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	- дата саенс
ML	- машинное обучение
RV	- случайная величина
CDF	- функция распределения случайной величины
CLT	- центральная предельная теорема

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

1.7 Bootstrap

1.8 Классический и байесовский подход