Введение в искусственный интеллект. Машинное обучение

Лекция 3. Вероятностный подход

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

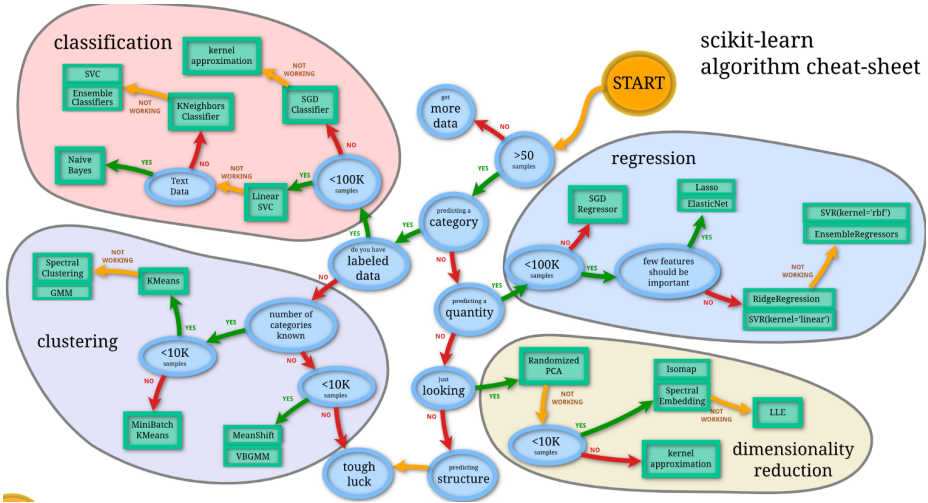
кафедра Математической Теории Интеллектуальных Систем

20 октября 2020г.

План лекции

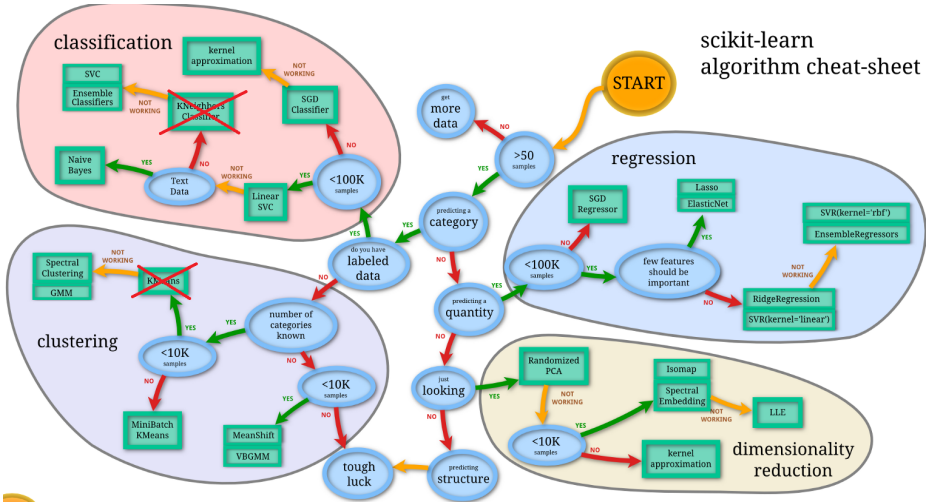
1. Вероятностная постановка задач машинного обучения
2. Оптимальный байесовский регрессор
3. Оптимальный байесовский классификатор
4. Наивный байесовский классификатор
5. Логистическая регрессия
6. Перекрестная энтропия (cross entropy)

Дорожная карта Scikit-Learn[1]



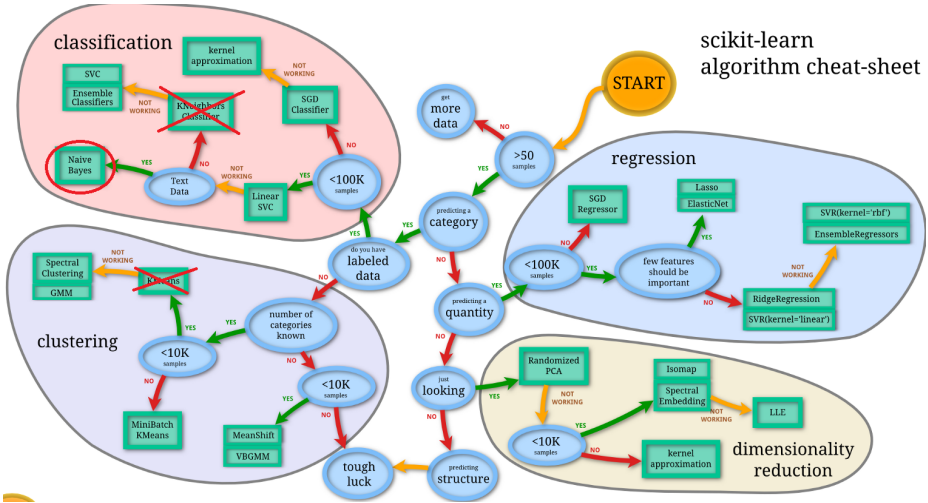
[1] https://scikit-learn.org/stable/tutorial/machine_learning_map/

Дорожная карта Scikit-Learn[1]



[1] https://scikit-learn.org/stable/tutorial/machine_learning_map/

Дорожная карта Scikit-Learn[1]



[1] https://scikit-learn.org/stable/tutorial/machine_learning_map/

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

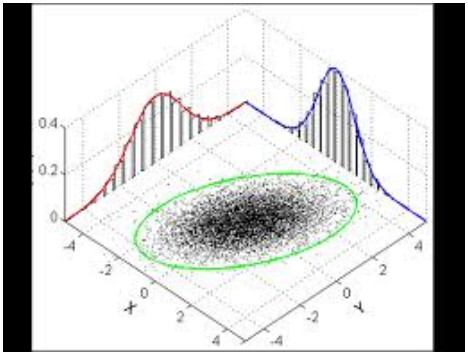
$P(x_i) = p_i$
плотности не существует

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



Математическое ожидание

Математическое ожидание (непрерывный случай)

Пусть $X \sim p(x)$. Тогда

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

Математическое ожидание

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 $\sum\limits_{i=0}^{+\infty} p_i = 1$. Тогда

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

Дисперсия

Дисперсия

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

Дисперсия

Дисперсия

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

Среднеквадратическое отклонение

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

Условная вероятность

Определение

$$p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$$

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

Условная вероятность

Определение

$$p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$$

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

$p(x|\theta)$ — правдоподобие (likelihood)
$p(\theta)$ — априорное распределение (prior)
$p(\theta|x)$ — постреориорное распределение (posterior)

Условная вероятность

Определение

$$p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$$

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

$p(x|\theta)$ — правдоподобие (likelihood)
$p(\theta)$ — априорное распределение (prior)
$p(\theta|x)$ — постреориорное распределение (posterior)

Теорема Байеса

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

Два вида оценивания параметров

Принцип максимального правдоподобия

$$\theta_{ML} = argmax\ p(x|\theta)$$

Два вида оценивания параметров

Принцип максимального правдоподобия

$$\theta_{ML} = argmax\ p(x|\theta)$$

Принцип максимума апостериорной вероятности

$$\theta_{MAP} = argmax\ p(\theta|x)$$

Предположение

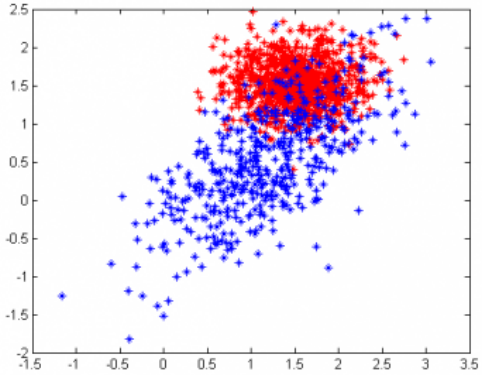
Предположение

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 среднего риска.

Квадратичная функция потерь

Теорема

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

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

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

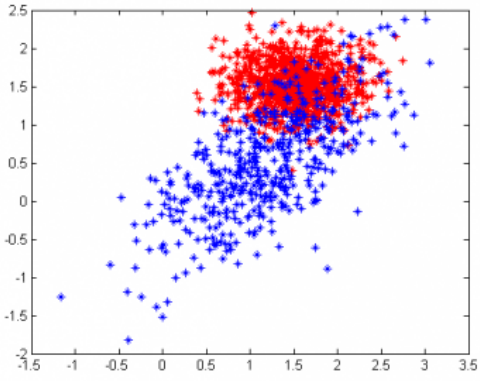
Доказательство

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

Время для вопросов

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

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

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 $R^* = \frac{1}{2}$, 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

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

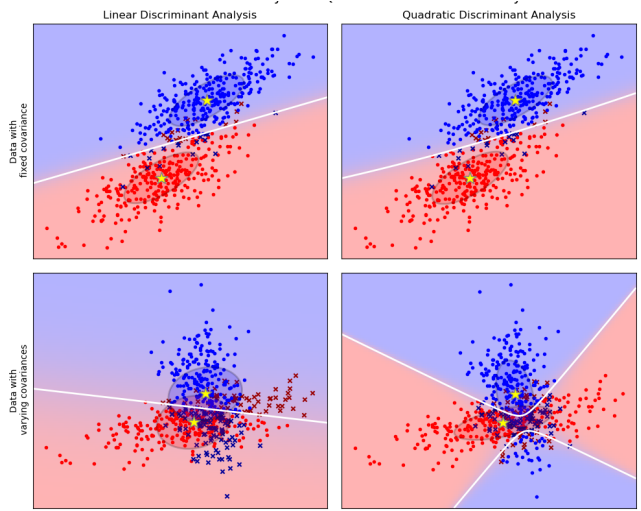
где $A = \Sigma_0^{-1} - \Sigma_1^{-1}$,
$w = \mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}$,
$b = \ln \frac{p_1}{p_0} + \frac{1}{2} \ln \frac{det\Sigma_0}{det\Sigma_1} - \mu_1^T \Sigma_1^{-1} \mu_1 + \mu_0^T \Sigma_0^{-1} \mu_0$.

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[2]



[2] 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.

Мультиномиальное распределение

Определение

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

$$P(X_1 = x_1, ..., X_m = x_m) := \frac{n!}{n_1!...n_m!} p_1^{x_1}...p_m^{x_m}$$

Задача

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

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:

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

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

$$\frac{n!}{n_1!...n_m!}p_{+1,1}^{x_1}...p_{+1,m}^{x_m}p(y = +1) = \frac{n!}{n_1!...n_m!}p_{-1,1}^{x_1}...p_{-1,m}^{x_m})p(y = -1)$$

Мультиномиальное распределение

Найдем разделяющую поверхность:

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

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

$$\frac{n!}{n_1!...n_m!}p_{+1,1}^{x_1}...p_{+1,m}^{x_m}p(y = +1) = \frac{n!}{n_1!...n_m!}p_{-1,1}^{x_1}...p_{-1,m}^{x_m})p(y = -1)$$

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

Мультиномиальное распределение

Найдем разделяющую поверхность:

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

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

$$\frac{n!}{n_1!...n_m!}p_{+1,1}^{x_1}...p_{+1,m}^{x_m}p(y = +1) = \frac{n!}{n_1!...n_m!}p_{-1,1}^{x_1}...p_{-1,m}^{x_m})p(y = -1)$$

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

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

Вывод 1

Разделяющая поверхность линейна

Мультиномиальное распределение

Вывод 1

Разделяющая поверхность линейна

Вывод 2

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

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

Экспонентные распределения

Рассмотрим задачу бинарной классификации $X \in \mathbb{R}^n$, $Y = \{-1, +1\}$, выборка $X^m = (x_i, y_i)_{i=1}^m$ - н.о.р. из распределения $p(x, y) = p(y|x)p(x)$.
Функции правдоподобия - **экспонентные**, т.е.
$p(x|y) = \exp\left(c_y(\delta)\langle\theta_y, x\rangle + b_y(\delta, \theta_y) + d(x, \delta)\right)$, где:

- $\theta_y \in \mathbb{R}^n$ – параметр сдвига,
- δ – параметр разброса,
- b_y, c_y, d – произвольные скалярные функции,
- параметры $d()$ и δ – не зависят от y.

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

Экспонентные распределения: пример

Многомерное нормальное распределение с $\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}$ является экспонентным с:

- параметр сдвига: $\theta = \Sigma^{-1}\mu$,
- параметр разброса: $\delta = \Sigma$.

$N(x; \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) =$
$\exp\left(\mu^T \Sigma^{-1} x - \frac{1}{2}\mu^T \Sigma^{-1} \Sigma \Sigma^{-1} \mu - \frac{1}{2} x^T \Sigma^{-1} x - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma|\right)$.
Если взять:

- $\mu^T \Sigma^{-1} x = \langle \theta, x \rangle$,
- $-\frac{1}{2}\mu^T \Sigma^{-1} \Sigma \Sigma^{-1} \mu = b(\delta, \theta)$,
- $-\frac{1}{2} x^T \Sigma^{-1} x - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| = d(x, \delta)$,

то получаем формулу из класса экспонентных распределений.

Линейность байесовского классификатора

Из предыдущего материала известно, что оптимальный байесовский бинарный классификатор определяется как:
$a(x) = \text{sign}(\lambda_+ p(y=+1|x) - \lambda_- p(y=-1|x)) = \text{sign}\left(\frac{p(y=+1|x)}{p(y=-1|x)} - \frac{\lambda_-}{\lambda_+}\right)$

Теорема о линейности байесовского классификатора

Если распределения $p(x|y)$ экспонентны, параметры $d(), \delta$ не зависят от y, и среди признаков $x_1, \ldots, x_n$ есть константа, то байесовский классификатор линеен:
$a(x) = \text{sign}(\langle w, x\rangle - w_0), w_0 = \ln \frac{\lambda_-}{\lambda_+}$;
при этом апостериорные вероятности классов $p(y|x) = \sigma(\langle w, x\rangle y)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$ – логистическая функция (сигмоид).

Линейность байесовского классификатора

Из предыдущего материала известно, что оптимальный байесовский бинарный классификатор определяется как:

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

Теорема о линейности байесовского классификатора

Если распределения $p(x|y)$ экспонентны, параметры $d(), \delta$ не зависят от y, и среди признаков $x_1, \ldots, x_n$ есть константа, то байесовский классификатор линеен:
$a(x) = \text{sign}(\langle w, x\rangle - w_0), w_0 = \ln \frac{\lambda_-}{\lambda_+}$;
при этом апостериорные вероятности классов $p(y|x) = \sigma(\langle w, x\rangle y)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$ – логистическая функция (сигмоид).

Определение логистической регрессии

Классификационная бинарная модель, в которой вероятность принадлежности к положительному классу задаётся сигмоидой от линейной функции по входу.

Доказательство теоремы о линейности байесовского классификатора

Подставим плотности классов $p(x|\pm) = \exp\left(c_\pm(\delta)\langle\theta_\pm, x\rangle + b_\pm(\delta, \theta_\pm) + d(x, \delta)\right)$ в формулу байесовского классификатора (с логарифмированием): $a(x) = \text{sign}\left(\ln\frac{p(+|x)}{p(-|x)} - \ln\frac{\lambda_-}{\lambda_+}\right)$.

Имеем $\ln\frac{p(+|x)}{p(-|x)} = \langle c_+(\delta)\theta_+ - c_-(\delta)\theta_-, x\rangle + b_+(\delta, \theta_+) - b_-(\delta, \theta_-)$.

Поскольку $c_+(\delta)\theta_+ - c_-(\delta)\theta_-$ и $b_+(\delta, \theta_+) - b_-(\delta, \theta_-)$ не зависят от x, а также в x есть константный признак (куда можно занести второй член), то

$\frac{p(y=+1|x)}{p(y=-1|x)} = e^{\langle w, x\rangle}$ для некоторого вектора весов w.

По формуле полной вероятности $p(y = +1|x) + p(y = -1|x) = 1$, имеем систему из двух уравнений на два неизвестных $p(\pm|x)$, решая которую, получим:

$p(y = +1|x) = \frac{1}{1+e^{-\langle w,x\rangle}}$, $p(y = -1|x) = \frac{1}{1+e^{\langle w,x\rangle}}$. Более компактно:

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

Т.о., разделяющая поверхность линейна:

$\lambda_+ p(y = +1|x) = \lambda_- p(y = -1|x) \Rightarrow \langle w, x\rangle - \ln\frac{\lambda_-}{\lambda_+} = 0$. Ч.т.д.

Время для вопросов

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

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

Логарифмическая функция потерь

- $L(w, X^m) = \log \prod_{i=1}^{m} p(x_i, y_i) \to \max_w$

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

- $L(w, X^m) = \sum_{i=1}^{m} \log \sigma(\langle w, x_i \rangle y_i) + const(w) \to \max_w$

Максимизация L эквивалентна минимизации аппроксимированного Э.Р. R:

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

Бинарная перекрестная энтропия

Бинарная кросс энтропия

Пусть $Y = \{0, 1\}$, $p_1 = \sigma(\langle w, x \rangle)$ и $p_0 = 1 - p_1$. 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 $Y = \{0, 1\}$, $p_1 = \sigma(\langle w, x \rangle)$ и $p_0 = 1 - p_1$. Тогда функция потерь логистической регрессии будет:

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

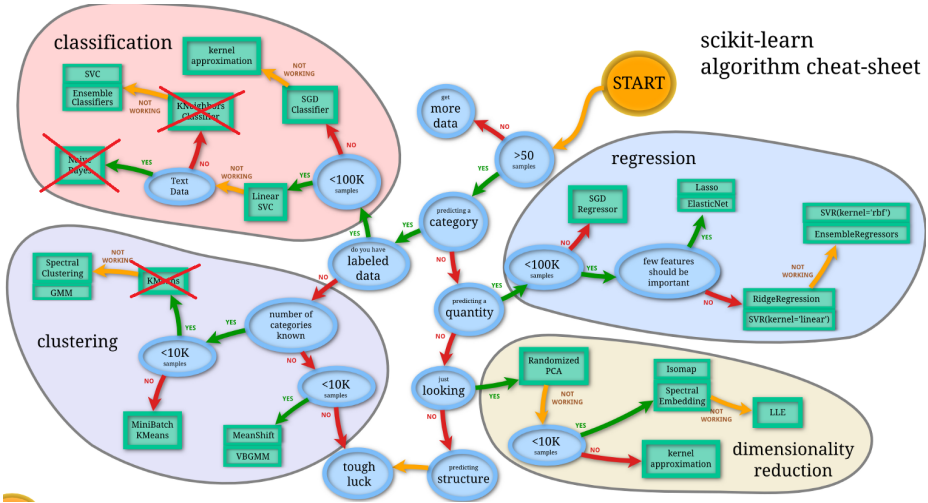
Замечание

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

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
- Логистическая регрессия — это однослойная нейронная сеть с активацией сигмоидой (или софтмакс) и лосс-функцией кросс энтропия

Дорожная карта Scikit-Learn[3]



[3] https://scikit-learn.org/stable/tutorial/machine_learning_map/

Время для вопросов

Источники

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