ML Handbook

Сергей Полянских

Оглавление

Предисловие

В данной книге описаны основные понятия, методы и подходы, широко используемые в современном DS и ML. Обычно, свободное владение этими понятиями необходимо для правильного понимания как основных, так и продвинутых методов ML и по умолчанию предполагается от DS специалиста.

Здесь собраны разные определения, встречавшиеся автору в научных статьях по ML и на собеседованиях. Охвачены: теория вероятностей, классическая и байесовская статистика, некоторые вопросы мат. анализа.

Освещение вопросов ни в коем случае не претендует на полноту и в некоторых случаях на строгость. Основная цель книги - составить расширенный глоссарий основных понятий и подходов, встретившихся автору в процессе работы в области ML.

Обозначения

DS	- наука о данных
ML	- машинное обучение
RV	- случайная величина
CDF	- функция распределения случайной величины
PDF	- плотность распределения случайной величины
CLT	- центральная предельная теорема
EX	- среднее случайной величины X
DX	- дисперсия случайной величины X
$X \sim Y$	- случайные величины X и Y одинаково распределены

Глава 1

Математика

В этой главе описаны основные математические понятия, необходимые для правильного понимания как основных, так и продвинутых методов ML. Охвачены: теория вероятностей, классическая и байесовская статистика, некоторые вопросы мат. анализа.

1.1 Случайная величина

Случайной величиной (RV) называется числовая функция X, определенная на некотором множестве элементарных исходов Ω (обычно подмножество $\mathbb{R}$ или $\mathbb{R}^n$),

$$X : \Omega \to \mathbb{R}.$$

С прикладной точки зрения на RV часто смотрят как на генераторы случайных чисел с заданным распределением.

Примеры:

- Рост людей, взятых из некоторой группы.
- Цвет фиксированного пикселя изображения, взятого из некоторого множества изображений.
- Некоторый признак из датасета ML задачи.

1.2 Распределение случайной величины

Если RV принимает дискретное множество значений $x_1, x_2, ...$, то она полностью определяется значениями их вероятностей: $p_k = \mathbb{P}(X = x_k)$.

Если множество значений RV не дискретно, то RV может быть описана своей функцией распределения (CDF, Cumulative distribution function): $F(x) = \mathbb{P}(X < x)$.

В большинстве прикладных случаев CDF оказывается дифференцируемой функцией. Производная от CDF называется плотностью распределения случайной величины (PDF, Probability density function): $f(x) = F'(x)$. Таким образом, по определению

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx.$$

1.3 Выборка

Выборкой объема n из генеральной совокупности X называется последовательность независимых и распределенных как X случайных величин:

$$X_1, X_2, ..., X_n, \quad X_k \sim X$$

На практике под выборкой понимают конкретные реализации величин X_k, то есть последовательность чисел $x_1, x_2, ..., x_n$.

1.4 Закон больших чисел

Закон больших чисел утверждает, что если $X_1, X_2, ..., X_n$ - выборка объема n из генеральной совокупности X, то ее среднее с ростом n стабилизируется к среднему значению X:

$$\frac{X_1 + X_2 + ... + X_n}{n} \approx EX, \quad n \to \infty.$$

1.5 Центральная предельная теорема

Центральная предельная теорема (CLT) является в некотором смысле уточнением закона больших чисел. В упрощенном варианте она утверждает, что если $X_1, X_2, ..., X_n$ - выборка объема n из генеральной совокупности X, то ее распределение ее среднего при больших n очень близко к нормальному,

$$\frac{X_1 + X_2 + ... + X_n}{n} \approx N(\mu, \sigma^2/n), \quad \mu = EX, \sigma^2 = DX, \quad n \to \infty.$$

Заметим, что если совокупность распределена нормально, $X \sim N(\mu, \sigma^2)$, то предыдущая формула обращается в точное равенство при любых n.

1.6 Статистики

Пусть $X_1, X_2, ..., X_n$ - выборка объема n. Статистикой называется произвольная RV, являющаяся функцией выборки:

$$T = T(X_1, X_2, ..., X_n).$$

Часто статистикой выборки называют конкретное значение $T(x_1, x_2, ..., x_n)$, полученное на данной реализации $x_1, x_2, ..., x_n$ выборки.

Примеры:

- $\bar{X} = (X_1 + X_2 + ... + X_n)/n$ - выборочное среднее.
- $X_{(n)} = \max(X_1, X_2, ..., X_n)$ - максимальное значение в выборке.
- медиана, перцентили.

1.7 Bootstrap

1.8 Классический и байесовский подход

1.9 Метод максимального правдоподобия

1.10 Доверительный интервал

1.11 Байесовский доверительный интервал

1.12 Основные дискретные распределения

1.13 Основные непрерывные распределения

1.14 Матричные разложения

...может разделить главу на части...

1.15 КЛ дивергенция

1.16 Энтропия

1.17 Квантили

1.18 Точечные оценки

1.19 Интервальные оценки

1.20 Проверка гипотез

1.21 Множественная проверка гипотез

1.22 Параметрические и непараметрические критерии, бутстреп

1.23 Ошибки I и II рода

1.24 Достигаемый уровень значимости

1.25 Мощность статистического критерия

1.26 Основные задачи статистики

...из лекций новосиба курсера...

Глава 2

Анализ данных

Анализ и предобработка данных - первая задача, успешное решение которой зачастую определяет успех в решении любых задач ML. В этой главе описываются основные подходы....

2.1 Типы данных

2.2 Предобработка данных

2.3 Понижение размерности

Глава 3

Общие вопросы

В этой главе приводятся основные понятия ML и DS.

3.1 Машинное обучение

Машинное обучение (ML) - область искусственного интеллекта, изучающая самообучающиеся модели, то есть решаюшие поставленную задачу не по заранее запрограммированному алгоритму, а предварительно настраивая свое поведение согласно имеющимся данным.

Обычно методы ML содержат свободные параметры, подбор которых наилучшим (в смысле имеющихся данных) образом и составляет процесс обучения алгоритма.

3.2 Основные классы задач

3.3 Обнаружение аномалий

3.4 Контроль качества

...оценка обобщающей способности...

3.5 Недообучение

3.6 Переобучение

3.7 Регуляризация

3.8 Отбор признаков

3.9 Параметры алгоритма

3.10 Подбора метапараметров

3.11 Основные типы алгоритмов

3.12 Многоклассовая классификация

3.13 Дисбаланс классов

3.14 Ансамбли алгоритмов

3.15 Метрики классификации

3.16 ROC-AUC метрика

3.17 Метрики регрессии

3.18 Метрики кластеризации

3.19 Разложение ошибки алгоритма

3.20 Кривые валидации

3.21 Кривые обучения

3.22 Метрические методы

3.23 Метод ближайших соседей

3.24 Линейные методы

3.25 Линейная регрессия

3.26 Логистическая регрессия

...отличие от линейной...

3.27 SVM

3.28 Ядра и спрямляющие пространства

3.29 Решаюшие деревья

3.30 Случайный лес

...отличие от беггинга над решающими деревьями...

3.31 Градиентный бустинг

3.32 Байесовские методы

Глава 4

Нейросети

В данной главе приводится обзор основных понятий и методов, связанных с нейросетями.