ML Handbook

s.pol

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 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

https : *//medium.com/@srowen/common−probability−distributions−*347*e*6*b*945*ce*4

1.13 Основные непрерывные распределения

1.14 Матричные разложения

...может разделить главу на части...

1.15 К-Л 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...

1.27 Проверка основных гипотез

1.28 Корреляция Пирсона

1.29 Корреляция Спирмена

1.30 Корреляция Метьюса

1.31 Корреляция Крамера

1.32 Z-тест Фишера

1.33 T-тест Стьюдента

1.34 Критерий Пирсона χ^2

1.35 Точный тест Фишера

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

...чем плохо... как бороться (over/undersampling/SMOTE)...

3.14 Ансамбли алгоритмов

3.15 Метрики классификации

Пусть некоторый алгоритм a решает задачу бинарной классификации с классами 0 (негативный) и 1 (позитивный). Тестирование алгоритма a проводится на n объектах, ответы y на которых известны. Пусть TP и TN - числа правильно классифицированных позитивных и негативных объектов соответственно. Аналогично, FP и FN - числа неправильно классифицированных позитивных и негативных объектов соответственно.

О качестве алгоритма a можно судить по матрице ошибок:

	y=1	y=0
a=1	TP	FP
a=0	FN	TN

Для оценки качества работы алгоритмов бинарной классификации обычно используются описанные далее основные метрики.

3.15.1 Accuracy

Точность (accuracy) - отношение числа правильных ответов к общему количеству,

$$accuracy = \frac{TP + TN}{n}.$$

Проста в использовании и интерпретации, но плоха для несбалансированных выборок, и потому довольно редко используемая напрямую.

3.15.2 Precision

Точность (precision) - отношение числа правильно классифицированных позитивных объектов к общему количеству позитивно классифицированных,

$$precision = \frac{TP}{TP + FP}.$$

Чем ближе значение к 1, тем меньше ложных срабатываний (FP).

3.15.3 Полнота (recall)

Полнота (recall) - вычисляется как отношение

$$recall = \frac{TP}{TP + FN}.$$

Чем ближе значение к 1, тем меньше ложных пропусков (FN).

3.15.4 F1-мера

F1-мера - среднее гармоническое точности и полноты,

$$F = \frac{2PR}{P + R}.$$

F1-мера усредняет точность и полноту, является неплохом компромиссом между обеими метриками.

3.15.5 F-мера

Обобщенная F-мера вычисляется как

$$F = (1 + \beta^2)\frac{PR}{\beta^2 P + R}.$$

F-мера усредняет точность и полноту, является неплохом компромиссом между обеими метриками, имеет настраиваемый параметр β.

3.15.6 ROC-AUC

ROC-AUC - площадь под ROC кривой. Принимает значения от 0 до 1. Значения близкие к 0.5 интерпретируются как самые худшие (случайное гадание), близкие к 1 - как хорошие. ROC-AUC более устойчива к дисбалансу классов, чем Accuracy, но не так хорошо, как PR-AUC. ROC-AUC также не учитывает уверенность алгоритма в своих предсказаниях (насколько близко распределены предсказания к 0 и 1).

Метрика ROC-AUC применяется к алгоритмам бинарной классификации, дающим вероятностноподобный вывод, $a \in [0, 1]$. Для таких алгоритов после их обучения необходимо выбрать порог дискретизации - число, ниже которого ответ будет считаться принадлежащим классу 0, а выше которого - классу 1. Подбор наилучшего значения порога - отдельная задача, однако, есть метрики качества, позволяющие оценить алгоритм "в целом без привязки к конкретному значению порога.

Сама ROC кривая строится в координатах $FPR = FP/(FP + TN)$, $TPR = TP/(TP + FN)$ следующим образом:

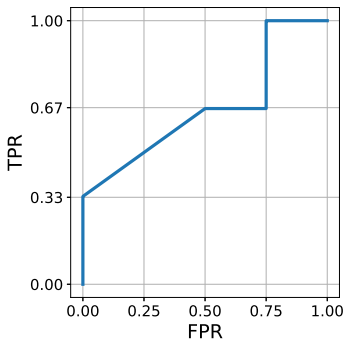
1. отрезки $[0, 1]$ по осям TPR и FPR разбиваются на $\#[y = 0]$ и $\#[y = 1]$ частей соответственно.

2. пары реальных ответов y_i упорядочиваются по убыванию соответствующих ответов алгоритма a_i

3. проходя по получившемуся после сортировки массиву значений y_i, строим ROC кривую, начиная от начала координат и делая шаг вправо, если $y_i = 0$ и вверх, если $y_i = 1$. Важный момент: если рядом по порядку оказались несколько a_i с одинаковыми значениями, то соответствующий им участок ROC кривой будет не ступенчатым, а прямолинейным (см. пример ниже).

Другой способ построения ROC кривой указан в 3.15.7.
Пример:

y	1	0	0	1	0	1	0
a	1.0	0.9	0.9	0.9	0.8	0.3	0.2



ROC-AUC как площадь под ROC кривой равна 2/3.

3.15.7 PR-AUC

PR-AUC - площадь под PR кривой. Как и ROC-AUC, применяется для алгоритмов бинарной классификации, дающих не бинарный, а вероятностноподобный вывод и также оценивает алгоритм "в целом без привязки к конкретному значению порога классификации.

Сама PR кривая строится в координатах $recall = TP/(TP+FN)$, $precision = TP/(TP+FP)$.

Аналигичны образом (в других координатах может быть построена и ROC кривая).

PR-AUC хороша для несбалансированных классов, так как учитывает FP - число негативных объектов, неверно классифицированных как позитивные. PR-AUC также не учитывает уверенность алгоритма в своих предсказаниях (насколько близко распределены предсказания к 0 и 1).

see..... `https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-p`

3.16 Метрики многоклассовой классификации

3.17 ROC-AUC метрика

3.18 Индекс Джини

3.19 Метрики регрессии

3.20 Метрики кластеризации

3.21 Разложение ошибки алгоритма

3.22 Кривые валидации

3.23 Кривые обучения

3.24 Метрические методы

3.25 Метод ближайших соседей

3.26 Линейные методы

3.27 Линейная регрессия

3.28 Логистическая регрессия

...отличие от линейной...

3.29 SVM

3.30 Ядра и спрямляющие пространства

3.31 Решаюшие деревья

3.32 Случайный лес

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

3.33 Градиентный бустинг

3.34 Байесовские методы

Глава 4

Нейросети

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