Машинное обучение

Семинар 3. Применение вероятностного аппарата в машинном обучении.

Applied Data Science Center

23 января 2018

Вероятностная задача

Возьмём какую-то болезнь. Она достаточно редкая — ей болеет примерно один человек на миллион. Есть аппарат, диагностирующий эту болезнь, который для здоровых людей с вероятностью 99.9% даёт отрицательный результат, а для больных с вероятностью 99.9% положительный. Во время ежегодного медицинского обследования человек проходит анализ на этом аппарате и получает положительный результат (диагностирована болезнь). Какова вероятность того, что он болен?

Формула Байеса

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \text{ где}$$

$P(A)$ — априорная вероятность события A;
$P(A \mid B)$ — вероятность события A при условии события B;
$P(B \mid A)$ — вероятность события B при условии события A;
$P(B)$ — априорная вероятность наступления события B.

Формула Байеса

Если есть полный набор событий $(\mathrm{A_i})_{\mathrm{i}=1}^{\mathrm{n}}$, то:

$$\mathrm{P(B)} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P(B \mid A_i)P(A_i)}$$

$$\mathrm{P(A_k \mid B)} = \frac{\mathrm{P(B \mid A_k)P(A_k)}}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P(B \mid A_i)P(A_i)}}$$

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(B \mid A)P(A)}}{\mathrm{P(B \mid A)P(A)} + \mathrm{P(B \mid \overline{A})P(\overline{A})}}$$

Формула Байеса

В нашей задаче A — человек болен, B — аппарат диагностировал болезнь, нужно найти $P(A \mid B)$.

$$P(A) = ?$$

$$P(B \mid A) = ?$$

Формула Байеса

В нашей задаче A — человек болен, B — аппарат диагностировал болезнь, нужно найти P(A | B).

$$P(A) = 10^{-6}$$

$$P(B \mid A) = 1 - 10^{-3}$$

Формула Байеса

В нашей задаче A — человек болен, B — аппарат диагностировал болезнь, нужно найти $P(A \mid B)$.

$$P(A) = 10^{-6}$$

$$P(B \mid A) = 1 - 10^{-3}$$

$$P(A \mid B) = \frac{\left(1 - 10^{-3}\right) 10^{-6}}{\left(1 - 10^{-3}\right) 10^{-6} + 10^{-3} \left(1 - 10^{-6}\right)} \approx 10^{-3}$$

Формула Байеса в повседневности

Если есть полный набор гипотез $(A_i)_{i=1}^n$, и вы оцениваете их шансы как

$$P(A_1) : P(A_2) : \ldots : P(A_n).$$

Далее вы наблюдаете событие B, из-за чего соотношение шансов гипотез меняется, по формуле Байеса оно будет таким:

$$P(A_1)P(B|A_1) : P(A_2)P(B|A_2) : \ldots : P(A_n)P(B|A_n).$$

То есть шаны умножаются на условные вероятности.

Стандартные понятия

Функция распределения случайной величины ξ — это функция $\mathrm{F}(\mathrm{x}) = \mathrm{P}(\xi \leqslant \mathrm{x})$.

Плотность случайной величины ξ — это функция $\mathrm{f}(\mathrm{x}) = \mathrm{F}'(\mathrm{x})$.

Математическое ожидание случайной величины ξ — $\mathrm{E}\xi = \int\limits_{-\infty}^{\infty} \mathrm{x}\, \mathrm{f}(\mathrm{x}) \mathrm{dx}$.

Дисперсия случайной величины ξ — $\mathrm{D}\xi = \mathrm{E}\xi^2 - (\mathrm{E}\xi)^2$

Когда это может пригодиться?

Вы делаете умную автобусную остановку. Она будет говорить сколько ещё осталось ждать автобуса. К сожалению, наладить расписание не удалось, поэтому автобусы приходят на остановку случайно, но с некоторой интенсивностью. Какие проблемы могут возникнуть при предсказании такой величины?

Когда это может пригодиться?

Вы делаете умную автобусную остановку. Она будет говорить сколько ещё осталось ждать автобуса. К сожалению, наладить расписание не удалось, поэтому автобусы приходят на остановку случайно, но с некоторой интенсивностью. Какие проблемы могут возникнуть при предсказании такой величины?

Время между автобусами будет распределено экспоненциально:

$$f(x) = \lambda e^{-\lambda x} I\{x \geqslant 0\}, \text{ где } \lambda \text{ — параметр отвечающий за интесивнос}$$

Для экспоненциально распределённой величины ξ значение $P(\xi - s \geqslant t \mid \xi \geqslant s)$ не зависит от s. Какие сложности это может вызвать?

Максимизация правдоподобия в задачах ML

Допустим, у нас есть выборка X, целевые переменные y и набор параметров θ. Логично, что нам бы хотелось выбрать наиболее вероятный вектор параметров, при условии того, что мы какие-то данные пронаблюдали:

$$\mathrm{P}(\theta \mid \mathrm{X}, \mathrm{y}) \to \max_{\theta}.$$

По формуле Байеса:

$$\mathrm{P}(\theta \mid \mathrm{X}, \mathrm{y}) \propto \mathrm{P}(\mathrm{y} \mid \mathrm{X}, \theta)\mathrm{P}(\theta \mid \mathrm{X}) = \mathrm{P}(\mathrm{y} \mid \mathrm{X}, \theta)\mathrm{P}(\theta)$$

Последнее равернство выполнено, так как априорные представления о параметрах не зависят от данных, которые мы пронаблюдали.

Максимизация правдоподобия в задачах ML

Таким образом, задача

$$\mathrm{P}(\theta \mid \mathrm{X}, \mathrm{y}) \to \max_{\theta}$$

эквивалентна

$$\mathrm{P}(\mathrm{y} \mid \mathrm{X}, \theta)\mathrm{P}(\theta) \to \max_{\theta}.$$

Поскольку все объекты выборки независимы, то

$$\mathrm{P}(\mathrm{y} \mid \mathrm{X}, \theta) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}(\mathrm{y_i} \mid \mathrm{x_i}, \theta).$$

Подставив в оптимизационную задачу и прологарифмировав получим:

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \log \mathrm{P}(\mathrm{y_i} \mid \mathrm{x_i}, \theta) + \log \mathrm{P}(\theta) \to \max_{\theta}.$$

Максимизация правдоподобия в задачах ML

Итоговая оптимизационная задача:

$$\sum_{i=1}^{n} \log P(y_i \mid x_i, \theta) + \log P(\theta) \to \max_{\theta}.$$

Сравните с минимизацией регуляризированного эмпирического риска:

$$\sum_{i=1}^{n} L(y_i, a(x_i)) + R(a) \to \min_{a}.$$

Таким образом, различные функции потерь и регуляризаторы приобретают вероятностный смысл условного и априорного распределений соответственно.

Максимизация правдоподобия в задачах ML

В линейных моделях, если положить $P(w) \sim N(0, \tau)$, то

$$\log P(w) = \log\left(\frac{1}{\sqrt{2\pi\tau}} e^{-\frac{||w||^2}{2\tau}}\right) = -\mathrm{const}(w) - \frac{1}{2\tau}||w||^2,$$

что при максимизации по w соответсвует l_2 регуляризации.

- Какое априорное распределение соответсвует l_1 регуляризации?
- Какое условное распределение соответсвует MSE функции потерь?
- Какое условное распределение соответсвует MAE функции потерь?
- Какая функция потерь будет в задаче про автобусную остановку?

Максимизация правдоподобия в задачах ML

В линейных моделях, если положить $P(w) \sim N(0, \tau)$, то

$$\log P(w) = \log\left(\frac{1}{\sqrt{2\pi\tau}} e^{-\frac{||w||^2}{2\tau}}\right) = -\mathrm{const}(w) - \frac{1}{2\tau}||w||^2,$$

что при максимизации по w соответсвует l_2 регуляризации.

- Распределение Лапласа
- $y_i \sim N(a(x_i), 1)$
- $y_i \sim \mathrm{Lap}(1, a(x_i))$
- $-\ln a(x_i) + a(x_i) y_i$

Байесовский классификатор

Дан вектор признаков $(x_j)_{j=1}^{d}$, по формуле Байеса:

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

То есть выбор наиболее вероятного класса эквивалентен задаче

$$P(x \mid y)P(y) \to \max_{y},$$

так как знаменатель не зависит от y.

Наивный Байесовский классификатор

Наивный Байесовский классификатор предполагает, что

$$P(x \mid y) = \prod_{j=1}^{d} P(x^j|y)$$

так как знаменатель не зависит от y. Таким образом, задача определения класса становится:

$$P(y) \prod_{j=1}^{d} P(x^j|y) \to \max_y,$$

Наивный Байесовский классификатор

Если по выборке оценить условные вероятности и априорную вероятность, то мы сможем решать данную оптимизационную задачу для новых объектов x, подставив в формулу вместо вероятностей их оценки:

$$\hat{P}(y)\prod_{j=1}^{d}\hat{P}(x^j|y) \to \max_y,$$

однако, как найти эти оценки?

- $\hat{P}(y)$ — частота класса y в обучающей выборке
- $\hat{P}(x^j|y)$ — зависит от того, как устроены признаки

Наивный Байес в Sklearn

MultinomialNB

- Признаки категориальные
- $\hat{P}(x^j \mid y) = \frac{N_{yj}+\alpha}{N_y+d\alpha}$, где $N_{yj} = \sum_{i=1}^{n} x_i^j$, $N_y = \sum_{j=1}^{d} N_{yj}$, а α — сглаживающий параметр
- Может интерпретироваться как линейная модель.

BernoulliNB

- Признаки бинарные
- Оценки аналогично MultinomialNB, но при примернении $P(x^j \mid y) = P(j \mid y)x^j + (1 - P(j \mid y))(1 - x^j)$

Наивный Байес в Sklearn

BernoulliNB

- Признаки бинарные
- Оценки аналогично MultinomialNB, но при примернении $P(x^j \mid y) = P(j \mid y)x^j + (1 - P(j \mid y))(1 - x^j)$

GaussianNB

- Признаки вещественные
- Предполагается, что признаки распределены нормально, но на практике это не критично
- $P(x^j \mid y) = \frac{1}{\sqrt{2\pi\sigma_{yj}^2}} e^{-\frac{(x^j - \mu_{yj})^2}{2\sigma_{yj}^2}}$, где μ_{yj} и σ_{yj}^2 — параметры нормального распределения, находятся методом максимального правдоподобия по обучающей выборке

Метод максимального правдоподобия и байесовские методы

Метод максимального правдоподобия для оценки параметра θ по выборке X:

$$\mathrm{P}(\theta \mid \mathrm{X}) \to \max_{\theta}.$$

То есть на самом деле мы ищем не всё распределение $\mathrm{P}(\theta \mid \mathrm{X})$, а только его моду. В байесовских методах ищут непосредственно всё распределение $\mathrm{P}(\theta \mid \mathrm{X})$. А уже потом, в зависимости от задачи, берут его моду, среднее или другую статистику. Это позволяет более детально учесть получаемую информацию, но вычислительно гораздо сложнее.

Байесовские методы

Применим формулу Байеса:

$$\mathrm{P}(\theta \mid \mathrm{X}) = \frac{\mathrm{P}(\mathrm{X} \mid \theta)\mathrm{P}(\theta)}{\int \mathrm{P}(\mathrm{X} \mid \theta)\mathrm{P}(\theta)\mathrm{d}\theta}.$$

Заметьте, что так как нас интересует точное распределение, то мы не может пренебречь знаменателем, как делали это раньше. С этим интегралом и связана основная проблема, так как числитель вычиляется легко в силу независимости объектов выборки.

Байесовские методы. Примеры

- Считаем счётчики при обучении MultinomialNB
- Подпросили монетку n раз, выпало k орлов, нужно оценить вероятность орла
- Оцениваем конверсию показа в клик рекламного блока

Во всех этих случаях, если даннных немного мы бы применяли сглаживание. На самом деле у этой операции есть байесовское обоснование.

Байесовские методы. Сопряжённое распределение

$$\mathrm{P}(\theta \mid \mathrm{X}) = \frac{\mathrm{P}(\mathrm{X} \mid \theta)\mathrm{P}(\theta)}{\int \mathrm{P}(\mathrm{X} \mid \theta)\mathrm{P}(\theta)\mathrm{d}\theta}.$$

В некоторых случаях интеграл в знаменателе можно найти аналитически. Для этого необходимо, чтобы априорное распределение на параметр θ было связано с условным распределением $\mathrm{P}(\mathrm{X} \mid \theta)$ определённым способом.

Если апостериорное распределение $\mathrm{P}(\theta|\mathrm{X})$ принадлежит тому же семейству вероятностных распределений, что и априорное распределение $\mathrm{P}(\theta)$ (т.е. имеет тот же вид, но с другими параметрами), то это семейство распределений называется сопряжённым семейству функций правдоподобия $\mathrm{p}(\mathrm{X}|\theta)$.

Байесовские методы. Пример сопряжённых распределений

Для случайной величины, распределённой по закону Бернулли (приведённые примеры) с неизвестным параметром q в качестве сопряжённого априорного распределения обычно выступает бета-распределение с плотностью вероятности:

$$\mathrm{f}(\mathrm{q}) = \frac{1}{\mathrm{B}(\alpha, \beta)} \mathrm{q}^{\alpha-1}(1-\mathrm{q})^{\beta-1},$$

где $\mathrm{B}(\alpha, \beta)$ — бета функция т.е. $\int\limits_0^1 \mathrm{q}^{\alpha-1}(1-\mathrm{q})^{\beta-1}\mathrm{dq}$

Байесовские методы. Пример сопряжённых распределений

Итак, $P(p)$ — бета распределение, а $P(x|q) = q^x(1-q)^{(1-x)}$. Тогда

$$P(X|q)P(q) = P(q)\prod_{i=1}^{n} P(x_i|q) =$$

$$= \frac{1}{B(\alpha,\beta)} q^{\alpha-1}(1-q)^{\beta-1}\prod_{i=1}^{n} q^{x_i}(1-q)^{(1-x_i)} =$$

$$= \frac{1}{B(\alpha,\beta)} q^{\alpha-1+\sum_{i=1}^{n} x_i}(1-q)^{\beta-1+n-\sum_{i=1}^{n} x_i}$$

$$P(X|q)P(q) = \frac{1}{B(\alpha,\beta)} q^{\alpha-1+\sum_{i=1}^{n} x_i}(1-q)^{\beta-1+n-\sum_{i=1}^{n} x_i}$$

Это означает, что

$$\int\limits_0^1 P(X|q)P(q)dq = \frac{1}{B(\alpha,\beta)} \int\limits_0^1 q^{\alpha-1+\sum_{i=1}^{n} x_i}(1-q)^{\beta-1+n-\sum_{i=1}^{n} x_i}dq =$$

$$= \frac{B(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)}{B(\alpha,\beta)}$$

Это означает, что $P(q \mid X)$ будет иметь Бета распределение с параметрами $\alpha + \sum\limits_{i=1}^{n} x_i$ и $\beta + n - \sum\limits_{i=1}^{n} x_i$

Байесовские методы. Пример сопряжённых распределений

По формуле сатематического ожидания Бета распределения имеем, что

$$\mathrm{E}(\mathrm{q} \mid \mathrm{X}) = \frac{\alpha + \sum\limits_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}}{\alpha + \beta + \mathrm{n}},$$

то есть фактически то сглаживание, которое мы делали.