Шпаргалка по базовым понятиям в ML / Концепции

Cheatsheet (XeLaTeX)

Краткий справочник

April 2, 2025

Contents

Что такое Машинное Обучение?

Машинное обучение (Machine Learning, ML) — это раздел искусственного интеллекта, который позволяет компьютерам "обучаться" на данных без явного программирования. Модель сама находит закономерности в данных и использует их для решения задач.
Аналогия: Представь, что ты учишься отличать яблоки и груши. Вместо того чтобы тебе дали четкие инструкции ("если зеленое и круглое - яблоко"), тебе показывают много примеров яблок и груш. Ты сам вырабатываешь правила.

1 Типы Машинного Обучения

Основные парадигмы ML

В зависимости от задачи и типа данных, выделяют несколько основных типов ML:

- **Обучение с учителем (Supervised Learning):**
  - **Задача:** Модель учится на размеченных данных, где для каждого входа (**признаки**, features) известен правильный выход (**метка**, label или target).
  - **Цель:** Предсказать метку для новых, невиданных ранее входных данных.
  - **Примеры:**
    - **Классификация (Classification):** Предсказание категориальной метки (e.g., спам/не спам, кошка/собака).
    - **Регрессия (Regression):** Предсказание непрерывного значения (e.g., цена дома, температура воздуха).
- **Обучение без учителя (Unsupervised Learning):**
  - **Задача:** Модель учится на неразмеченных данных, пытаясь найти скрытую структуру или закономерности. Правильных ответов нет.
  - **Примеры:**
    - **Кластеризация (Clustering):** Группировка похожих объектов (e.g., сегментация клиентов).
    - **Снижение размерности (Dimensionality Reduction):** Уменьшение количества признаков с сохранением важной информации (e.g., PCA, t-SNE для визуализации).
- **Обучение с подкреплением (Reinforcement Learning):**
  - **Задача:** Агент учится взаимодействовать со средой, совершая действия и получая награды или штрафы, с целью максимизировать итоговую награду.
  - **Примеры:** Обучение игровых ботов, робототехника, системы рекомендаций.

На собеседованиях чаще всего спрашивают про **Supervised Learning**.

2 Процесс Разработки и Разделение Данных

Этапы ML проекта (очень упрощенно)

1. Постановка задачи.
2. Сбор и подготовка данных (*часто самый трудоемкий этап!*).
3. Выбор и обучение модели.
4. Оценка качества модели.
5. Развертывание и мониторинг.

Зачем делить данные? Train / Validation / Test

Чтобы честно оценить способность модели **обобщать** (generalize) на новых данных, исходный набор данных делят на три части:

- **Обучающая выборка (Train Set):** Используется непосредственно для обучения модели — подбора её внутренних параметров (весов). *Аналогия: Домашние задания, на которых студент учится.*
- **Валидационная выборка (Validation Set):** Используется для настройки **гиперпараметров** модели (e.g., скорость обучения, глубина дерева, параметр регуляризации) и/или **выбора наилучшего алгоритма** из нескольких кандидатов. Модель не обучается на этих данных напрямую, но мы используем её результаты для принятия решений о структуре/настройках модели. *Аналогия: Пробные экзамены, по результатам которых студент корректирует свою подготовку или выбирает стратегию.*
- **Тестовая выборка (Test Set):** Используется **только один раз** в самом конце для финальной, непредвзятой оценки качества лучшей выбранной и настроенной модели. Эти данные модель никогда не "видела" ни при обучении, ни при настройке. *Аналогия: Финальный экзамен, который показывает реальные знания студента.*

Важно: Никогда не используйте тестовую выборку для настройки модели! Это приведет к завышенной (нереалистичной) оценке качества.

3 Переобучение и Недообучение

Ключевая проблема: Баланс Модели

При обучении модели мы всегда сталкиваемся с риском двух крайностей:

- **Недообучение (Underfitting):** Модель слишком простая, она не может уловить основные закономерности в данных. Плохо работает как на обучающих, так и на новых данных. Характеризуется высоким **смещением (Bias)**. *Аналогия: Студент, который почти не готовился и плохо сдает даже тесты по пройденному материалу.*
- **Переобучение (Overfitting):** Модель слишком сложная, она "вызубрила" обучающие данные, включая случайный шум. Отлично работает на обучающих данных, но плохо обобщает на новые, невиданные данные. Характеризуется высоким **разбросом (Variance)**. *Аналогия: Студент, который вызубрил ответы на конкретные билеты, но не понял суть и "плывет" на похожих вопросах.*

Цель: Построить модель, которая находит золотую середину — хорошо улавливает общие закономерности, но игнорирует шум, и хорошо работает на новых данных.

4 Дилемма Смещения-Разброса (Bias-Variance Tradeoff)

Анатомия Ошибки Модели

Общая ожидаемая ошибка модели на новых данных может быть разложена на три компонента:

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

- **Смещение (Bias):** Систематическая ошибка модели. Насколько предсказания *в среднем* отклоняются от истинных значений. Высокое смещение (High Bias) означает, что модель слишком простая и не улавливает зависимости (Недообучение). (*Модель систематически ошибается / не попадает в цель*). *Аналогия: Стрелок, который всегда целится левее центра мишени. Пули ложатся кучно, но не туда.*
- **Разброс (Variance):** Чувствительность модели к изменениям в обучающей выборке. Насколько сильно будут различаться модели, обученные на разных подмножествах данных. Высокий разброс (High Variance) означает, что модель слишком сложная и подстраивается под шум (Переобучение). (*Модель нестабильна / сильно реагирует на данные*). *Аналогия: Стрелок, у которого сильно дрожат руки. Он целится в центр, но пули ложатся с большим разбросом вокруг него.*
- **Неустранимая ошибка (Irreducible Error):** Минимально возможная ошибка, обусловленная шумом в самих данных, который не может быть устранен никакой моделью.

Дилемма (Tradeoff)

Часто попытка уменьшить смещение (усложняя модель) приводит к увеличению разброса, и наоборот, попытка уменьшить разброс (упрощая модель, добавляя регуляризацию) может увеличить смещение.
Задача Data Scientist'а — найти модель с такой сложностью, которая обеспечивает наилучший компромисс между смещением и разбросом, минимизируя *общую ошибку* на **новых** данных (обычно оценивается на валидационной выборке).

- **Простые модели** (e.g., Линейная регрессия): Low Variance, High Bias.
- **Сложные модели** (e.g., Глубокие деревья решений, нейросети без регуляризации): High Variance, Low Bias.

5 Диагностика: Кривые Обучения

Анализ Кривых Обучения

Кривые обучения — это графики, показывающие метрику качества (например, ошибку MSE или долю правильных ответов Accuracy) на **обучающей** (`Train`) и **валидационной** (`Valid`) выборках в зависимости от некоторого параметра (чаще всего — размера обучающей выборки или номера эпохи обучения).

Анализ кривых:

- **Признак Недообучения (High Bias):**
  - Ошибка на `Train` и `Valid` высокая.
  - Кривые быстро сходятся и выходят на плато.
  - Разрыв между кривыми маленький.
  - *Что делать?* Усложнять модель (больше слоев/нейронов, полиномиальные признаки), добавить новые релевантные признаки, провести **Feature Engineering**, уменьшить регуляризацию. *Добавление данных скорее всего не поможет.*
- **Признак Переобучения (High Variance):**
  - Ошибка на `Train` низкая, а на `Valid` значительно выше.
  - Большой разрыв между кривыми `Train` и `Valid`.
  - *Что делать?* Собрать больше данных, использовать регуляризацию (L1, L2, Dropout), упростить модель (меньше глубина дерева), использовать **Feature Selection**, использовать ансамбли (Bagging).
- **Хороший баланс:**
  - Обе кривые сходятся к низкому значению ошибки.
  - Разрыв между кривыми небольшой.

Умение читать кривые обучения — важный навык для диагностики и улучшения ML моделей!