Введение в искусственный интеллект. Машинное обучение

Лекция 4. Вероятностный подход

МаТИС

15 марта 2019г.

Определения в одномерном случае

- Пусть дана некоторая вероятностная мера P
- X — случайная величина
- $F(x) = F_X(x) := P(X < x)$ — функция распределения
- $p(x) = p_X(x) := \frac{d}{dx} F_X(x)$ — плотность распределения

Дискретный случай

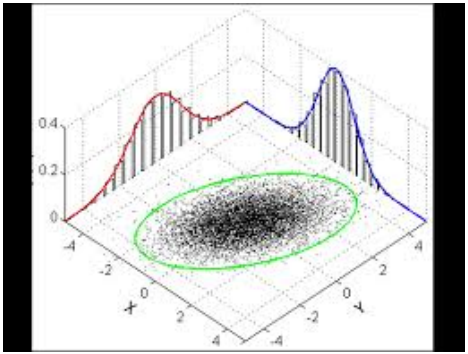
$P(x_i) = p_i$
плотности не существует

Непрерывный случай

$P(x_i) = 0$, но если рассмотреть окрестность, то вероятность уже не нулевая
$p(x_i) \geq 0$

Определения в многомерном случае

- Пусть дана некоторая вероятностная мера P
- $X = (X_1, ..., X_n)$ — многомерная случайная величина
- $F(x_1, ..., x_n) = F_X(x) := P(X_i < x_i \text{ для всех i})$ — функция распределения
- $p(x) = p_X(x) := \frac{\partial^n}{\partial x_1 ... \partial x_n} F_X(x)$ — плотность распределения



Вероятностная постановка задач машинного обучения

Предположения

Пусть известно совместное распределение $p(x, y)$ на $X \times Y$

Пусть задана функция потерь $L(a(x), y)$

Определение

Средняя величина потерь для алгоритма $a(x)$

$$R(a) = \iint L(a(x), y) dP(x, y) = \iint L(a(x), y) p(x, y) dx dy$$

Задача

Найти такой $a^*(x)$, что $a^*(x) = \arg\min_a R(a)$.

Будем называть модель a^* оптимальной и R^* — значение среднего риска.

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Лемма

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x))^2|x\right) + E\left((a(x) - E(y|x))^2|x\right)$$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Лемма

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x))^2|x\right) + E\left((a(x) - E(y|x))^2|x\right)$$

Доказательство

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x) + E(y|x) - a(x))^2|x\right) =$$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Лемма

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x))^2|x\right) + E\left((a(x) - E(y|x))^2|x\right)$$

Доказательство

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x) + E(y|x) - a(x))^2|x\right) =$$
$$= E\left((y - E(y|x))^2|x\right) + E\left((a(x) - E(y|x))^2|x\right) - 2E\left(y - E(y|x)|x\right) E\left(a(x) - E(y|x)|x\right)$$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Лемма

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x))^2|x\right) + E\left((a(x) - E(y|x))^2|x\right)$$

Доказательство

$$E\left((y - a(x))^2|x\right) = E\left((y - E(y|x) + E(y|x) - a(x))^2|x\right) =$$
$$= E\left((y - E(y|x))^2|x\right) + E\left((a(x) - E(y|x))^2|x\right) - 2E\left(y - E(y|x)|x\right) E\left(a(x) - E(y|x)|x\right)$$

Последнее слагаемое равно нулю, так как

$$E\left(y - E(y|x)|x\right) = E(y|x) - E(E(y|x)|x) = E(y|x) - E(y|x) = 0.$$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Доказательство

$R(a) = \int\int L(a(x), y)p(x, y)dydx = \int\int (a(x) - y)^2 p(x, y)dydx =$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

Доказательство

$R(a) = \iint L(a(x), y)p(x, y)dydx = \iint (a(x) - y)^2 p(x, y)dydx =$
$= (a(x) - y)^2 p(y|x)dyp(x)dx = \int E((y - a(x))^2|x)p(x)dx$

Квадратичная функция потерь

Теорема

Если $L(a(x), y) = (a(x) - y)^2$, то величина средних потерь минимальна при

$$a^* = E(y|x)$$

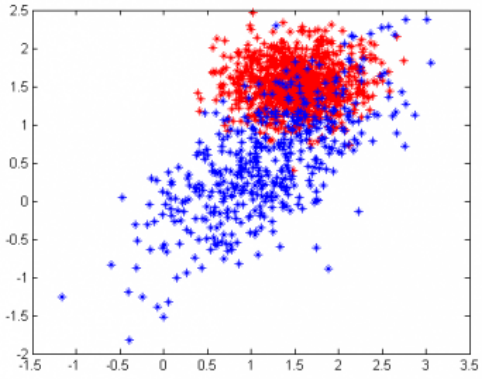
Доказательство

$R(a) = \iint L(a(x), y)p(x, y)dydx = \iint (a(x) - y)^2 p(x, y)dydx =$
$= (a(x) - y)^2 p(y|x)dyp(x)dx = \int E((y - a(x))^2|x)p(x)dx$ Применяя лемму, получаем:
$R(a) = \int E((y - a(x))^2|x)p(x)dx = \int E\left((y - E(y|x))^2|x\right)p(x)dx +$
$\int E\left((a(x) - E(y|x))^2|x\right)p(x)dx \geq \int E\left((y - E(y|x))^2|x\right)p(x)dx$, что и требовалось доказать.

Принцип максимума апостериорной вероятности

Вопрос

Как разделить объекты из этих двух плотностей при известном совместном распределении $p(x, y)$?



Оптимальный байесовский классификатор

Функция потерь

Если $L(a(x), y) = \lambda_y \geq 0$, если $a(x) \neq y$

Теорема

Минимум средних потерь при функции потерь $L(a(x), y)$ достигается байесовским классификатором

$$a(x) = \arg\max_y \lambda_y p(y|x) = \arg\max_y \lambda_y p(y) p(x|y)$$

Оптимальный байесовский классификатор

Функция потерь

Если $L(a(x), y) = \lambda_y \geq 0$, если $a(x) \neq y$

Теорема

Минимум средних потерь при функции потерь $L(a(x), y)$ достигается байесовским классификатором

$$a(x) = \arg\max_y \lambda_y p(y|x) = \arg\max_y \lambda_y p(y) p(x|y)$$

Следствие

Оптимальное правило классификации при одинаковых штрафах за ошибку максимизирует апостериорную вероятность класса

Недостатки байесовского подхода и методы их устранения

Недостатки байесовского подхода и методы их устранения

- Распределения в реальной жизни никогда не известны

Недостатки байесовского подхода и методы их устранения

- Распределения в реальной жизни никогда не известны
- В реальной жизни у нас есть лишь обучающая выборка, то есть сэмплы распределений

Недостатки байесовского подхода и методы их устранения

- Распределения в реальной жизни никогда не известны
- В реальной жизни у нас есть лишь обучающая выборка, то есть сэмплы распределений

Основные подходы

- Восстановить плотность распределения по входным данным

Недостатки байесовского подхода и методы их устранения

- Распределения в реальной жизни никогда не известны
- В реальной жизни у нас есть лишь обучающая выборка, то есть сэмплы распределений

Основные подходы

- Восстановить плотность распределения по входным данным
- Сделать предположение о параметрическом семействе функции распределения и по данным настроить параметры

Недостатки байесовского подхода и методы их устранения

- Распределения в реальной жизни никогда не известны
- В реальной жизни у нас есть лишь обучающая выборка, то есть сэмплы распределений

Основные подходы

- Восстановить плотность распределения по входным данным
- Сделать предположение о параметрическом семействе функции распределения и по данным настроить параметры
- Уменьшать эмпирический риск в надежде, что средний риск тоже будет уменьшен

Метод ближайшего соседа

Теорема (Cover-Hart inequality)

1. Для задачи двуклассовой классификации с функцией потерь $L(a(x), y) = [a(x)! = y]$ и непрерывной функцией $\eta(x) = P(y = 1|x)$ выполнено неравенство:

$$R^* \leq R^{1-NN}(\infty) \leq 2R^*(1 - R^*),$$

где $R^{1-NN}(n) = E\ R^n(x)$ — математическое ожидание эмпирического риска метода одного ближайшего соседа для выборки размера n, а $R^{1-NN}(\infty) = \lim\limits_{n\to\infty} R^{1-NN}(n)$.

Метод ближайшего соседа

Теорема (Cover-Hart inequality)

1. Для задачи двуклассовой классификации с функцией потерь $L(a(x), y) = [a(x)! = y]$ и непрерывной функцией $\eta(x) = P(y = 1|x)$ выполнено неравенство:

$$R^* \leq R^{1-NN}(\infty) \leq 2R^*(1 - R^*),$$

где $R^{1-NN}(n) = E\ R^n(x)$ — математическое ожидание эмпирического риска метода одного ближайшего соседа для выборки размера n, а $R^{1-NN}(\infty) = \lim\limits_{n\to\infty} R^{1-NN}(n)$.

2. В аналогичных условия для многоклассовой (M классов) классификации выполнено

$$R^* \leq R^{1-NN}(\infty) \leq R^*(2 - \frac{M}{M-1}R^*).$$

Метод ближайшего соседа

Теорема (Cover-Hart inequality)

1. Для задачи двуклассовой классификации с функцией потерь $L(a(x), y) = [a(x)! = y]$ и непрерывной функцией $\eta(x) = P(y = 1|x)$ выполнено неравенство:

$$R^* \leq R^{1-NN}(\infty) \leq 2R^*(1 - R^*),$$

где $R^{1-NN}(n) = E\ R^n(x)$ — математическое ожидание эмпирического риска метода одного ближайшего соседа для выборки размера n, а $R^{1-NN}(\infty) = \lim\limits_{n\to\infty} R^{1-NN}(n)$.

2. В аналогичных условия для многоклассовой (M классов) классификации выполнено

$$R^* \leq R^{1-NN}(\infty) \leq R^*(2 - \frac{M}{M-1}R^*).$$

Следствие

Если $R^* = 0$ или $R^1 = \frac{1}{2}$, то $R^{1-NN}(\infty) = R^*$.

Классификация двух многомерных нормальных распределений

Распределения

Пусть $Y = \{0, 1\}$, $X = \mathbb{R}^n$ и

$$p(x|y) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma_y)}} exp\left(-\frac{1}{2}(x - \mu_y)^T \Sigma_y^{-1}(x - \mu_y)\right),$$

где μ_y — вектор математического ожидания в классе y, а Σ_y — ковариационная матрица распределения x в классе y

Разделяющая поверхность

$$0 = ln\frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} = \ln\frac{p_1}{p_0} + \ln\frac{\frac{1}{\sqrt{(2\pi)^n det(\Sigma_1)}} exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1)\right)}{\frac{1}{\sqrt{(2\pi)^n det(\Sigma_0)}} exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma_0^{-1}(x - \mu_0)\right)} =$$

Классификация двух многомерных нормальных распределений

Распределения

Пусть $Y = \{0, 1\}$, $X = \mathbb{R}^n$ и

$$p(x|y) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma_y)}} exp\left(-\frac{1}{2}(x - \mu_y)^T \Sigma_y^{-1}(x - \mu_y)\right),$$

где μ_y — вектор математического ожидания в классе y, а Σ_y — ковариационная матрица распределения x в классе y

Разделяющая поверхность

$$0 = \ln \frac{p_1}{p_0} + \frac{1}{2} \ln \frac{detK_0}{detK_1} + \frac{1}{2}(x - \mu_0)^T \Sigma_0^{-1}(x - \mu_0) - \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1)$$

Квадратичный дискриминант и линейный дискриминант

Разделяющая поверхность в общем случае

$$a(x) = \frac{1}{2}x^T Ax + (w, x) - b = 0,$$

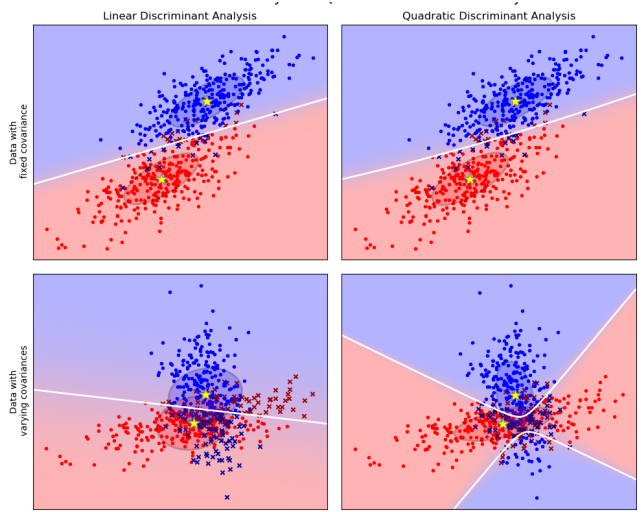
где $A = \Sigma_0^{-1} - \Sigma_1^{-1}$,
$w = \mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}$,
$b = \ln \frac{p_1}{p_0} + \frac{1}{2} \ln \frac{det\Sigma_0}{det\Sigma_1} - \mu_1^T \Sigma_1^{-1} \mu_1 + \mu_0^T \Sigma_0^{-1} \mu_0$.

Разделяющая поверхность при $\Sigma_0 = \Sigma_1$

$$a(x) = (w, x) - b = 0,$$

где $w = (\mu_1 - \mu_0)^T \Sigma^{-1}$,
$b = \ln \frac{p_1}{p_0} - \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_0 + \mu_1)$.

Квадратичный дискриминант и линейный дискриминант[1]



[1] https://scikit-learn.org/stable/auto_examples/classification/plot_lda_qda.html

Наивный байесовский классификатор

Предположение

Все признаки являются независимыми случайными величинами $p(x|y) = \prod_i p_i(x_i|y)$

Восстановление одномерной плотности гораздо более простая задача, чем восстановление многомерной.

Принцип максимума правдоподобия

Задача

Пусть $p(x) = p(x|\theta)$ — параметрическая модель распределения

Принцип максимума правдоподобия

Задача

Пусть $p(x) = p(x|\theta)$ — параметрическая модель распределения

Принцип максимума правдоподобия

$$L(\theta, X_{train}) = \prod_i p(x_i|\theta) \to \max_\theta$$

Принцип максимума правдоподобия

Задача

Пусть $p(x) = p(x|\theta)$ — параметрическая модель распределения

Принцип максимума правдоподобия

$$L(\theta, X_{train}) = \prod_i p(x_i|\theta) \to \max_\theta$$

Необходимое условие максимума

$$\frac{\partial}{\partial\theta} L(\theta, X_{train}) = 0$$

Takeaways

- В некоторых случаях при известном распределении оптимальный классификатор может быть вычислен аналитически

Takeaways

- В некоторых случаях при известном распределении оптимальный классификатор может быть вычислен аналитически
- Для разделения двух гауссиан достаточно квадратичной модели, а иногда и линейной

Takeaways

- В некоторых случаях при известном распределении оптимальный классификатор может быть вычислен аналитически
- Для разделения двух гауссиан достаточно квадратичной модели, а иногда и линейной
- Наивный байесовский классификатор довольно простая модель, которая работает

Takeaways

- В некоторых случаях при известном распределении оптимальный классификатор может быть вычислен аналитически
- Для разделения двух гауссиан достаточно квадратичной модели, а иногда и линейной
- Наивный байесовский классификатор довольно простая модель, которая работает
- Принцип максимума правдоподобия — рабочий инструмент для подбора параметров, если плотность задана некоторым параметрическим семейством

Источники

На основе материалов сайта http://www.machinelearning.ru.