

Машинное обучение (Machine Learning)

Неопределенность и калибровка в машинном обучении

Уткин Л.В.

Санкт-Петербургский политехнический университет Петра Великого



Неопределенность

Основные понятия и типы неопределенности

Aleatoric uncertainty

- Aleatoric uncertainty (AU) - стохастическая неопределенность - представляет собой объективную случайность, присущую задаче.
- Примеры - перекрытие классов, шум данных, разброс, неизвестные факторы.
- Она же - вариабельность, изменчивость.
- AU принципиально неустранима и не зависит от познающего субъекта.

Aleatoric uncertainty (пример)

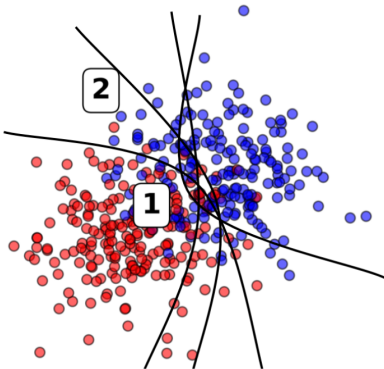
- Типичный пример - подбрасывание монеты: процесс генерации данных в экспериментах такого типа имеет стохастический компонент, который не может быть уменьшен какой-либо дополнительной информацией.
- Следовательно, даже лучшая модель этого процесса сможет дать только вероятности двух возможных исходов, орла и решки, но не даст однозначного ответа.

Aleatoric uncertainty в машинном обучении

- В МО AU интерпретируется как неопределенность из-за неоднозначности или шума в данных.
- В задаче классификации это означает наличие перекрывающихся классов.
- AU НЕ может быть уменьшена наблюдением большего числа примеров из одного источника, а только добавлением дополнительных признаков, улучшением качества признаков и другими процедурами, делающими классы разделимыми.
- Увеличение данных не приводит к снижению AU.
- В медицинских моделях часто сталкиваемся с определенной степенью AU, так как у нас нет полной информации для предсказания будущих событий.

Aleatoric uncertainty (иллюстрация)

область 1 - имеется перекрытие классов



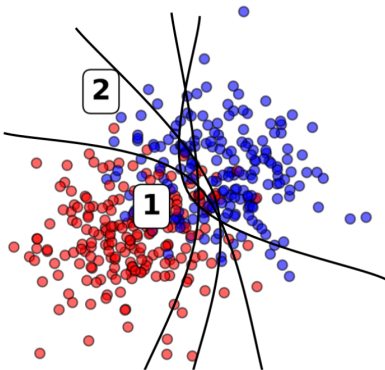
- Epistemic uncertainty (EU) эпистемическая неопределенность обусловлен фактором субъективным - недостаточностью (неполнотой, неточностью, неоднозначностью) имеющихся знаний о свойствах изучаемого объекта.
- EU можно интерпретировать как неопределенность из-за отсутствия знаний об оптимальной модели, вызванное отсутствием данных наблюдения.
- EU высока в регионах, где отсутствуют обучающие данные.

Epistemic uncertainty (пример)

- Эпистемическую неопределенность можно разделить на
 - неопределенность в отношении параметров модели
 - неопределенность в отношении структуры модели (или класса гипотез).
- Ее можно устранить путем наблюдения большего количества данных, поэтому ее также называют уменьшаемой неопределенностью.
- Пока о пациенте не известно ничего важного, врач не будет знать истинного диагноза. Собирая все больше и больше информации в виде медицинских анализов и т. д., это незнание будет исчезать шаг за шагом.

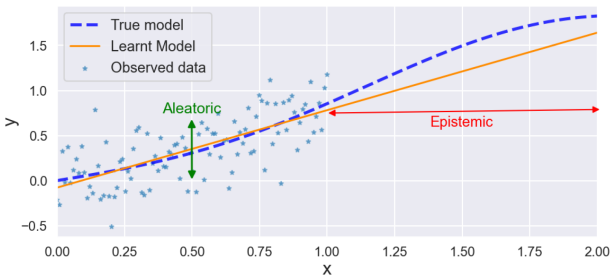
Epistemic uncertainty в машинном обучении (иллюстрация)

2 - неопределенность из-за отсутствия знаний об оптимальной модели (разделяющей функции).



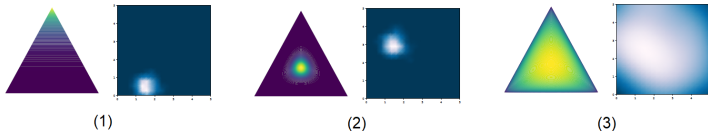
Два типа неопределенности (пример)

- AU - разброс наблюдаемых данных о модели, которая является хорошим приближением к истинной модели в области, в которой мы наблюдали данные.
- EU - модель становится худшим приближением к истинной функции по мере того, как мы удаляемся от наблюдаемых данных при $x > 1$



Два типа неопределенности (еще пример)

Классификация (симплекс) и регрессия (квадрат)



1 - AU и EU низкие (достоверные прогнозы с низкой дисперсией)

2 - AU высока, EU низка (предсказания концентрируются вокруг центра симплекса или областей с большой дисперсией)

3 - AU и EU высоки

Пример с монетой (1)

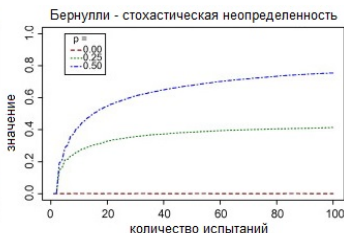
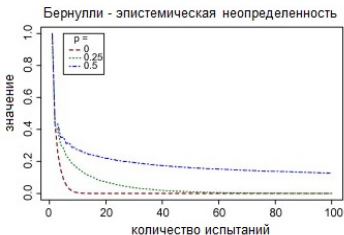
- Последовательность экспериментов Бернулли: монета с неизвестной вероятностью орла $p \in [0; 1]$ подбрасывается много раз и после броска задача в том, чтобы предсказать исход следующего броска.
- Предсказание следующего исхода является неопределенным:
 - вначале это EU, т.к. о p ничего не известно;
 - с течением времени о p узнаем все больше и больше, так что EU становится все меньше;
 - в пределе бесконечного объема выборки EU полностью исчезнет, так как p можно оценить по частоте событий с произвольной точностью;
 - оставшаяся неопределенность - AU.

Пример с монетой (2)

- После N испытаний и K “орлов” функция правдоподобия

$$L(p) = \binom{N}{K} \cdot p^K \cdot (1 - p)^{N-K}.$$

- Неопределенность двух типов (средняя по большому кол-ву повторений) в зависимости от N .



Пример с монетой (3)

- Чем ближе p к $1/2$, тем медленнее исчезает EU и тем больше AU, которая в конечном итоге остается.
- Случай $p = 1/2$ особенный, так как он соответствует “полной неопределенности”. Вначале полностью EU становится полностью AU при $N \rightarrow \infty$.
- Важно, общая величина неопределенности не уменьшается (кривая AU медленно сходится к 1): даже точное знание p не помогает предсказать исход следующего испытания.
- Для $p \neq 1/2$ иначе, так как даже приблизительное знание p поможет сделать лучше, чем случайное угадывание.

Достоверность

Понятие достоверности предсказания

Достоверность предсказания (1)

- Классификация: есть обучающие данные (\mathbf{X}, \mathbf{y}) , $\mathbf{x}_i \in \mathbf{X}$ - вектор признаков, $y_i \in \mathbf{Y}$ - класс.
- Цель: найти функцию $\psi(\mathbf{X})$ (классификатор, модель), которая отображает новый пример \mathbf{x}_0 в оценку \hat{y}_0 , т.е. оценить y_0 для \mathbf{x}_0 с учетом $\psi(\mathbf{X})$ из обучающих данных \mathbf{X} .
- Для \mathbf{x}_0 оцениваем достоверность предсказания как

$$C(\hat{y}_0 = y_0 \mid \mathbf{x}_0, \psi(\mathbf{x}_0), \mathbf{X}) = \hat{p}_0,$$

где \hat{p}_0 - оценка достоверности предсказания \hat{y}_0 .

Достоверность предсказания (2)

- Для \mathbf{x}_0 оцениваем достоверность предсказания как

$$C(\hat{y}_0 = y_0 \mid \mathbf{x}_0, \psi(\mathbf{x}_0), \mathbf{X}) = \hat{\rho}_0,$$

где $\hat{\rho}_0$ - оценка достоверности предсказания \hat{y}_0 .

- Модель оценивания достоверности учитывает насколько информативными являются обучающие данные при попытке классифицировать пример \mathbf{x}_0 . Пример учета этого - ядро $K(\mathbf{x}_0, \mathbf{X})$ как мера сходства между точкой \mathbf{x}_0 и \mathbf{X} .
- Если \mathbf{x}_0 не похож на то, что было в \mathbf{X} , мы не можем быть уверены в \hat{y}_0 .

Калибровка (1)

- Калибровка модели оценки достоверности C - это степень, с которой оценки соответствуют эмпирической точности классификатора, т.е. если модель C оценивает вероятность в 75%, это должно быть правильным примерно в 75% случаев.
- Модель C *идеально калибрована*, если это верно для всех оцениваемых вероятностей, $p \in [0, 1]$. Если это не так, то мы говорим, что модель либо *слишком достоверна* (*over-confident*), либо *недостаточно достоверна* (*under-confident*): т.е. если модель C оказывается верной реже, чем \hat{p}_0 , то говорят, что она *over-confident*, и наоборот.
- Формально:

$$\Pr(\hat{y}_i = y_i \mid \hat{p}_i = p_i) = p \in [0, 1].$$

Калибровка (2)

- Т.к. истинная вероятность является неизвестной случайной величиной, обычно невозможно точно ее вычислить, поэтому стандартный метод заключается в преобразовании непрерывного доверительного пространства в набор из n дискретных интервалов, B_n (например, $[0, 5, 0, 6, 0, 7, 0, 8, 0, 9, 1.]$) и сравнении средней оценки достоверности \bar{p} каждого интервала с эмпирической точностью интервала, т.е.

$$\text{Acc}(B_n) = \frac{1}{|B_n|} \sum_{i \in B_n} \delta_{\hat{y}_i, y_i},$$

где

$$\delta_{\hat{y}_i, y_i} = \begin{cases} 1, & \hat{y}_i = y_i, \\ 0, & \text{иначе,} \end{cases} \quad C(B_n) = \frac{1}{|B_n|} \sum_{i \in B_n} \bar{p}_i$$

Калибровка (3)

- Модель C идеально калибрована, если $C(B_n) = Acc(B_n)$ для каждого B_n .
- Показатель ошибки калибровки: ожидаемая ошибка калибровки (ECE). Эта метрика измеряет разницу между прогнозируемыми оценками достоверности и эмпирической точностью каждого интервала. Эти остатки затем объединяются в сумму, взвешенную по количеству баллов в каждой ячейке:

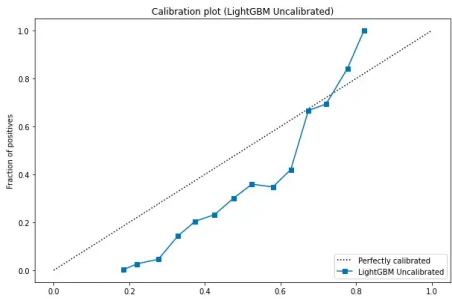
$$ECE = \frac{1}{N} \sum_{i=1}^n |C(B_i) - Acc(B_i)| \cdot |B_i|$$

Кривые достоверности (1)

- Кривая достоверности (надежности) - визуальный метод, позволяющий определить, откалибрована ли наша модель.
- Разбиваем $[0; 1]$ на интервалы (пусть 0.1).
- Пусть есть 5 примеров в первом интервале, т. е. есть 5 точек $(0.05, 0.05, 0.02, 0.01, 0.02)$, чей диапазон предсказания модели лежит между 0 и 0.1.
- По оси X откладываем среднее значение этих прогнозов, т.е. 0.03, а по оси Y откладываем эмпирические вероятности, т.е. $\text{Acc}(B_1)$. Если из 5 точек для одной точки $\delta_{\hat{y}_i, y_i} = 1$, то $\text{Acc}(B_1) = 1/5$. Следовательно, координаты нашей первой точки равны $(0.03, 0.2)$.

Кривые достоверности (2)

- Повторяем процедуру для всех интервалов и соединяем точки, чтобы сформировать линию.
- Сравниваем с линией $y = x$ и оцениваем калибровку:
 - когда точки находятся выше этой линии, модель занижает истинную вероятность,
 - если ниже линии, модель завышает истинную вероятность.



Методы калибровки

Методы калибровки

Гистограммный метод для бинарной классификации (1)

- Все неоткалиброванные прогнозы \hat{p}_i делятся на интервалы B_1, \dots, B_M .
- Каждому интервалу присваивается калиброванная оценка θ_m , т.е., если \hat{p}_i ставится в соответствие B_m , то $\hat{q}_i = \theta_m$.
- Во время тестирования, если предсказание \hat{p}_{te} попадает в B_m , то откалиброванное предсказание $\hat{q}_{te} = \theta_m$.
- Точнее, для подходящего выбора M сначала определяем границы интервалов $0 = a_1 \leq a_2 \leq \dots \leq a_M + 1 = 1$, где B_m определяется интервалом $(a_m, a_m + 1]$.

Гистограммный метод для бинарной классификации (2)

- Значения θ_i выбираются так, чтобы минимизировать квадратичные потери по интервалам:

$$\min_{\theta_1, \dots, \theta_M} \sum_{m=1}^M \sum_{i=1}^n \mathbf{1}(a_m \leq \hat{p}_i < a_m + 1) (\theta_m - y_i)^2.$$

- При фиксированных границах интервалов решение приводит к θ_m , которые соответствуют среднему количеству “правильных” примеров класса в интервале B_m .

Изотонная регрессия для бинарной классификации

- Идея - найти кусочно-постоянную функцию f :
 $\hat{q}_i = f(\hat{p}_i)$ или минимизировать потери
 $\sum_{i=1}^n (f(\hat{p}_i) - y_i)^2$.
- Это соответствует

$$\min_{\substack{\theta_1, \dots, \theta_M \\ a_1, \dots, a_{M+1}}} \sum_{m=1}^M \sum_{i=1}^n \mathbf{1}(a_m \leq \hat{p}_i < a_{m+1}) (\theta_m - y_i)^2.$$

при ограничениях

$$\begin{aligned} 0 &= a_1 \leq a_2 \leq \dots \leq a_{M+1} = 1, \\ \theta_1 &\leq \theta_2 \leq \dots \leq \theta_M. \end{aligned}$$

- Здесь $\theta_1, \dots, \theta_M$ - значения функции f

Масштабирование Платта (Platt et al., 1999)

- Параметрический подход к калибровке, в отличие от других подходов.
- Невероятные предсказания классификатора используются в качестве признаков модели логистической регрессии, которая обучается на тестовом наборе для получения вероятностей.
- В нейронных сетях масштабирование Платта определяет скалярные параметры $a, b \in \mathbb{R}$ и выдает $\hat{q}_i = \sigma(az_i + b)$ в качестве калиброванной вероятности, где $z_i \in \mathbb{R}$ невероятный выход сети.

Гистограммный метод (обобщение на $K > 2$)

- Один из способов обобщения - рассматривать задачу как K задача “один против всех”.
- Для $k = 1, \dots, K$ формируем задачу бинарной калибровки, где метка $\mathbf{1}$ ($y_i = k$) и предсказанная вероятность

$$\sigma_{Soft \max}(\mathbf{z}_i)^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_{j=1}^K \exp(z_i^{(j)})}$$

- Это дает нам K калибровочных моделей, каждая для определенного класса. При тестирования мы получаем ненормированный вектор вероятностей $[\hat{q}_i^{(1)}, \dots, \hat{q}_i^{(K)}]$, $\hat{q}_i^{(k)}$ - калиброванная вероятность для класса k .

Гистограммный метод (обобщение на $K > 2$)

- Это дает нам K калибровочных моделей, каждая для определенного класса. При тестирования мы получаем ненормированный вектор вероятностей $[\hat{q}_i^{(1)}, \dots, \hat{q}_i^{(K)}]$, $\hat{q}_i^{(k)}$ - калиброванная вероятность для класса k .
- Предсказание нового класса \hat{y}'_i - argmax вектора, а новая вероятность \hat{q}'_i - максимальное значение вектора, нормализованного делением на $\hat{q}_i^{(1)} + \dots + \hat{q}_i^{(K)}$.
- Это обобщение может быть применено к другим методам

Масштабирование температуры

- Простейшее обобщение масштабирования Платта
- Использует один скалярный параметр (температура) $T > 0$ для всех классов.
- Если дан вектор \mathbf{z}_i , новое доверительное предсказание равно

$$\hat{q}_i = \sigma_{\text{Soft max}}(\mathbf{z}_i / T)^{(k)}.$$

- Если $T \rightarrow \infty$, то $\hat{q}_i \rightarrow 1/K$. Если $T = 1$, то $\hat{q}_i = \hat{p}_i$.
Если $T \rightarrow 0$, то $\hat{q}_i = 1$.
- Т.к. параметр T не изменяет максимум функции softmax , предсказание класса \hat{y}'_i остается неизменным, т.е., масштабирование температуры не влияет на точность модели.

Вопросы

?