Байесовская классификация

Интеллектуальные системы и технологии, 2016

План лекции

1. Задача классификации
   - Постановка задачи
   - Примеры
2. Тервер и теорема Байеса
3. Наивный байесовский классификатор
4. Проверочная работа

1.1 Задача классификации

Дано:

X – множество объектов $\{X_1,\ldots,X_l\}$ с атрибутами $\{F_1,\ldots,F_n\}$,

Y – множество ответов $\{Y_1,\ldots,Y_m\}$.

Найти:

A = f(***X***) -> ***Y*** – решающий алгоритм, который сможет сопоставить любому объекту из *X* ответ *Y*.

Для задачи классификации:

- Бинарная классификация ***Y***={0,1}
- Многоклассовая классификация ***Y*** = $\{Y_1,\ldots,Y_m\}$

1.2 Пример задачи. Классификация спама

Дано:

X – множество писем/SMS-сообщений/комментариев $\{X_1,\ldots,X_n\}$

Y – множество ответов {*спам/не спам*}.

Найти:

A = f(***X***) -> ***Y*** – решающий алгоритм, который сможет классифицировать сообщение как спам или не спам.

Особенности:

- Как правило, используется достаточно большое количество признаков для решения задачи;
- Используются алгоритмы обработки текстов (для текстового спама)

1.2 Пример задачи. Медицинская диагностика

Дано:

X – множество пациентов $\{X_1,\ldots,X_n\}$

Y – множество ответов {*патология1/патология2/…*}.

Найти:

A = f(***X***) -> ***Y*** – решающий алгоритм, который сможет классифицировать состояние пациента.

Особенности:

- Имеются пропуски данных;
- Признаки измерены в разных шкалах (бинарных, количественных, ранговых)

1.3 Задача классификации и машинное обучение

Как правило:

- Весь массив исходных данных делится на два множества:
  - Обучающую выборку для построения алгоритма A
  - Тестовую выборку для проверки качества алгоритма A
- Существуют разные стратегии как делить массив исходных данных на обучающую и тестовую выборки
- Существует множество различных метрик качества полученного алгоритма классификации A

2. Теория вероятностей и теорема Байеса

2.1 Типы вероятностей

Априорная вероятность $P(A)$ – вероятность, что произойдет событие при отсутствии какой-либо другой информации.

Апостериорная вероятность $P(A|B)$ – вероятность, что произойдет событие A, если известно, что произошло B.

2.1 Типы вероятностей. Пример

A – выдать кредит (известна статистика: одобряют 8 из 10 заявок)

Априорная вероятность:

$P(A) = 0.8$

Апостериорная вероятность:

$P(A|\text{клиент безработный}) < 0.8$

$P(A|\text{клиент владеет недвижимостью}) > 0.8$

2.2 Теорема (формула) Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) - апостериорная вероятность гипотезы *A* при наступлении события *B*

P(A) - априорная вероятность гипотезы *A*

P(B|A) - вероятность наступления события *B* при истинности гипотезы *A (функция правдоподобия)*

P(B) - априорная вероятность наступления события *B*

2.2 Теорема (формула) Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ = P(A,B)

P(A|B) - апостериорная вероятность гипотезы *A* при наступлении события *B*

P(A) - априорная вероятность гипотезы *A*

P(B|A) - вероятность наступления события *B* при истинности гипотезы *A (правдоподобие)*

P(B) - априорная вероятность наступления события *B*

2.2 Теорема Байеса. Пример 1

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Событие A - в баке нет бензина, событие B - машина не заводится.

Вероятность P(B|A) того, что машина не заведется, если в баке нет бензина, равняется единице.

Тем самым, вероятность P(A) того, что в баке нет бензина, равна произведению вероятности P(B) того, что машина не заводится, на вероятность P(A|B) того, что причиной события B стало именно отсутствие бензина (событие A), а не, к примеру, разряженный аккумулятор.

2.2 Теорема Байеса. Пример 2

Состоится ли матч, если будет солнечно?

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

2.2 Теорема Байеса. Пример 2

Состоится ли матч, если будет солнечно?

A=Yes – матч состоится, B=Sunny – будет солнечно

Здесь мы имеем следующие значения:

P(Sunny | Yes) = 3 / 9 = 0,33

P(Sunny) = 5 / 14 = 0,36

P(Yes) = 9 / 14 = 0,64

Теперь рассчитаем *P(Yes | Sunny)*:

*P(Yes | Sunny) = 0,33 * 0,64 / 0,36 = 0,60*

Значит, при солнечной погоде более вероятно, что матч состоится.

3.1 Вероятностная постановка задачи классификации

X — объекты, ***Y*** — ответы, ***X*** × ***Y*** — с плотностью p(x, y);

Дано:

$X^\ell = (x_i , y_i)^\ell$ — простая выборка;

Найти:

классификатор a : X → Y с минимальной вероятностью ошибки.

Временное допущение: пусть известна совместная плотность
p(x, y) = p(x) P(y|x) = P(y)p(x|y).

$P(y) \equiv P_y$ — априорная вероятность класса y;

$p(x|y) \equiv p_y(x)$ — функция правдоподобия класса y;

P(y|x) — апостериорная вероятность класса y;

Принцип максимума апостериорной вероятности:

$$a(x) = \arg\max P(y|x) = \arg\max P_y p_y(x)$$

3.2 Байесовский классификатор

Байесовский классификатор — класс алгоритмов классификации, основанный на принципе максимума апостериорной вероятности. Для классифицируемого объекта вычисляются функции правдоподобия каждого из классов, по ним вычисляются апостериорные вероятности классов. Объект относится к тому классу, для которого апостериорная вероятность максимальна.

3.3 Наивный байесовский классификатор (naive bayes) НБК

Предположение:

Все признаки, описывающие объекты выборки $F_1,\ldots F_n$, строго независимы!

(На практике это не так, но это не мешает НБК давать хорошие результаты)

3.3 Наивный байесовский классификатор (naive bayes) НБК

Предположение:

Все признаки, описывающие объекты выборки $F_1,\dots F_n$, строго независимы!

(На практике это не так, но это не мешает НБК давать хорошие результаты)

Тогда:

$$P(Y|F_1,\dots,F_n) = \frac{P(Y)P(F_1,\dots,F_n|Y)}{P(F_1,\dots,F_n)}$$

3.3 Наивный байесовский классификатор (naive bayes) НБК

Предположение:

Все признаки, описывающие объекты выборки $F_1,\dots F_n$, строго независимы!

(На практике это не так, но это не мешает НБК давать хорошие результаты)

Тогда:

$$P(Y|F_1,\dots,F_n) = \frac{P(Y)P(F_1,\dots,F_n|Y)}{P(F_1,\dots,F_n)}$$

Или:

$$P(Y,F_1,\dots,F_n) = P(Y)P(F_1,\dots,F_n|Y)$$

3.3 Наивный байесовский классификатор (naive bayes) НБК

Распишем с учетом формулы условной вероятности:

$$P(Y, F_1, \dots, F_n)$$

$$= P(Y)P(F_1, \dots, F_n|Y)$$

$$= P(Y)P(F_1|Y)P(F_2, \dots, F_n|Y, F_1)$$

$$= P(Y)P(F_1|Y)P(F_2|Y, F_1)P(F_3, \dots, F_n|Y, F_1, F_2)$$

3.3 Наивный байесовский классификатор (naive bayes) НБК

И воспользуемся «наивностью» классификатора о независимости признаков:

$$P(F_i|C, F_j) = P(F_i|C)$$

Тогда:

$$P(Y, F_1, \dots, F_n) = P(Y) \prod_{i=1}^{n} p(F_i|Y)$$

3.3 Наивный байесовский классификатор (naive bayes) НБК

Параметры модели:

$P(Y)$, $P(F_i,Y)$

В момент обучения:

Аппроксимируются относительными частотами из набора данных обучения. Это оценки максимального правдоподобия вероятностей.

3.4 Вспоминаем пример с погодой

Состоится ли матч, если будет солнечно?

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Как вычислить правдоподобие? Для Sunny/No, Rainy/Yes

3.5 Плюсы и минусы НБК

Плюсы:

- Классификация, в том числе многоклассовая, выполняется легко и быстро;
- превосходит другие алгоритмы в случае независимости признаков. Но даже, если это не выполняется, может давать очень хорошие результаты;
- Потребляет мало памяти;

Минусы:

- Если в тестовом наборе данных присутствует некоторое значение признака, которое не встречалось в обучающем наборе данных, тогда модель присвоит нулевую вероятность этому значению и не сможет сделать прогноз. Это явление известно под названием «нулевая частота». Данную проблему можно решить с помощью сглаживания, например, по Лапласу.
- Допущение о независимости признаков. В реальности наборы полностью независимых признаков встречаются крайне редко.

3.6 Сферы применения

- **Классификация в режиме реального времени.** НБК очень быстро обучается, поэтому его можно использовать для обработки данных в режиме реального времени.
- **Многоклассовая классификация.** Это позволяет прогнозировать вероятности для множества значений целевой переменной.
- **Классификация текстов, фильтрация спама, анализ тональности текста.** При решении задач, связанных с классификацией текстов, НБК превосходит многие другие алгоритмы. Благодаря этому, данный алгоритм находит широкое применение в области фильтрации спама (идентификация спама в электронных письмах) и анализа тональности текста (анализ социальных медиа, идентификация позитивных и негативных мнений клиентов).
- **Рекомендательные системы.** Наивный байесовский классификатор в сочетании с коллаборативной фильтрацией позволяет реализовать рекомендательную систему.

Проверочная работа

Наличие слова	Спам/не спам
продажа	Yes
котики	No
купить	Yes
котики	Yes
продажа	No
котики	No
скидка	Yes
скидка	No
купить	No
котики	Yes
продажа	Yes

1) Посчитайте P(Спам), P(спам|котики), P(Не спам|скидка)
2) Поясните, в чем заключается наивность байесовского классификатора.