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

$https://medium.com/@srowen/common-probability-distributions-347e6b945ce4$

1.13 Основные непрерывные распределения

1.14 Матричные разложения

...может разделить главу на части...

1.15 К-Л дивергенция

1.16 Энтропия

1.17 Кросс-энтропия

1.18 Квантили

1.19 Точечные оценки

1.20 Интервальные оценки

1.21 Проверка гипотез ***

Пусть $x_1, x_2, ..., x_n$ - выборка из генеральной совокупности X, распределение которой заранее неизвестно. Требуется проверить некоторое утверждение о распределении генеральной совокупности X исходя из имеющейся выборки.

Общий подход к решению таких задач состоит в следующем:

1. Формулируется основная гипотеза H_0 о распределении X и некоторая альтернативная гипотеза H_1, которая может являться полным отрицанием H_0 (двусторонняя альтернатива), но не обязательно (односторонние и др. альтернативы).

2. Выбирается некоторая статистика T исходя из условия, что нулевая гипотеза H_0 верна тогда и только тогда, когда распределение T известно: $T \sim F(t|H_0)$ (нулевое распределение статистики).

3. Вычисляется значение t^* статистики T на имеющейся выборке. По известному распределению F можно судить, насколько вероятно получить значения t^*.

4. Выбирается уровень значимости α.

5. Вычисляется достигаемый уровень значимости p_{value} и сравнивается с α. Если $p_{value} > \alpha$, то нулевая гипотеза H_0 не может быть отвергнута, а если $p_{value} \leqslant \alpha$, гипотеза H_0 отвергается в пользу альтернативной H_1.

Пример: TODO

Обычно нулевая гипотеза H_0 означает, что "ничего интересного не происходит а альтернативная, напротив, говорит о том, что "что-то произошло".

1.22 p_{value} ***

Пусть статистика T приняла на выборке $x_1, x_2, ..., x_n$ значение t^*. Так как нулевое распределение статистики известно, по значению t^* можно судить, насколько оно характерно для данного распределения. Именно, для случая правосторонней альтернативной определим значение

$$p_{value} = \mathbb{P}(Tt^*|H_0).$$

TODO картинка

Если бы H_0 была справедлива, то значение t^* вероятно оказалось бы около матожидания распределения T и, как следствие, p_{value} было бы велико. В противном случае, если t^* оказалось далеко правее среднего, p_{value} мало, и нулевую гипотезу следует отвергнуть в пользу правосторонней альтернативы H_1.

Число p_{value} называется достигаемым уровнем значимости. Оно сравнивается с заранее заданным уровнем значимости α, и если $p_{value} > \alpha$, то нулевая гипотеза H_0 не может быть отвергнута, а если $p_{value} \leqslant \alpha$, гипотеза H_0 отвергается в пользу альтернативной H_1.

Число p_{value} также равно вероятности ошибки первого рода, то есть вероятности отвергнуть верную нулевую гипотезу.

Аналогичные рассуждения справедливы для случаев лево- и двусторонней альтернативы.

1.23 Множественная проверка гипотез

1.24 Параметрические и непараметрические критерии, бутстреп

1.25 Ошибки I и II рода

1.26 Достигаемый уровень значимости

1.27 Мощность статистического критерия

1.28 Основные задачи статистики

...из лекций новосиба курсера...

1.29 Проверка основных гипотез

1.30 Корреляция Пирсона

1.31 Корреляция Спирмена

1.32 Корреляция Метьюса

1.33 Корреляция Крамера

1.34 Z-тест Фишера

1.35 T-тест Стьюдента

1.36 Критерий Пирсона χ^2

1.37 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 и задачи) образом и составляет процесс обучения алгоритма. После обучения алгоритм можно использовать на новых данных, которые не были представлены алгоритму на стадии обучения.

3.2 Основные классы задач

3.2.1 Обучение с учителем

3.2.2 Обучение без учителя

3.2.3 Частичное обучение

3.2.4 Обучение с подкреплением

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 и функции потерь

Метрика - величина, обычно диктуемая бизнесом, оптимизация (максимизация или минимизация) которой вполне очевидным образом свидетельствует об улучшении качества работы модели.

Функция потерь - величина, более удобная для оценки/оптимизации модели, уменьшение которой, вообще говоря, приводит к оптимизации метрики задачи.

Иными словами, улучшение метрики - конечная цель процесса обучения алгоритма, но достигается это зачастую оптимизацией именно некоторой функции потерь, с которой может быть удобнее работать. Метрики и функции потерь - близкие понятия, когда речь идет об оценки качества алгоритма, и их довольно часто смешивают.

Пример: Пусть в задаче бинарной классификации основной метрикой является accuracy - доля правильных ответов. Эта метрика не дифференцируема, поэтому ее оптимизация напрямую методами гладкой оптимизации невозможна. В качестве функции потерь выберем MSE - среднеквадратичную ошибку. Это уже гладкая функция своих аргументов,и ее минимизация скорее всего приведет к увеличению доли правильных ответов, то есть к конечной цели.

3.16 Метрики бинарной 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.

3.16.1 Accuracy

Точность (accuracy) - доля правильных ответов,

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

Проста в использовании и интерпретации, но плоха для несбалансированных выборок. Кроме того, не дифференцируема и потому не может быть использована напрямую в качестве функции потерь для алгоритмов гладкой оптимизации.

3.16.2 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). Проста в использовании и интуитивна, то не использует информацию о негативно классифицированных объектах и, кроме того, не является дифференцируемой.

3.16.3 Полнота (recall)

Полнота (recall) - вычисляется как отношение

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

Чем ближе значение к 1, тем меньше ложных пропусков (FN). Проста в использовании и интуитивна, то не использует TN, FP и, кроме того, не является дифференцируемой.

3.16.4 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. Проста в использовании, но плохо интерпретируема и не является дифференцируемой.

3.16.5 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. Проста в использовании, но плохо интерпретируема и не является дифференцируемой.

3.16.6 ROC кривая

ROC кривая - характеристика качества алгоритмов бинарной классификации, дающих вероятностноподобный вывод, $a \in [0, 1]$. ROC кривая строится в координатах

$$FPR = \frac{FP}{FP + TN}, \quad TPR = \frac{TP}{TP + FN}.$$

Каждая точка кривой - значение (FPR, TPR), полученное для некоторого порога дискретизации алгоритма (см. 3.16.8).

Более простой способ построения ROC кривой состоит в следующем:

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).

ROC кривая идеального алгорима проходит через точки $(0,0)$, $(0,1)$, $(1,1)$; для случайного гадания - проходит вблизи прямой $FPR = TPR$. Наилучшим значением порога дискретизации алгоритма может считаться порог, соответствующий точке на ROC кривой, ближайшей к $(0,1)$, либо точке, наиболее удаленной от прямой случайного гадания $TPR = FPR$.

Пример: для алгоритма, дающего вывод как в таблице ниже, график ROC кривой выглядит следующим образом

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

../out/pdf-images/roc-curve-eps-converted-to.p

3.16.7 ROC-AUC

ROC-AUC - площадь под ROC кривой. Применяется к алгоритмам бинарной классификации, дающим вероятностноподобный вывод, $a \in [0,1]$, позволяя оценить алгоритм "в целом без привязки к конкретному значению порога дискретизации алгоритма.

ROC-AUC 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). Не является дифференцируемой.

ROC-AUC для примера 3.16.6 равна 2/3.

3.16.8 PR кривая

PR кривая - характеристика качества алгоритмов бинарной классификации, дающих вероятностноподобный вывод, $a \in [0,1]$. PR кривая строится в координатах

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP}.$$

Каждая точка кривой - значение $(recall, precision)$, полученное для некоторого порога дискретизации алгоритма.

Способ построения PR кривой состоит в следующем:

1. вычисляются пороги h - всевозможные значения ответов алгоритма a.

2. ответы a_i дискретизируются для каждого значения порога и вычисляются значения *recall* и *precision*. При этом ордината первой точки кривой, соответствующей порогу $h > 1$, не определена, так как знаменатель *precision* обращается в ноль. В качестве ординаты берется ордината второй точки.

3. по полученным точкам строится график PR кривой.

PR кривая идеального алгорима проходит через точки $(0, 1)$, $(1, 1)$, $(1, \#[y = 1]/n)$; для случайного гадания - проходит вблизи прямой $precision = \#[y = 1]/n$.

Пример: для алгоритма, дающего вывод как в таблице ниже, график PR кривой выглядит следующим образом

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

../out/pdf-images/pr-curve-eps-converted-to.p

3.16.9 PR-AUC

PR-AUC - площадь под PR кривой. Применяется к алгоритмам бинарной классификации, дающим вероятностноподобный вывод, $a \in [0, 1]$, позволяя оценить алгоритм "в целом без привязки к конкретному значению порога дискретизации алгоритма.

PR-AUC - площадь под PR кривой. Принимает значения от 0 до 1. Значения близкие к 1 интерпретируются как хорошие, близкие к $\#[y = 1]/n$ - как самые худшие (случайные гадания). PR-AUC более устойчива к дисбалансу классов, чем ROC-AUC, однако, не учитывает уверенность алгоритма в своих предсказаниях (насколько близко распределены предсказания к 0 и 1). Не является дифференцируемой.

PR-AUC для примера 3.16.8 равна $11/15 \approx 0.73$.

3.16.10 Бинарная кросс-энтропия (logloss)

Пусть y - истинная метка объекта (0 или 1), а a - ответы некоторого алгоритма (число из $[0, 1]$). Бинарная кросс-энтропия (logloss) вычисляется как

$$L(y, a) = -y \log_2 a - (1 - y) \log_2 (1 - a).$$

Слагаемые с нулевым множителем при логарифме (соответствующие $y = 0$ и $y = 1$) полагаются равными нулю.

Полная кросс-энтропия на множестве ответов определяется усреднением значений по всем объектам.

Бинарная кросс-энтропия имеет следующую вероятностную интерпретацию. Пусть метка i-му объекту назначается по схеме Бернулли, т.е. метка полагается равной $y_i = 1$ с вероятностью a_i и $y_i = 0$ с вероятностью $1 - a_i$. Тогда вероятность получить истинные ответы y_i равна

$$\prod_{i=1}^{n} a_i^{y_i}(1 - a_i)^{1-y_i}.$$

Логарифмируя, получаем правдоподобие, совпадающее с бинарной кросс-энтропией с точностью до знака. Таким образом, нахождение ответов a_i с позиции минимизации бинарной кросс-энтропии равносильно максимизации правдоподобия.

https://en.wikipedia.org/wiki/Loss$_{f}unctions_{f}or_{c}lassification$

3.17 Метрики многоклассовой классификации

3.17.1 Категориальная кросс-энтропия (logloss)

3.18 Индекс Джини

3.19 Метрики регрессии

3.19.1 Среднеквадратичная ошибка (MSE)

Пусть y - истинная метка объекта, а a - ответы некоторого алгоритма. Квадратичная ошибка вычисляется как

$$L(y, a) = (y - a)^2.$$

Полная среднеквадратичная ошибка на множестве ответов определяется усреднением значений по всем объектам. Наилучшим константным предсказанием для MSE является выборочное среднее:

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i$$

MSE дифференцируема и проста в использовании, но плохо интрепретируема, так как дает ненормированный ни к чему результат, который трудно с чем-либо сравнить. Кроме того, MSE чувствительна к выбросам в выборке.

3.19.2 Среднеабсолютная ошибка (MAE)

Пусть y - истинная метка объекта, а a - ответы некоторого алгоритма. Абсолютная ошибка вычисляется как

$$L(y, a) = |y - a|.$$

Полная среднеквадратичная ошибка на множестве ответов определяется усреднением значений по всем объектам. Наилучшим константным предсказанием для MSE является выборочная медиана.

MAE проста в использовании, но недифференцируема и плохо интрепретируема, так как дает ненормированный ни к чему результат, который трудно с чем-либо сравнить. MAE менее чувствительна к выбросам в выборке, чем MSE.

3.19.3 Коэффициент детерминации (R^2)

Пусть y_i - истинные метки объектов x_i, а a_i - ответы некоторого алгоритма. Коэффициент детерминации R^2 вычисляется как

$$R^2(y,a) = 1 - \frac{\sum_{i=1}^n (y_i - a_i)^2}{\sum_{i=1}^n (y_i - \hat{y})^2}, \quad \hat{y} = \frac{1}{n}\sum_{i=1}^n y_i.$$

R^2 показывает долю дисперсии y_i, объясняемую моделью. Является по сути линейной функцией от MSE, но более интерпретируема в силу нормировки к результату с константным прогнозом $\hat{y}$. Минусом является увеличение R^2 при увеличении числа признаков, что далеко не всегда свидетельствует о увеличении качества модели.

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.