Практическая работа №6: «Наивный Байесовский классификатор»

Оглавление

Цель работы

Получить практические навыки использования наивного Байесовского классификатора и логистической регрессии.

Задачи работы

1. Научиться решать задачу классификации с помощью библиотеки sklearn, используя наивный Байесовский классификатор.
2. Научиться решать задачу классификации с помощью библиотеки sklearn, используя логистическую регрессию.
3. Научиться сравнивать результаты работы моделей классификации, используя F1-меру.

Перечень обеспечивающих средств

1. ПК.
2. Учебно-методическая литература.
3. Задания для самостоятельного выполнения.

Общие теоретические сведения

Логистическая регрессия

Данные: $(x_{1i}, x_{2i}, \dots, x_{Ki}, y_i)$, где $i = 1, \dots, N$; $y_i \in \{0,1\}$.

Задача: Найти такие значения a_k, где *k* = 1, …, *K*, чтобы функция

$f(z) = \frac{1}{1+e^{-z}}$,

где $z(x_1, x_2, \dots, x_K) = a_1 x_1 + a_2 x_2 + \cdots + a_K x_K + b$,

аппроксимировала вероятность того, что $y_i = 1$.

Метрика производительности:

$$\mathrm{P}\{y|x\} = \prod_i \mathrm{P}\{y_i|x_i\} = \prod_i f(z_i)^{y_i}(1 - f(z_i))^{1-y_i} \to \max$$

$$\log \mathrm{P}\{y|x\} = \sum_i y_i \log f(z_i) + (1 - y_i)\log(1 - f(z_i)) \to \max$$

$$\mathrm{logloss} = \sum_i -y_i \log f(z_i) - (1 - y_i)\log(1 - f(z_i)) \to \min$$

Наивный байесовский классификатор

Данные: элементы $(x_{i1}, x_{i2}, \dots, x_{iK}, y_i)$,

где $y_i \in \{Y_1, \dots, Y_M\}$,

$i = 1, \dots, N$ (размер набора данных), K – количество параметров, описывающих входные данные, M – количество классов.

Задача: Найти такую функцию $f(x)$, чтобы $f(x_i) \approx y_i$ для всех *i*.

Вероятность того, что данный входной вектор *x* относится к данному классу Y_i:

$$\mathrm{P}(Y_i|x) = \frac{\mathrm{P}(x|Y_i)\mathrm{P}(Y_i)}{\mathrm{P}(x)}$$

$\mathrm{P}(Y_i)$ – априорные вероятности классов,

$\mathrm{P}(x|Y_i)$ – функции правдоподобия.

«Наивность» байесовского классификатора:

$$\mathrm{P}(x|Y_i) = \mathrm{P}(x_1|Y_i)\mathrm{P}(x_2|Y_i) \dots \mathrm{P}(x_K|Y_i) = \prod_{k=1}^{K} \mathrm{P}(x_k|Y_i)$$

Все параметры независимы, их порядок не имеет значения.

Сравнение двух алгоритмов

Интересующие нас условные вероятности принадлежности к классу для имеющихся значений параметров:

- Наивный байесовский классификатор:
  Вычисляются на основании вероятностей этих значений параметров (генеративный подход).

- Логистическая регрессия:
  Вычисляются напрямую с помощью минимизации ошибки (дискриминативный подход).

Задание

Пояснение

Для сохранения результатов данной работы вам понадобится два файла: doc/docx – для текста и ipynb – для кода. Назовите их одинаково: «*Фамилия* – задание 6».

Часть 1

- Обновите свой репозиторий, созданный в практической работе №1, из оригинального репозитория: https://github.com/mosalov/Notebook_For_AI_Main.

Часть 2

- Откройте свой репозиторий в Binder (https://mybinder.org/).
- Откройте файл «task6.ipynb».
- Изучите, при необходимости – выполните повторно, приведённый в файле код.
- По аналогии с изученным выполните задание, приведённое в последней ячейке.
- Сохраните код в ipynb-файле. Необходимые пояснения опишите в своём docx/doc-файле.

Требования к отчету

Оба файла (doc/docs и ipynb) загрузите в свой репозиторий, созданный в практическом задании №1 по пути: «Notebook_For_AI_Main/2020 Осенний семестр/Практическое задание 6/» и сделайте пул-реквест.

Литература

- https://neurohive.io/ru/osnovy-data-science/kak-primenjat-teoremu-bajesa-dlja-reshenija-realnyh-zadach/
- https://habr.com/ru/post/170545/
- https://science.wikia.org/ru/wiki/Байесовская_вероятность
- https://dyakonov.org/2018/07/30/байесовский-подход/
- http://bazhenov.me/blog/2012/06/11/naive-bayes.html
- https://scikit-learn.org/stable/modules/naive_bayes.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://scikit-

learn.org/stable/modules/generated/sklearn.metrics.f1_score.html