

# ML Handbook

Сергей Полянских

# Оглавление

<b>1</b>	<b>Математика</b>	<b>4</b>
1.1	Случайная величина . . . . .	4
1.2	Распределение случайной величины . . . . .	4
1.3	Выборка . . . . .	5
1.4	Закон больших чисел . . . . .	5
1.5	Классический и байесовский подход . . . . .	5

# 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. 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

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

### 1.1 Случайная величина

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

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

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  $x_1, x_2, \dots$ , 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 называется последовательность RV:  $X_1, X_2, \dots, X_n$ . Предполагается, что все  $X_k$  попарно независимы и имеют одно и то же распределение:  $X_k \sim X$ . В этом случае говорят о выборке из генеральной совокупности  $X$ .

На практике выборкой являются конкретные реализации величин  $X_k$ , то есть последовательность чисел  $x_1, x_2, \dots, x_n$ .

### 1.4 Закон больших чисел

### 1.5 Статистики

### 1.6 Bootstrap

### 1.7 Классический и байесовский подход