Машинное обучениее. Вводная лекция.

Московский физико-технический институт, МФТИ

Москва

Этапы развития области искусственный интелект

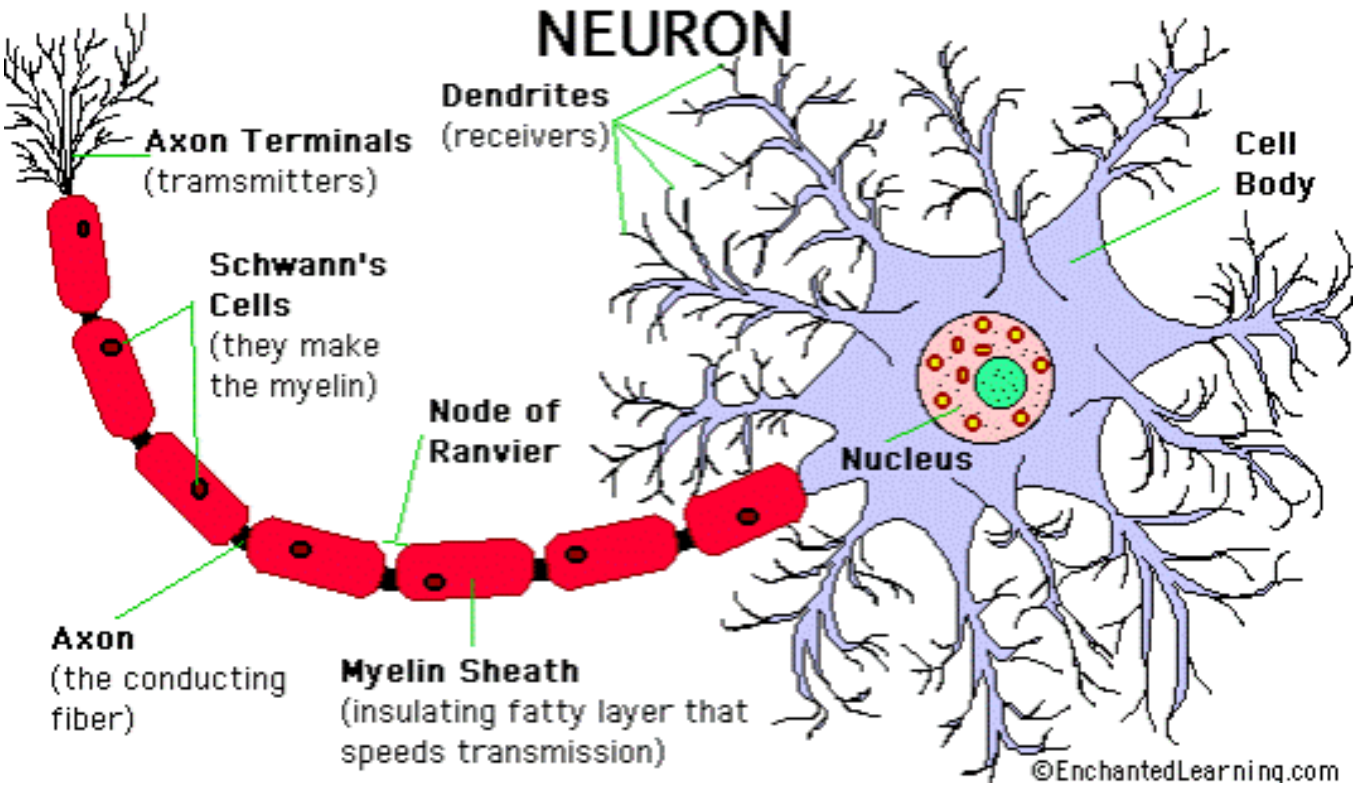
- Предыстория (мечты об ИИ и теоретические основы алгоритмистики);
- Появление самообучающихся машин конец 50х, начало 60х;
- Системы основанные на знаниях 70е;
- Первые нейронные сети (в основе алгоритм обратного распространения ошибки) начало 80х;
- Прорыв связанный с применением глубокого обучения в 2006г;

Причины бурного развития ИИ: **глубокое обучение**, **вычислительные мощности**, **большие наборы качественных данных**.

Модель МакКаллока-Питтса

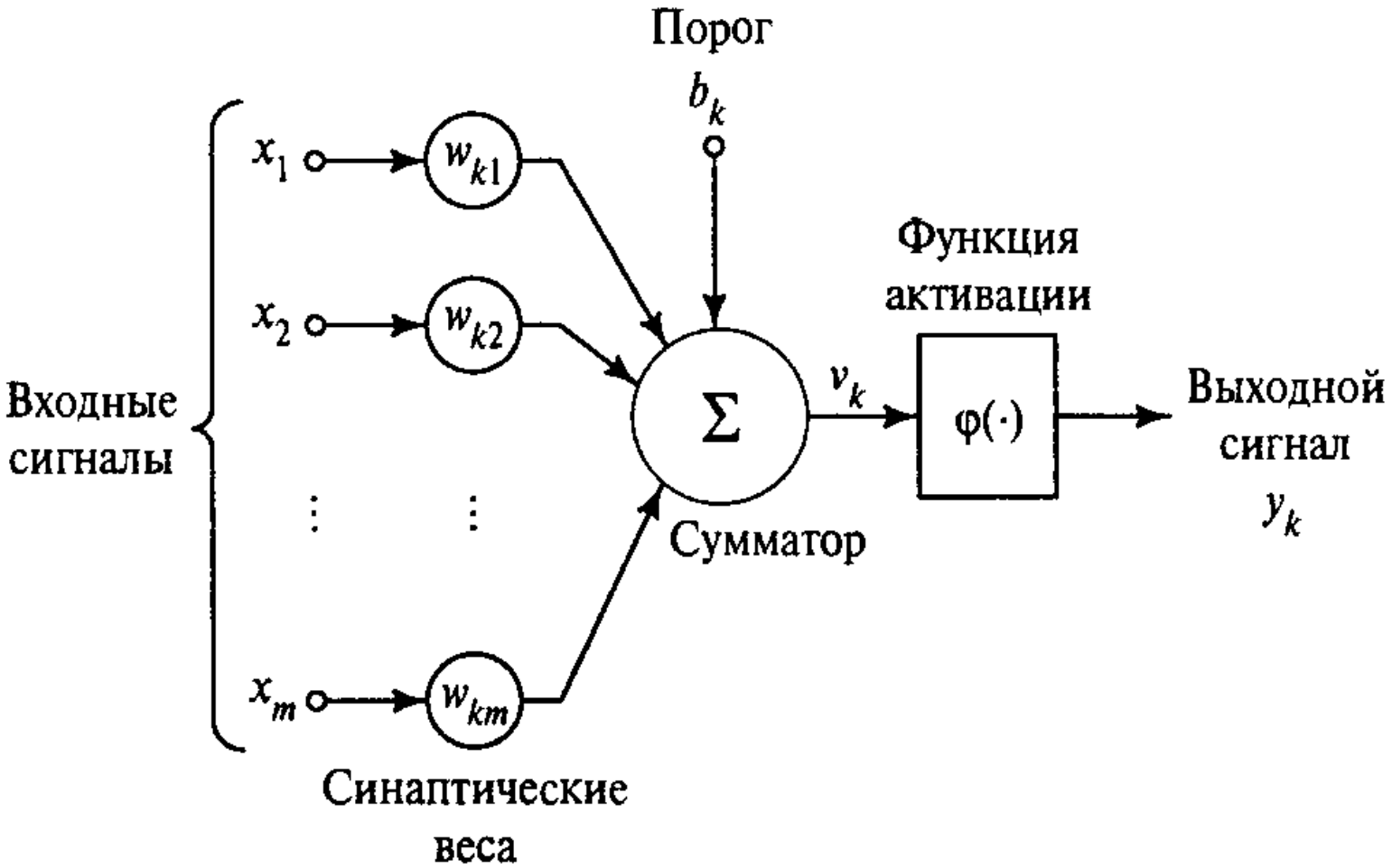
Линейный классификатор или персептрон является простейшей математической моделью нервной клетки- нейрона. Нейрон имеет множество разветвлённых отростков - дендритов, и одно длинное тонкое волокно - аксон, на конце которого находятся синапсы, примыкающие к дендритам других нервных клеток. Нервная клетка может находиться в двух состояниях: обычном и возбуждённом. Клетка возбуждается, когда в ней накапливается достаточное количество положительных зарядов. В возбуждённом состоянии клетка генерирует электрический импульс величиной около 100 мВи длительностью около 1 мс, который проходит по аксону до синапсов.

Картинка нейрона



Особенности работы мозга: асинхронность, параллелизм, пластичность.

Мат модель нейрона



От нейрона к нейросетям

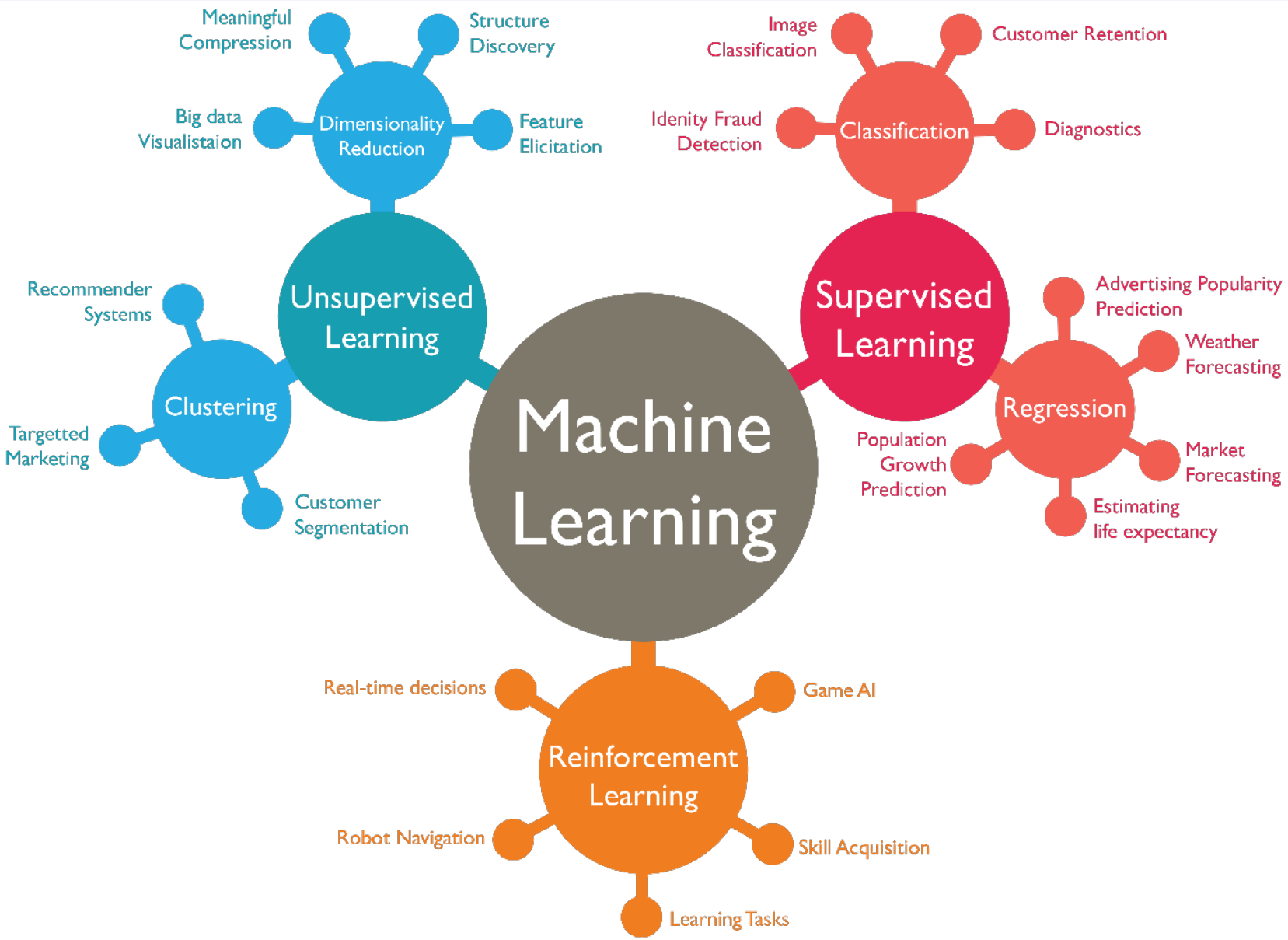
Тринадцатая проблема Гилберта решена Колмогоровым в 1957 году:

Теорема

Любая непрерывная n аргументов на единичном кубе $[0,1]$ *представима в виде суперпозиции непрерывных функций одного аргумента и операции сложения:*

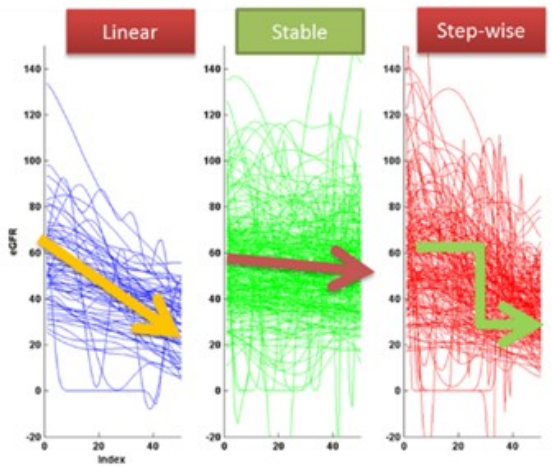
$f(x^1, x^2, ..., x^n) = \sum_{k=1}^{2n+1} h_k \left(\sum_{i=1}^{n} \varphi_{ik}(x^i) \right)$, *где* h_k, φ_{ik} *непрерывные функции, причем* φ_{ik} *не зависят от выбора* f.

Нетрудно видеть, что записанное здесь выражение имеет структуру нейронной сети с одним скрытым слоем из 2n+1 нейронов. Таким образом, двух слоёв уже достаточно, чтобы вычислять произвольные непрерывные функции, и не приближённо, а точно.

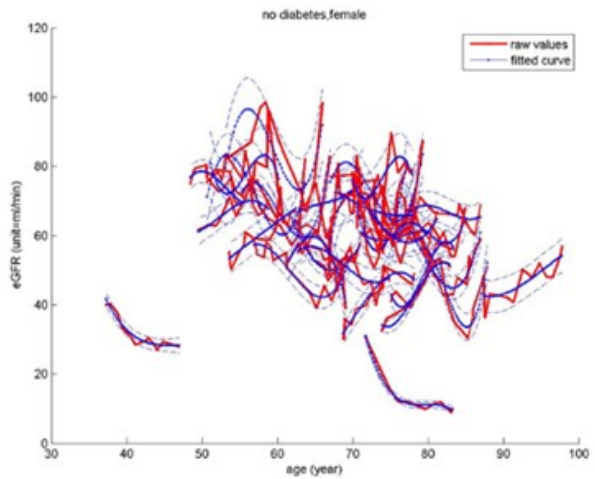

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating life expectancy
Population Growth Prediction
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Решаемые задачи

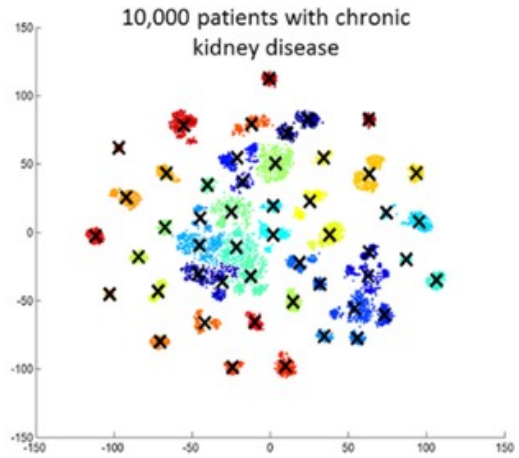
Classification



Regression



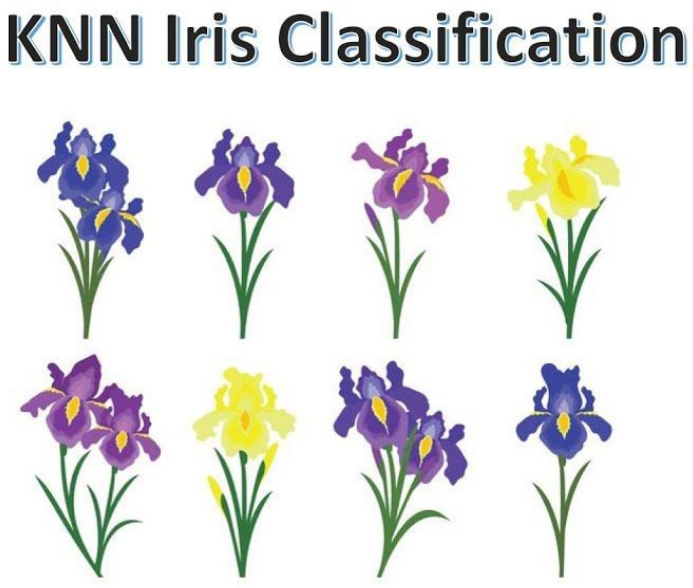
Clustering



Классификация

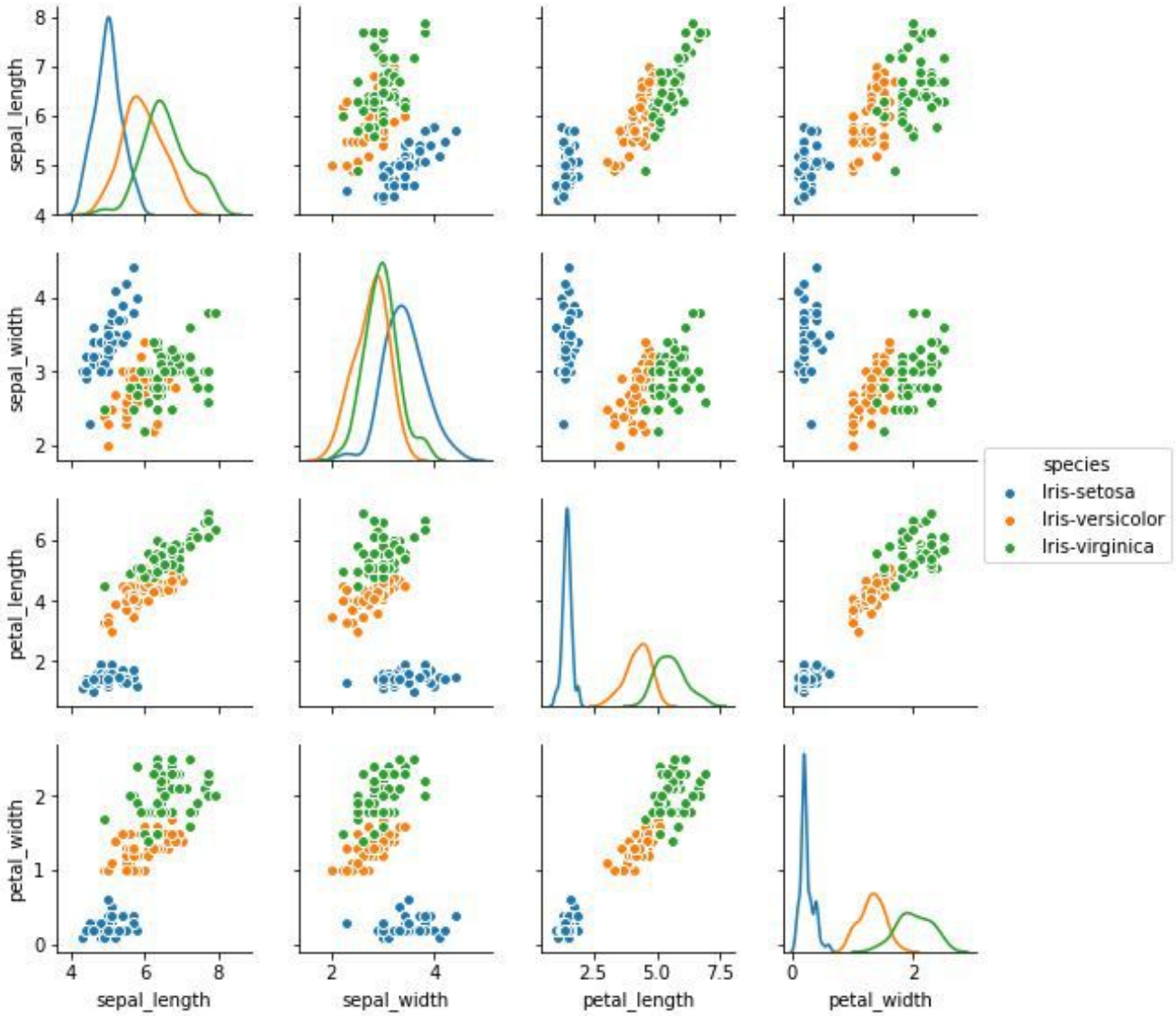
- байсовский метод (например наивный байсовский классификатор);
- метод опорных векторов (SVM);
- kNN (k -ближайших соседей);
- деревья решений;
- Random forest;

Параметры цветков



1. Длина наружной доли околоцветника (англ. sepal length);
2. Ширина наружной доли околоцветника (англ. sepal width);
3. Длина внутренней доли околоцветника (англ. petal length);
4. Ширина внутренней доли околоцветника (англ. petal width).

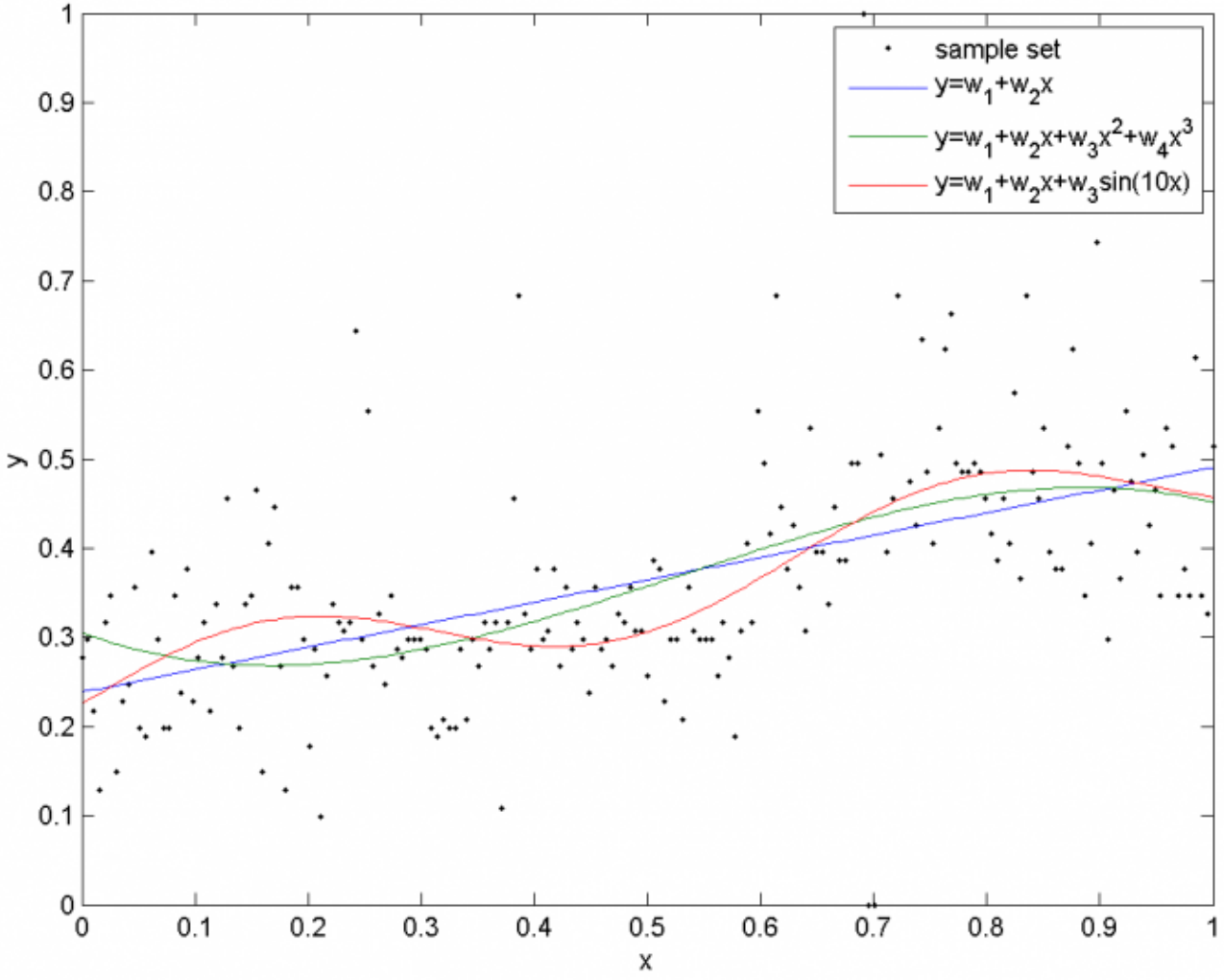
Иллюстрация классификации цветков ириса



Регрессия

Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка.

Регрессия. Иллюстрация.



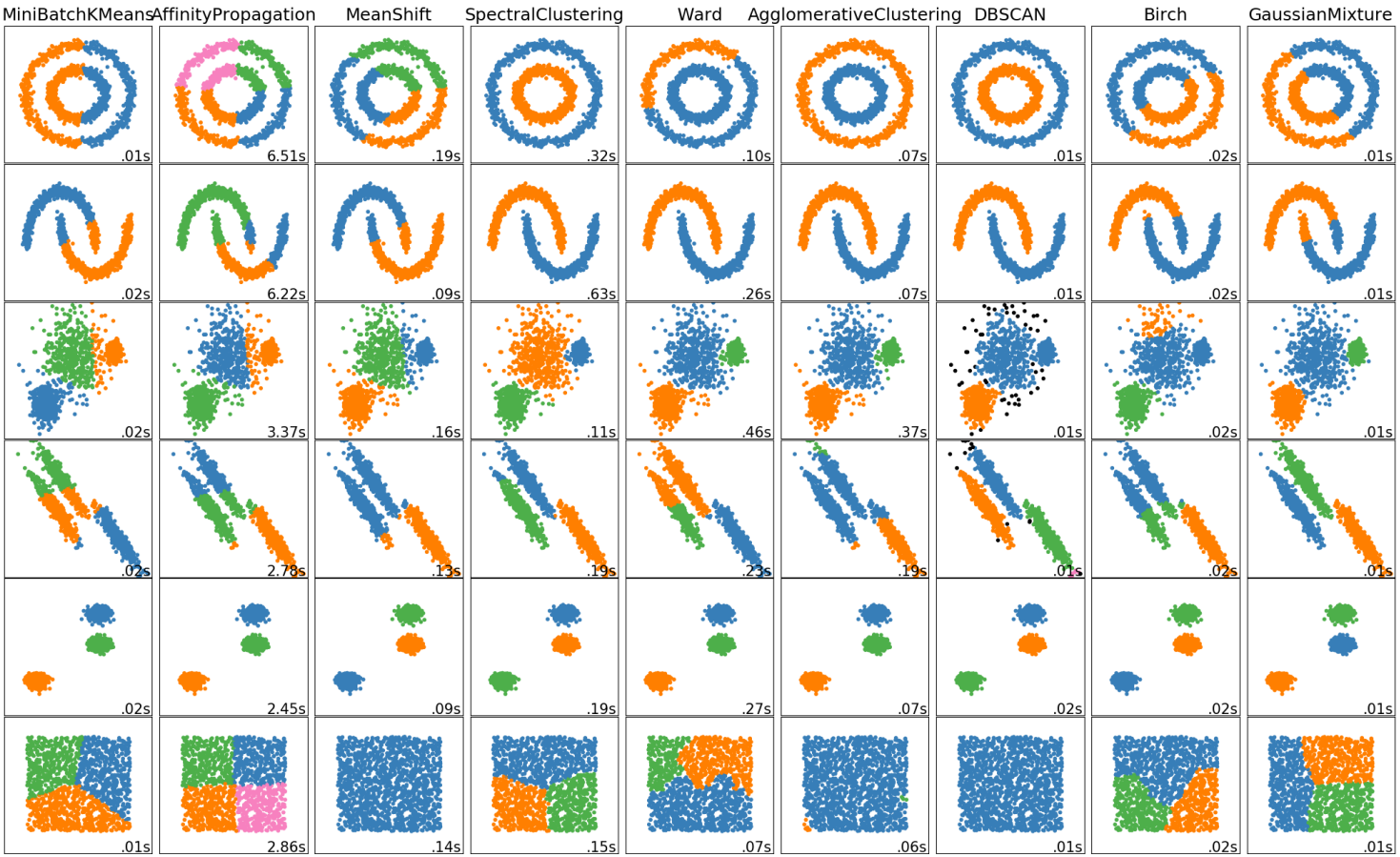
Кластеризация

Задача кластеризации (или обучение без учителя) заключается в следующем. Имеется обучающая выборка $X^l = x_1, .., x_l \subset X$ и функция расстояния между обьектами $\rho(x, x')$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а обьекты разных классов существенно различались. При этом каждому обьекту $x_i \in X^l$ приписывается метка (номер) кластера y_i.

Кластеризация

- Сгущения: внутрикластерные расстояния, как правило меньше межкластерных;
- Ленты: для любого обьекта найдется близкий ему обьект того же кластера, в то же время существуют не близкие другу другу обьекты кластера;
- Кластеры с центром
- Кластеры могут соединяться перемычками;
- На фоне кластеров могут присутствовать лишние обьекты;
- Перекрытие кластеров;
- Кластеры образованные по спцифическим признакам;
- Отсутствие кластеров;

Кластеризация. Иллюстрация.



Области применения кластеризации

- Биология и биоинформатика: группировка сложных геномных последовательностей, выделение пространственных и временных сообдеств;
- Маркетинг: выделение типичных групп покупателей, создание персоанльных предложений;
- Интернет: группировка веб-сайтов по смысловым значениям поисковых запросов;
- Компьютерные науки: определение границ - распознавание обектов;

Математические обозначения

Пусть задано множество обьектов X и множество допустимых ответов Y. Исследуемый процесс описыватся целевой функцией $y^* : X \to Y$, задано конечное множество обьектов $\{x_1, .., x_l\} \subset X$. Пары обьектов $X^l = (x_i, y_i)_{i=1}^l$ представляют собой обучающую выборку.

Основная задача состоит в восстановлении зависимости y^*, а именно в построенни **решающей функции** $a : X \to Y$, причем на всем множестве X.

Решаются следующие вопросы

- как задаются обьекты и какими могут быть ответы (представление данных);
- в какомы смысле $a(x)$ приближает $y(x)$ (выбор нормы);
- правило постороения $a(x)$;

Пространство признаков

- $\mathbb{Y} = \{0, 1\}$ бинарная классификация;
- $\mathbb{Y} = \{1, .., K\}$ - многоклассовая классификация;
- $\mathbb{Y} = \{0, 1\}^K$ - многоклассовая классификация с пересекающимися классами;

Методика обучения

В качестве **модели алгоритма** рассмотрим функцию $g(x, \theta) : X \times \Theta \to Y$, где Θ множество допутимых значений параметра.

$$g(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x), \; Y = \mathbb{R}$$

$$g(x, \theta) = \text{sign} \sum_{j=1}^{n} \theta_j f_j(x), \; Y = \{-1, +1\}$$

Метод обучения - это отображение $\mu : (X \times Y)^l \to A$, которое произвольной коеченой выборке ставит некоторый алгоритм.

Функционал качества

Функционал качества

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(a, x_i).$$

Разновидности функции потерь (loss function):

- $\mathcal{L}(a, x_i) = [a(x) \neq y^*(x)]$ - индикатор ошибки,
- $\mathcal{L}(a, x_i) = |a(x) - y^*(x)|$ - отклонение от правильного ответа;
- $\mathcal{L}(a, x_i) = (a(x) - y^*(x))^2$ - квадратичная функция потерь;

Иными словами задача восстановлени регрессии есть не что иное, как метод наименьших квадратов

$$\mu(X^l) = \arg\min_{\theta} \sum_{i=1}^{l} (g(x_i, \theta) - y_i)^2,$$

Принцип максимума правдоподобия

Вместо модели алгоритма $g(x, \theta)$, аппроксимирующей $y^*(x)$ зададим модель совместной плотности распределения обьектов и ответов $\varphi(x, y, \theta)$.

В случае независимой последовательности наблюдений $p(X^l) = p(x_1, y_1)..p(x_l, y_l)$, получаем функию правдоподобия

$$L(\theta, X^l) = \prod_{i=1}^{l} \varphi(x_i, y_i, \theta).$$

Задача найти θ при котором функция правдоподобия максимальна. [Пример с нормальным распределением и средней дисперсией]

Оценки обучающей способности

- Эмпирический риск на тестовых данных

$$HO(\mu, X^l, X^k) = Q(\mu(X^l), X^k) \to \min,$$

- Кросс-проверка, $L = l + k$, $X^L = X_n^l \sqcup X_n^k$:

$$CV(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^l), X_n^k) \to \min,$$

- Эмпирическая оценка вероятности переобучения:

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \Big[Q(\mu(X_n^l), X_n^k) - Q(\mu(X_n^l), X_n^l) \geq \varepsilon \Big] \to \min$$

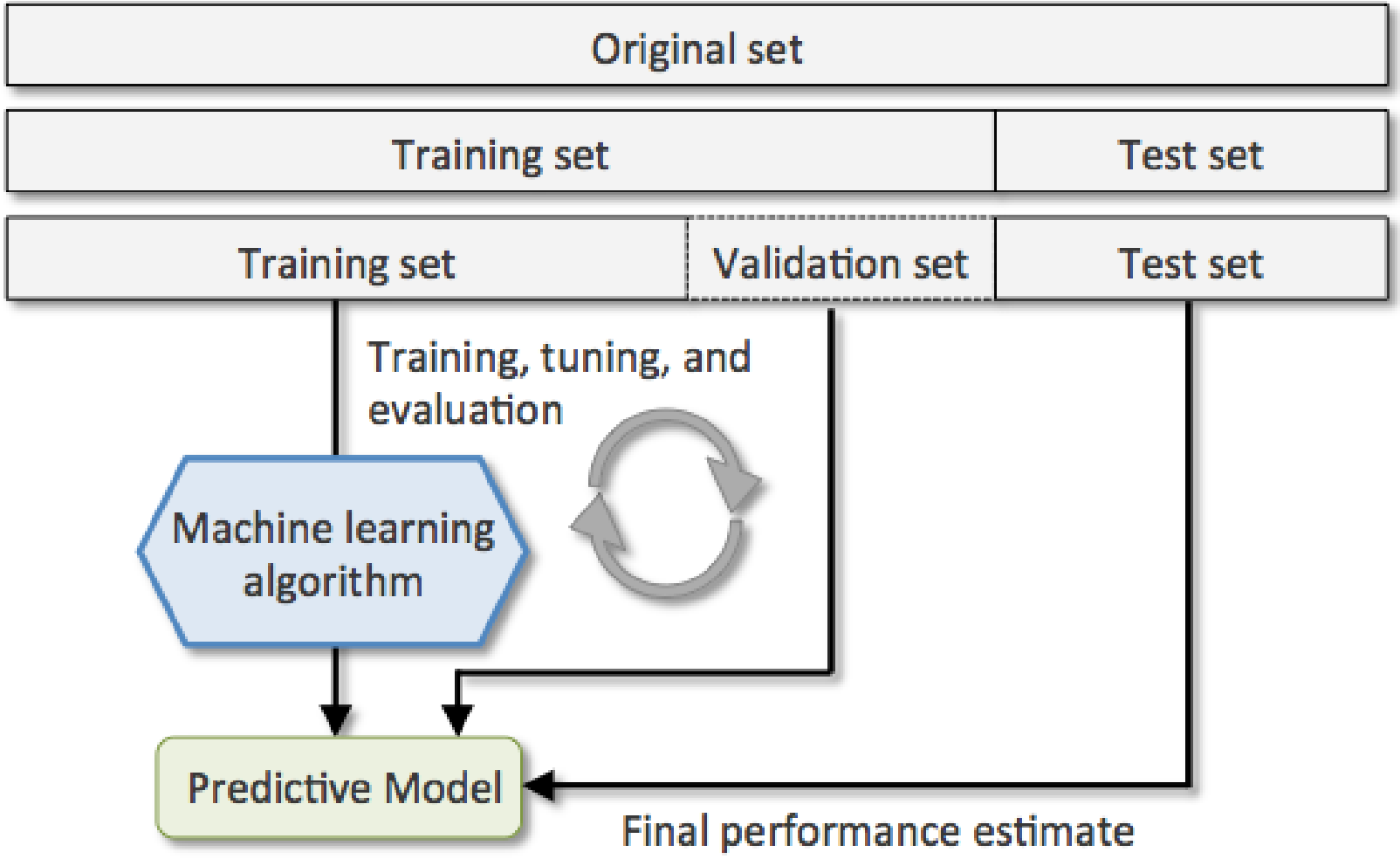
Кросс-проверка

Выборку $X^L = (x_i, y_i)_{i=1}^l$ разобьем N различными способами.

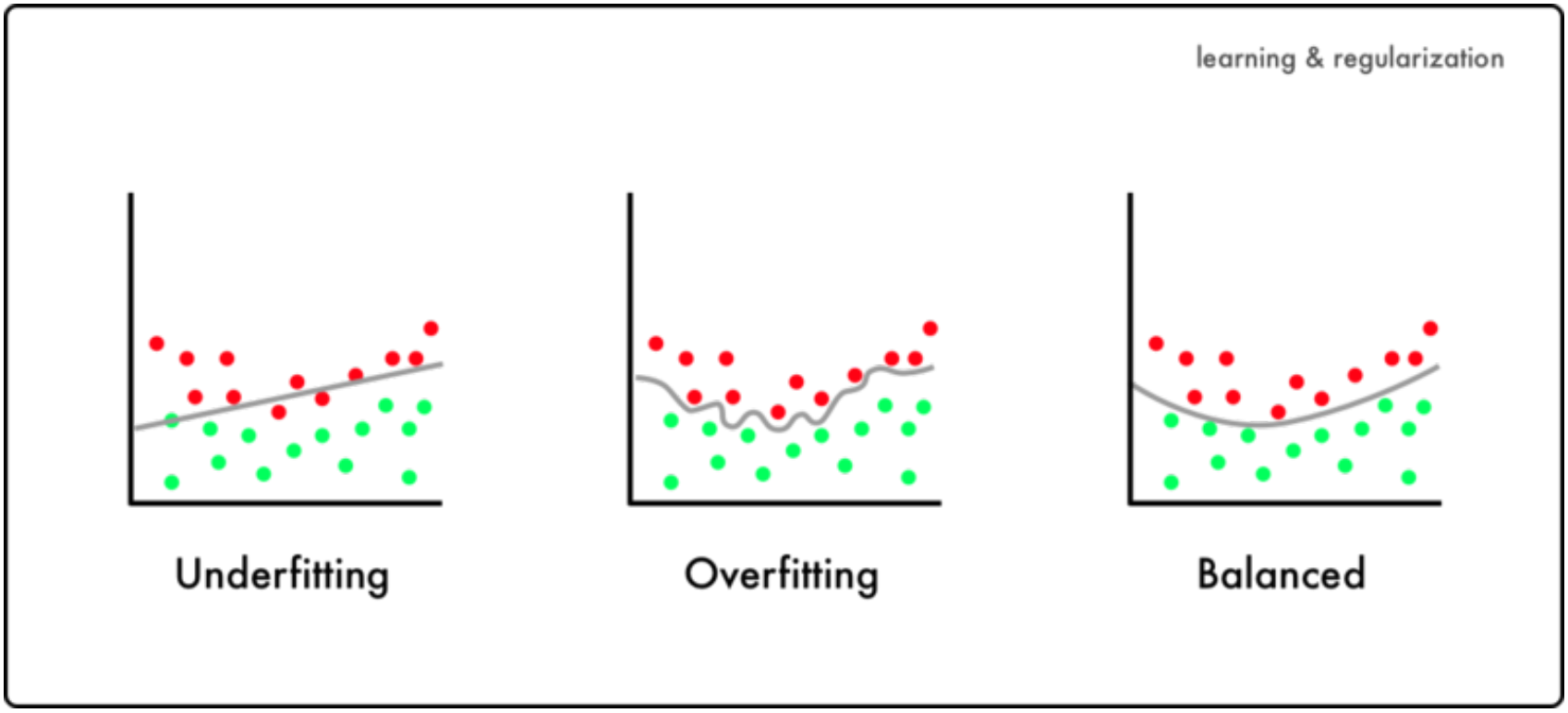
Для каждого разбиение $n = 1..N$ построим алгоритм $a_n = \mu(X_n^l)$ и вычислим среднее значение ошибки $Q_n = Q(a_n, X_n^k)$. Среднее арифитическое невязок всех указанных алгоритмов называется **скользящим контролем**

$$CV(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{l} Q(\mu(X_n^l), X_n^l).$$

Скользящий контроль. Иллюстрация.



Переобучение



Еще раз регрессия

Задача линейной регрессии $g(x, \alpha) = \sum_{j=1}^{n} \alpha_j x_j$ сводится к следующей системе уравнений

$$F^T F \alpha = F^t y,$$

так что решение $\alpha^* = F^+ y$ ищется через псевдообратную матрицу $F^+ = (F^T F)^{-1} F^T$.

Мультиколлинеарность

В случае когда $\Sigma = F^T F$ имеет полный ранг, но близка к матрице с неполным рангом. Обьекты выборки сосредоточены вблизи линейного подпространства меньшей размерности. Или малеенькие собств значения. Определим число обусловленности

$$\mu(\Sigma) = \frac{\max_{||u||=1} ||\Sigma u||}{\min_{||u||=1} ||\Sigma u||} = \frac{\lambda_{max}}{\lambda_{min}},$$

происходит увеличение погрешности при умножении на обратную матрицу

$$\frac{||\partial z||}{||z||} \leq \frac{||\partial u||}{||u||}.$$

Сглаживающий функционал (гребневая регрессия, L2 регулрязиация), эффективная размерность, выбор константы регуляризации, L1 - регуляризация (лассо тибширани).

ДЗЗ изучения растительного покрова

Особое место среди попиксельной классификации занимают методы машинного обучения: классификация с использованием самоорганизующихся нейронных сетей, деревьев решений, опорных векторов. **Нейронные сети** более всегод подходят для обьектов с нечеткими границами.
Деревья решений здесь в отличие от "черного ящика"нейронных сетей можно указать, какие показатели использваны для разделения на классы (так называемые "ключи")

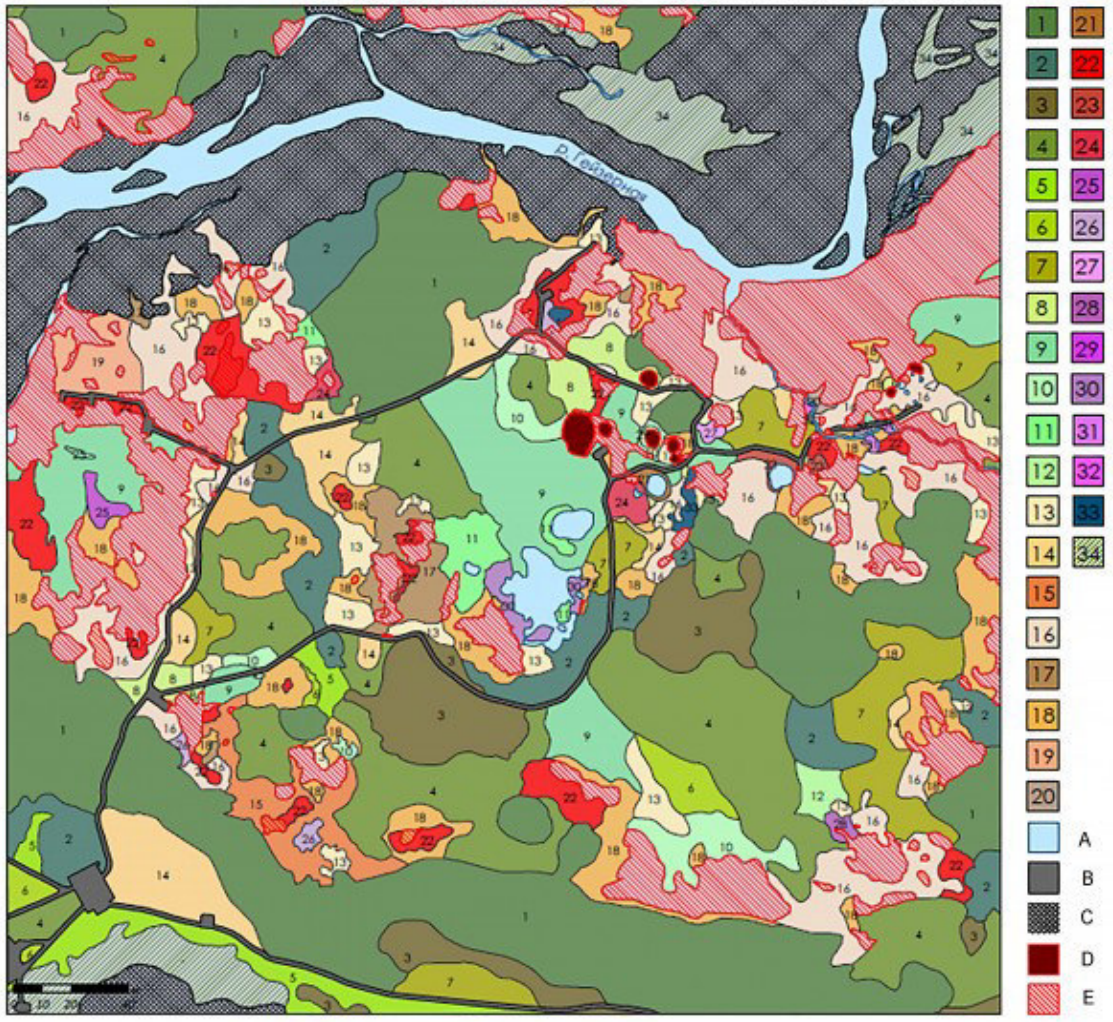
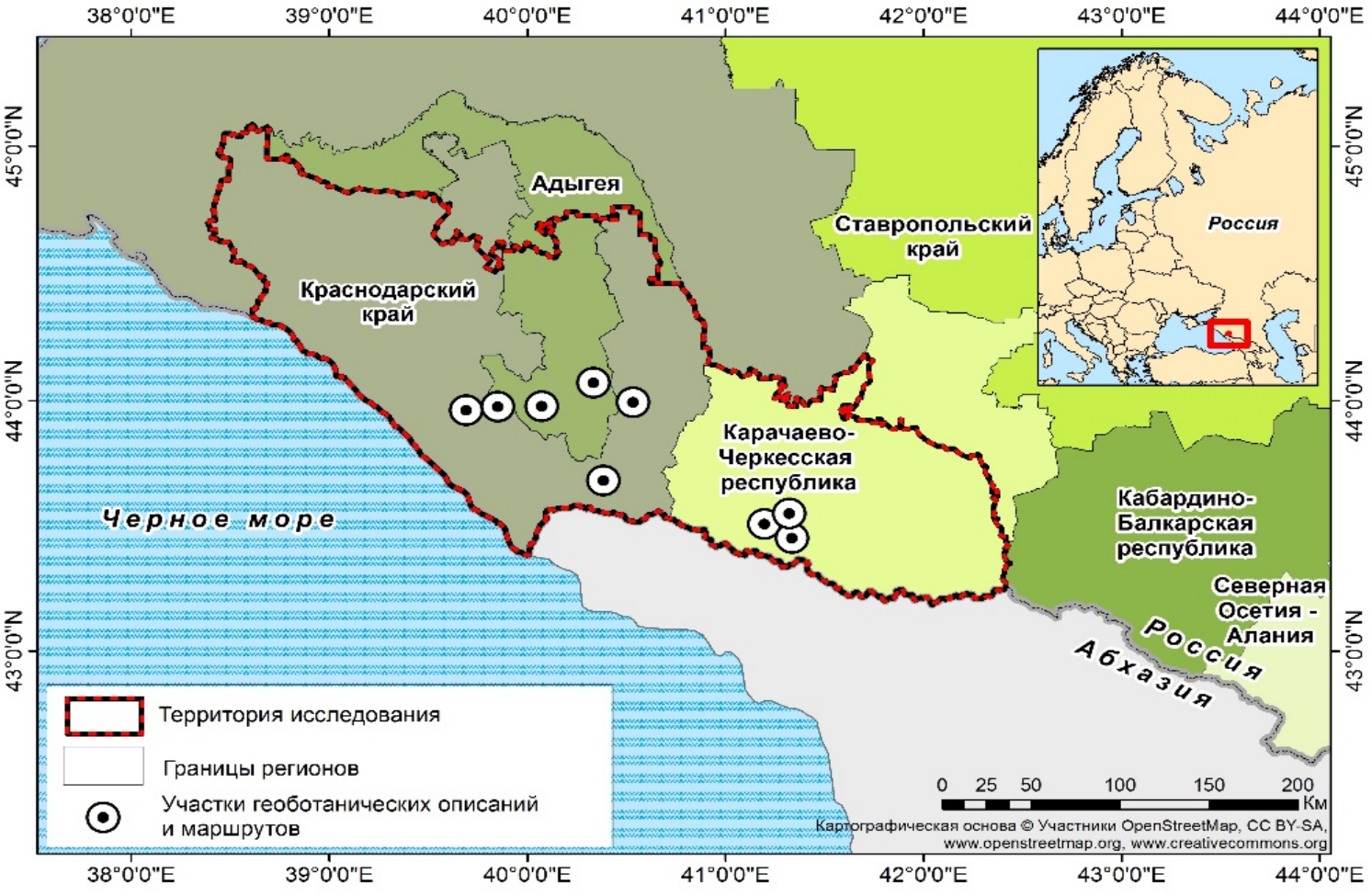


Рис. 1. Растительность Долины р. Гейзерной:
1 – каменноберезняк разнотравный, 2 – ивняк шеломайниковый, 3 – сообщества ольхового стланика, 4 –крупнотравно-шеломайниковые луговые сообщества, 5 – высокотравные луговые сообщества с преобладанием лабазника и бодяка, 6 – высокотравные луговые сообщества с преобладанием вол-


38°0'0"E
39°0'0"E
40°0'0"E
41°0'0"E
42°0'0"E
43°0'0"E
44°0'0"E
45°0'0"N
44°0'0"N
43°0'0"N
Адыгея
Краснодарский край
Ставропольский край
Карачаево-Черкесская республика
Кабардино-Балкарская республика
Северная Осетия - Алания
Черное море
Россия
Абхазия
Россия
Территория исследования
Границы регионов
Участки геоботанических описаний и маршрутов
0 25 50 100 150 200
Км
Картографическая основа © Участники OpenStreetMap, CC BY-SA, www.openstreetmap.org, www.creativecommons.org

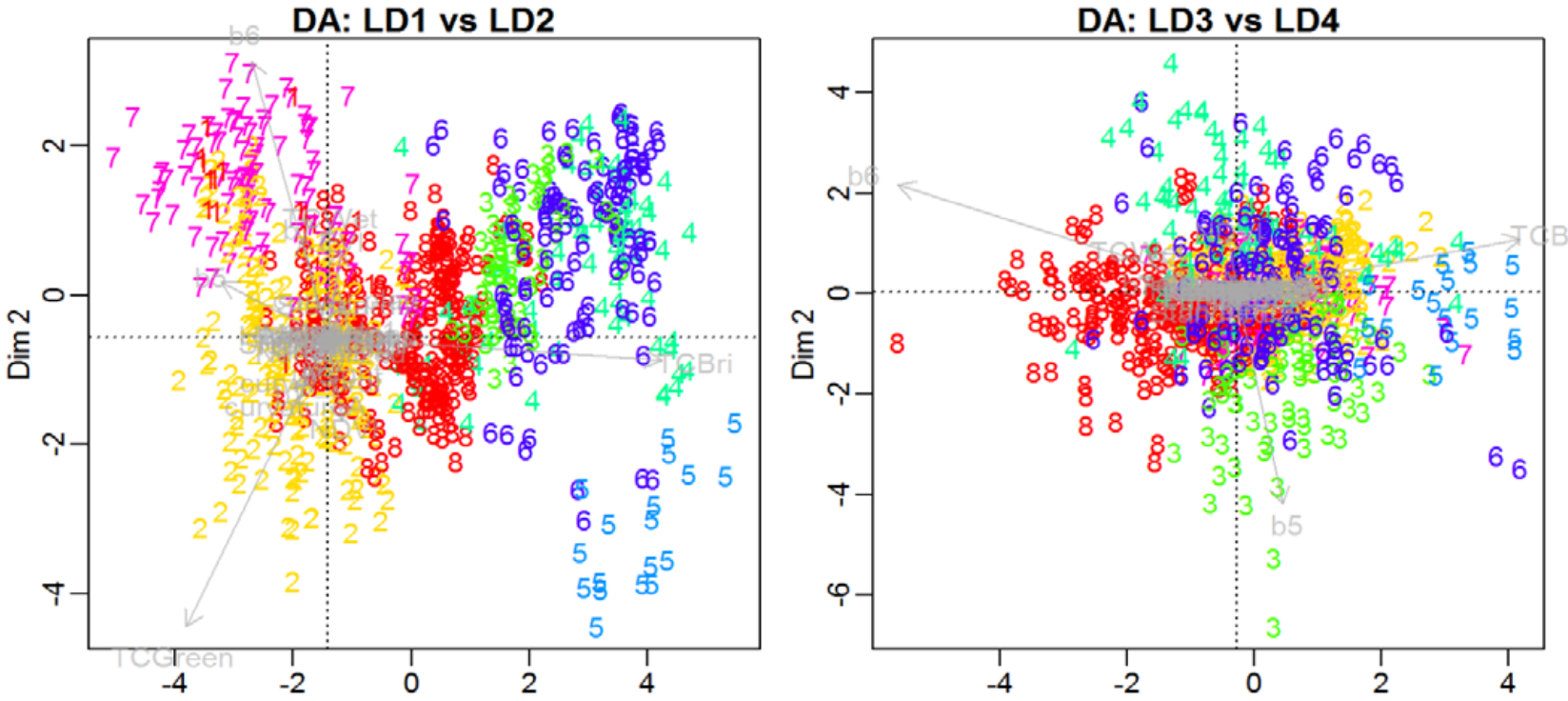


Рисунок 5. Варианты темнохвойных лесов хорошо разделяются в пространстве спектральных и морфометрических признаков методом дискриминантного анализа (цифрами обозначены синтаксоны в соответствии с таблицей 1)

Таблица 1. Сообщества темнохвойных лесов Северо-Западного Кавказа. В правом столбце – нумерация синтаксонов, используемая в дальнейшем обсуждении

Cl. ***Carpino-Fagetea sylvaticae*** Jakucs ex Passarge 1968	
Or. ***Rhododendro pontici - Fagetalia orientalis*** (Soó 1964) Pass. 1981	
Al. ***Abieti-Fagion orientalis*** Korotkov et Belonovskaja 1987	
Subal. ***Abieti-Fagenion orientalis*** Korotkov et Belonovskaja 1987	
Ass. ***Ilici colchicae – Abietetum nordmannianae*** Korotkov et Belonovskaja 1987	
Subass. ***I. c. – A. n. vaccinietosum arctostaphyli*** Belonovskaja et al 1990	2
Subass. ***I. c. – A. n. rhododendretosum pontici*** Belonovskaja et al 1990	1
Ass. ***Abieti – Fagetum orientalis*** Korotkov et Belonovskaja 1987	
Subass. ***A. – F. o. piceetosum orientalis*** Belonovskaja et al 1990	
Var. ***Luzula sylvatica***	3
Var. ***Calamagrostis arundinacea***	4
Var. ≡ Ass. ***Petasito albae-Abietetum nordmannianae*** Frantsuzov 2006	5
Var. ***Salvia glutinosa***	6
Subass. ***A. – F. o. athyrietosum filici-feminae*** Belonovskaja et al 1990	
Var. ***Galeobdolon luteum***	7
Var. ***Sanicula europaea***	8

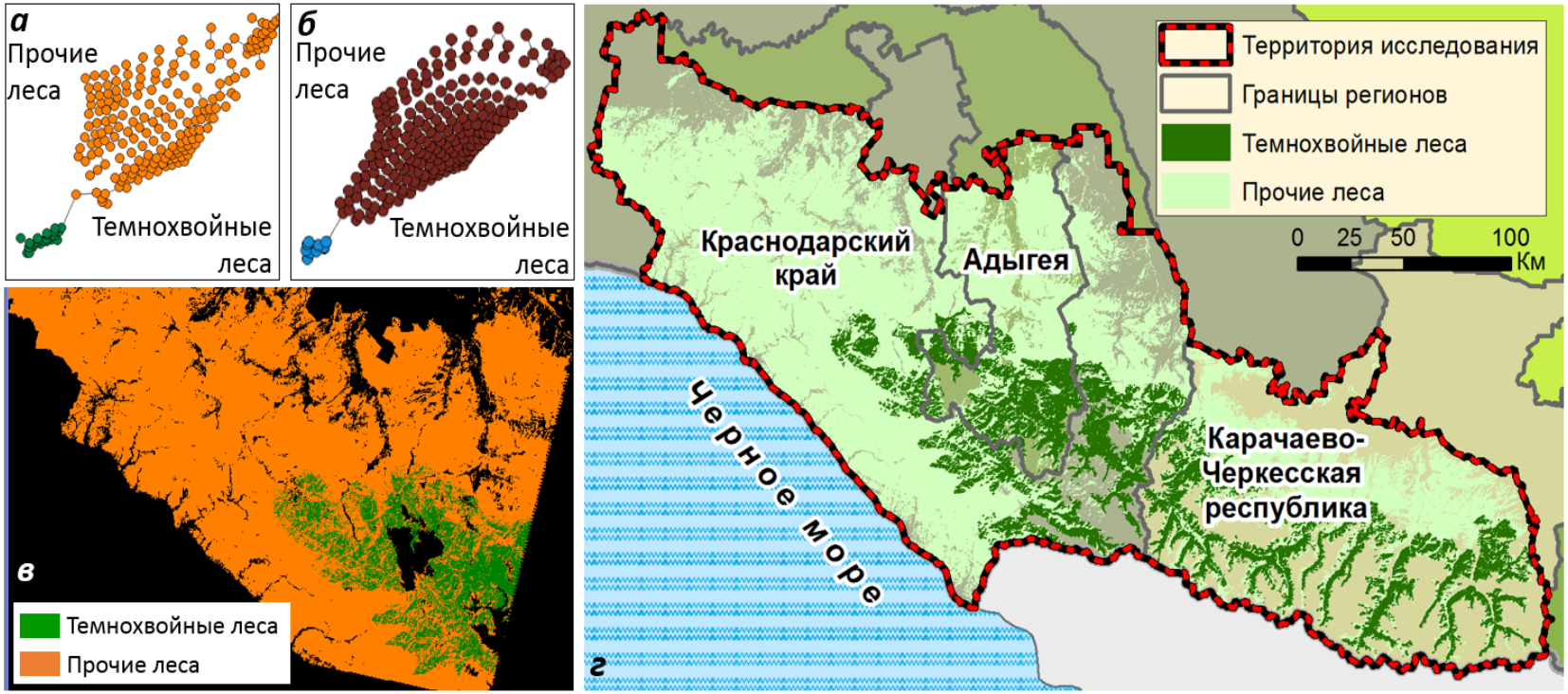


Рисунок 8. Картографирование темнохвойных лесов. а, б – структура нейронных сетей, построенных независимо по разным снимкам; в – результат классификации снимка нейронной сетью; г – конечный результат

Заключение

1. Анализ предметной области;
2. Коррекция данных, формализация признаков;
3. Построение прямой мат модели;
4. Подбор невязки, решение задачи оптимизации;
5. Поиск оптимального решения, регуляризация решения;
6. Оценки качества полученной модели;
7. Тестирование модельное и аппробация;