СПЕЦКУРС
Логический анализ данных в распознавании
(Logical data analysis in recognition)

лектор д.ф.-м.н. Елена Всеволодовна Дюкова

Спецкурс посвящён вопросам применения аппарата дискретной математики в задачах интеллектуального анализа данных. Излагаются общие принципы, лежащие в основе логического подхода к задачам машинного обучения. Описываются методы конструирования процедур классификации по прецедентам с использованием понятий теории булевых функций и теории покрытий булевых матриц. Рассматриваются основные модели логических процедур классификации, вопросы сложности их реализации и качества решения прикладных задач.

Спецкурс для бакалавров 2-4 курсов ВМК МГУ им. М.В. Ломоносова.

По спецкурсу издано учебное пособие:
http://www.ccas.ru/frc/papers/djukova03mp.pdf

Лекция 6
Использование аппарата логических функций для конструирования дискретных процедур распознавания в случае целочисленной информации

- Пусть $\boldsymbol{E_k^n}$, $\boldsymbol{k \geq 2}$, - множество наборов вида $\boldsymbol{(\alpha_1, \dots, \alpha_n)}$, где $\boldsymbol{\alpha_i \in \{0, 1, \dots, k-1\}}$. Пусть переменная $\boldsymbol{x}$ принимает значения из множества $\boldsymbol{\{0, 1, \dots, k-1\}}$, $\boldsymbol{\sigma \in \{0, 1, \dots, k-1\}}$. Положим

$$\boldsymbol{x^{\sigma}} = \begin{cases} \boldsymbol{1}, \text{если } \boldsymbol{x = \sigma}, \\ \boldsymbol{0}, \text{если } \boldsymbol{x \neq \sigma}. \end{cases}$$

- *Элементарной конъюнкцией* (ЭК) над переменными $\boldsymbol{x_1, \dots, x_n}$ назовём функцию вида $\boldsymbol{x_{j_1}^{\sigma_1}}$ & … & $\boldsymbol{x_{j_r}^{\sigma_r}}$, где $\boldsymbol{\sigma_i \in \{0, 1, \dots, k-1\}}$, $\boldsymbol{x_{j_i} \in \{x_1, \dots, x_n\}}$ при $\boldsymbol{i = 1, 2, \dots, r}$ и $\boldsymbol{x_{j_q} \neq x_{j_t}}$ при $\boldsymbol{t, q \in \{1, 2, \dots, r\}}$, $\boldsymbol{t \neq q}$. Также, как и в случае $\boldsymbol{k = 2}$, для краткости знак & опускается.

• Нетрудно видеть, что ЭК $\boldsymbol{B = x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}}$ обращается в 1 на тех и только тех наборах $(\boldsymbol{\alpha_1, \dots, \alpha_n})$, для которых $\boldsymbol{\alpha_{j_i} = \sigma_i}$, $\boldsymbol{i = 1, 2, \dots, r}$, т.е. ЭК принимает значение 1 на наборе из $\boldsymbol{E_k^n}$, если каждый её сомножитель принимает значение 1 на этом наборе. Интервал истинности ЭК $\boldsymbol{B}$ будем обозначать через $\boldsymbol{N_B}$.

• Пусть $\boldsymbol{f(x_1, \dots, x_n)}$ – функция, частично определённая на наборах из $\boldsymbol{E_k^n}$ и принимающая значения из $\boldsymbol{\{0, 1\}}$. $\boldsymbol{N_f}$ и $\boldsymbol{N_{\bar{f}}}$ - соответственно множество единиц и множество нулей функции $\boldsymbol{f}$.

• Определения почти допустимой, допустимой, неприводимой и максимальной конъюнкций, данные в лекции 4 для случая всюду определённой булевой функции, полностью переносятся на случай двузначной частичной функции $\boldsymbol{f}$, определённой на наборах из $\boldsymbol{E_k^n}$.

- Рассмотрим ситуацию, когда объекты из исследуемого множества $\boldsymbol{M}$ описаны признаками, каждый из которых принимает значения из множества $\boldsymbol{\{0, 1, \ldots, k - 1\}}$.

- Эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \ldots, \boldsymbol{\sigma_r})$, порождённому набором признаков $\boldsymbol{H} = \{\boldsymbol{x_{j_1}}, \ldots, \boldsymbol{x_{j_r}}\}$, поставим в соответствие ЭК $\boldsymbol{B} = \boldsymbol{x_{j_1}^{\sigma_1}} \ldots \boldsymbol{x_{j_r}^{\sigma_r}}$.

- Напомним (см. лекцию 3), что близость объекта $\boldsymbol{S} = (\boldsymbol{a_1}, \ldots, \boldsymbol{a_n})$ из $\boldsymbol{M}$ и элементарного классификатора $(\boldsymbol{\sigma}, \boldsymbol{H})$ оценивается величиной $\boldsymbol{B}(\boldsymbol{\sigma}, \boldsymbol{S}, \boldsymbol{H})$, равной $\boldsymbol{1}$, если $\boldsymbol{a_{j_t}} = \boldsymbol{\sigma_t}$ при $\boldsymbol{t} = \boldsymbol{1}, \boldsymbol{2}, \ldots, \boldsymbol{r}$, и раной $\boldsymbol{0}$ в противном случае.

- Очевидно, $\boldsymbol{B}(\boldsymbol{\sigma}, \boldsymbol{S}, \boldsymbol{H}) = \boldsymbol{1}$ тогда и только тогда, когда $(\boldsymbol{a_1}, \ldots, \boldsymbol{a_n}) \in \boldsymbol{N_B}$.

- Рассмотрим основные модели логических классификаторов, а именно, алгоритмы голосования по представительным наборам, по антипредставительным наборам, по покрытиям классов и тестам. Покажем, что в каждой из этих моделей построение множества эл.кл. для класса $\boldsymbol{K}$ сводится к построению допустимых или максимальных конъюнкций двухзначной логической функции, частично или всюду определённой на наборах из $\boldsymbol{E_k^n}$. Такая функция на обучающих объектах из $\boldsymbol{K}$ и $\overline{\boldsymbol{K}} = \{\boldsymbol{K_1}, \dots , \boldsymbol{K_l}\} \setminus \boldsymbol{K}$ принимает разные значения и называется *характеристической функцией* класса $\boldsymbol{K}$.

- Процедура голосования по построенной ЭК $\boldsymbol{B}$ заключается в проверке принадлежности набора $(\boldsymbol{a_1}, \dots, \boldsymbol{a_n})$ интервалу $\boldsymbol{N_B}$.

1. Алгоритм голосования по представительным наборам

- В данном случае характеристическая функция класса $\boldsymbol{K}$ – частичная логическая функция $\boldsymbol{f_K(x_1, \dots, x_n)}$, принимающая значение 1 на описаниях обучающих объектов из класса $\boldsymbol{K}$, значение 0 на описаниях обучающих объектов из $\boldsymbol{\bar{K}}$ и не определённая на остальных наборах из $\boldsymbol{E_k^n}$.

- Представительному набору класса $\boldsymbol{K}$ соответствует допустимая конъюнкция функции $\boldsymbol{f_K}$, тупиковому представительному набору соответствует максимальная конъюнкция функции $\boldsymbol{f_K}$.

- Допустимая (максимальная) конъюнкция $\boldsymbol{B}$ функции $\boldsymbol{f_K}$ голосует за принадлежность объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$, если $\boldsymbol{(a_1, \dots, a_n) \in N_B}$.

2. Алгоритм голосования по антипредставительным наборам

• Характеристическая функция класса $\boldsymbol{K}$ – частичная логическая функция $\boldsymbol{f_{\bar{K}}(x_1, \ldots, x_n)}$, принимающая значение 0 на описаниях обучающих объектов из класса $\boldsymbol{K}$, значение 1 на описаниях обучающих объектов из $\boldsymbol{\bar{K}}$ и не определённая на остальных наборах из $\boldsymbol{E_k^n}$.

• Антипредставительному набору класса $\boldsymbol{K}$ соответствует допустимая конъюнкция функции $\boldsymbol{f_{\bar{K}}}$, тупиковому антипредставительному набору соответствует максимальная конъюнкция функции $\boldsymbol{f_{\bar{K}}}$.

• Допустимая (максимальная) конъюнкция $\boldsymbol{B}$ функции $\boldsymbol{f_{\bar{K}}}$ голосует за принадлежность объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$, если $\boldsymbol{(a_1, \ldots, a_n) \notin N_B}$.

3. Алгоритмы голосования по покрытиям класса

• Характеристическая функция класса $\boldsymbol{K}$ – всюду определённая логическая функция $\boldsymbol{F_{\bar{K}}(x_1, \dots, x_n)}$, принимающая значение 0 на описаниях обучающих объектов из класса $\boldsymbol{K}$ и значение 1 на остальных наборах из $\boldsymbol{E_k^n}$.

• Покрытию класса $\boldsymbol{K}$ соответствует допустимая конъюнкция функции $\boldsymbol{F_{\bar{K}}}$, тупиковому покрытию – максимальная конъюнкция для $\boldsymbol{F_{\bar{K}}}$.

• Допустимая (максимальная) конъюнкция $\boldsymbol{B}$ функции $\boldsymbol{F_{\bar{K}}}$ голосует за принадлежность объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$, если $\boldsymbol{(a_1, \dots, a_n) \notin N_B}$.

4. Алгоритм голосования по тестам

- Характеристическая функция класса $\boldsymbol{K}$ определяется так же, как и для алгоритмов голосования по представительным наборам.

- Множество всех почти допустимых конъюнкций функции $\boldsymbol{f_K}$, порождаемое набором признаков $\boldsymbol{H}$, обозначим через $\boldsymbol{Q(H,K)}$.

- Очевидно, набор признаков $\boldsymbol{H}$ является тестом тогда и только тогда, когда для каждого $\boldsymbol{t} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{l}\}$, каждая конъюнкция из $\boldsymbol{Q(H, K_t)}$ является допустимой для $\boldsymbol{f_{K_t}}$.

- Очевидно также, что тест $\boldsymbol{H} = \{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}}\}$ является тупиковым тогда и только тогда, когда для каждого $\boldsymbol{i}$, $\boldsymbol{i} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{r}\}$, в $\{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{l}\}$ можно указать $\boldsymbol{t_i}$ такое, что $\boldsymbol{Q\left(H \setminus \{x_i\}, K_{t_i}\right)}$ содержит конъюнкцию, не являющуюся допустимой для соответствующей характеристической функции.

- Построение требуемого множества конъюнкций (ДНФ, реализующей характеристическую функцию) может быть осуществлено на основе преобразования нормальных форм.

- Случай, когда $\boldsymbol{k = 2}$ рассмотрен в лекции 5. Описанные в лекции 5 алгоритмы построения сокращённой ДНФ булевой функции, заданной множеством нулей, могут быть обобщены на рассматриваемый нами общий случай $\boldsymbol{(k \geq 2)}$.

- Пусть множество нулей характеристической функции $\boldsymbol{f}$ состоит из наборов $(\boldsymbol{\beta_{11}, \dots, \beta_{1n}})$, $(\boldsymbol{\beta_{21}, \dots, \beta_{2n}})$, …, $(\boldsymbol{\beta_{u1}, \dots, \beta_{un}})$.

- Если $\boldsymbol{k = 2}$ и $\boldsymbol{f}$ всюду определена, то нужно построить КНФ $\boldsymbol{K = D_1}$& … & $\boldsymbol{D_u}$, где $\boldsymbol{D_i = x_1^{\overline{\beta_{i_1}}} \vee \dots \vee x_n^{\overline{\beta_{i_n}}}}$, $\boldsymbol{i = 1, 2, \dots, u}$, реализующую функцию $\boldsymbol{f}$. Затем преобразовать эту КНФ в (сокращённую) ДНФ функции $\boldsymbol{f}$. Для этого можно воспользоваться одним из описанных в лекции 5 способов преобразования нормальных форм булевой функции.

- Построение требуемого множества конъюнкций (ДНФ, реализующей характеристическую функцию) может быть осуществлено на основе преобразования нормальных форм.

- Случай, когда $\boldsymbol{k = 2}$ рассмотрен в лекции 5. Описанные в лекции 5 алгоритмы построения сокращённой ДНФ булевой функции, заданной множеством нулей, могут быть обобщены на рассматриваемый нами общий случай $\boldsymbol{(k \geq 2)}$.

- Пусть множество нулей характеристической функции $\boldsymbol{f}$ состоит из наборов $(\boldsymbol{\beta_{11}, \dots, \beta_{1n}})$, $(\boldsymbol{\beta_{21}, \dots, \beta_{2n}})$, …, $(\boldsymbol{\beta_{u1}, \dots, \beta_{un}})$.

- Если $\boldsymbol{k = 2}$ и $\boldsymbol{f}$ всюду определена, то нужно построить КНФ $\boldsymbol{K = D_1}$& … & $\boldsymbol{D_u}$, где $\boldsymbol{D_i = x_1^{\overline{\beta_{i_1}}} \vee \dots \vee x_n^{\overline{\beta_{i_n}}}}$, $\boldsymbol{i = 1, 2, \dots, u}$, реализующую функцию $\boldsymbol{f}$. Затем преобразовать эту КНФ в (сокращённую) ДНФ функции $\boldsymbol{f}$. Для этого можно воспользоваться одним из описанных в лекции 5 способов преобразования нормальных форм булевой функции.

• Если же $\boldsymbol{k = 2}$ и $\boldsymbol{f}$ не всюду определена, то КНФ $\boldsymbol{K}$ реализует всюду определённую булеву функцию $\boldsymbol{F}$ с тем же множеством нулей. В этом случае необходимо удалить из (сокращённой) ДНФ функции $\boldsymbol{F}$ те конъюнкции, которые не являются допустимыми для $\boldsymbol{f}$.

• В случае $\boldsymbol{k > 2}$ вместо $\boldsymbol{D_i = x_1^{\overline{\beta_{i_1}}} \vee \ldots \vee x_n^{\overline{\beta_{i_n}}}}$ необходимо взять дизъюнкцию вида $\boldsymbol{\overline{x_1^{\beta_{i_1}}} \vee \ldots \vee \overline{x_n^{\beta_{i_n}}}}$ и воспользоваться равенством $\boldsymbol{\overline{x^{\beta}} = \vee_{\alpha \neq \beta} x^{\alpha}}$. Тогда преобразуемая КНФ примет вид

$$\boldsymbol{D_1^* \& D_2^* \& \ldots \& D_n^*},$$

где $\boldsymbol{D_i^* = \vee_{\alpha \neq \beta_{i_1}} x_1^{\alpha} \vee_{\alpha \neq \beta_{i_2}} x_2^{\alpha} \vee \ldots \vee_{\alpha \neq \beta_{i_n}} x_n^{\alpha}}$.

• После перемножения логических скобок в получившейся ДНФ следует сделать упрощения, пользуясь тем, что $\boldsymbol{x^{\alpha} \cdot x^{\beta} = 0}$ при $\boldsymbol{\alpha \neq \beta}$, $\boldsymbol{x \cdot x = x}$, $\boldsymbol{x \vee x = x}$, $\boldsymbol{x \vee xx' = x}$. Остальные рассуждения полностью совпадают с рассуждениями, приведёнными в лекции 5.

- Построение требуемого множества тестов может быть осуществлено на основе преобразования КНФ, не содержащей отрицаний переменных (реализующей монотонную булеву функцию), в ДНФ.

- Действительно, рассмотрим множество $\boldsymbol{P}$ всех неупорядоченных пар наборов $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ таких, что $\boldsymbol{\alpha} \in \boldsymbol{N}_{\boldsymbol{f}_{\boldsymbol{k}_u}}$, $\boldsymbol{\beta} \in \boldsymbol{N}_{\boldsymbol{f}_{\boldsymbol{k}_v}}$, $\boldsymbol{u} \neq \boldsymbol{v}$, $\boldsymbol{u}, \boldsymbol{v} \in \{\boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{l}\}$. Для каждой пары $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ из $\boldsymbol{P}$ построим дизъюнкцию $\boldsymbol{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{x}_{\boldsymbol{p}_1} \vee \dots \vee \boldsymbol{x}_{\boldsymbol{p}_q}$, где $\boldsymbol{p}_1, \dots, \boldsymbol{p}_q$ - номера разрядов, в которых различаются наборы $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$.

- Нетрудно убедится в том, что для построения множества тестов нужно преобразовать КНФ $\&_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in \boldsymbol{P}} \boldsymbol{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ в ДНФ, состоящую из допустимых конъюнкций функции, реализуемой этой КНФ. Для построения множества тупиковых тестов нужно преобразовать КНФ $\&_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in \boldsymbol{P}} \boldsymbol{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ в ДНФ, состоящую из максимальных конъюнкций функции, реализуемой этой КНФ.

УПРАЖНЕНИЯ

1. Пусть двузначная логическая функция $\boldsymbol{F(x_1, \dots, x_n)}$ определена на $\boldsymbol{E_k^n}$ и принимает значение 0 на описаниях обучающих объектов из класса $\boldsymbol{K}$, а значение 1 на остальных наборах из $\boldsymbol{E_k^n}$. Пусть конъюнкция $\boldsymbol{x_{j_1}^{\sigma_1}} \& \dots \& \boldsymbol{x_{j_r}^{\sigma_r}}$ является неприводимой для $\boldsymbol{F}$. Является ли эл.кл. $\boldsymbol{(\sigma, H)}$, где $\boldsymbol{\sigma = (\sigma_1, \dots, \sigma_r)}$, $\boldsymbol{H = (x_{j_1}, \dots, x_{j_r})}$, тупиковым покрытием класса $\boldsymbol{K}$?