1. Уменьшение размерности (dimensionality reduction) в машинном обучении - это процесс сокращения количества входных переменных (признаков) в модели, сохраняя при этом максимум полезной информации. Это делается для упрощения модели, повышения её эффективности и предотвращения проблемы избыточности признаков (curse of dimensionality).

2. Методы Feature Selection, используемые для выбора информативных признаков, включают:

* Фильтрующие методы (Filter methods), которые оценивают важность признаков независимо от модели, например, корреляция, информационный критерий, дисперсионный анализ.
* Обёртывающие методы (Wrapper methods), которые используют модель машинного обучения для оценки подмножеств признаков, например, последовательный отбор признаков (forward/backward selection).
* Встроенные методы (Embedded methods), которые совмещают фильтрацию и обёртку, встраивая отбор признаков в процесс обучения модели, например, LASSO, гребневая регрессия.

3. Принцип работы метода Главных Компонент (PCA) заключается в преобразовании исходного набора признаков в новый набор, называемый главными компонентами, которые являются линейными комбинациями исходных признаков. Главные компоненты упорядочены по убыванию объясняемой дисперсии, что позволяет выбрать наиболее информативные из них и уменьшить размерность данных.

4. Главная компонента (Principal Component) - это новый признак, полученный в результате преобразования исходных признаков методом PCA. Главные компоненты являются ортогональными векторами, максимизирующими дисперсию данных.

5. Ансамбли (Ensemble methods) в машинном обучении - это методы, объединяющие несколько моделей (базовых алгоритмов) в одну более сильную модель. Идея заключается в том, что комбинирование нескольких "слабых" моделей может дать более надежный и точный результат, чем использование одной "сильной" модели.

6. Основные ансамблевые алгоритмы:

* Бэггинг (Bagging) - создает множество моделей на разных подвыборках данных и усредняет их предсказания.
* Бустинг (Boosting) - последовательно обучает модели, фокусируясь на ошибках предыдущих моделей.
* Стекинг (Stacking) - использует выходы нескольких базовых моделей в качестве входов для метамодели, обучающей их комбинацию.

7. Метод Random Forest (Случайный Лес) - это ансамблевый алгоритм, использующий бэггинг в комбинации с деревьями решений. Он обучает множество деревьев решений на различных подвыборках данных и признаков, а затем объединяет их предсказания путем голосования (для классификации) или усреднения (для регрессии). Таким образом, алгоритм ансамблирования, используемый в Random Forest, - это бэггинг.

1. Обучение с учителем (supervised learning) и обучение без учителя (unsupervised learning):

* **Обучение с учителем** - модель обучается на размеченных данных, где известны правильные ответы (например, классифицировать изображения как "кошка" или "собака"). Алгоритмы: линейная регрессия, логистическая регрессия, деревья решений.
* **Обучение без учителя** - модель ищет скрытые закономерности в неразмеченных данных. Алгоритмы: кластеризация, ассоциативные правила, понижение размерности.

**2. Задачи классификации, регрессии и кластеризации:**

* **Классификация** - определение принадлежности объекта к одному из заранее известных классов (например, распознавание рукописных цифр, определение, является ли email спамом или нет).
* **Регрессия** - предсказание численного значения (например, прогнозирование цен на недвижимость, предсказание продаж).
* **Кластеризация** - группировка объектов в кластеры на основе сходства (например, сегментация покупателей в маркетинге, обнаружение аномалий в данных).

3**. Алгоритмы для задач:**

* **Классификация**: логистическая регрессия, деревья решений, SVM, нейронные сети.
* **Регрессия**: линейная регрессия, деревья решений, случайный лес.
* **Кластеризация**: K-means, DBSCAN, иерархическая кластеризация.

4. **Переобучение и способы его выявления:**

* Переобучение - когда модель слишком хорошо подстраивается под обучающие данные, но плохо работает на новых, неизвестных данных.
* Признаки переобучения: большая разница между точностью на обучающих и тестовых данных, сложность модели не соответствует размеру и сложности данных.
* Способы борьбы: регуляризация, перекрестная проверка, ранняя остановка обучения.

5. Понижение размерности:

* Уменьшение количества признаков (переменных) в данных без потери важной информации.
* Применяется для снижения вычислительной сложности, визуализации данных высокой размерности, устранения мультиколлинеарности.
* Методы: PCA, t-SNE, LDA.

6. Регуляризация:

* Способ борьбы с переобучением путем добавления штрафа за сложность модели.
* Помогает избежать чрезмерного подгонки к обучающим данным, улучшает обобщающую способность.
* Виды регуляризации: L1 (Lasso), L2 (Ridge), эластичная сеть.

7. Генетические алгоритмы:

* Оптимизационные алгоритмы, основанные на принципах эволюции и естественного отбора.
* Используются для решения сложных оптимизационных задач, где невозможно перебрать все возможные решения.
* Применяются в задачах планирования, расписания, маршрутизации, оптимального проектирования.