

СПАМ-ФИЛЬТРЫ И НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР

ЗАДАЧА ФИЛЬТРАЦИИ СПАМА

- › Задача классификации писем на 2 класса — спам (spam) и не спам (ham):

$$Y = \{\text{spam}, \text{ham}\}$$

- › Первые спам-фильтры использовали наивный байесовский классификатор
- › Придем к нему, оттолкнувшись от задачи фильтрации спама

ПРИМЕРЫ СПАМА

- › Hi! :) Purchase Exclusive Tabs Online <http://...>
- › We Offer Loan At A Very Low Rate Of 3%. If Interested, Kindly Contact Us, Reply by this email@hotmail.com
- › Купите специализацию Машинное обучение и анализ данных от МФТИ и Яндекса с супер-скидкой 0,99%! Станьте Data Scientist за 5 месяцев!

ФИЛЬТРУЕМ СПАМ: ОБУЧЕНИЕ

- › Посчитать для каждого слова w из коллекции текстов количество писем с ним n_{ws} в спаме (spam) и количество писем с ним n_{wh} в «не спаме» (ham)

ФИЛЬТРУЕМ СПАМ: ОБУЧЕНИЕ

- › Посчитать для каждого слова w из коллекции текстов количество писем с ним n_{ws} в спаме (spam) и количество писем с ним n_{wh} в «не спаме» (ham)
- › Оценить вероятность появления каждого слова w в спамном и в неспамном тексте:

$$P(w \mid \text{spam}) = n_{ws}/n_s$$

$$P(w \mid \text{ham}) = n_{wh}/n_s$$

ФИЛЬТРУЕМ СПАМ: ПРИМЕНЕНИЕ

- › Получив текст письма, для которого нужно определить, относится оно к спаму или нет, мы можем:
 1. Оценить вероятность появления всего текста в классе «спам» и в классе «не спам» просто произведением вероятностей слов:
$$P(\text{text} \mid \text{spam}) = P(w_1 \mid \text{spam})P(w_2 \mid \text{spam})\dots P(w_N \mid \text{spam})$$
$$P(\text{text} \mid \text{ham}) = P(w_1 \mid \text{ham})P(w_2 \mid \text{ham})\dots P(w_N \mid \text{ham})$$

ФИЛЬТРУЕМ СПАМ: ПРИМЕНЕНИЕ

- › Получив текст письма, для которого нужно определить, относится оно к спаму или нет, мы можем:
 1. Оценить вероятность появления всего текста в классе «спам» и в классе «не спам» просто произведением вероятностей слов:
$$P(\text{text} \mid \text{spam}) = P(w_1 \mid \text{spam})P(w_2 \mid \text{spam})\dots P(w_N \mid \text{spam})$$
$$P(\text{text} \mid \text{ham}) = P(w_1 \mid \text{ham})P(w_2 \mid \text{ham})\dots P(w_N \mid \text{ham})$$

«Наивная» гипотеза: входления слов в текст — независимые события

ФИЛЬТРУЕМ СПАМ: ПРИМЕНЕНИЕ

› Получив текст письма, для которого нужно определить, относится оно к спаму или нет, мы можем:

2. Выбрать тот класс, в котором вероятность возникновения этого текста больше:

$$a(\text{text}) = \operatorname{argmax}_y P(\text{text} | y)$$

› Это **почти** правильный алгоритм

УТОЧНЯЕМ АЛГОРИТМ

- › На самом деле

$$\cancel{a(\text{text}) = \operatorname{argmax}_y P(\text{text} | y)}$$

- › Нам нужна вероятность $P(y | \text{text})$,
а не $P(\text{text} | y)$

УТОЧНЯЕМ АЛГОРИТМ

- › На самом деле

$$\cancel{a(\text{text}) = \operatorname{argmax}_y P(\text{text} | y)}$$

- › Нам нужна вероятность $P(y | \text{text})$,
а не $P(\text{text} | y)$
- › Значит, правильный алгоритм должен
выбирать тот класс, для которого больше
вероятность $P(y | \text{text})$:

$$a(\text{text}) = \operatorname{argmax}_y P(y | \text{text})$$

НО КАК ЖЕ ЕЁ ОЦЕНИТЬ?



ТЕОРЕМА БАЙЕСА

$$P(y|\text{text}) = P(\text{text}|y)P(y)/P(\text{text})$$

› Т.к. $P(\text{text})$ одинакова для обоих классов:

$$\operatorname{argmax}_y P(y|\text{text}) = \operatorname{argmax}_y P(\text{text}|y)P(y)$$

СПАМ-ФИЛЬТР НА НАИВНОМ БАЙЕСОВСКОМ КЛАССИФИКАТОРЕ

» Обучение:

$$P(w | \text{spam}) = n_{ws}/n_s$$

$$P(w | \text{ham}) = n_{wh}/n_s$$

» Применение:

$$\begin{aligned} & P(\text{new text} | \text{spam}) = \\ & = P(w_1 | \text{spam})P(w_2 | \text{spam})...P(w_N | \text{spam}) \end{aligned}$$

$$\begin{aligned} & P(\text{new text} | \text{ham}) = \\ & = P(w_1 | \text{ham})P(w_2 | \text{ham})...P(w_N | \text{ham}) \end{aligned}$$

Отнести текст к такому классу y (ham или spam),
для которого будет больше величина

$$P(y)P(\text{new text} | y)$$

ФИЛЬТРАЦИЯ СПАМА: ЧТО ЕЩЁ НЕ УЧЛИ

- › Работали только с текстом письма
- › Если в тексте есть слово w , которое ни разу не встретилось в спаме в нашей обучающей выборке, то $P(w | \text{spam}) = 0$
- › В тексте может быть как слово, никогда не встречавшееся в спаме, так и слово, никогда не встречавшееся не в спаме

РЕЗЮМЕ

- › Как решали задачу фильтрации спама на заре развития Интернета
- › Идея наивного байесовского классификатора
- › «Наивный байес» прост в реализации и быстро работает
- › Далее мы рассмотрим байесовский классификатор в более общем виде

БАЙЕСОВСКИЙ КЛАССИФИКАТОР

БАЙЕСОВСКИЙ КЛАССИФИКАТОР

- › По известному вектору признаков x алгоритм относит объект к классу $a(x)$ по правилу:

$$a(x) = \operatorname{argmax}_y P(y|x)$$

БАЙЕСОВСКИЙ КЛАССИФИКАТОР

$$a(x) = \operatorname{argmax}_y P(y|x)$$



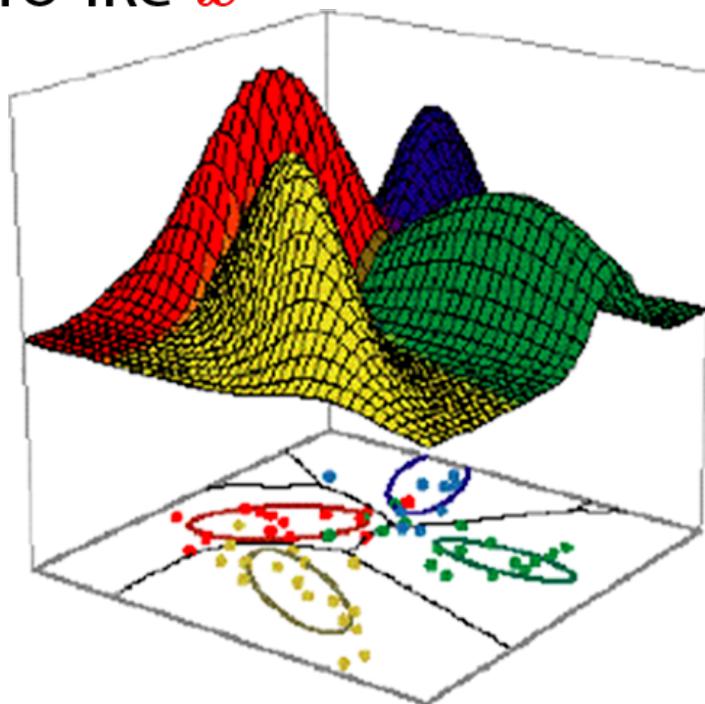
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$a(x) = \operatorname{argmax}_y P(x|y)P(y)$$

БАЙЕСОВСКИЙ КЛАССИФИКАТОР

$$a(x) = \operatorname{argmax}_y P(x|y)P(y)$$

Если $P(y)$ одинаковы для всех классов – мы просто выбираем класс, плотность которого больше в точке x



ЗАЧЕМ НАМ ПОНАДОБИЛАСЬ ТЕОРЕМА БАЙЕСА?

- › $P(y|x)$ — вероятность класса y при признаках x
- › X часто из вещественных чисел и признаков часто очень много
- › Всевозможных значений признаков так много, что скорее всего каждый вектор x встретится только один или несколько раз
- › Этого недостаточно для оценки $P(y|x)$

ЧТО ОЦЕНИВАЕТСЯ ПО ОБУЧАЮЩЕЙ ВЫБОРКЕ

- › $P(x|y)$ — вероятность увидеть набор признаков x в классе y , если x дискретный
- › Если координаты вектора x — вещественные, $P(x|y)$ — плотность распределения x
- › Именно эту величину и можно оценивать по обучающей выборке
- › А затем подставлять в классификатор:

$$a(x) = \operatorname{argmax}_y P(x|y)P(y)$$

ПРОБЛЕМА НЕХВАТКИ ДАННЫХ

- › Пример: в обучающей выборке **100 000** объектов с **10 000** признаков
- › **100 000** точек в пространстве размерности **10 000** — очень мало
- › Например, если x — бинарный, то у него может быть 2^{10000} значений, что сильно больше **100 000**
- › Поэтому восстановить $P(x|y)$ как функцию от признаков x довольно трудно

РЕЗЮМЕ

- › Байесовский классификатор:

$$a(x) = \operatorname{argmax}_y P(x|y)P(y)$$

- › Обучение: оценить по выборке $P(x|y)$ и $P(y)$
- › Оценивать $P(x|y)$ как функцию многих переменных затруднительно — нужно много данных

ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ (ЧАСТЬ 1)

ПРОБЛЕМА НЕХВАТКИ ДАННЫХ

- › Пример: в обучающей выборке **100 000** объектов с **10 000** признаков
- › **100 000** точек в пространстве размерности **10 000** — очень мало
- › Например, если x — бинарный, то у него может быть 2^{10000} значений, что сильно больше **100 000**
- › Поэтому восстановить $P(x|y)$ как функцию от признаков x довольно трудно

РЕШЕНИЕ 1 — «НАИВНЫЙ БАЙЕС»

- › Свести задачу восстановления $P(x|y)$ от оценки функции многих переменных к оценке функций одной переменной

НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР

- » Байесовский классификатор:

$$a(x) = \operatorname{argmax}_y P(y|x) = \operatorname{argmax}_y P(x|y)P(y)$$

- » С «наивной» гипотезой:

$$P(x|y) = P(x_{(1)}|y)P(x_{(2)}|y)\dots P(x_{(N)}|y)$$

$x_{(k)}$ — k -ый признак объекта x

ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ

Оцениваем $P(y)$ и $P(x_{(k)}|y)$

ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ

Оцениваем $P(y)$ и $P(x_{(k)}|y)$

- › $P(y)$ можно оценить как долю объектов класса y в обучающей выборке:

$$P(y) \approx \frac{\ell_y}{\ell}$$

ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ

Оцениваем $P(y)$ и $P(x_{(k)}|y)$

- » $P(x_{(k)}|y)$ можно оценить как долю объектов с данным значением признака $x_{(k)}$ среди объектов класса y :

$$P(x_{(k)}|y) \approx \frac{1}{\ell_y} \#(x_{(k)}, y)$$

- » Т.е. для бинарных признаков:

$$P(x_{(k)} = 1|y) \approx \frac{1}{\ell_y} \#(x_{(k)} = 1, y)$$

$$P(x_{(k)} = 0|y) \approx \frac{1}{\ell_y} \#(x_{(k)} = 0, y)$$

ПРИМЕР: КЛАССИФИКАЦИЯ ТЕКСТОВ

- › Составим словарь по обучающей выборке и в качестве признаков будем использовать вхождения слов из словаря в текст: **1**, если входит и **0**, если нет

ПРИМЕР: КЛАССИФИКАЦИЯ ТЕКСТОВ

- › Составим словарь по обучающей выборке и в качестве признаков будем использовать вхождения слов из словаря в текст: **1**, если входит и **0**, если нет
- › Восстановим распределения:

$$P(x_{(k)} = 1|y) \approx \frac{1}{\ell_y} \#(x_{(k)} = 1, y)$$

$$P(x_{(k)} = 0|y) \approx \frac{1}{\ell_y} \#(x_{(k)} = 0, y)$$

ПРИМЕР: КЛАССИФИКАЦИЯ ТЕКСТОВ

- › Составим словарь по обучающей выборке и в качестве признаков будем использовать вхождения слов из словаря в текст: **1**, если входит и **0**, если нет
- › Восстановим распределения:

$$P(x_{(k)} = 1|y) \approx \frac{1}{\ell_y} \#(x_{(k)} = 1, y)$$

$$P(x_{(k)} = 0|y) \approx \frac{1}{\ell_y} \#(x_{(k)} = 0, y)$$

- › Применим наивный байесовский классификатор

СГЛАЖИВАНИЕ ВЕРОЯТНОСТЕЙ

- › Если значение признака $x_{(k)}$ никогда не встречалось в классе y : $P(x_{(k)}|y) = 0$
- › Тогда из-за одного признака все произведение
$$P(x_{(1)}|y)P(x_{(2)}|y)...P(x_{(N)}|y) = 0$$
- › Выход – сглаживание вероятностей:

$$P(x_{(k)} = 1|y) \approx \frac{\#(x_{(k)}=1,y)+a}{\ell_y+a+b}$$

$$P(x_{(k)} = 0|y) \approx \frac{\#(x_{(k)}=0,y)+b}{\ell_y+a+b}$$

РЕЗЮМЕ

- › Проблема нехватки данных и наивный байесовский классификатор
- › Сведение задачи к восстановлению N одномерных распределений
- › В случае бинарных признаков распределения можно восстанавливать простыми частотными оценками

ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ (ЧАСТЬ 2)

ПРОБЛЕМА НЕХВАТКИ ДАННЫХ

- › Пример: в обучающей выборке **100 000** объектов с **10 000** признаков
- › **100 000** точек в пространстве размерности **10 000** — очень мало
- › Например, если x — бинарный, то у него может быть 2^{10000} значений, что сильно больше **100 000**
- › Поэтому восстановить $P(x|y)$ как функцию от признаков x довольно трудно

РЕШЕНИЕ 1 — «НАИВНЫЙ БАЙЕС»

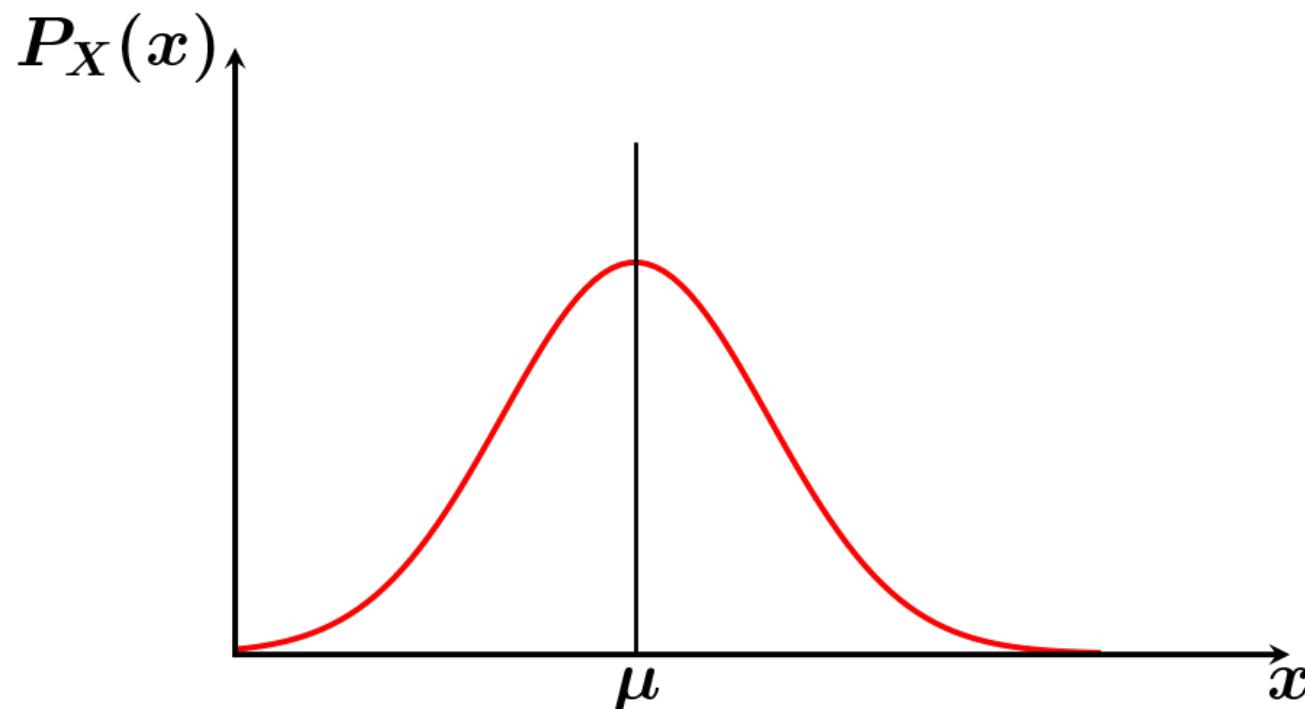
- › Свести задачу восстановления $P(x|C)$ от оценки функции многих переменных к оценке функций одной переменной
- › В предыдущем видео мы рассмотрели случай бинарных признаков, знакомый нам по примеру со спам-фильтром

ПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ РАСПРЕДЕЛЕНИЙ

- › Признаки бывают еще и вещественными
- › Тогда наши формулы для восстановления $P(x_{(k)}|y)$ уже не подойдут
- › Можем предположить, что распределение признаков похоже на какое-то стандартное — пуассоновское, экспоненциальное, нормальное
- › И попробовать восстановить его

ПРИМЕР: НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

$$X \sim \mathcal{N}(\mu, \sigma^2)$$



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ПРИМЕР: НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

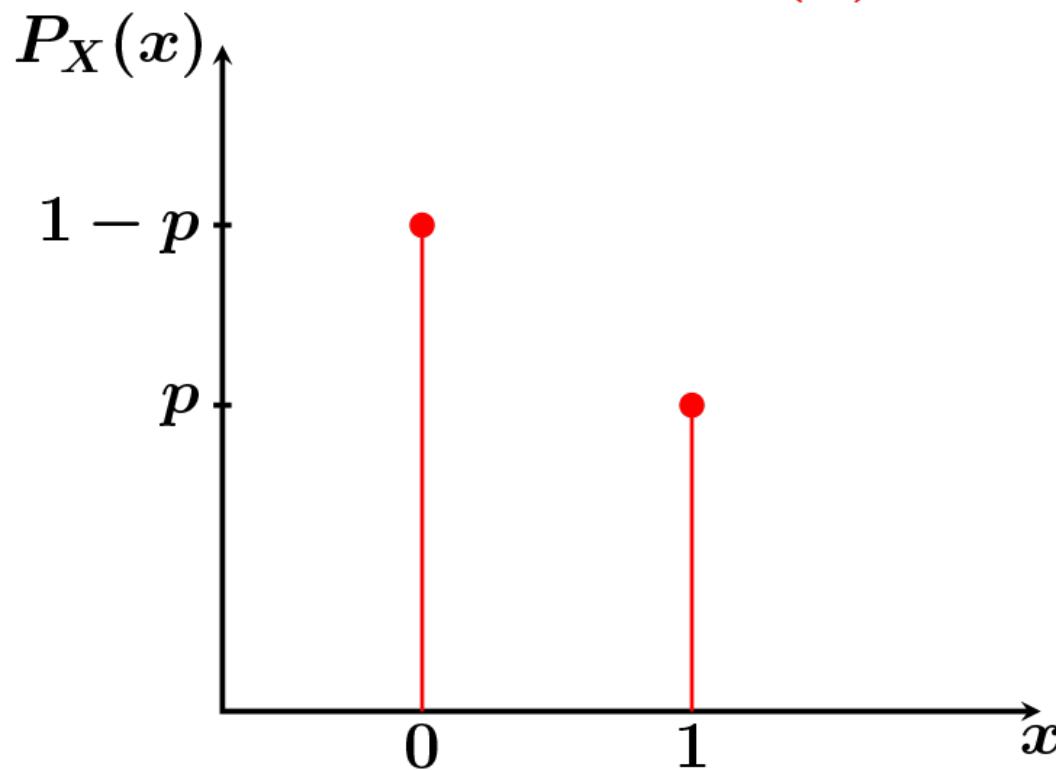
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Несмешённый вариант оценки для дисперсии:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

ДРУГОЙ ПРИМЕР: РАСПРЕДЕЛЕНИЕ БЕРНУЛЛИ

$$X \sim \text{Bernoulli}(p)$$



$$P(x = 1) = p$$

$$P(x = 0) = 1 - p$$

ДРУГОЙ ПРИМЕР: РАСПРЕДЕЛЕНИЕ БЕРНУЛЛИ

- › p можно оценить долей случаев, в которых случайная величина равнялась 1:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^N [x_i = 1]$$

- › Получаем рассмотренные нами ранее оценки распределения бинарных признаков

РЕКОМЕНДАЦИИ ПО ВЫБОРУ РАСПРЕДЕЛЕНИЙ

- › Данные с разреженными дискретными признаками — мультиномиальное распределение
- › Непрерывные признаки с маленьким разбросом — нормальное распределение
- › Непрерывные признаки с выбросами в обучающей выборке — можно попробовать более «размазанные» распределения

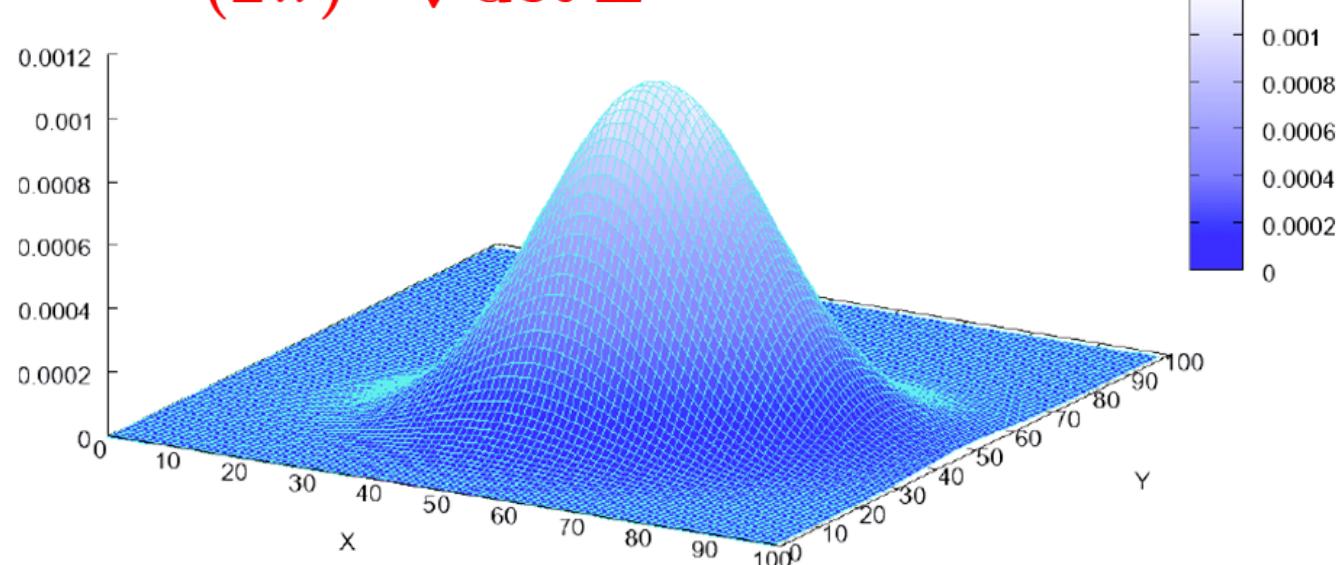
ПАРАМЕТРИЧЕСКАЯ ОЦЕНКА МНОГОМЕРНОГО РАСПРЕДЕЛЕНИЯ

- › Применить для восстановления $P(x|y)$ параметрический подход, но восстанавливать сразу многомерное распределение
- › Т.к. приближаем $P(x|y)$ распределением из некоторого узкого класса, число параметров может быть приемлемым

ПАРАМЕТРИЧЕСКАЯ ОЦЕНКА МНОГОМЕРНОГО РАСПРЕДЕЛЕНИЯ

- › Пример: многомерное нормальное распределение

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$



- › Параметры: вектор средних μ и матрица ковариаций Σ

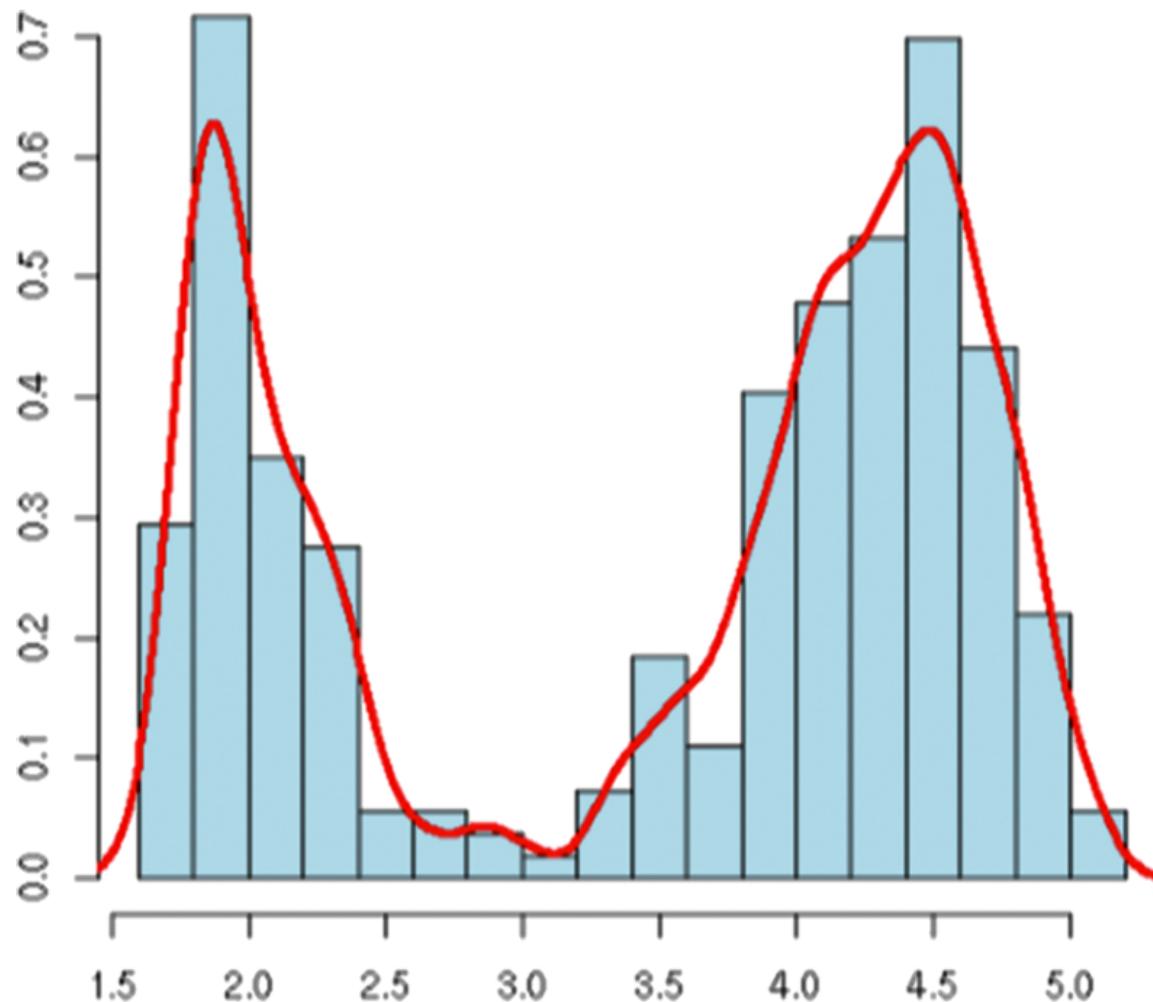
НЕДОСТАТКИ ПОДХОДА

- › Возникает больше параметров, чем в «наивном» подходе
- › Для нормального распределения: n средних и n дисперсий в наивном подходе против вектора средних размерности n и матрицы ковариаций $n \times n$
- › Оценка параметров может получиться неправильной из-за нехватки данных
- › Часто требуется выполнять «неустойчивые» операции — например, обращение матриц, которые почти вырождены

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ

- › Оценивать $P(x|y)$ можно не в точке, а в точке и ее окрестности
- › Примеры, которые ближе к точке — учитывать с большим весом, те, которые дальше — с меньшим

НЕПАРАМЕТРИЧЕСКОЕ ВОССТАНОВЛЕНИЕ ПЛОТНОСТИ



РЕЗЮМЕ

- › Проблема нехватки данных для восстановления распределений
- › Наивный байесовский классификатор
- › Параметрическая оценка многомерной плотности
- › Непараметрическая оценка многомерной плотности

МИНИМИЗАЦИЯ РИСКА

БАЙЕСОВСКИЙ КЛАССИФИКАТОР

$$a(x) = \operatorname{argmax}_y P(y)P(x|y)$$

x — признаковое описание

y — класс

БАЙЕСОВСКАЯ РЕГРЕССИЯ

$$a(x) = \operatorname{argmax}_y P(y)P(x|y)$$

x — признаковое описание

y — прогнозируемая величина

БАЙЕСОВСКАЯ РЕГРЕССИЯ

$$\cancel{a(x) = \operatorname{argmax}_y P(y)P(x|y)} ?$$

x — признаковое описание

y — прогнозируемая величина

Вряд ли получится восстановить $P(x|y)$

БАЙЕСОВСКАЯ РЕГРЕССИЯ

$$\cancel{a(x) = \operatorname{argmax}_y P(y)P(x|y)} ?$$

x — признаковое описание

y — прогнозируемая величина

Вряд ли получится восстановить $P(x|y)$

$a(x) = \operatorname{argmax}_y P(y|x)$ тоже сомнительный
вариант для регрессии

ШТРАФЫ ЗА ОШИБКИ

В классификации

- › Разные ошибки классификации могут быть в разной степени критичны
- › Пример: классификация мест, в которых может быть обнаружено месторождение нефти на классы «есть нефть» и «нет нефти».
- › Можем захотеть назначить разные штрафы за разные ошибки

ШТРАФЫ ЗА ОШИБКИ

В регрессии

› Зависимость в любом случае восстанавливается неточно

› Квадратичные потери:

$$\text{MSE} = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - a(x_i))^2$$

› Сумма модулей отклонений:

$$\text{MAE} = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - a(x_i)|$$

БОЛЕЕ ОБЩИЙ ПОДХОД

- › Для объекта x мы делаем прогноз $a(x)$
- › Правильный ответ на этом объекте y
- › Величина ошибки алгоритма $L(y, a(x))$
(функцию выбираем сами)
- › Пример 1: для классификации можно взять
 $L(y, a(x)) = [y \neq a(x)]$
- › Пример 2: для регрессии подойдет
 $L(y, a(x)) = (y - a(x))^2$

ФУНКЦИОНАЛ РИСКА

$$R(a(x), x) = \mathbb{E}(L(y, a(x))|x)$$

Можно давать на объекте x ответ, который минимизирует ожидаемую ошибку:

$$a(x) = \operatorname{argmin}_s R(s, x)$$

ОПТИМАЛЬНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР

» Для классификации:

$$\begin{aligned} R(a(x), x) &= \mathbb{E}(L(y, a(x))|x) = \\ &= \sum_{y \in Y} L(y, a(x))P(y|x) \end{aligned}$$

$$a(x) = \underset{s}{\operatorname{argmin}} R(s, x) =$$

$$= \underset{s}{\operatorname{argmin}} \sum_{y \in Y} L(y, s)P(y|x) =$$

$$= \boxed{\underset{s}{\operatorname{argmin}} \sum_{y \in Y} L(y, s)P(y)P(x|y)}$$

» Реальный классификатор в точности оптимальным не будет из-за погрешности восстановления плотностей

ОПТИМАЛЬНЫЙ БАЙЕСОВСКИЙ РЕГРЕССОР

» Для регрессии:

$$\begin{aligned} R(a(x), x) &= \mathbb{E}(L(y, a(x))|x) = \\ &= \int_{y \in Y} L(y, a(x))p(y|x)dy \end{aligned}$$

$$\begin{aligned} a(x) &= \underset{s}{\operatorname{argmin}} R(s, x) = \\ &= \boxed{\underset{s}{\operatorname{argmin}} \int_{y \in Y} L(y, s)p(y|x)dy} \end{aligned}$$

ФУНКЦИОНАЛ СРЕДНЕГО РИСКА

» Можно рассмотреть $R(a) = \mathbb{E}_x R(a(x), x)$
(по всем x из X)

» Для определённости рассмотрим случай классификации объектов с дискретными признаками:

$$R(a) = \sum_{x \in X} R(a(x), x) P(x)$$

» Можно оценить $R(a)$ снизу:

$$\sum_{x \in X} R(a(x), x) P(x) \geq \sum_{x \in X} P(x) \min_s R(s, x)$$

ФУНКЦИОНАЛ СРЕДНЕГО РИСКА

- › Если $a(x)$ — оптимальный байесовский, он минимизирует $R(a(x), x)$
- › Значит, оценка достигается и $R(a)$ он тоже минимизирует

РЕЗЮМЕ

- › Проблемы обобщения байесовской классификации на случай регрессии и учёта различных штрафов за различные ошибки в прогнозах
- › Функционал риска и байесовский классификатор в более общем виде
- › Байесовская регрессия
- › Минимизация среднего риска

МИНИМИЗАЦИЯ РИСКА И АНАЛИЗ ФУНКЦИИ ПОТЕРЬ

ОПТИМАЛЬНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР

$$a(x) = \operatorname{argmin}_s \sum_{y \in Y} L(s, y) P(y) P(x|y)$$

ЧАСТНЫЙ СЛУЧАЙ

» Если $L(s, y) = [y \neq s]$:

$$\sum_{y \in Y} L(s, y) P(y|x) \rightarrow \min$$

ЧАСТНЫЙ СЛУЧАЙ

» Если $L(s, y) = [y \neq s]$:

$$\sum_{y \in Y} L(s, y) P(y|x) \rightarrow \min$$

$$\sum_{y \in Y \setminus \{s\}} P(y|x) = \left(\sum_{y \in Y} P(y|x) \right) - P(s|x) \rightarrow \min$$

ЧАСТНЫЙ СЛУЧАЙ

» Если $L(s, y) = [y \neq s]$:

$$\sum_{y \in Y} L(s, y) P(y|x) \rightarrow \min$$

$$\sum_{y \in Y \setminus \{s\}} P(y|x) = \left(\sum_{y \in Y} P(y|x) \right) - P(s|x) \rightarrow \min$$
$$P(s|x) \rightarrow \max$$

ЧАСТНЫЙ СЛУЧАЙ

$$a(x) = \operatorname{argmin}_y P(y|x) = \operatorname{argmin}_y P(y)P(x|y)$$

КВАДРАТИЧНАЯ ФУНКЦИЯ ПОТЕРЬ В РЕГРЕССИИ

$$\int_{\mathbf{Y}} (t - y)^2 p(y|x) dy \rightarrow \min_t$$

КВАДРАТИЧНАЯ ФУНКЦИЯ ПОТЕРЬ В РЕГРЕССИИ

$$\int_{\mathbf{Y}} (t - y)^2 p(y|x) dy \rightarrow \min_t$$

$$\frac{\partial}{\partial t} \int_{\mathbf{Y}} (t - y)^2 p(y|x) dy = 2 \int_{\mathbf{Y}} (t - y) p(y|x) dy =$$

КВАДРАТИЧНАЯ ФУНКЦИЯ ПОТЕРЬ В РЕГРЕССИИ

$$\int_{\mathbf{Y}} (t - y)^2 p(y|x) dy \rightarrow \min_t$$

$$\begin{aligned} \frac{\partial}{\partial t} \int_{\mathbf{Y}} (t - y)^2 p(y|x) dy &= 2 \int_{\mathbf{Y}} (t - y) p(y|x) dy = \\ &= 2 \left(\int_{\mathbf{Y}} tp(y|x) dy - \int_{\mathbf{Y}} yp(y|x) dy \right) = 0 \end{aligned}$$

КВАДРАТИЧНАЯ ФУНКЦИЯ ПОТЕРЬ В РЕГРЕССИИ

$$\int_{\mathbf{Y}} (t - y)^2 p(y|x) dy \rightarrow \min_t$$

$$\begin{aligned} \frac{\partial}{\partial t} \int_{\mathbf{Y}} (t - y)^2 p(y|x) dy &= 2 \int_{\mathbf{Y}} (t - y) p(y|x) dy = \\ &= 2 \left(\int_{\mathbf{Y}} tp(y|x) dy - \int_{\mathbf{Y}} yp(y|x) dy \right) = 0 \end{aligned}$$

$$a(x) = t = \int_{\mathbf{Y}} yp(y|x) dy = E(y|x)$$

АБСОЛЮТНОЕ ОТКЛОНЕНИЕ

$$\int_{\mathbf{Y}} |t - y| p(y|x) dy \rightarrow \min_t$$

АБСОЛЮТНОЕ ОТКЛОНЕНИЕ

$$\int_{\mathbf{Y}} |t - y| p(y|x) dy \rightarrow \min_t$$

$$\frac{\partial}{\partial t} \int_{\mathbf{Y}} |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{\mathbf{Y} \setminus \{t\}} |t - y| p(y|x) dy =$$

АБСОЛЮТНОЕ ОТКЛОНЕНИЕ

$$\int_Y |t - y| p(y|x) dy \rightarrow \min_t$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \text{sign}(t - y) p(y|x) dy =$$

$$= \int_{\{t > y\}} p(y|x) dy - \int_{\{t < y\}} p(y|x) dy =$$

АБСОЛЮТНОЕ ОТКЛОНЕНИЕ

$$\int_Y |t - y| p(y|x) dy \rightarrow \min_t$$

$$\begin{aligned} \frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy &= \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy = \\ &= \int_{Y \setminus \{t\}} \text{sign}(t - y) p(y|x) dy = \\ &= \int_{\{t > y\}} p(y|x) dy - \int_{\{t < y\}} p(y|x) dy = \\ &= P(\{t > y\}|x) - P(\{t < y\}|x) = 0 \end{aligned}$$

АБСОЛЮТНОЕ ОТКЛОНЕНИЕ

$$\int_Y |t - y| p(y|x) dy \rightarrow \min_t$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \text{sign}(t - y) p(y|x) dy =$$

$$= \int_{\{t > y\}} p(y|x) dy - \int_{\{t < y\}} p(y|x) dy =$$

$$= P(\{t > y\}|x) - P(\{t < y\}|x) = 0$$

$$P(\{t = y\}|x) = 0$$

АБСОЛЮТНОЕ ОТКЛОНЕНИЕ

$$\int_Y |t - y| p(y|x) dy \rightarrow \min_t$$

$$\frac{\partial}{\partial t} \int_Y |t - y| p(y|x) dy = \frac{\partial}{\partial t} \int_{Y \setminus \{t\}} |t - y| p(y|x) dy =$$

$$= \int_{Y \setminus \{t\}} \text{sign}(t - y) p(y|x) dy =$$

$$= \int_{\{t > y\}} p(y|x) dy - \int_{\{t < y\}} p(y|x) dy =$$

$$= P(\{t > y\}|x) - P(\{t < y\}|x) = 0$$

$$P(\{t = y\}|x) = 0$$

$$\Rightarrow P(\{t \leq y\}|x) = P(\{t > y\}|x) = \frac{1}{2}$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$Y = \{0; 1\}$$

$$L(y, a(x)) = -y \ln a(x) - (1 - y) \ln (1 - a(x))$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$Y = \{0; 1\}$$

$$L(y, a(x)) = -y \ln a(x) - (1 - y) \ln (1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln (1 - t)) P(y|x) \rightarrow \min_t$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$Y = \{0; 1\}$$

$$L(y, a(x)) = -y \ln a(x) - (1 - y) \ln (1 - a(x))$$

$$\sum_{y \in Y} (-y \ln t - (1 - y) \ln (1 - t)) P(y|x) \rightarrow \min_t$$

$$P(1|x) = p$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$P(1|x) = p$$

$$-(1-p) \ln (1-t) - p \ln t \rightarrow \min_t$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$P(1|x) = p$$

$$-(1-p) \ln (1-t) - p \ln t \rightarrow \min_t$$

$$\frac{\partial}{\partial t}(-(1-p) \ln (1-t) - p \ln t) = \frac{1-p}{1-t} - \frac{p}{t} =$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$P(1|x) = p$$

$$-(1-p) \ln (1-t) - p \ln t \rightarrow \min_t$$

$$\frac{\partial}{\partial t}(-(1-p) \ln (1-t) - p \ln t) = \frac{1-p}{1-t} - \frac{p}{t} =$$

$$= \frac{(1-p)t - p(1-t)}{(1-t)t} = \frac{t-p}{(1-t)t} = 0$$

ОЦЕНКА ВЕРОЯТНОСТИ

$$P(1|x) = p$$

$$-(1-p) \ln (1-t) - p \ln t \rightarrow \min_t$$

$$\frac{\partial}{\partial t}(-(1-p) \ln (1-t) - p \ln t) = \frac{1-p}{1-t} - \frac{p}{t} =$$

$$= \frac{(1-p)t - p(1-t)}{(1-t)t} = \frac{t-p}{(1-t)t} = 0 \Rightarrow t = p$$

ПОЧЕМУ ЭТО ВСЁ РАБОТАЕТ

- › Средний риск:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x))$$

ПОЧЕМУ ЭТО ВСЁ РАБОТАЕТ

› Средний риск:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x))$$

Ошибка на обучающей выборке:

$$Q = \frac{1}{l} \sum_{i=1}^l L(y_i, a(x_i)) \approx \mathbb{E}_{x,y} L(y, a(x))$$

РЕЗЮМЕ

- › Принцип минимизации функционала среднего риска
- › Анализ функции потерь
- › Квадратичная функция потерь — для оценки матожидания
- › Абсолютные отклонения — для оценки 1/2 квантили
- › Log loss — для оценки вероятностей
- › Понимание неудачного выбора функции потерь