**1. Что означает уменьшение размерности в машинном обучении?**

Уменьшение размерности (или сокращение размерности) — это процесс уменьшения количества входных переменных в наборе данных. Это делается для того, чтобы:

* Снизить сложность модели.
* Уменьшить объем вычислений.
* Избежать проблемы "проклятия размерности", когда данные становятся слишком разбросанными в пространстве высокой размерности.
* Улучшить визуализацию данных.
* Повысить производительность модели за счет удаления шума и неинформативных признаков.

**2. Какие методы входят в Feature Selection? Как они работают?**

Методы отбора признаков (Feature Selection) включают:

* **Filter Methods (Фильтрационные методы):**
  + **Univariate Selection (Одномерный отбор):** Оценивают каждый признак по отдельности, используя статистические тесты, и отбирают признаки с наивысшими оценками.
  + **Mutual Information (Взаимная информация):** Измеряет зависимость между признаками и целевой переменной.
* **Wrapper Methods (Методы-обертки):**
  + **Forward Selection (Пошаговый отбор):** Начинают с пустого набора признаков, добавляя по одному признаку, который наиболее улучшает модель.
  + **Backward Elimination (Пошаговое исключение):** Начинают с полного набора признаков, удаляя по одному признаку, который наименее важен для модели.
  + **Recursive Feature Elimination (Рекурсивное исключение признаков):** Постепенно удаляют наименее важные признаки, используя модель (обычно линейную или деревья решений).
* **Embedded Methods (Встроенные методы):**
  + **Regularization Methods (Регуляризационные методы):** Такие как LASSO (L1-регуляризация), которая уменьшает коэффициенты неинформативных признаков до нуля.
  + **Tree-based Methods (Методы на основе деревьев):** Такие как Random Forest и Gradient Boosting, которые оценивают важность признаков на основе улучшения критерия разбиения (например, прирост информации).

**3. Расскажите принцип работы метода PCA.**

Метод главных компонент (PCA, Principal Component Analysis) работает следующим образом:

1. **Центрирование данных:** Сначала все данные центрируются, т.е. из каждого признака вычитается его среднее значение.
2. **Ковариационная матрица:** Вычисляется ковариационная матрица признаков.
3. **Собственные вектора и значения:** Находятся собственные значения и соответствующие собственные вектора ковариационной матрицы.
4. **Сортировка:** Собственные значения сортируются по убыванию, и выбираются соответствующие им собственные вектора.
5. **Проекция:** Данные проецируются на выбранные собственные вектора (главные компоненты), создавая новый набор данных с уменьшенной размерностью.

**4. Что означает понятие главная компонента?**

Главная компонента — это новая переменная, полученная в результате проекции исходных данных на оси, которые максимизируют дисперсию данных. Первая главная компонента объясняет наибольшую часть дисперсии в данных, вторая — вторую по величине и т.д. Главные компоненты ортогональны друг другу и независимы.

**5. Что означает термин Ансамбли в контексте машинного обучения?**

Ансамбли (ensembles) — это методы, которые объединяют несколько моделей для улучшения общей производительности по сравнению с использованием одной модели. Идея состоит в том, чтобы комбинировать слабые модели (которые могут быть плохими предсказателями по отдельности) в сильную модель, уменьшая вероятность ошибки и повышая устойчивость.

**6. Как работают алгоритмы стекинг, бэггинг, бустинг?**

* **Бэггинг (Bagging, Bootstrap Aggregating):**
  + Сначала создаются несколько выборок данных с повторением (bootstrap).
  + На каждой выборке обучается модель.
  + Предсказания всех моделей объединяются (например, путем усреднения или голосования).
* **Бустинг (Boosting):**
  + Модели обучаются последовательно.
  + Каждая следующая модель пытается исправить ошибки предыдущей, уделяя больше внимания ошибкам.
  + Финальное предсказание — взвешенная сумма предсказаний всех моделей.
* **Стекинг (Stacking):**
  + Несколько моделей (базовые обучающие модели) обучаются на одних и тех же данных.
  + Предсказания этих моделей используются как новые признаки для модели-мета-алгоритма (метамодель).
  + Мета-модель обучается на этих предсказаниях, чтобы сделать финальное предсказание.

**7. Объясните метод Random Forest. Какой алгоритм ансамблирования в нем используется?**

Random Forest — это ансамблевый метод, основанный на бэггинге. Он работает следующим образом:

1. **Создание выборок:** Генерируются несколько подвыборок данных с повторением (bootstrap).
2. **Обучение деревьев:** На каждой подвыборке обучается дерево решений.
3. **Случайный отбор признаков:** При каждом разбиении узла дерева рассматривается случайное подмножество признаков, что увеличивает разнообразие деревьев и снижает корреляцию между ними.
4. **Объединение предсказаний:** Предсказания всех деревьев объединяются (например, путем голосования для классификации или усреднения для регрессии).

Random Forest улучшает устойчивость и точность модели, уменьшая переобучение по сравнению с отдельными деревьями решений.