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 7
Поиск элементарных классификаторов на основе построения покрытий булевой и целочисленной матриц. Связь задач построения покрытий булевой матрицы и преобразования нормальных форм булевой функции

• При реализации логических классификаторов для поиска элементарных классификаторов чаще используются построения, в основе которых лежит поиск покрытий булевой или целочисленной матрицы.

• Пусть L – булева матрица. Набор столбцов H матрицы L называется *покрытием*, если каждая строка матрицы L в пересечении хотя бы с одним из столбцов, входящих в H, дает **1.** Покрытие называется *неприводимым,* если никакое его собственное подмножество не является покрытием.

• При построении неприводимых покрытий обычно используется следующий ***критерий***.

• **Утверждение 1.** *Набор* $\boldsymbol{H}$ *из* $\boldsymbol{r}$ *различных столбцов матрицы* $\boldsymbol{L}$ *является неприводимым покрытием, если выполнены два условия: 1) подматрица* $\boldsymbol{L^H}$ *матрицы* $\boldsymbol{L}$ *образованная столбцами набора* $\boldsymbol{H}$ *, не содержит строк вида* **(0, 0, …, 0)** ; *2)* $\boldsymbol{L^H}$ *содержит каждую из строк вида каждую из строк вида* **(1, 0, …, 0), (0, 1, …, 0), … , (0, 0, …, 1),** т.е. $\boldsymbol{L^H}$ содержит единичную подматрицу порядка $\boldsymbol{r}$.

• Первое условие называется условием *покрываемости*, а второе условием *совместимости*.

• Как возникают задачи построения покрытий и неприводимых покрытий при построении (тупиковых) тестов и (тупиковых) представительных наборов?

• Материал обучения обычно представляется в виде таблицы $\boldsymbol{T}$ с $\boldsymbol{n}$ столбцами (таблицы обучения), в которой столбец с номером $\boldsymbol{j}$ соответствует признаку $\boldsymbol{x_j}$, а каждая строка есть набор значений признаков, описывающих один из обучающих объектов. Для нахождения искомого множества элементарных классификаторов строится специальная булева матрица (матрица сравнения таблицы $\boldsymbol{T}$). Обозначим её $\boldsymbol{L_T}$. Каждая строка этой матрицы образуется в результате сравнения пары строк таблицы $\boldsymbol{T}$, описывающих объекты из разных классов. При этом в столбце матрицы $\boldsymbol{L_T}$ с номером $\boldsymbol{j}$ ставится **1**, если сравниваемые строки различаются в разряде с номером $\boldsymbol{j}$, и **0** в противном случае.

• Обозначим через $\boldsymbol{L_T^{(i)}}$, $\boldsymbol{i = 1, \dots, m}$, подматрицу матрицы $\boldsymbol{L_T}$, которая образована сравнением обучающего объекта $\boldsymbol{S_i}$ со всеми обучающими объектами, не принадлежащими тому же классу, что и объект $\boldsymbol{S_i}$.

- Очевидными являются следующие два утверждения.

- **Утверждение 2.** *Набор признаков* $\{x_{j_1}, \dots, x_{j_r}\}$ *является (тупиковым) тестом тогда и только тогда, когда набор столбцов матрицы* L_T *с номерами* $j_1, \dots, j_r$ *является (неприводимым) покрытием.*

- **Утверждение 3.** *Эл.кл.* $(\sigma, \{x_{j_1}, \dots, x_{j_r}\})$, *является (тупиковым) представительным набором для класса* K *тогда и только тогда, когда* $B(\sigma, S_i, \{x_{j_1}, \dots, x_{j_r}\}) = 1$ *для некоторого обучающего объекта* S_i *из* K *и набор столбцов матрицы* $L_T^{(i)}$ *с номерами* $j_1, \dots, j_r$ *является (неприводимым) покрытием.*

- Задача построения неприводимых покрытий булевой матрицы L размера $u \times n$ может быть сформулирована как задача преобразования КНФ монотонной булевой функции в сокращенную ДНФ.

• Действительно, строке с номером $\boldsymbol{i}$ поставим в соответствие дизъюнкцию $\boldsymbol{D_i = x_{p_1} \vee \cdots \vee x_{p_q}}$, где $\boldsymbol{p_1, \ldots, p_q}$ – номера тех столбцов, которые в пересечении с этой строкой дают **1**. Пусть $\boldsymbol{f}$-монотонная булева функция реализуемая КНФ $\boldsymbol{D_1}$& ... &$\boldsymbol{D_u}$.

• Пользуясь утверждением 2 из лекции 5 нетрудно доказать

• **Утверждение 4.** *ЭК* $\boldsymbol{x_{j_1} \ldots x_{j_r}}$ *является допустимой для* $\boldsymbol{f}$ *тогда и только тогда, когда набор столбцов* $\boldsymbol{H}$ *матрицы* $\boldsymbol{L}$ *с номерами* $\boldsymbol{j_1, \ldots, j_r}$ *является покрытием.*

• Пользуясь утверждением 4 из лекции 5 нетрудно доказать

• **Утверждение 5.** *ЭК* $\boldsymbol{x_{j_1} \ldots x_{j_r}}$ *является неприводимой для* $\boldsymbol{f}$ *тогда и только тогда, когда набор столбцов матрицы* $\boldsymbol{L}$ *с номерами* $\boldsymbol{j_1, \ldots, j_r}$ *содержит единичную подматрицу порядка* **r**.

- Из утверждений 4 и 5 следует
- **Утверждение 6.** ЭК $x_{j_1} \dots x_{j_r}$ *является максимальной для* f *тогда и только тогда, когда набор столбцов матрицы* L *с номерами* $j_1, \dots, j_r$ *является неприводимым покрытием.*

- Из последних трех утверждений следует, что алгоритмы построения неприводимых покрытий булевой матрицы можно применять для построения максимальных конъюнкций монотонной булевой функции, заданной КНФ, и наоборот. В теории алгоритмической сложности рассматриваемая задача преобразования нормальных форм монотонной булевой функции называется *монотонной дуализацией* и относится к числу труднорешаемых дискретных задач. О существующих подходах к её решению будет рассказано в следующей лекции. В силу вычислительной сложности монотонной дуализации на практике отказываются от поиска всех тупиковых тестов и тупиковых представительных наборов. В основном используются модели с ограничением на длину эл.кл. и стохастические алгоритмы, в которых строится достаточно представительная выборка из искомого множества элементарных классификаторов.

- Можно избежать построения вспомогательной булевой матрицы $\boldsymbol{L_T}$, если ввести понятие покрытия более общего вида.
- Действительно, пусть $\boldsymbol{L}$ – целочисленная матрица размера $\boldsymbol{u \times n}$ с элементами из $\{\mathbf{0}, \mathbf{1}, \dots, \boldsymbol{k-1}\}, \boldsymbol{k \geq 2}$; $\boldsymbol{E_k^r, r \leq n}$, - множество наборов вида $\boldsymbol{\sigma = (\sigma_1, \dots, \sigma_r)}$, где $\boldsymbol{\sigma_i \in \{0, 1, \dots, k-1\}}$.
- Пусть далее $\boldsymbol{\sigma \in E_k^r, \sigma = (\sigma_1, \dots, \sigma_r)}$.
- Набор $\boldsymbol{H}$ из $\boldsymbol{r}$ различных столбцов матрицы $\boldsymbol{L}$ называется $\boldsymbol{\sigma}$–*покрытием*, если в подматрице $\boldsymbol{L^H}$ матрицы $\boldsymbol{L}$, образованной столбцами набора $\boldsymbol{H}$, нет строки $(\boldsymbol{\sigma_1, \dots, \sigma_r})$. Набор $\boldsymbol{H}$ из $\boldsymbol{r}$ различных столбцов матрицы $\boldsymbol{L}$, являющийся $\boldsymbol{\sigma}$–покрытием, называется *тупиковым* $\boldsymbol{\sigma}$–*покрытием*, если $\boldsymbol{L^H}$ содержит подматрицу, имеющую с точностью до перестановки строк вид

$$\begin{pmatrix} \boldsymbol{\beta_1} & \boldsymbol{\sigma_2} & \boldsymbol{\sigma_3} & \dots & \boldsymbol{\sigma_{r-1}} & \boldsymbol{\sigma_r} \\ \boldsymbol{\sigma_1} & \boldsymbol{\beta_2} & \boldsymbol{\sigma_3} & \dots & \boldsymbol{\sigma_{r-1}} & \boldsymbol{\sigma_r} \\ & & & \dots & & \\ \boldsymbol{\sigma_1} & \boldsymbol{\sigma_2} & \boldsymbol{\sigma_3} & \dots & \boldsymbol{\sigma_{r-1}} & \boldsymbol{\beta_r} \end{pmatrix},$$

где $\boldsymbol{\beta_p \neq \sigma_p}$ при $\boldsymbol{p = 1, 2, \dots, r}$. Такая подматрица называется $\boldsymbol{\sigma}$-*подматрицей*.

• Если $\boldsymbol{k = 2}$ и $\boldsymbol{\sigma = (0, \ldots, 0)}$, то понятие (тупикового) $\boldsymbol{\sigma}$*–покрытия* совпадает с понятием (неприводимого) покрытия. Аналогом единичной подматрицы в этом случае является $\boldsymbol{\sigma}$ -подматрица.

• Таблицу обучения $\boldsymbol{T}$ можно рассматривать как пару матриц $\boldsymbol{L_1}$ и $\boldsymbol{L_2}$, где $\boldsymbol{L_1}$ - матрица, состоящая из описаний обучающих объектов из класса $\boldsymbol{K}$, $\boldsymbol{L_2}$ - матрица, состоящая из описаний остальных обучающих объектов. Тогда, очевидно, эл.кл. $\boldsymbol{(\sigma, \{x_{j_1}, \ldots, x_{j_r}\})}$, $\boldsymbol{\sigma = (\sigma_1, \ldots, \sigma_r)}$, является (тупиковым) представительным набором для класса $\boldsymbol{K}$, если набор столбцов матрицы $\boldsymbol{L_1}$ с номерами $\boldsymbol{j_1, \ldots, j_r}$ не является $\boldsymbol{(\sigma_1, \ldots, \sigma_r)}$-покрытием, а набор столбцов матрицы $\boldsymbol{L_2}$ с номерами $\boldsymbol{j_1, \ldots, j_r}$ является (тупиковым) $\boldsymbol{(\sigma_1, \ldots, \sigma_r)}$ -покрытием.

• Нетрудно видеть, что эл.кл. $\boldsymbol{(\sigma, \{x_{j_1}, \ldots, x_{j_r}\})}$, $\boldsymbol{\sigma = (\sigma_1, \ldots, \sigma_r)}$, является (тупиковым) покрытием класса $\boldsymbol{K}$, если набор столбцов матрицы $\boldsymbol{L_1}$ с номерами $\boldsymbol{j_1, \ldots, j_r}$ является (тупиковым) $\boldsymbol{(\sigma_1, \ldots, \sigma_r)}$ -покрытием.

• Нетрудно также видеть, что эл.кл. вида $\left(\boldsymbol{\sigma}, \{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}}\}\right)$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r})$, является (тупиковым) антипредставительным набором для класса $\boldsymbol{K}$, если набор столбцов матрицы $\boldsymbol{L_2}$ с номерами $\boldsymbol{j_1}, \dots, \boldsymbol{j_r}$ не является $(\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r})$ -покрытием, а набор столбцов матрицы $\boldsymbol{L_1}$ с номерами $\boldsymbol{j_1}, \dots, \boldsymbol{j_r}$ является (тупиковым) $(\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r})$ -покрытием.

• Через $\boldsymbol{R(\sigma)}$ обозначим множество наборов $(\boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_r})$ в $\boldsymbol{E_k^r}$ таких, что $\boldsymbol{\beta_j \neq \sigma_j}$ при $\boldsymbol{j = 1, 2, \dots, r}$. Набор столбцов $\boldsymbol{H}$ матрицы $\boldsymbol{L}$ назовем $\boldsymbol{R(\sigma)}$ -*покрытием*, если в подматрице $\boldsymbol{L^H}$ матрицы $\boldsymbol{L}$, образованной столбцами набора $\boldsymbol{H}$, нет ни одной строки из $\boldsymbol{R(\sigma)}$. Набор столбцов $\boldsymbol{H}$ матрицы $\boldsymbol{L}$, являющийся $\boldsymbol{R(\sigma)}$-покрытием, назовем *тупиковым* $\boldsymbol{R(\sigma)}$ -*покрытием*, если $\boldsymbol{L^H}$ содержит подматрицу, имеющую с точностью до перестановки строк вид

$$\begin{pmatrix} \boldsymbol{\sigma_1} & \boldsymbol{\beta_2} & \boldsymbol{\beta_3} & \dots & \boldsymbol{\beta_{r-1}} & \boldsymbol{\beta_r} \\ \boldsymbol{\beta_1} & \boldsymbol{\sigma_2} & \boldsymbol{\beta_3} & \dots & \boldsymbol{\beta_{r-1}} & \boldsymbol{\beta_r} \\ & & & \dots & & \\ \boldsymbol{\beta_1} & \boldsymbol{\beta_2} & \boldsymbol{\beta_3} & \dots & \boldsymbol{\beta_{r-1}} & \boldsymbol{\sigma_r} \end{pmatrix},$$

где $\boldsymbol{\beta_p \neq \sigma_p}$ при $\boldsymbol{p = 1, 2, \dots, r}$. Такая подматрица называется $\boldsymbol{R(\sigma)}$ -*подматрицей*.

- Если $\boldsymbol{k = 2}$ и $\boldsymbol{\sigma = (1, \ldots, 1)}$, то понятие (тупикового) $\boldsymbol{R(\sigma)}$ -покрытия совпадает с понятием (неприводимого) покрытия. Аналогом единичной подматрицы является $\boldsymbol{R(\sigma)}$ -подматрица.

- Пусть далее $\boldsymbol{L}$ – булева матрица размера $\boldsymbol{u \times n}$, $\boldsymbol{\sigma \in E_2^r}$.
- Через $\boldsymbol{C(L, \sigma)}$ и $\boldsymbol{B(L, \sigma)}$ обозначим соответственно совокупность всех $\boldsymbol{\sigma}$ -покрытий и всех тупиковых $\boldsymbol{\sigma}$ -покрытий матрицы $\boldsymbol{L}$. Положим

- $\boldsymbol{C(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \epsilon E_2^r} C(L, \sigma)}$, $\boldsymbol{B(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \epsilon E_2^r} B(L, \sigma)}$.

- Связь между задачами построения множеств $\boldsymbol{C(L)}$, $\boldsymbol{B(L)}$ матрицы $\boldsymbol{L}$ и задачей преобразования нормальных форм булевой функции устанавливается следующим образом.

- Пусть $\boldsymbol{(\sigma_{i1}, \ldots, \sigma_{in})}$ – строка матрицы $\boldsymbol{L}$ с номером $\boldsymbol{i}$, $\boldsymbol{i \in \{1, 2, \ldots, u\}}$. Этой строке ставится в соответствие дизъюнкция $\boldsymbol{D_i^{(1)} = x_1^{\overline{\sigma_{i1}}} \vee \cdots \vee x_n^{\overline{\sigma_{in}}}}$

• Пусть $\boldsymbol{F_1}$ – булева функция, реализуемая КНФ $\boldsymbol{D_1^{(1)} \& \ldots \& D_u^{(1)}}$. Используя утверждения 2 и 4 из лекции 5 нетрудно доказать приведенные ниже утверждения 7 , 8, из которых сразу следует утверждение 9.

• **Утверждение 7.** ЭК $\boldsymbol{x_{j_1}^{\sigma_1} \& \ldots \& x_{j_r}^{\sigma_r}}$ *является допустимой для* $\boldsymbol{F_1}$ *тогда и только тогда, когда набор столбцов матрицы* $\boldsymbol{L}$ *с номерами* $\boldsymbol{j_1, \ldots, j_r}$ *является* $\boldsymbol{(\sigma_1, \ldots, \sigma_r)}$ *-покрытием.*

• **Утверждение 8.** ЭК $\boldsymbol{x_{j_1}^{\sigma_1} \& \ldots \& x_{j_r}^{\sigma_r}}$ *является неприводимой для* $\boldsymbol{F_1}$ *тогда и только тогда, когда набор столбцов матрицы* $\boldsymbol{L}$ *с номерами* $\boldsymbol{j_1, \ldots, j_r}$ *содержит* $\boldsymbol{(\sigma_1, \ldots, \sigma_r)}$ *-подматрицу.*

• **Утверждение 9.** ЭК $\boldsymbol{x_{j_1}^{\sigma_1} \& \ldots \& x_{j_r}^{\sigma_r}}$ *является максимальной для* $\boldsymbol{F_1}$ *тогда и только тогда, когда набор столбцов матрицы* $\boldsymbol{L}$ *с номерами* $\boldsymbol{j_1, \ldots, j_r}$ *является тупиковым* $\boldsymbol{(\sigma_1, \ldots, \sigma_r)}$ *-покрытием.*

- Если строке матрицы $\boldsymbol{L}$ с номером $\boldsymbol{i}$, $\boldsymbol{i} \in \{\boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{u}\}$, поставить в соответствие дизъюнкцию $\boldsymbol{D_i^{(2)} = x_1^{\sigma_{i1}} \vee \dots \vee x_n^{\sigma_{in}}}$ и рассмотреть булеву функцию $\boldsymbol{F_2}$, реализуемую КНФ $\boldsymbol{D_1^{(2)} \& \dots \& D_u^{(2)}}$, то задачи построения допустимых и максимальны конъюнкций функции $\boldsymbol{F_2}$ могут быть сформулированы соответственно как задачи построения $\boldsymbol{R(\sigma)}$ -покрытий и тупиковых $\boldsymbol{R(\sigma)}$ -покрытий матрицы $\boldsymbol{L}$.

- Таким образом, алгоритмы построения покрытий общего вида булевой матрицы можно применять для построения ДНФ булевой функции, заданной КНФ, и наоборот. Аналогичное утверждение имеет место и для целочисленной матрицы. Алгоритмы построения покрытий общего вида целочисленной матрицы можно применять для ДНФ двузначной логической функции $\boldsymbol{k}$ -значной логики, заданной КНФ.

УПРАЖНЕНИЯ

- Пусть две строки булевой матрицы $\boldsymbol{L}$ имеют вид $(\boldsymbol{b_1}, \dots, \boldsymbol{b_n})$ и $(\boldsymbol{c_1}, \dots, \boldsymbol{c_n})$, где $\boldsymbol{c_j} \geq \boldsymbol{b_j}$ при $\boldsymbol{j = 1, 2, \dots, n}$. Будем говорить, что вторая строка охватывает первую. Показать, что при удалении охватывающих строк из матрицы $\boldsymbol{L}$ множество её покрытий не меняется.
- 2. Пусть булева $\boldsymbol{F}$ функция задана КНФ $(\boldsymbol{x_1} \vee \boldsymbol{x_2})(\boldsymbol{x_2} \vee \boldsymbol{x_3})(\boldsymbol{x_1} \vee \boldsymbol{x_3})$. Построить сокращенную ДНФ функции $\boldsymbol{F}$ сведением задачи к задаче построения неприводимых покрытий булевой матрицы.
- 3. Пусть булева $\boldsymbol{F}$ функция задана множеством наборов, на которых она равна $\boldsymbol{0}$, а именно, $\boldsymbol{N_{\overline{F}}} = \{(\boldsymbol{1,1,1}), (\boldsymbol{0,1,1}), (\boldsymbol{1,0,1})\}$. Построить сокращенную ДНФ функции $\boldsymbol{F}$ двумя способами: а) сведением к задаче построения тупиковых $\boldsymbol{\sigma}$ -покрытий булевой матрицы; б) сведением к задаче построения тупиковых $\boldsymbol{R(\sigma)}$ -покрытий булевой матрицы.
- 4. Доказать утверждения $4 - 9$.