Введение в искусственный интеллект. Машинное обучение

Семинар 3. Вероятностный подход

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем

20 октября 2020г.

План семинара

1. Разбор предыдущего задания
2. Выдача домашнего задания
3. Наивный байесовский классификатор
4. Разбор пройденных методов в scikit-learn
5. Решение задач

Домашнее задание

- Первое домашнее задание доступна на гитхабе курса
- Дедлайн: 07 ноября 23:59:59 (после этого срока баллы будут умножаться на 0.5)
- Отправлять на почту курса mlcoursemm@gmail.com с темой [ML2020:theory01]

Наивный байесовский классификатор

Оптимальный байесовский классификатор

$$a(x) = \arg\max_y p(y|x) = \arg\max_y p(y)p(x|y)$$

Наивное предположение

Все признаки являются независимыми случайными величинами $p(x|y) = \prod_i p_i(x_i|y)$

Наивный байесовский классификатор

$$a(x) = \arg\max_y p(y|x) = \arg\max_y p(y) \prod_i p_i(x_i|y)$$

Гауссовский наивный байесовский классификатор

Наивное предположение

Все признаки являются независимыми случайными величинами $p(x|y) = \prod_i p_i(x_i|y)$

Будем предполагать, что $p_i(x_i|y) \sim N(\mu_y, \sigma_y)$, то есть

$$p_i(x_i|y) = \frac{1}{\sigma_y\sqrt{2\pi}} exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

Параметры μ_y и σ_y настраиваются по данным.

Область применения

Часто используется как бейзлайн модель

Используется в обработке текстов

Мультиномиальный наивный байесовский классификатор

Наивное предположение

Все признаки являются независимыми случайными величинами $p(x|y) = \prod_i p_i(x_i|y)$

Определение

Пусть $X = (X_1, ..., X_m)$ и $n_1 + ...n_m = n$, а $p_{y,1}, ..., p_{y,m} \geq 0$ и $\sum p_i = 1$.

$$P(X_1 = x_1, ..., X_m = x_m|y) = \frac{n!}{n_1!...n_m!} p_{y,1}^{x_1} ... p_{y,m}^{x_m}$$

Для настроки параметров применяют формулу $\hat{p_{y,i}} = \frac{N_{yi}+\alpha}{N_y+\alpha m}$, где *alpha* — неотрицательный коэффициент сглаживания

Область применения

Используется в обработке текстов

Наивный байесовский классификатор Бернулли

Наивное предположение

Все признаки являются независимыми случайными величинами $P(x|y) = \prod_i P_i(x_i|y)$

Определение

$$P(x_i|y) = p_{i,y}x_i + (1 - p_{i,y})(1 - x_i)$$

Область применения

Метод требует бинарного представления данных

Категориальный наивный байесовский классификатор

Наивное предположение

Все признаки являются независимыми случайными величинами $P(x|y) = \prod_i P_i(x_i|y)$

Определение

$P(x_i|y)$ — любое дискретное распределение (с конечным носителем)

Область применения

Подходит для категориальных данных

Обобщение

Идея

Не обязательно использовать одно семейство распределений для всех переменных

Задача

Задача

Запрограммировать один из предложенных наивных байесовских классификаторов. Реализовать методы fit, predict, predict_proba.