Экзаменационные вопросы по курсу «Введение в машинное обучение»

Лазар В. И. и Козлова Е. Р.

Лекция 1. Введение в машинное обучение

1. Дать определение задачи машинного обучения. Что такое обучающая выборка, модель и функция потерь?
2. Сформулировать математическую постановку задачи обучения с учителем: выборка $D = \{(x_i, y_i)\}_{i=1}^n$, поиск модели $f(x)$, минимизирующей ошибку на данных.
3. Перечислить основные типы задач машинного обучения (обучение с учителем, обучение без учителя, обучение с подкреплением) и привести по одному примеру для каждого типа.
4. Объяснить различие между задачами классификации и регрессии. Какие метрики качества используются в этих задачах (Accuracy, Precision, Recall, F1, MAE, RMSE, R^2)?
5. Объяснить смысл разбиения данных на обучающую, валидационную и тестовую выборки. Какую задачу решает каждая из этих частей?
6. Что такое переобучение и недообучение модели? Как они проявляются на графике зависимости сложности модели от ошибки на обучающей и валидационной выборках?
7. Дать определение признаков (features) и целевой переменной (label/target). Что такое предобработка признаков и feature engineering? Привести примеры.
8. Описать алгоритм k-ближайших соседей для задачи регрессии: формула предсказания при равномерных весах и при весах, зависящих от расстояния, роль метрики и гиперпараметра k.
9. Записать правило обновления параметров при градиентном спуске $\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$. Пояснить смысл шага обучения η и градиента $\nabla_\theta L(\theta)$.

Лекция 2. Линейная регрессия, Ridge и Lasso

1. Записать модель линейной регрессии в векторной форме $\hat{y} = w^\top x + b$. Объяснить смысл параметров w и b.
2. Сформулировать цель обучения линейной регрессии: какие параметры ищутся и по какому критерию они выбираются.

3. Записать функционал Ridge-регрессии (L2-регуляризации) и объяснить, как L2-штраф влияет на значения коэффициентов и устойчивость модели.

4. Записать функционал Lasso-регрессии (L1-регуляризации) и объяснить, почему L1-штраф часто приводит к появлению нулевых коэффициентов (разреженности модели).

5. Сравнить Ridge и Lasso-регрессии: в каких ситуациях предпочтителен каждый из методов, что такое Elastic Net и как он сочетает L1 и L2-штрафы.

6. Объяснить, зачем перед применением регуляризованных моделей стандартизуют признаки. Как на практике подбирают параметр регуляризации λ?

Лекция 3. Логистическая регрессия

1. Сформулировать задачу бинарной классификации в логистической регрессии. Что считается входом и выходом модели?

2. Объяснить, почему обычная линейная регрессия плохо подходит для решения задач бинарной классификации.

3. Записать выражение для вероятности наблюдения $y \in \{0, 1\}$ при фиксированном x и параметрах модели. Записать правдоподобие выборки как произведение по объектам.

4. Вывести или привести выражение для функции потерь логистической регрессии (log loss)

5. Пояснить, как и зачем добавляются L1- и L2-регуляризации в логистическую регрессию. Как регуляризация влияет на переобучение и интерпретируемость модели?

Лекция 4. Решающие деревья и ансамбли (бэггинг, случайные леса, стэкинг)

1. Описать структуру решающего дерева: что такое внутренний узел, лист, рёбра, глубина дерева. Чем дерево классификации отличается от дерева регрессии по типу предсказания в листе?

2. Объяснить пошагово, как решающее дерево делает предсказание для одного объекта.

3. Описать жадный алгоритм построения дерева сверху вниз. Какие критерии останова используются (максимальная глубина, минимальное число объектов в листе, минимальное уменьшение нечистоты)?

4. Объяснить, как деревья работают с числовыми и категориальными признаками, как можно обрабатывать пропуски и почему деревья обычно не требуют масштабирования признаков.

5. Рассказать о причинах переобучения решающих деревьев и способах борьбы с ним: пред-обрезка и пост-обрезка.

6. Перечислить основные плюсы и минусы решающих деревьев как модели для табличных данных.

7. Объяснить идею бэггинга (Bootstrap Aggregating): как формируются бутстрап-выборки, как усредняются предсказания, что такое OOB-оценка качества.

8. Дать определение случайного леса. Чем он отличается от обычного бэггинга деревьев и какую роль играет ограничение числа признаков при поиске сплита (параметр `max_features`)?

9. Объяснить идею стэкинга: что такое базовые модели и мета-модель, как формируются признаки второго уровня. Что такое OOF-предсказания и зачем они используются?

Лекция 5. Бустинг и градиентный бустинг над деревьями

1. Сформулировать общую идею бустинга. Чем бустинг принципиально отличается от бэггинга и стэкинга?

2. Записать общее обновление модели в градиентном бустинге.

3. Пояснить точку зрения градиентного спуска в пространстве функций: как задаётся функционал ошибки $L(F)$ и какую роль играют псевдо-остатки (антиградиенты) на обучающих объектах.

4. Записать шаг алгоритма градиентного бустинга над деревьями: вычисление псевдо-остатков, обучение регрессионного дерева по этим значениям, поиск оптимальных сдвигов по листам и обновление ансамбля.

5. Перечислить основные способы регуляризации в градиентном бустинге: малая скорость обучения ν, ограничение глубины деревьев, минимальный размер листа, субсемплинг объектов (`subsample`), субсемплинг признаков, ранняя остановка по валидационной выборке.

6. Объяснить, как по поведению метрик на обучающей и валидационной выборках диагностировать переобучение бустинговой модели и как выбирать оптимальное число итераций.

Лекция 6. Кластеризация, SVD и PCA

1. Дать определение обучения без учителя. Какие задачи обычно относят к обучению без учителя? Привести примеры.

2. Описать алгоритм K-Means. Записать целевую функцию, которую минимизирует K-Means.

3. Перечислить основные преимущества и недостатки K-Means. Как влияет масштабирование признаков и наличие выбросов на работу алгоритма?

4. Объяснить идею алгоритма DBSCAN. Каковы роли параметров ε (eps) и `min_samples`? Чем отличаются ядровые, пограничные точки и шум?

5. Сравнить DBSCAN и K-Means по форме находящихся кластеров, устойчивости к выбросам, необходимости заранее задавать число кластеров и чувствительности к плотности данных.

6. Описать идею агломеративной иерархической кластеризации: начальное состояние, правило слияния кластеров, возможные варианты связи (single, complete, average, Ward) и роль дендрограммы.

7. Перечислить достоинства и недостатки иерархической кластеризации, в том числе с точки зрения вычислительной сложности и работы с большими выборками.

8. Сформулировать основные цели снижения размерности и назвать несколько наиболее популярных алгоритмов.