

Введение в искусственный интеллект. Машинное обучение

Тема: Вероятностный подход к классификации

Бабин Д.Н., Иванов И.Е.

кафедра Математической Теории Интеллектуальных Систем



- 1 Вероятностная постановка задач машинного обучения
- 2 Оптимальный байесовский классификатор
- 3 Наивный байесовский классификатор
- 4 Логистическая регрессия
- 5 Перекрестная энтропия (cross entropy)

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

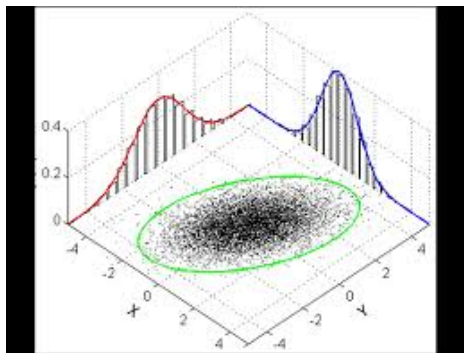
Непрерывный случай

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

$$p(x_i) \geq 0$$

Определения в многомерном случае

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



Математическое ожидание (непрерывный случай)

Пусть $X \sim p(x)$. Тогда

$$EX := \int x dF(x) = \int x p(x) dx$$



Математическое ожидание (непрерывный случай)

Пусть $X \sim p(x)$. Тогда

$$EX := \int x dF(x) = \int x p(x) dx$$

Математическое ожидание (дискретный случай)

Пусть $P(X = x_i) = p_i$ и $\sum_{i=0}^{+\infty} p_i = 1$. Тогда

$$EX := \sum_{i=0}^{+\infty} p_i x_i$$



Дисперсия

$$DX := E(X - EX)^2$$



Дисперсия

$$DX := E(X - EX)^2$$

Среднеквадратическое отклонение

$$\sigma = \sqrt{DX}$$

Определение

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) := \frac{p(x, y)}{p(y)}$$

Определение

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) := \frac{p(x, y)}{p(y)}$$

Формула полной вероятности

$$p(x) = \int_Y p(x|y)p(y)dy \text{ или } p(x) = \sum_{y \in Y} p(x|y)P(y)$$



Определение

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) := \frac{p(x, y)}{p(y)}$$

Формула полной вероятности

$$p(x) = \int_Y p(x|y)p(y)dy \text{ или } p(x) = \sum_{y \in Y} p(x|y)P(y)$$

Теорема Байеса

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int_X p(y|x)p(x)dx}$$

Предположение

Предположение

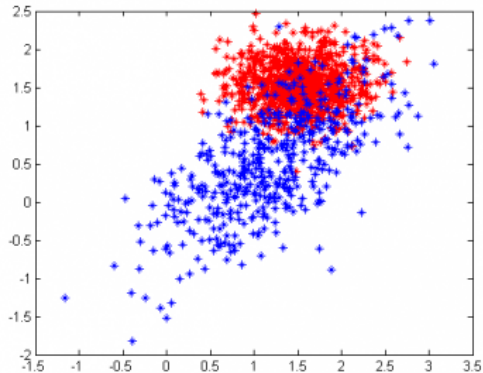
Пусть известно совместное распределение $p(x, y)$ на $X \times Y$.



Предположение

Предположение

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



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 оптимального среднего риска.

Вопрос

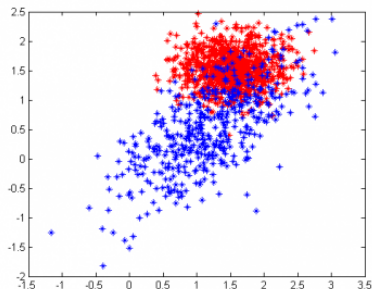
Каким образом задаётся совместное распределение $p(x, y)$?

Вопрос

Каким образом задаётся совместное распределение $p(x, y)$?

Вопрос

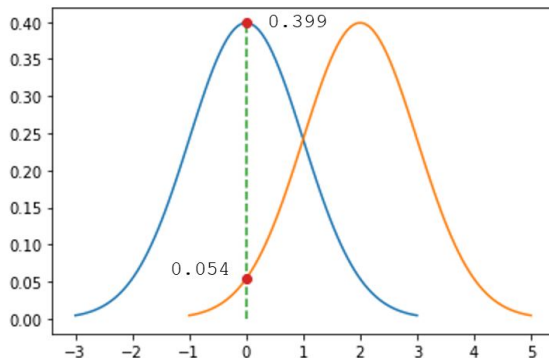
Как разделить объекты разных классов при известном совместном распределении $p(x, y)$?



Бинарная классификация

Дано: $p(x|y = -1) \sim N(\mu = 0, \sigma = 1)$, $p(x|y = 1) \sim N(\mu = 2, \sigma = 1)$

К какому классу отнести объект $x = 0$?

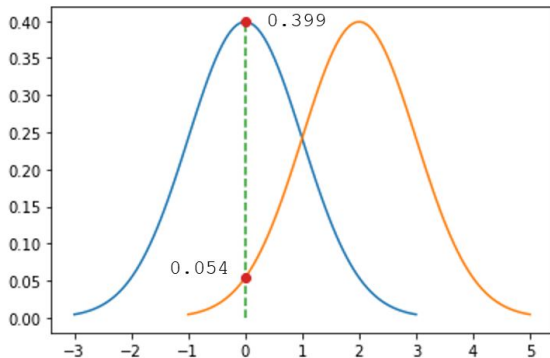


Пример

Бинарная классификация

Дано: $p(x|y = -1) \sim N(\mu = 0, \sigma = 1)$, $p(x|y = 1) \sim N(\mu = 2, \sigma = 1)$

К какому классу отнести объект $x = 0$?



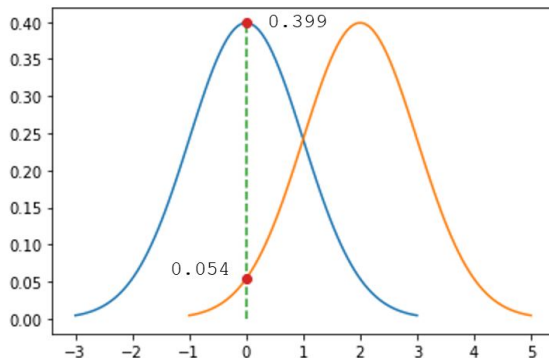
Выводы

- Для того, чтобы задать распределение $p(x, y)$ необходимо задать $P(y)$.

Бинарная классификация

Дано: $p(x|y = -1) \sim N(\mu = 0, \sigma = 1)$, $p(x|y = 1) \sim N(\mu = 2, \sigma = 1)$

К какому классу отнести объект $x = 0$?



Выводы

- Для того, чтобы задать распределение $p(x, y)$ необходимо задать $P(y)$.
- Функция потерь тоже может сильно влиять на решающее правило

Вывод

Рассмотрим простейшую функцию потерь индикатор ошибки $L(a(x), y) = [a(x) \neq y]$ и запишем средний риск

$$R(a) = \iint L(a(x), y) p(x, y) dx dy = \int \sum_Y [a(x) \neq y] p(x|y) P(y) dx =$$

Вывод

Рассмотрим простейшую функцию потерь индикатор ошибки $L(a(x), y) = [a(x) \neq y]$ и запишем средний риск

$$R(a) = \iint L(a(x), y) p(x, y) dx dy = \int \sum_Y [a(x) \neq y] p(x|y) P(y) dx =$$

$$= \int \sum_Y (1 - [a(x) = y]) p(x|y) P(y) dx = \int \sum_Y p(x|y) P(y) dx - \int \sum_Y [a(x) = y] p(x|y) P(y) dx$$

Вывод

Рассмотрим простейшую функцию потерь индикатор ошибки $L(a(x), y) = [a(x) \neq y]$ и запишем средний риск

$$R(a) = \iint L(a(x), y) p(x, y) dx dy = \int \sum_Y [a(x) \neq y] p(x|y) P(y) dx =$$

$$= \int \sum_Y (1 - [a(x) = y]) p(x|y) P(y) dx = \int \sum_Y p(x|y) P(y) dx - \int \sum_Y [a(x) = y] p(x|y) P(y) dx$$

Откуда получаем, что

$$\arg \min_a R(a) = \arg \max_a \int \sum_Y [a(x) = y] p(x|y) P(y) dx = \arg \max_y p(x|y) P(y)$$

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

Следствие

Для бинарного классификатора при $Y = \{-1, 1\}$ разделяющая поверхность может быть записана в следующем виде:

$$\lambda_+ P(y = +1|x) = \lambda_- P(y = -1|x),$$

а сам классификатор:

$$a(x) = \text{sign}(\lambda_+ P(y = +1|x) - \lambda_- P(y = -1|x)) = \text{sign} \left(\frac{P(y = +1|x)}{P(y = -1|x)} - \frac{\lambda_-}{\lambda_+} \right)$$



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





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

$$0 = \ln \frac{p_1}{p_0} + \frac{1}{2} \ln \frac{\det K_0}{\det K_1} + \frac{1}{2} (x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1)$$



Квадратичный дискриминант и линейный дискриминант

Разделяющая поверхность в общем случае

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

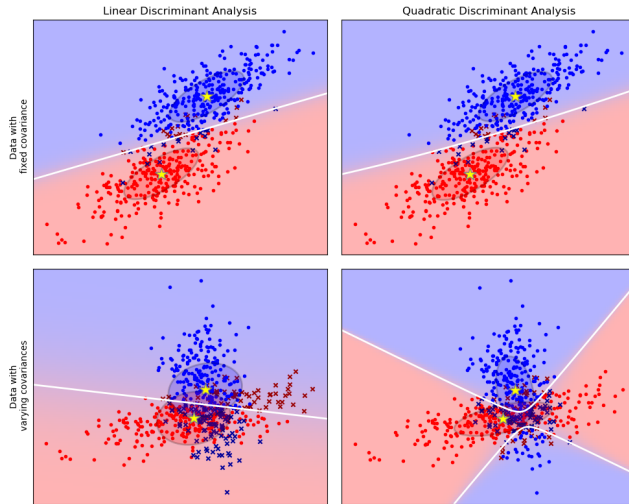
где $A = \Sigma_0^{-1} - \Sigma_1^{-1}$,
 $w = \mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}$,
 $b = \ln \frac{p_1}{p_0} + \frac{1}{2} \ln \frac{\det \Sigma_0}{\det \Sigma_1} - \mu_1^T \Sigma_1^{-1} \mu_1 + \mu_0^T \Sigma_0^{-1} \mu_0$.

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 дискриминант¹



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

Наивный байесовский классификатор

Предположение

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

Наивный байесовский классификатор

$$a(x) = \arg \max_{y \in Y} P(y) \prod_i p(x_i|y)$$

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

Наивный байесовский гауссовский классификатор

Наивный байесовский классификатор

$$a(x) = \arg \max_{y \in Y} P(y) \prod_i p(x_i|y)$$

Дополнительное предположение

$$p_i(x_i|y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$



Наивный байесовский гауссовский классификатор

Наивный байесовский классификатор

$$a(x) = \arg \max_{y \in Y} P(y) \prod_i p(x_i|y)$$

Дополнительное предположение

$$p_i(x_i|y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Настройка параметров

$P(y)$ и параметры распределений μ и σ настраиваются по обучающему множеству



Другие реализации наивного байесовского классификатора в scikit-learn

- BernoulliNB
- CategoricalNB
- MultinomialNB

Мультиномиальное распределение

Определение

Пусть $X = (X_1, \dots, X_m)$ и $n_1 + \dots + n_m = n$, а $p_1, \dots, p_m \geq 0$ и $\sum p_i = 1$.

$$P(X_1 = x_1, \dots, X_m = x_m) := \frac{n!}{x_1! \dots x_m!} p_1^{x_1} \dots p_m^{x_m}$$

Задача

Найдем оптимальный байесовский классификатор для двух классов в случае, когда $p(x|y) \sim \text{Poly}(n, p_1^y, \dots, p_m^y)$



Мультиномиальное распределение

Найдем разделяющую поверхность:

$$p(y = +1|x) = p(y = -1|x)$$

Мультиномиальное распределение

Найдем разделяющую поверхность:

$$p(y = +1|x) = p(y = -1|x)$$

$$p(x|y = +1)p(y = +1) = p(x|y = -1)p(y = -1)$$



Мультиномиальное распределение

Найдем разделяющую поверхность:

$$p(y = +1|x) = p(y = -1|x)$$

$$p(x|y = +1)p(y = +1) = p(x|y = -1)p(y = -1)$$

$$\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1) = \frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)$$



Мультиномиальное распределение

Найдем разделяющую поверхность:

$$p(y = +1|x) = p(y = -1|x)$$

$$p(x|y = +1)p(y = +1) = p(x|y = -1)p(y = -1)$$

$$\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1) = \frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)$$

$$p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1) = p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)$$

Мультиномиальное распределение

Найдем разделяющую поверхность:

$$p(y = +1|x) = p(y = -1|x)$$

$$p(x|y = +1)p(y = +1) = p(x|y = -1)p(y = -1)$$

$$\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1) = \frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)$$

$$p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1) = p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)$$

$$x_1 \ln p_{+1,1} + \dots + x_m \ln p_{+1,m} + \ln p(y = +1) = x_1 \ln p_{-1,1} + \dots + x_m \ln p_{-1,m} + \ln p(y = -1)$$

Вывод 1

Разделяющая поверхность линейна

Мультиномиальное распределение

$$\frac{p(y = +1|x)}{p(y = -1|x)} = \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1)}{\frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)} =$$

Мультиномиальное распределение

$$\begin{aligned}\frac{p(y = +1|x)}{p(y = -1|x)} &= \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1)}{\frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)} = \\ &= \frac{p(y = +1)}{p(y = -1)} \exp(x_1 \ln p_{+1,1} + \dots + x_m \ln p_{+1,m} - x_1 \ln p_{-1,1} + \dots + x_m \ln p_{-1,m}) =\end{aligned}$$

Мультиномиальное распределение

$$\begin{aligned}\frac{p(y = +1|x)}{p(y = -1|x)} &= \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1)}{\frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)} = \\&= \frac{p(y = +1)}{p(y = -1)} \exp(x_1 \ln p_{+1,1} + \dots + x_m \ln p_{+1,m} - x_1 \ln p_{-1,1} + \dots + x_m \ln p_{-1,m}) = \\&= \frac{p(y = +1)}{p(y = -1)} \exp\left(\sum_{i=1}^m x_i \ln \frac{p_{+1,i}}{p_{-1,i}}\right) = \exp\left(\sum_{i=1}^m x_i \ln \frac{p_{+1,i}}{p_{-1,i}} + \ln \frac{p(y = +1)}{p(y = -1)}\right) = e^{(w,x)},\end{aligned}$$

где $x = (x_1, \dots, x_m, 1)$, $w = (\ln \frac{p_{+1,1}}{p_{-1,1}}, \dots, \ln \frac{p_{+1,m}}{p_{-1,m}}, \ln \frac{p(y=+1)}{p(y=-1)})$.



Мультиномиальное распределение

$$\begin{aligned}\frac{p(y = +1|x)}{p(y = -1|x)} &= \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1)}{\frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)} = \\&= \frac{p(y = +1)}{p(y = -1)} \exp(x_1 \ln p_{+1,1} + \dots + x_m \ln p_{+1,m} - x_1 \ln p_{-1,1} + \dots + x_m \ln p_{-1,m}) = \\&= \frac{p(y = +1)}{p(y = -1)} \exp\left(\sum_{i=1}^m x_i \ln \frac{p_{+1,i}}{p_{-1,i}}\right) = \exp\left(\sum_{i=1}^m x_i \ln \frac{p_{+1,i}}{p_{-1,i}} + \ln \frac{p(y = +1)}{p(y = -1)}\right) = e^{(w,x)},\end{aligned}$$

где $x = (x_1, \dots, x_m, 1)$, $w = (\ln \frac{p_{+1,1}}{p_{-1,1}}, \dots, \ln \frac{p_{+1,m}}{p_{-1,m}}, \ln \frac{p(y=+1)}{p(y=-1)})$.

Учитывая, что $p(y = +1|x) + p(y = -1|x) = 1$, получаем, что $\frac{p(y=+1|x)}{1-p(y=+1|x)} = e^{(w,x)}$.



Мультиномиальное распределение

$$\begin{aligned}\frac{p(y = +1|x)}{p(y = -1|x)} &= \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{\frac{n!}{x_1! \dots x_m!} p_{+1,1}^{x_1} \dots p_{+1,m}^{x_m} p(y = +1)}{\frac{n!}{x_1! \dots x_m!} p_{-1,1}^{x_1} \dots p_{-1,m}^{x_m} p(y = -1)} = \\&= \frac{p(y = +1)}{p(y = -1)} \exp(x_1 \ln p_{+1,1} + \dots + x_m \ln p_{+1,m} - x_1 \ln p_{-1,1} + \dots + x_m \ln p_{-1,m}) = \\&= \frac{p(y = +1)}{p(y = -1)} \exp\left(\sum_{i=1}^m x_i \ln \frac{p_{+1,i}}{p_{-1,i}}\right) = \exp\left(\sum_{i=1}^m x_i \ln \frac{p_{+1,i}}{p_{-1,i}} + \ln \frac{p(y = +1)}{p(y = -1)}\right) = e^{(w,x)},\end{aligned}$$

где $x = (x_1, \dots, x_m, 1)$, $w = (\ln \frac{p_{+1,1}}{p_{-1,1}}, \dots, \ln \frac{p_{+1,m}}{p_{-1,m}}, \ln \frac{p(y=+1)}{p(y=-1)})$.

Учитывая, что $p(y = +1|x) + p(y = -1|x) = 1$, получаем, что $\frac{p(y=+1|x)}{1-p(y=+1|x)} = e^{(w,x)}$. Откуда

$$p(y = +1|x) = \frac{1}{1 + e^{-(w,x)}}$$



$$p(y = +1|x) = \frac{1}{1 + e^{-(w,x)}}$$



$$p(y = +1|x) = \frac{1}{1 + e^{-(w,x)}}$$

$$p(y = -1|x) = 1 - p(y = +1|x) = \frac{1}{1 + e^{(w,x)}}$$



$$p(y = +1|x) = \frac{1}{1 + e^{-(w,x)}}$$

$$p(y = -1|x) = 1 - p(y = +1|x) = \frac{1}{1 + e^{(w,x)}}$$

Вывод 2

$$p(y|x) = \sigma((w, x)y),$$

где $\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоида



Мультиномиальное распределение: логистическая регрессия

Вывод 1

Разделяющая поверхность линейна

Вывод 2

$$p(y|x) = \sigma((w, x)y),$$

где $\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоида

Мультиномиальное распределение: логистическая регрессия

Вывод 1

Разделяющая поверхность линейна

Вывод 2

$$p(y|x) = \sigma((w, x)y),$$

где $\sigma(x) = \frac{1}{1+e^{-x}}$ — сигмоида

Определение

Классификационная бинарная модель, в которой вероятность принадлежности к положительному классу задаётся сигмоидой от линейной функции по входу называется **логистической регрессией**

Экспонентное семейство распределений ²

Определение

Будем говорить, что распределение принадлежит экспонентному семейству распределений, если плотность распределения может быть записана в следующем виде:

$$p(x|\theta) = h(x)g(\theta)\exp(\eta(\theta)T(x))$$

Примеры экспонентных распределений: равномерное, нормальное, гипергеометрическое, пуассоновское, биномиальное, Г-распределение и др.

²https://en.wikipedia.org/wiki/Exponential_family

Линейность байесовского классификатора

Предположения

- 1 $T(x) = x$
- 2 $p(x|y) = h(x)g_y(\theta_y)\exp(\eta_y(\theta_y)x)$

Линейность байесовского классификатора

Предположения

- 1 $T(x) = x$
- 2 $p(x|y) = h(x)g_y(\theta_y)\exp(\eta_y(\theta_y)x)$

Теорема о линейности байесовского классификатора

Если для бинарной классификации плотности распределений имеют следующий вид

$$p(x|y) = h(x)g_y(\theta_y)\exp(\eta_y(\theta_y)x)$$

и среди признаков есть константа, то выполнено:

- 1 Разделяющая поверхность линейна $(w, x) = \ln \frac{\lambda_-}{\lambda_+}$
- 2 $p(y|x) = \sigma(\langle w, x \rangle)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$ – логистическая функция (сигмоид)

Доказательство теоремы о линейности байесовского классификатора

$$\frac{p(y = +1|x)}{p(y = -1|x)} = \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = 1)} = \frac{p(y = +1)h(x)g_+(\theta_+)\exp(\eta_+(\theta_+)x)}{p(y = -1)h(x)g_-(\theta_-)\exp(\eta_-(\theta_-)x)} =$$

Доказательство теоремы о линейности байесовского классификатора

$$\frac{p(y = +1|x)}{p(y = -1|x)} = \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{p(y = +1)h(x)g_+(\theta_+)\exp(\eta_+(\theta_+)x)}{p(y = -1)h(x)g_-(\theta_-)\exp(\eta_-(\theta_-)x)} =$$

(выражение перед экспонентой можно внести в скалярное произведение, так как среди признаков есть константа)

$$= \frac{p(y = +1)g_+(\theta_+)}{p(y = -1)g_-(\theta_-)} \exp(\eta_+(\theta_+)x - \eta_-(\theta_-)x) = e^{(w,x)}$$



Доказательство теоремы о линейности байесовского классификатора

$$\frac{p(y = +1|x)}{p(y = -1|x)} = \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{p(y = +1)h(x)g_+(\theta_+)\exp(\eta_+(\theta_+)x)}{p(y = -1)h(x)g_-(\theta_-)\exp(\eta_-(\theta_-)x)} =$$

(выражение перед экспонентой можно внести в скалярное произведение, так как среди признаков есть константа)

$$= \frac{p(y = +1)g_+(\theta_+)}{p(y = -1)g_-(\theta_-)} \exp(\eta_+(\theta_+)x - \eta_-(\theta_-)x) = e^{\langle w, x \rangle}$$

Из полученного выражения и того, что $p(y = +1|x) + p(y = -1|x) = 1$ получаем, что $p(y|x) = \sigma(\langle w, x \rangle)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$.



Доказательство теоремы о линейности байесовского классификатора

$$\frac{p(y = +1|x)}{p(y = -1|x)} = \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = 1)} = \frac{p(y = +1)h(x)g_+(\theta_+)\exp(\eta_+(\theta_+)x)}{p(y = -1)h(x)g_-(\theta_-)\exp(\eta_-(\theta_-)x)} =$$

(выражение перед экспонентой можно внести в скалярное произведение, так как среди признаков есть константа)

$$= \frac{p(y = +1)g_+(\theta_+)}{p(y = -1)g_-(\theta_-)} \exp(\eta_+(\theta_+)x - \eta_-(\theta_-)x) = e^{(w,x)}$$

Из полученного выражения и того, что $p(y = +1|x) + p(y = -1|x) = 1$ получаем, что $p(y|x) = \sigma(\langle w, x \rangle)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$. Для бинарной классификации разделяющая поверхность оптимального байесовского классификатора имеет вид:
 $\frac{p(y=+1|x)}{p(y=-1|x)} - \frac{\lambda_-}{\lambda_+} = e^{(w,x)} - \frac{\lambda_-}{\lambda_+} = 0$, что и завершает доказательство.





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

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

Принцип максимума правдоподобия

Задача

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

Принцип максимума правдоподобия

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

Необходимое условие максимума

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



Логарифмическая функция потерь

$$L = \log \prod_{i=1}^m p(x_i, y_i) \rightarrow \max_w$$

Логарифмическая функция потерь

$$L = \log \prod_{i=1}^m p(x_i, y_i) \rightarrow \max_w$$

Подставим в формулу выражение для логистической регрессии
 $p(x, y) = p(y|x) \cdot p(x) = \sigma(\langle w, x \rangle) \cdot p(x)$:

$$L = \sum_{i=1}^m \log \sigma(\langle w, x_i \rangle y_i) + p(x_i) \rightarrow \max_w$$



Логарифмическая функция потерь

$$L = \log \prod_{i=1}^m p(x_i, y_i) \rightarrow \max_w$$

Подставим в формулу выражение для логистической регрессии $p(x, y) = p(y|x) \cdot p(x) = \sigma(\langle w, x \rangle) \cdot p(x)$:

$$L = \sum_{i=1}^m \log \sigma(\langle w, x_i \rangle y_i) + p(x_i) \rightarrow \max_w$$

Максимизация L эквивалентна следующей задаче минимизации:

$$R = \sum_{i=1}^m \log(1 + \exp(-\langle w, x_i \rangle y_i)) \rightarrow \min_w$$



Бинарная кросс энтропия

Пусть $Y = \{0, 1\}$, $p_i = \sigma(\langle w, x_i \rangle)$. Тогда функция потерь логистической регрессии будет:

$$ce = - \sum_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$



Бинарная перекрестная энтропия

Бинарная кросс энтропия

Пусть $Y = \{0, 1\}$, $p_i = \sigma(\langle w, x_i \rangle)$. Тогда функция потерь логистической регрессии будет:

$$ce = - \sum_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

Замечание

Однослойная нейронная сеть с функцией активации сигмоида и лосс-функцией кросс энтропия — логистическая регрессия.

- 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



