Лабораторная работа №2 – Наивный Баейсовский классификатор

Теоретическая часть

Наивный баейсовский классификатор – семейство простых вероятностыз классфикаторов, которые основываются на теореме Байеса.

Классификатор использует решающее правило MAP (maximum a posteriori), которое ставит в соответствие объекту наиболее вероятную для него метку и описывается формулой:

$$y = \underset{c \in C}{\operatorname{argmax}} P(C) \prod_{i=1}^{n} P(x_i|C)$$

$P(C)$ – априорная вероятнасть принадлежности объекта к классу C. В рамках лабораторной рассчитывается как доля объектов обучающей выборки, принадлежащик к классу.

$P(x_i|C)$ – правдоподобие принадлежности объекта к классу C, исходя из значения аттрибута x_i.

x_i – аттрибут объекта.

При работе с непрерывными аттрибутами используется предположение, что аттрибуты выбираются из независимых непрерывных нормальных распределений. Таким образом задача обучения классификатора заключается в нахождении параметров распределения – математического ожидания и дисперсии, для каждого из аттрибутов, что позволит вычислять $P(x_i|C)$.

$$P(x_i = v|C) = \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} e^{-\frac{(v-\overline{x_{ic}})^2}{2\sigma_{ic}^2}}$$

$\overline{x_{ic}}$ – среднее значение величины аттрибута, рассчитанное для объектов принадлжещих классу C.

σ_{ic}^2 – выборочная дисперсия значания аттрибута объектов из класс C.

$$\sigma_{ic}^2 = \frac{1}{n-1} \sum_{x \in C} (x_i - \overline{x_{ic}})^2$$

Задание

1. В результате выполнения лабораторной работы должно получится консольное приложение, которое строит байесовский классификатор и классифицирует объекты.
2. Реализовать парсинг csv файла. Обеспечить возможность настройки параметров формата файла – возможность выбирать разделители колонки, игнорировать первую строку (заголовок), первую колонку (там может быть порядковый номер объекта). Реализовывать на базе одной из следующих библиотек – iteratees, conduit, pipes. Фукнции, непосредственно разделяющие csv на колонки использовать нельзя (необходимо

реализовать самостоятельно). Парсить колонки csv в числа можно библиотечными функциями. Так же нельзя читать весь файл сразу в память.

3. В качестве одного из аргументов программы необходимо реализовать процент разделения данных файла. На основе этого процента файл делится случайным образом на тренировочную и тестовую выборки. Например, если задано 80/20 – на случайных 80% объектов классификатор обучается, на 20% - тестируется. Релизовать возможность задания количества попыток. Если задано более одной попытки – найти классификатор, для которого меньше ошибка классификации на тестовой выборке.
4. Сохранить лучший классификатор в текстовый файл в виде (для вывода использовать туже библиотеку, что и для парсинга):
   метка класса – индекс аттрибута1(мат ожидание;дисперсия) - индекс аттрибута2(мат ожидание;дисперсия)
   индексы объектов обучающей выборки

   Например:

   Iris-setosa – 1(0.3;0.15) – 2(0.5;0.25)
   Iris-verginica – 1(0.5;0.25) – 2(0.3, 0.11)
   1, 2, 3, 4, 10, 15, 20

5. В качестве параметра должна быть возможность задать файл, в который должны быть записаны результаты кластеризации. Если он не задан – результат выводится на консоль.
6. Параметры командной строки должны быть описаны в файле readme. Там же должны быть описаны любые особенности компиляции, даже если программа собирается при помощи cabal.
7. Обязательно использовать хотя бы одну из монад Reader или State (или соответствующие монадные трансформеры)
8. Результат выполнения работы (исходники) должен быть загружен на GitHub, а ссылка на него прислана на stasshiray@gmail.com с темой вида группа – ФИО – номер работы.

При написании лабораторно работы можно использовать любые библиотеки с Hackage, если они не реализуют алгоритм и не парсят csv.

К условию прилагаются 3 файла с исходными данными. В каждом из них последняя колонка – метка класса.