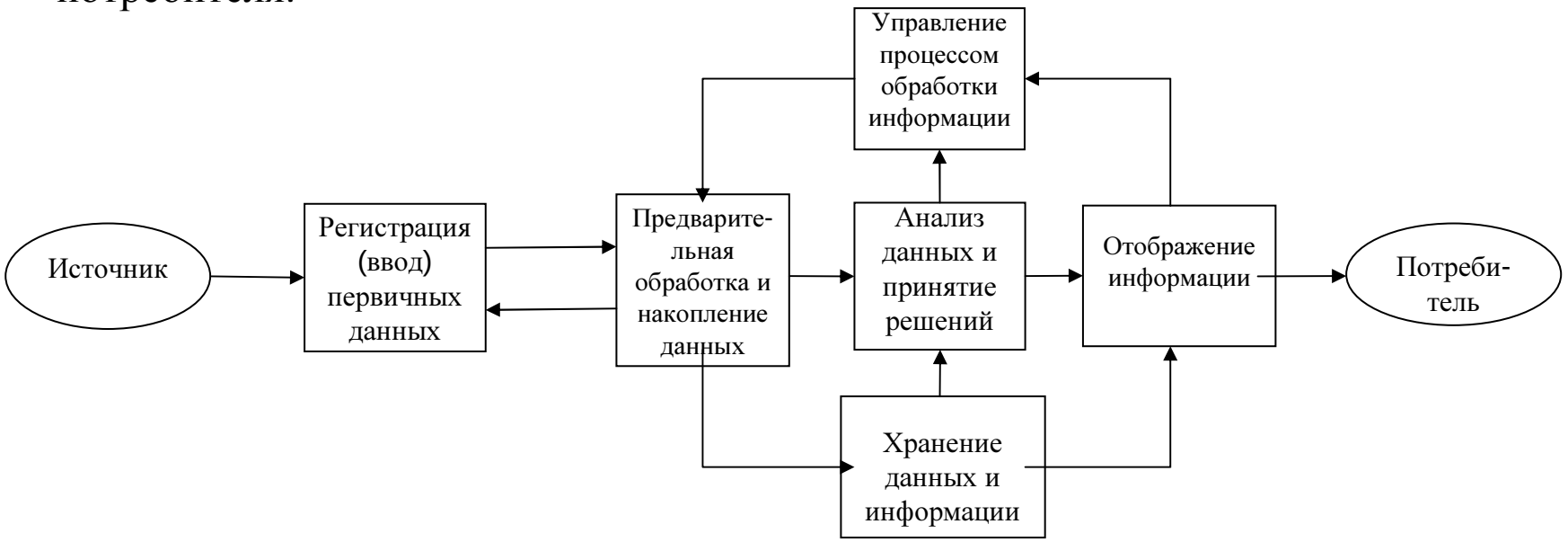
Лекция 1. Обработка информации, анализ данных, машинное обучение. Основные понятия и определения

- **Данные** – факты, характеризующие объекты, явления или процессы некоторой предметной области и зафиксированные на каком-либо материальном (физическом) носителе в виде необработанных результатов измерений и наблюдений.
- **Замечание:** главной особенностью данных является их первичный характер и их востребованность потребителем (пользователем): они всегда получаются (добываются) целенаправленно и являются «сырьем» для последующего анализа и обработки.
- **Информация** – сведения об объектах, явлениях или процессах некоторой предметной области, получаемые путем анализа и обработки данных и пригодные для принятия решений потребителем.
- Информацию можно рассматривать как «вторичные», преобразованные данные, имеющие определенный уровень обобщения и смыслового содержания.
- **Обработка информации (ОИ)** – систематическое выполнение операций над данными, представляющими предназначенную для обработки информацию **(ГОСТ 15971-90)**

- **Сигнал –** определенным образом структурированный физический носитель данных и информации при их передаче от одного объекта к другому или от объекта к потребителю.
- **Замечание:** физическая природа сигналов может быть различной (электромагнитное излучение, акустическое или механическое воздействие и т.п.).
- **Информационная система (ИС)** – сложная человеко-машинная система, целевое назначение, элементный состав и структура которой ориентированы на различного рода преобразования данных и информации в интересах обеспечения потребностей пользователей (физических лиц, организаций, органов управления и т.п.).
- **Информационный процесс** – целенаправленно организованный процесс изменения информационных состояний системы, в ходе которого осуществляется последовательность операций по преобразованию информации, в результате которой она может изменять свою форму и/или содержание в пространстве и/или во времени.
- **Информационная технология** (как способ реализации ИП) – совокупность приемов, способов, мероприятий, обеспечивающих организацию и реализацию данного информационного процесса с использованием компьютерной техники, средств сетевого взаимодействия, других технических средств, а также программных средств и информационных ресурсов.

- **Технология обработки информации –** совокупность приемов, способов, мероприятий, обеспечивающих организацию и реализацию процессов извлечения информации о состоянии объектов при бесконтактном или контактном взаимодействии с использованием компьютерной техники, средств сетевого взаимодействия и других технических средств, а также программных средств и информационных ресурсов.
- **Операции преобразования данных (информации) –** элементарные действия в каждом процессе, являющиеся типовыми звеньями в общей последовательности выполняемых изменений информационных состояний системы.
- Для реализации каждой операции используется конкретный **алгоритм обработки данных (информации),** а для реализации информационного процесса в целом **– алгоритм функционирования информационной системы.**
- **Алгоритм –** инструкция о том, в какой последовательности нужно выполнить действия при переработке исходного материала в требуемый результат (в данном случае таким материалом является информация).
- Алгоритм всегда определяет представленную в стандартной форме совокупность предписаний, понятных человеку и/или компьютеру, детально описывающая последовательность действий, направленных на достижение конкретного результата.

- **Система обработки информации (СОИ)** - совокупность технических средств и программного обеспечения, а также методов обработки информации и действий персонала, обеспечивающая выполнение автоматизированной обработки информации (системы технической диагностики, системы аналитической отчетности, OLAP- системы и т.п.).
- **Целевое назначение СОИ –** проведение анализа и обобщения большого объема данных, получаемых от источника, на основе решения вычислительной задачи, либо задачи принятия информационных решений.
- В результате решения задачи формируется новая информация относительно прошлого, текущего или будущего состояния объекта, которая интересует потребителя.



- **Анализ данных** – область науки, занимающаяся построением и исследованием математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (наблюдаемых) данных; процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.
- Анализ данных имеет несколько аспектов, основан на применении множества и подходов и охватывает разнообразные методы в различных областях науки и деятельности.

Отражением этого является то, что часто используются ряд близких и сопутствующих понятий. К ним относятся такие понятия, как *искусственный интеллект, машинное обучение, Data Mining.*

- **Машинное обучение** (machine learning) - методология получения новых знаний на основе алгоритмов и программ, реализованных с использованием современных средств вычислительной техники.
- **Машинное обучение** – наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время своей работы» (Т.М. Митчелл)

Наше понимание терминологии

- **Обработка информации** - наиболее широкое понятие, определяющее все возможные типы преобразований, систематически выполняемых над данными первичных измерений и наблюдений в интересах автоматизации решения задач в науке, технике и технологии.
- **Анализ данных** - более узкое понятие, отражающее такие преобразования, которые непосредственно направлены на интеллектуальную обработку первичных измерительных данных в интересах извлечения знаний об объектах и формирования информации (вторичных обработанных данных), пригодной для принятия решений.
- **Машинное обучение –** базовый подход, реализуемый в ходе анализа данных, который основан на построении математических моделей объектов в контексте решаемой задачи и синтезировании на этой основе компьютерных алгоритмов принятия решений, обладающих способностью к совершенствованию.
- **Замечание**: для анализа данных не всегда необходимо прибегать к машинному обучению, так как модель объекта, используемая для построения алгоритма принятия решения, может быть изначально задана или постулирована.

Основные задачи анализа данных

1.Задача распознавания (классификации).

2.Задача кластеризации.

3. Задача оценивания.

4.Задача регрессии.

5.Задача установления ассоциаций.

Вспомогательные задачи анализа данных

1. Задача отбора информативных признаков.
2. Задача визуализации

Классификация задач

- Задачи АД делят на **описательные (descriptive)** и **предсказательные (predictive).**
- **Описательные задачи** (кластеризация, ассоциация) основное внимание уделяют пониманию обрабатываемых данных.
- **Предсказательные задачи** (классификация, регрессия и оценивание) направлены на предсказание состояний объектов.
- Задачи АД также делят на задачи, реализующие **обучение с учителем** (*supervised learning*), и задачи, реализующие **обучение без учителя или самообучение** (*unsupervised learning*).

Терминологий необходимая для постановки задач АД

•**Наблюдения** – первичные данные, отражающие характеристики объектов, получаемые в процессе измерения и регистрации параметров сигналов (любых носителей информации) на основе используемых в СОИ физических датчиков в ходе бесконтактного или контактного информационного взаимодействия. Наблюдения, как правило, организованы в виде вектора

$$z = (z_1, ..., z_q)^T \in R^q$$

•**Признаки (дескрипторы)** – особым образом отобранные первичные данные (наблюдения) или данные, полученные в результате их преобразования, которые используются для составления информативного описания объекта с точки зрения решения конкретной задачи обработки информации. Признаки, как правило, организованы в виде вектора признаков (дескрипторов)

$$x = (x_1, ..., x_n)^T \in R^n$$

•**Замечание:** обычно размерность пространства признаков существенно меньше, чем размерность пространства наблюдений, хотя не исключается ситуация, когда наблюдения и признаки совпадают .

• **Задача отбора информативных признаков** является отдельной задачей, которая может решаться на различных этапах процесса обработки информации.

- **Состояния** -– совокупность непосредственно не наблюдаемых параметров объекта, характеризующих интересующие нас существенные свойства объекта, присущие ему в текущий момент времени и, возможно, изменяющиеся в другие моменты времени. В процессе информационного взаимодействия ненаблюдаемые параметры объекта могут быть представлены в виде вектора

$$s = (s_1, \ldots, s_m)^T \in \mathbf{R}^m$$

- **Замечание:** состояние объекта часто рассматривается как некоторое сообщение, которое требуется извлечь в процессе обработки наблюдений, тем или иным образом связанных с состояниями.
- **Образ –** формализованное описание конкретного объекта в пространстве используемых признаков $x = (x_1, \ldots, x_n)^T \in R^n$
- **Класс образо**в **–** совокупность объектов, имеющих определенное сходство, общие свойства, проявляющееся при их описании в виде образов и, соответственно, отличающихся по эти свойствам от объектов, включаемых в другие классы.
- **Множество классов** будем далее обозначать $\Omega = \{\omega_1, \ldots, \omega_M\}$
- **Замечание**: обозначение $x \in \omega_i$ свидетельствует о том, что данный образ принадлежит *i* - му классу.

- **Эталонное описание класса** – априорные сведения и характеристики класса в пространстве используемых признаков, опирающиеся на используемую для анализа математическую модель объектов данного класса.
- **Обучение –** процедура получения эталонных описаний классов в рамках системы выбранных признаков, основанная на использовании совокупности обучающих (опытных) данных и априорных сведений относительно физической природы анализируемых объектов.
- **Замечание**: в широком смысле обучение можно охарактеризовать как получение математической модели объектов, описывающей закономерности данных.
- **Принятие решения** – процедура отнесения объекта, представленного своим образом, к заданному классу или процедура определения неизвестных параметров объекта (состояний), представленного наблюдениями, основанная на использовании ранее полученной математической модели класса объектов в контексте решаемой задачи обработки информации.
- **Замечание**: в основе принятия решений всегда лежит решающее правило, или алгоритм принятия решений, определяющий конкретную последовательность преобразования входных данных в итоговый результат.

- **Обучающая выборка** – совокупность реализаций (прецедентов) вектора признаков, описывающая конкретное множество объектов в системе этих признаков и используемая при обучении

$$X^N = \{x^{(1)},...,x^{(N)}\},\ x^{(i)} = (x_1^{(i)},...,x_n^{(i)})^T \in \mathbf{R}^n$$

- **Выделяют ситуацию**, когда обучающие данные изначально проиндексированы (помечены), т.е. смешанной совокупности X^N дополнительно придается совокупность индексов $D^N = \{d^{(1)},...,d^{(N)}\}$, $d^{(i)} \in D$, где D^N - конечное множество целых чисел , определяющих принадлежность объектов конкретным классам.
- **В другой ситуации** задается обучающая выборка, в которой смешанной совокупности X^N дополнительно придается совокупность «ответов» $Y^N = \{y^{(1)},...,y^{(N)}\}, y^{(i)} \in \mathbf{R}^h,\ i = \overline{1,N}$, каждый из которых является вещественным вектором, т.е. принадлежит конечномерному вещественному векторному пространству (например, евклидову).

- **В задачах *unsupervised learning*** при обучении D^N или Y^N изначально не заданы и имеется только совокупность входных данных X^N без указания принадлежности образов конкретных классом объектов или определения необходимого ответа.

- **Тестирующая выборка** – совокупность реализаций вектора признаков, описывающая конкретное множество объектов в системе признаков и используемая для проведения тестирования (контрольного эксперимента) и оценки качества ранее полученного (синтезированного) алгоритма обработки информации.
- **Замечание: для тестирующей выборки** данные обязательно должны быть либо проиндексированы (помечены), с помощью совокупности индексов, определяющих реальную принадлежность каждого объекта одному из возможных классов, либо должны быть снабжены контрольными ответами в виде вещественных переменных (в противном случае качество используемого алгоритма невозможно оценить).

- **Опираясь на введенную терминологию, можно перейти к более детальному определению содержания ранее перечисленных типовых задач обработки информации.**

- **Задача распознавания** (классификации, различения, узнавания) состоит в отнесении некоторого объекта, описываемого совокупностью характеристик, или признаков, определяющих «образ» этого объекта, к одному из ранее выделенных классов.
- **При постановке задачи** распознавания считается, что количество классов и их эталонные описания известны или могут быть определены в процессе обучения.
- **Во многих случаях** эталонное описание классов может быть получено без обучения на основе постулирования математической модели образов. Например, в задачах распознавания радиосигналов часто используются вероятностные модели, основанные на использовании гауссовского распределения шумов, описывающих вариации анализируемых образов в пределах каждого класса.
- **В других случаях**, перед тем, как система сможет выполнять данную функцию, предполагается её обучение на множестве прецедентов обучающей выборки объектов распознавания, в результате чего формируются необходимые эталонные описания.
- **Замечание**: вносится следующее уточнение: распознавание с обучением на основе имеющегося множества прецедентов называют классификацией с обучением (с учителем).

- **Задача кластеризации** состоит в разбиении некоторого множества объектов, представленных своими образами, на классы (кластеры, группы), в каждый из которых помещаются в известном смысле (по степени сходства) «близкие» образы, в то время как образы, помещаемые в различные классы, имеют существенные «отличия».

- **Разделяют случаи**, когда число классов известно, и случай, когда числа классов неизвестно. Последняя задача еще называется автоматической классификацией, или классификацией без обучения.

- **Задача классификации** при не заданном количестве классов является наиболее сложной для решения; высокое качество разбиения образов на классы достигается только при достаточно большей степени различия образов, объективно относящихся к разным классам.

- **Задача оценивания** состоит в определении неизвестных характеристик объекта на основе анализа (обработки) совокупности первичных данных, представленных либо непосредственно как наблюдения, либо в виде признаков (отобранных наблюдений).
- **В качестве оцениваемых характеристик** могут выступать самые различные величины. К ним, например, может относиться один из параметров состояния объекта, либо вектор параметров, характеризующих полное состояние объекта $s \in \mathbf{R}^m$
- **Во многих случаях** задача оценивания направлена на определение неизвестных параметров статистического распределения (плотности распределения вероятностей, функции распределения вероятностей) известного аналитического представления или на определение вида закона распределения в целом, если его аналитическая форма неизвестна.
- Такие задачи называются, соответственно, задачей **параметрического оценивания** и задачей **непараметрического оценивания** статистического распределения.
- В классе задач оценивания выделяют задачу **фильтрации**, которая состоит в оценивании непосредственно ненаблюдаемой реализации векторного случайного процесса (случайной функции времени) или реализации случайного поля (случайной функции двух и более переменных), описывающих изменения состояния объекта во времени или пространстве.

- **Задача регрессии** состоит в установлении функциональной зависимости между ожидаемыми значениями зависимых (выходных) переменных и значениями независимых (входных) переменных.
- **Решение задачи** базируется на использовании особым образом организованной обучающей выборки данных, в которой задается совокупность значений вектора независимых переменных $X^N = \{x^{(1)}, \ldots, x^{(N)}\}$ и соответствующая ей совокупность значений зависимой переменной – «ответов» обучающей выборки $Y^N = \{y^{(1)}, \ldots, y^{(N)}\}$, где каждый элемент определен на множестве вещественных чисел.
- **В более широкой постановке** может рассматриваться несколько зависимых переменных, которые объединяются в вектор.
- **Полученная функциональная зависимость** используется для предсказания ожидаемого значения зависимой переменной для нового набора независимых (входных переменных).
- **Замечание**: если значение зависимой переменной определяется по отношению к будущему моменту времени, которое присутствует в перечне компонентов вектора независимых переменных, то задача регрессии называется **задачей прогнозирования.**

- **Задача ассоциаций** состоит в нахождении значимых связей (ассоциаций) между объектами или событиями.
- **Ассоциации** представляются в виде определенных правил, которые могут быть использованы для объяснения природы анализируемых процессов и предсказания появления новых событий.
- **Поиск ассоциативных правил** является одним популярных приложений Data Minning, при реализации которого ищутся часто встречающиеся наборы объектов среди большого множества таким наборов.
- **Замечание:** по сути, данная задача является частным случаем задачи классификации. Поэтому в дальнейшем, в связи с ограниченным объемом книги, данная задача отдельно рассматриваться не будет.

- **Общее замечание:** с точки зрения понимания изложенного материала и, прежде всего, условности введенных понятий и определений, весьма важно подчеркнуть связь между различными задачами анализа данных.
- **В системах обработки информации** эти задачи часто решаются в комплексе и образуют цепочку функциональных операций, выполняемых как последовательно, так и параллельно, и обеспечивающих достижение общего результата.

- **Принципиальным моментом** при обосновании алгоритмов анализа данных в СОИ является выбор подходящего математического аппарата и используемого метода для решения той или иной задачи, реализация наиболее удобной технологии, объединяющей различные подходы и приемы выполнения базовых операций преобразования информации.
- В этом плане обычно выделяют два базовых подхода: **статистический и детерминистский**.
- **Статистический подход** основан на использовании для анализируемых объектов и процессов вероятностных моделей данных: плотностей и функций распределения вероятностей, моментов распределений, достаточных статистик.
- **Алгоритм анализа данны**х в данном случае реализует статистическое решающее правило.
- **Качество или эффективность** обработки определяется с использованием статистических характеристик (вероятностей ошибок при принятии решений, дисперсий или матриц ковариации ошибок и т.п.).
- **В рамках статистического подхода** реализуются различные методы решения задач, которые условно можно разделить на два класса: байесовские и небайесовские.

- **Детерминистский подход** определяет группу разнообразных подходов и методов, в основе которых лежит отказ от использования вероятностных моделей данных. К ним можно отнести: метрический подход, нейросетевой подход, структурно-лингвистический подход и др.
- **Метрический подход** базируется на использовании при построении алгоритмов АД различных метрик (функций расстояния), определяющих степень «близости» и «различия» анализируемых данных в многомерных пространствах, а также – на выдвижении определенных гипотез относительно свойств анализируемых объектов, представленных точками в многомерных пространствах (гипотеза компактности и однородности классов образов).
- **Нейросетевой подход** предполагает представление алгоритма АД в виде универсального преобразователя информации «вход» – «выход». В основе архитектуры преобразователя лежит использование ансамбля однотипных элементарных вычислительных элементов – искусственных нейронов,. Применение нейросетевых методов и алгоритмов анализа данных в СОИ предполагает обучение (тренировку) некоторой универсальной вычислительной среды для решения поставленной задачи.
- **Структурно-лингвистический подход** основан на использовании так называемых непроизводных элементов (подобразов) в качестве признаков (наблюдений) и анализе их отношений. Для этого формируется «грамматика» образов в виде иерархической структуры подобразов как лингвистических элементов.
- Это позволяет применять для решений задачи обработки информации теорию формальных языков.

Вопросы?