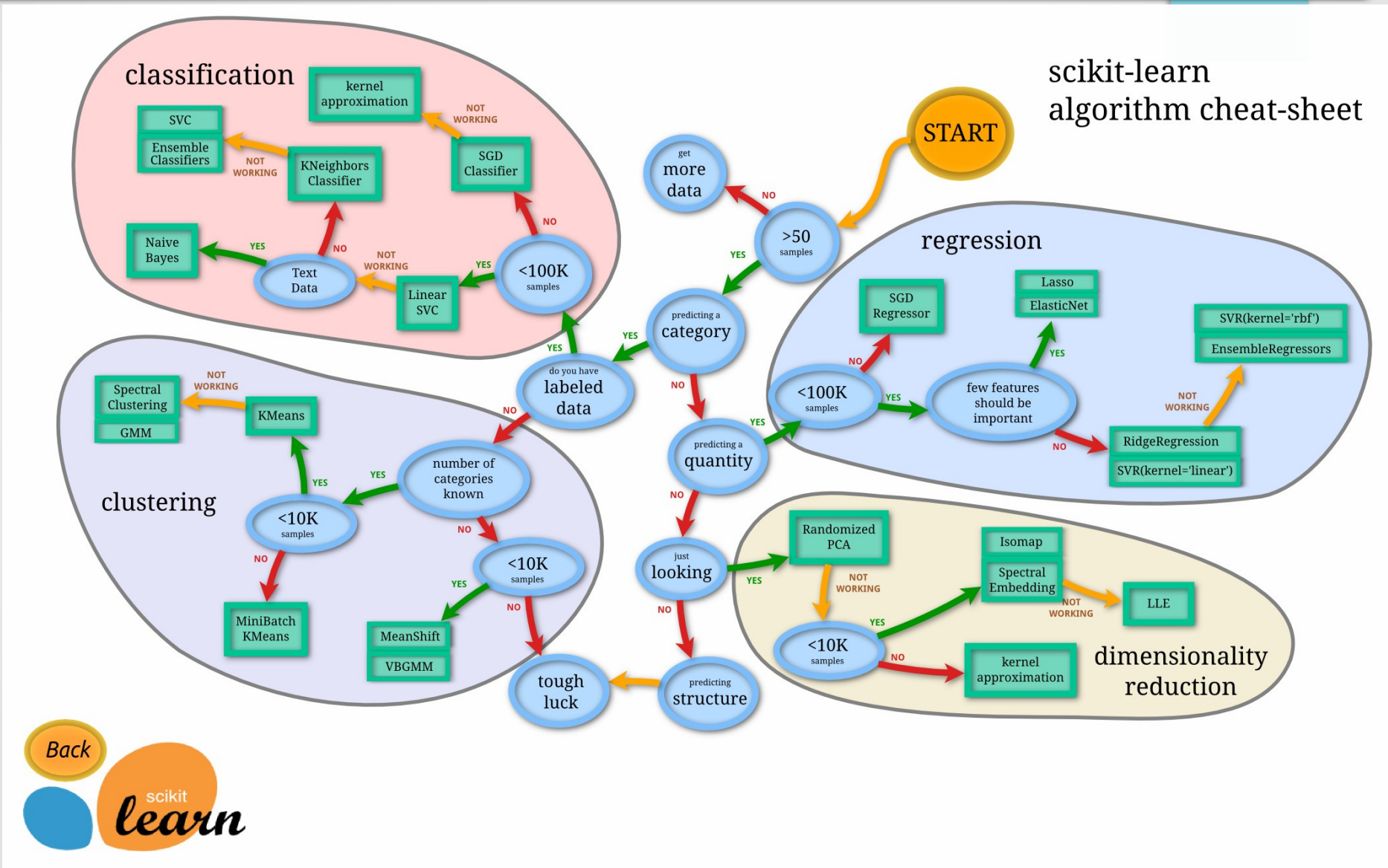
Александр Куменко

Как я победил в конкурсе BigData от Beeline

7 ноября 2015

https://habr.com/post/270367/

схема применения методов ML



https://scikit-learn.org/stable/tutorial/machine_learning_map/

схема применения методов ML

Литература

git clone *https://github.com/mechanoid5/ml_lectorium.git*

К.В. Воронцов Обобщающая способность. Методы отбора признаков. - курс "Машинное обучение" ШАД Яндекс 2014

Александр Дьяконов Поиск аномалий https://dyakonov.org

http://www.machinelearning.ru