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 3
Общая схема конструирования дискретных (логических) процедур распознавания с использованием понятия элементарного классификатора. Модели алгоритмов голосования по антипредставительным наборам и по покрытиям класса.

- В лекции 2 рассматривались модели классических логических алгоритмов распознавания. В этих алгоритмах используется процедура голосования по фрагментам описаний обучающих объектов, т.е. по таким наборам значений признаков, которые встречаются в описаниях прецедентов. В последнее время построены новые модели, в которых при вычислении оценки за класс K используется процедура голосования по наборам значений признаков, не встречающихся в описаниях прецедентов из K .
- Для того, чтобы описать общую схему конструирования логических процедур распознавания нам понадобится понятие элементарного классификатора.

• Пусть $\boldsymbol{H}$ – набор из $\boldsymbol{r}$ различных признаков вида $\boldsymbol{H} = \{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}}\}$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r})$, $\boldsymbol{\sigma_i}$ – допустимое значение признака $\boldsymbol{x_{j_i}}$, $i = 1, 2, \dots, r$. Пару $(\boldsymbol{\sigma}, \boldsymbol{H})$ назовем элементарным классификатором (эл.кл.).

• Близость объекта $\boldsymbol{S} = (\boldsymbol{a_1}, \dots, \boldsymbol{a_n})$ из M и эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r})$, $\boldsymbol{H} = (\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}})$, будем оценивать величиной $\boldsymbol{B}(\boldsymbol{\sigma}, \boldsymbol{S}, \boldsymbol{H})$ равной 1, если $\boldsymbol{a_{j_t}} = \boldsymbol{\sigma_t}$ при $\boldsymbol{t = 1, 2, \dots, r}$, и равной 0 в противном случае.

• Если $\boldsymbol{B}(\boldsymbol{\sigma}, \boldsymbol{S}, \boldsymbol{H}) = \boldsymbol{1}$, то будем говорить, что объект $\boldsymbol{S}$ содержит эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$.

• Нетрудно видеть, что фрагмент описания обучающего объекта вида $(\boldsymbol{S'}, \boldsymbol{H})$, где $\boldsymbol{S'} = (\boldsymbol{a'_1}, \dots, \boldsymbol{a'_n})$, $\boldsymbol{H} = \{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}}\}$, порождает эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$, где $\boldsymbol{\sigma} = (\boldsymbol{a'_{j_1}}, \dots, \boldsymbol{a'_{j_r}})$.

• Множество всех эл.кл. обозначим через $\boldsymbol{C}$. Итак, $\boldsymbol{C} = \{(\boldsymbol{\sigma}, \boldsymbol{H})\}$, где $\boldsymbol{H} \subseteq (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$, $\boldsymbol{H} = (\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}})$, $\boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r})$, $\boldsymbol{\sigma_i}$ – допустимое значение признака $\boldsymbol{x_{j_i}}$, $\boldsymbol{i = 1, 2, \dots, r}$.

• Будем считать, что каждый распознающий алгоритм $\boldsymbol{A}$ для каждого класса $\boldsymbol{K}$, $\boldsymbol{K} \in \{\boldsymbol{K_1}, \dots, \boldsymbol{K_l}\}$, строит некоторое подмножество $\boldsymbol{C^A(K)}$ множества $\boldsymbol{C}$. Обозначим

$$\boldsymbol{C^A} = \bigcup_{\boldsymbol{j=1}}^{\boldsymbol{l}} \boldsymbol{C^A(K_j)}.$$

• Распознавание объекта $\boldsymbol{S}$ осуществляется на основе вычисления величины $\boldsymbol{B(\sigma, S, H)}$ для каждого эл.кл. $(\boldsymbol{\sigma, H})$ множества $\boldsymbol{C^A}$, т.е. по каждому элементу множества $\boldsymbol{C^A}$ осуществляется процедура голосования. В результате для каждого класса $\boldsymbol{K}$, $\boldsymbol{K} \in \{\boldsymbol{K_1}, \dots, \boldsymbol{K_l}\}$, вычисляется оценка $\boldsymbol{\Gamma(S, K)}$ принадлежности объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$. Таким образом, алгоритм $\boldsymbol{A}$ из рассматриваемого семейства распознающих алгоритмов определяется множеством эл.кл. $\boldsymbol{C^A}$. Далее мы увидим, что алгоритмы будут отличаться и способом вычисления оценки $\boldsymbol{\Gamma}(\mathbf{S}, \mathbf{K})$, которая получается на основе голосования по эл.кл. из $\boldsymbol{C^A(K)}$.

- В рассмотренных в лекции 2 моделях число голосов, поданных парами из $\boldsymbol{C^A(K)}$ за принадлежность объекта $\boldsymbol{S}$ к классу $\boldsymbol{K}$, в простейших модификациях вычисляется по формуле

$$\boldsymbol{\Gamma_1(S,K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} P_{(\sigma,H)} B(\sigma,S,H),}$$

где $\boldsymbol{P_{(\sigma,H)}}$ – вес эл.кл. $\boldsymbol{(\sigma,H)}$. В качестве $\boldsymbol{P_{(\sigma,H)}}$ обычно берётся число обучающих объектов из $\boldsymbol{K}$, содержащих $\boldsymbol{(\sigma,H)}$.

- Например, если $\boldsymbol{A}$ – алгоритм вычисления оценок, то множество $\boldsymbol{C^A}$ состоит из таких эл.кл., которые порождаются опорными множествами алгоритма $\boldsymbol{A}$. Если же $\boldsymbol{A}$ – алгоритм голосования по представительным наборам, то $\boldsymbol{C^A(K)}$ порождается некоторым подмножеством множества представительных наборов класса $\boldsymbol{K}$. В обоих случаях оценка $\boldsymbol{\Gamma(S,K)}$ получается на основе суммирования величин $\boldsymbol{P_{(\sigma,H)}}$ $\boldsymbol{B(\sigma,S,H)}$, где $\boldsymbol{(\sigma,H)} \in \boldsymbol{C^A(K)}$. Далее объект $\boldsymbol{S}$ относится к тому классу, для которого оценка принадлежности наибольшая (если таких классов несколько, то происходит отказ от распознавания).

• Будем говорить, что элементарный классификатор $(\boldsymbol{\sigma}, \boldsymbol{H})$ является ***корректным для класса*** $\boldsymbol{K}$, если нельзя указать пару обучающих объектов $\boldsymbol{S'}$ и $\boldsymbol{S''}$ таких, что $\boldsymbol{S'} \in \boldsymbol{K}$, $\boldsymbol{S''} \notin \boldsymbol{K}$ и $\boldsymbol{B}(\boldsymbol{\sigma}, \boldsymbol{S''}, \boldsymbol{H}) = \boldsymbol{B}(\boldsymbol{\sigma}, \boldsymbol{S''}, \boldsymbol{H}) = \mathbf{1}$. Нетрудно видеть, что если $\boldsymbol{A}$ – тестовый алгоритм или алгоритм голосования по представительным наборам, то множество $\boldsymbol{C^A}$ состоит только из корректных элементарных классификаторов.

• В общем случае элементарный классификатор $(\boldsymbol{\sigma}, \boldsymbol{H})$, по отношению к классу $\boldsymbol{K}$ может обладать одним из следующих трех свойств:

• 1) каждый обучающий объект $\boldsymbol{S'}$ из класса $\boldsymbol{K}$ содержит $(\boldsymbol{\sigma}, \boldsymbol{H})$;

• 2) не все, а лишь некоторые обучающие объекты $\boldsymbol{S'}$ из класса $\boldsymbol{K}$ содержат $(\boldsymbol{\sigma}, \boldsymbol{H})$;

• 3) ни один обучающий объект $\boldsymbol{S'}$ из класса $\boldsymbol{K}$ не содержит $(\boldsymbol{\sigma}, \boldsymbol{H})$.

- Первая ситуация встречается крайне редко, поэтому работать с эл.кл., для которых выполняется свойство 1, не представляется возможным.
- Существенное различие в информативности следующих двух свойств заключается в том, что свойство 2 характеризует лишь некоторое подмножество обучающих объектов из класса ***K***, а свойство 3 все объекты из ***K***.
- Следовательно, в случае, когда важно рассматривать класс ***K*** изолированно от других классов, напрашивается вывод о большей информативности таких эл.кл., для которых выполнено свойство 3. В указанном случае аргументом за отнесение распознаваемого объекта ***S*** в класс ***K*** более естественно считать ситуацию, когда набор значений признаков не присутствует у всех объектов из класса ***K*** и не присутствует у объекта ***S***.

- Классические логические процедуры классификации (тестовые алгоритмы, алгоритмы голосования по представительным наборам) основаны на построении корректных эл.кл., обладающих свойством 2. В этой лекции мы рассмотрим корректные процедуры классификации, основанные на построении корректных эл.кл., обладающих свойством 3. Этими моделями являются **модель голосования по покрытиям класса** и **модель голосования по антипредставительным наборам.** В ряде случаев указанные модели позволяют повысить качество распознавания и требуют меньших вычислительных затрат.
- Описываемые ниже модели (модель голосования по покрытиям класса и модель голосования по антипредставительным наборам класса) основаны на построении для каждого класса только таких эл.кл., которые не содержатся в описаниях ни одного объекта класса. Если распознаваемый объект ***S*** также не содержит подобный эл.кл., то считается, что этот объект близок к рассматриваемому классу. Таким образом, ищутся закономерности, присущие всем обучающим объектам рассматриваемого класса, т.е. каждый эл.кл. характеризует весь класс целиком. Использование данных моделей позволяет несколько снизить вычислительные затраты в случае большого числа классов.

• Эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$ называется ***покрытием класса*** $\boldsymbol{K}$, если любой обучающий объект из $\boldsymbol{K}$ не содержит $(\boldsymbol{\sigma}, \boldsymbol{H})$. Покрытие класса $\boldsymbol{K}$ называется тупиковым, если любое его собственное подмножество не является покрытием класса $\boldsymbol{K}$.

• В моделях голосования по (тупиковым) покрытиям класса множество $\boldsymbol{C^A(K)}$ состоит из таких эл.кл., которые являются (тупиковыми) покрытиями класса $\boldsymbol{K}$. В отличие от классического алгоритма здесь эл.кл. из $\boldsymbol{C^A(K)}$ голосует за принадлежность распознаваемого объекта классу $\boldsymbol{K}$, если этот эл.кл. не встречается в описании рассматриваемого объекта. Принадлежность объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$ (в простейшей модификации) оценивается величиной

$$\boldsymbol{\Gamma_2(S,K) = \frac{1}{|C^A(K)|}\sum_{(\sigma,H)\in C^A(K)}(1 - B(\sigma,S,H))}.$$

• Переформулируем понятие представительного набора класса $\boldsymbol{K}$, используя понятие эл.кл. Пусть $\overline{\boldsymbol{K}} = \{\boldsymbol{K_1}, \dots, \boldsymbol{K_l}\} \setminus \boldsymbol{K}$. Будем рассматривать $\overline{\boldsymbol{K}}$ как отдельный класс, т.е. будем считать, что у нас всего два класса $\boldsymbol{K}$ и $\overline{\boldsymbol{K}}$.

• Эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$ называется ***представительным набором для класса*** $\boldsymbol{K}$, если ни один обучающий объект из $\overline{\boldsymbol{K}}$ не содержит $(\boldsymbol{\sigma}, \boldsymbol{H})$ и хотя бы один обучающий объект из $\boldsymbol{K}$ содержит $(\boldsymbol{\sigma}, \boldsymbol{H})$. Таким образом, (тупиковый) представительный набор $(\boldsymbol{\sigma}, \boldsymbol{H})$ является (тупиковым) покрытием для $\overline{\boldsymbol{K}}$ и не является покрытием для $\boldsymbol{K}$.

- Рассмотрим теперь модель с антипредставительными наборами.
- Эл.кл. $(\boldsymbol{\sigma}, \boldsymbol{H})$ называется **(тупиковым) антипредставительным набором**, если $(\boldsymbol{\sigma}, \boldsymbol{H})$ является (тупиковым) покрытием класса $\boldsymbol{K}$ и хотя бы один обучающий объект из $\overline{\boldsymbol{K}}$ содержит $(\boldsymbol{\sigma}, \boldsymbol{H})$, т.е. $(\boldsymbol{\sigma}, \boldsymbol{H})$ не является покрытием для $\overline{\boldsymbol{K}}$. Принадлежность объекта $\boldsymbol{S}$ классу $\boldsymbol{K}$ (в простейшей модификации) оценивается величиной

$$\boldsymbol{\Gamma_3(S, K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} P_{(\sigma,H)}(1 - B(\sigma, S, H))},$$

где $\boldsymbol{P_{(\sigma,H)}}$ – число обучающих объектов и $\overline{\boldsymbol{K}}$, содержащих $(\boldsymbol{\sigma}, \boldsymbol{H})$.

- Заметим, что представительный набор для класса $\boldsymbol{K}$ является антипредставительным для $\overline{\boldsymbol{K}}$. Поэтому нетрудно показать, в случае двух классов при использовании обеих моделей объект $\boldsymbol{S}$ будет отнесен к одному и тому же классу.

• В самом деле, пусть $\boldsymbol{A_1}$ – алгоритм голосования по всем представительным наборам, $\boldsymbol{A_2}$ – алгоритм голосования по всем антипредставительным наборам. Как уже было сказано, $\boldsymbol{C^{A_1}(K_1) = C^{A_2}(K_2) = C_1}$, $\boldsymbol{C^{A_1}(K_2) = C^{A_2}(K_1) = C_2}$. Пусть в распознаваемом объекте $\boldsymbol{S}$ с представительными наборами класса $\boldsymbol{K_1}$ совпадают $\boldsymbol{q_1}$ фрагментов, а с представительными наборами класса $\boldsymbol{K_2 - q_2}$ фрагментов. Пусть далее $\boldsymbol{\Gamma_{A_i}(S, K_j)}$, $\boldsymbol{i \in \{1, 2\}}$, $\boldsymbol{j \in \{1, 2\}}$ – результат голосования для алгоритма $\boldsymbol{A_i}$ за класс $\boldsymbol{K_j}$. Тогда имеем следующие оценки

$$\boldsymbol{\Gamma_{A_1}(S, K_1) = \frac{q_1}{|C^{A_1}(K_1)|} = \frac{q_1}{C_1}}, \quad \boldsymbol{\Gamma_{A_1}(S, K_2) = \frac{q_2}{|C^{A_1}(K_2)|} = \frac{q_2}{C_2}},$$

$$\boldsymbol{\Gamma_{A_2}(S, K_1) = \frac{|C^{A_2}(K_1)| - q_2}{C^{A_2}(K_1)} = \frac{C_2 - q_2}{C_2} = 1 - \frac{q_2}{C_2}},$$

$$\boldsymbol{\Gamma_{A_2}(S, K_2) = \frac{|C^{A_2}(K_2)| - q_1}{C^{A_2}(K_2)} = 1 - \frac{q_1}{C_1}}$$

Очевидно, если $\boldsymbol{\Gamma_{A_1}(S, K_1) > \Gamma_{A_1}(S, K_2)}$, то и $\boldsymbol{\Gamma_{A_2}(S, K_1) > \Gamma_{A_2}(S, K_2)}$.

- Рассмотрим теперь случай, когда классов больше двух. Очевидно, что представительный набор для класса $\boldsymbol{K_i}$, $\boldsymbol{i \in \{1, 2, \dots, l\}}$, является антипредставительным для каждого из остальных классов. Однако антипредставительный набор для $\boldsymbol{K_i}$ может не являться представительным набором для любого из остальных классов. В самом деле, если некий фрагмент описания встречается в описаниях объектов из классов $\boldsymbol{K_{i_1}}$ и $\boldsymbol{K_{i_2}}$, но не встречается в $\boldsymbol{K_i}$, тогда он является антипредставительным набором для K_i, но не является представительным ни для $\boldsymbol{K_{i_1}}$, ни для $\boldsymbol{K_{i_2}}$. Таким образом, если $\boldsymbol{A_1}$ - алгоритм голосования по представительным наборам, $\boldsymbol{A_2}$ - алгоритм голосования по антипредставительным наборам, то справедливо

$$\boldsymbol{C^{A_1} \subseteq C^{A_2}}$$

- Наиболее трудоемким этапом при построении рассматриваемых алгоритмов является нахождение требуемого множества покрытий. В классических моделях покрытия строятся для $\boldsymbol{\overline{K}}$, в новых моделях покрытия строятся для $\boldsymbol{K}$. Таким образом, при большом числе классов новые модели требуют меньших вычислительных затрат.

УПРАЖНЕНИЯ

- 1. В обучающей выборке

$\{\boldsymbol{S_1 = (0,1,1)},\ \boldsymbol{S_2 = (1,2,0)}\}$ – класс $\boldsymbol{K_1}$

$\{\boldsymbol{S_3 = (1,2,1)},\ \boldsymbol{S_4 = (1,1,0)}\}$ – класс $\boldsymbol{K_2}$

найти все корректные эл.кл. для каждого из классов $\boldsymbol{K_1}$ и $\boldsymbol{K_2}$, порождаемые одним признаком.

- 2. Пусть $\boldsymbol{A_1}$ – алгоритм голосования по представительным наборам, $\boldsymbol{A_3}$ – алгоритм голосования по покрытиям классов. Верно ли хотя бы одно из включений: 1) $\boldsymbol{C^{A_1} \subseteq C^{A_3}}$ 2) $\boldsymbol{C^{A_3} \subseteq C^{A_1}}$