Вероятности, Байес и регрессия

Машинное обучение, 20!7

Спасибо К. В. Воронцову, МФТИ, Data Factory Яндекса и кофеину.

Малютин Е. А.

Содержание

Сегодня в программе:

- вероятностная постановка задачи машинного обучения
- чутка Байесовской статистики
- очень наивный Байесовский классификатор
- практика
- прекрасный мир логистической и не-логистической регрессий

Вероятности и Байес

Вероятностная постановка задачи ML:

Пусть X – множество объектов, Y – конечное множество имён классов, множество $X \times Y$ является вероятностным пространством с плотностью распределения $p(x, y) = P(y)p(x|y)$. Вероятности появления объектов каждого из классов $P_y = P(y)$ называются априорными вероятностями классов. Плотности распределения $p_y(x) = p(x|y)$ называются функциями правдоподобия классов

Вероятности и Байес

Итак, есть две задачи:

- Имеется простая выборка $X^l = (xi_, y_i)_{i=1}^l$ из неизвестного распределения $p(x, y) = P_y p_y(x)$. Требуется построить эмпирические оценки априорных вероятностей P_y и функций правдоподобия $p^y(x)$ для каждого из классов $y \in Y$.
- По известным плотностям распределения $p_y(x)$ и априорным вероятностям P_y всех классов $y \in Y$ построить алгоритм $a(x)$, минимизирующий вероятность ошибочной классификации.

Вероятности и Байес

Функционал среднего риска:

$$P(\Omega|y) = \int_{\Omega} p_y(x)dx, \quad \Omega \in X$$

– событие $x \in \Omega$ при условии, что x принадлежит к классу y

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} P_y P(A_s|y) - -\text{функционал среднего риска;}$$

где λ_{ys} – величина потери при отнесении объекта класса y к классу s

Вероятности и Байес

Оптимальное байесовское решающее правило

Если известны априорные вероятности P_y и функции правдоподобия $p_y(x)$, то минимум среднего риска $R(a)$ достигается алгоритмом:

$$a(x) = arg \min_{s \in Y} \sum_{y \in Y} \lambda_{ys} P_y p_y(x)$$

А в случае $\lambda_{ys} \triangleq \lambda_y$:

$$a(x) = arg \max_{y \in Y} \lambda_y P_y p_y(x)$$

Вероятности и Байес

Ещё немного магии

- **Разделяющая поверхность между s и t:** ГМТ таких точек, что при отнесении как к классу s, так и к классу t – достигается максимум в $a(x)$.
- **Апостериорная вероятность класса y для объекта x** – это условная вероятность $P(y|x)$. Вычисляется по Байесу:

$$P(y|x) = \frac{p(x,y)}{p(x)} = \frac{p_y(x) * P_y}{\sum_{s \in Y} P_s}$$

- **Величина ожидаемых потерь на объекте x :**

$$R(x) = \sum_{y \in Y} \lambda_y P(y|x)$$

Наивный Байес

Восстановление плотности распределений

Требуется оценить, какой могла бы быть плотность вероятностного распределения $p(x, y) = P_y p_y(x)$, сгенерировавшего выборку X_l. Обозначим подвыборку прецедентов класса y через $X_y^l = \{(x_i, y_i)_{i=1}^l \mid y_i = y\}$;
Оценка априорных вероятностей: $\hat{P}_l = \frac{l_y}{l}$, $l_y = |X_y^l|$;

Восстановление плотности вероятностей

Задано множество объектов $X_m = \{x_1, \dots, x_m\}$, выбранных случайно и независимо согласно неизвестному распределению $p(x)$. Требуется построить эмпирическую оценку плотности — функцию $\hat{p}(x)$, приближающую $p(x)$ на всём X.

Наивный Байес

Наивность:

Признаки $f_1(x), ..., f_n(x)$ являются независимыми случайными величинами.
Следовательно, функции правдоподобия классов представимы в виде:
$p_y(x) = p_{y1}(\xi_1)...p_{yn}(\xi_n), y \in Y$, где $p_{y_j}(\xi_j)$ — плотность распределения значений j-го признака для класса y.
В итоге:

$$a(x) = arg \max_{y \in Y}(ln \lambda_y \hat{P}_y + \sum_{j=1}^{n} ln \hat{p}_{yj}(\xi_j))$$

Наивный Байес

Подробнее:

1. $a(x) = argmax\ P(y|x) = argmax\ P(x|y) * P(y)$
2. $P(x|y) = P(x_1|y)P(x_2|y)...P(x_k|y)$
3. $P(x_k|y) = \frac{1}{l}\#(x_k, y)$ – доля объектов с данным значением признака k среди объектов класса y

Наивный Байес

Подробнее:

1. $a(x) = argmax\ P(y|x) = argmax\ P(x|y) * P(y)$
2. $P(x|y) = P(x_1|y)P(x_2|y)...P(x_k|y)$
3. $P(x_k|y) = \frac{1}{l}\#(x_k, y)$ – доля объектов с данным значением признака k среди объектов класса y
4. $ln(P(x|y) = ln(P(x_1|y)P(x_2|y)...P(x_k|y)) = \sum ln(P(x_i|y)$

Наивный Байес

Подробнее:

1. $a(x) = argmax\ P(y|x) = argmax\ P(x|y) * P(y)$
2. $P(x|y) = P(x_1|y)P(x_2|y)...P(x_k|y)$
3. $P(x_k|y) = \frac{1}{l}\#(x_k, y)$ – доля объектов с данным значением признака k среди объектов класса y
4. $ln(P(x|y) = ln(P(x_1|y)P(x_2|y)...P(x_k|y)) = \sum ln(P(x_i|y)$
5. $a(x) = argmax\ ln(\ P(x|y) * P(y)) =$ $argmax(lnP_y + \sum ln(P(x_1|y)P(x_2|y)...P(x_k|y)) = argmax(ln\ P_y + \sum ln(P(x_i|y)))$

Наивный Байес

В чем проблема?

- Классифицируем текст по вопросу - спам/не-спам
- Никогда не видели слов "веагра"
- $P(x|y) = 0$ – никуда не относим

Наивный Байес

В чем проблема?

- Классифицируем текст по вопросу - спам/не-спам
- Никогда не видели слов "веагра"
- $P(x|y) = 0$ – никуда не относим

Сглаживание вероятностей:

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

a, b – кросс-валидация

Все ещё наивный Байес

Что делать с вещественными признаками?

Использовать предположения о распределении этих признаков, например:

- Нормальное распределение: $p(x) = \frac{1}{\sqrt{(2\pi\sigma)}} e^{-\frac{-(x-\mu)^2}{2\sigma^2}}$
- Мультиномиальное распределение: $p(x) = \frac{n}{y_1 \dots y_k} p_1^{y_1} p_2^{y_2} \dots$
- Гауссово, Бернулли, Пуассона и пр.

Что выбрать?

- Разреженные дискретные – мультиномиальное
- Непрерывные признаки с маленьким разбросом – нормальное
- Непрерывные, с выбросами – взять более "размазанное"распределение

By Practice

$$P(spam|penis, viagra)$$

$$= \frac{P(penis|spam) * P(viagra|spam) * P(spam)}{P(penis) * P(viagra)}$$

$$= \frac{\frac{24}{30} * \frac{20}{30} * \frac{30}{74}}{\frac{25}{74} * \frac{51}{74}} = 0.928$$

Что делать?!

Практика

- Качаем спам https://archive.ics.uci.edu/ml/datasets/sms+spam+collection
- устанавливаем себе текстовые предобработчики на python: nltk, gensim, – токенезируем, чистим
- написать в простом виде Байеса – руками(!) (проявите фантазию)
- посчитать k-fold validation
- Сравнить результаты работы с результатами работы multinomialNB