Е.Р. Горяинова
НИУ ВШЭ, Москва
Т.И. Слепнёва
МАИ, Москва

Методы бинарной классификации объектов с номинальными показателями

Рассматривается задача классификации респондентов на классы принимающих и не принимающих участие в благотворительных пожертвованиях. Построено оптимальное (в байесовском смысле) решающее дискриминантное правило разделения объектов на два класса в условиях, когда все признаки наблюдаемых объектов измеряются в номинальной шкале и между этими признаками есть зависимости. Методами ROC-анализа проведено сравнение разработанного классификатора с «наивным» байесовским классификатором, классификатором, основанным на методе опорных векторов, а также с реализованными в пакете SPSS классификаторами, использующими бинарную логистическую регрессию и дискриминантное правило Фишера. Результаты ROC-анализа показали, что предложенное правило имеет более высокое качество классификации респондентов, чем все перечисленные выше правила.

Ключевые слова: *дискриминантный анализ, решающее правило, байесовское решение, линейное правило Фишера, бинарная логистическая регрессия, метод опорных векторов, ROC-кривая, показатель AUC.*

Классификация JEL: C38, Z13.

Введение

Задача дискриминантного анализа – это задача классификации объектов, каждый из которых характеризуется многомерным показателем X по двум (или нескольким) классам. При этом требуется, чтобы каждый класс был представлен своими типичными объектами. Набор показателей типичных представителей классов называется *обучающей выборкой*. Используя ее, надо построить правило дискриминации, которое позволяет распознавать, к какому классу относится объект, не вошедший в обучающую выборку. Такие задачи актуальны для экономических исследований, например при классификации стран или предприятий по их экономическим показателям. Известно об успешном решении таких задач в медицине при классификации пациентов на классы здоровых и больных (Zweig, Campbell, 1993; Schlisterman, Perkins, Li, 2005), в политологии при классификации сенаторов по фракциям по итогам их голосования (Ким, Мьюллер, Клекка, 1989, с. 84–87) и в психологии. В последнее время методы дискриминантного анализа стали активно применяться в кредитном скоринге при классификации предполагаемых заемщиков банка на кредитоспособных и некредитоспособных (Паклин, Орешков, 2010).

Известно оптимальное (в минимаксном и байесовском смысле) решение такой задачи в случае, когда закон распределения совокупности показателей X для каждого класса является многомерным гауссов-

Журнал НЭА
№2 (14), 2012
С. 27–49.

ским (Ивченко, Медведев, 2010, глава 7; Дронов, 2003, глава 9). Этим решением является прогностическое правило Фишера. Несмотря на то что данное правило оптимально для нормальной модели, его часто используют и в ситуациях, когда семейство распределений показателей X неизвестно.

Однако если такое правило можно реализовать в случаях, когда распределение показателей X «похоже» на гауссовское, то в ситуациях, когда наблюдения характеризуются показателями, измеренными в номинальной шкале, это правило неприемлемо. Это связано с тем, что переменные, измеренные в номинальной шкале (например, регион проживания, пол или род деятельности респондента), несравнимы. Важно отметить, что любые показатели, которые измеряются в количественной шкале, могут быть переведены в номинальную шкалу. Например, возраст респондента можно измерять не годами, а категориями: младшая (18–34 года), средняя (35–54 года) и старшая (55 и более лет) возрастные группы. Более того, часто возникают ситуации, когда респонденты не готовы или отказываются указать такой количественный показатель, как размер доходов семьи, однако готовы сказать, к какой категории, описывающей материальное положение семьи, они могут себя отнести. Поскольку при проведении социологических и психологических исследований большая часть показателей измеряется в качественной шкале, то в этой ситуации оптимальное решающее дискриминантное правило будет отличаться от правила Фишера.

В работе (Ивченко, Медведев, 2010, глава 7) показано, что построение оптимального решения в случае произвольного распределения показателя X представляется достаточно сложным. Однако показатели, измеренные в номинальной шкале, могут быть описаны многомерным полиномиальным распределением. Это обстоятельство позволяет найти оптимальное в байесовском смысле решение задачи. Байесовское правило, построенное в предположении независимости компонент вектора показателей X, получило в теории машинного обучения название «наивного» байесовского классификатора (Mirkin, 2011). В задаче, которая рассматривается в данной работе, некоторые компоненты вектора X являются зависимыми.

Помимо байесовского оптимального решения следует отметить еще два подхода к решению задачи классификации – метод опорных векторов Вапника (Vapnik, 1995) и метод логистической регрессии, позволяющий оценивать условную вероятность принадлежности объекта к классу при заданных значениях показателя X (Hosmer, Lemeshov, 2000).

Все обсуждаемые в этой работе методы относятся к линейным алгоритмам классификации.

Работа имеет следующую структуру. В разд. 1 описана задача разделения респондентов на классы принимающих и не принима-

Журнал НЭА
№2 (14), 2012
С. 27–49.

ющих участие в благотворительных пожертвованиях (по данным Центра исследований гражданского общества и некоммерческого сектора НИУ ВШЭ). В разд. 2 дана математическая постановка описанной социологической задачи, кратко представлены такие известные линейные классификаторы, как дискриминантное правило Фишера, классификатор, использующий бинарную логистическую модель, метод опорных векторов Вапника и «наивный» байесовский классификатор. Затем в разд. 2.6 построено оптимальное (в байесовском смысле) решающее правило классификации объектов с номинальными зависимыми признаками. В разд. 3 решена задача классификации респондентов с помощью описанных в разд. 2 решающих правил, обсуждены достоинства и недостатки рассматриваемых классификаторов в данной задаче. В разд. 4 методами ROC-анализа проведен сравнительный анализ качества классификации построенного правила, учитывающего зависимость компонент вектора показателей, с правилами, рассмотренными в разд. 2.

1. Классификация респондентов на группы участвующих и не участвующих в благотворительных пожертвованиях

Центром исследований гражданского общества и некоммерческого сектора НИУ ВШЭ была разработана анкета (Мерсиянова, Якобсон, 2009), состоящая из более чем ста вопросов, характеризующих социально-демографическое положение и качество жизни респондента, его отношение к благотворителям и благотворительной деятельности. Затем была составлена репрезентативная выборка из 1600 респондентов и проведен социологический опрос. Теперь по результатам опроса необходимо построить такое классифицирующее правило, которое по социально-демографическим характеристикам респондента позволило бы определить его принадлежность (или непринадлежность) к группе людей, принимающих участие в благотворительных пожертвованиях. Отметим, что аналогичную классификацию можно провести, используя в качестве характеристик респондента показатели качества его жизни (например, удовлетворенность жизнью в целом, семьей, работой, досугом, друзьями). Однако смешение в одной модели объективных социально-демографических характеристик и субъективных характеристик качества жизни нецелесообразно.

Обозначим через K_1 – класс респондентов, участвующих в благотворительных пожертвованиях, а через K_2 – не участвующих. В качестве ключевой переменной, определяющей разделение респондентов на классы, выберем переменную Y, которая будет принимать значение «класс K_1», если респондент на вопрос анкеты: «Как часто за последние 2–3 года вам приходилось делать благотворительные пожертвования, лично давать незнакомым нуждающимся людям, включая просящих милостыню, деньги?» – давал ответ: 1) «занимаюсь этим посто-

Журнал НЭА
№2 (14), 2012
С. 27–49.

янно»; 2) «занимался этим несколько раз»; 3) «занимался этим один раз»; и значение «класс K_2», если респондент давал ответ «не делал» или «затрудняюсь ответить».

Первый важный шаг при проведении классификации – это отбор показателей, которые являются наиболее специфичными для качественного разделения на классы. Очевидно, что включение в дискриминантное решающее правило малоинформативных переменных не только усложняет вычисления, но может и заметно ухудшить качество классификации, так как каждый малоинформативный показатель несет в себе достаточно большую долю «шума». В качестве информативных показателей целесообразно выбрать переменные, распределение которых в классах K_1 и K_2 сильно различается. Другими словами, эти переменные имеют значимую статистическую связь с переменной Y, определяющей принадлежность респондента к классу. Для выявления зависимости между номинальными переменными применим критерий χ^2. Используя процедуру пакета SPSS, получаем, что на уровне значимости менее 0,001 отвергнута гипотеза о независимости переменной Y с переменными X_1 (пол респондента), X_2 (тип населенного пункта, в котором проживает респондент), X_3 (федеральный округ проживания), X_4 (возраст респондента), X_5 (образование респондента).

Интересно отметить, что гипотезу о независимости Y и переменной «материальное положение» можно отвергнуть только на уровне значимости 0,102. Поэтому можно сказать, что материальное положение не является той характеристикой, которая определяет желание респондента совершать благотворительные пожертвования. Результаты теста χ^2 представлены в табл. 1. В последнем столбце приведены коэффициенты Крамера, измеряющие силу связи показателей с переменной Y.

Таблица 1

Результаты теста χ^2

Признаки	Значение статистики	Число степеней свободы, l	Значение квантили $\chi^2_{0,95}(l)$	Уровень значимости, которому соответствует вычисленное значение статистики	Значение коэффициента Крамера
Пол респондента	11,684	1	3,84	< 0,001	0,085
Тип населенного пункта	18,293	4	9,49	< 0,001	0,107
Федеральный округ	32,213	6	12,59	< 0,001	0,142
Возраст	24,418	2	5,99	< 0,001	0,124
Образование	24,906	3	7,81	< 0,001	0,125

Журнал НЭА
№2 (14), 2012
С. 27–49.

Итак, в результате отбора признаков классификация респондентов будет проводиться по следующим переменным, градации (категории) которых указаны в скобках:

1) пол респондента (мужской, женский);
2) тип населенного пункта (Москва или Санкт-Петербург; город с населением более 500 тыс. человек; город с населением 100–500 тыс. человек; город с населением менее 100 тыс. человек; село);
3) федеральный округ проживания (Центральный федеральный округ; Северо-Западный федеральный округ; Южный федеральный округ; Приволжский федеральный округ; Уральский федеральный округ; Сибирский федеральный округ; Дальневосточный федеральный округ);
4) возраст (младшая возрастная группа 18–35 лет; средняя возрастная группа 36–54 года; старшая возрастная группа 55 лет и старше);
5) образование (ниже среднего; среднее общее; среднее специальное; высшее).

Отметим также, что помимо определения дискриминантных признаков при помощи критерия χ^2 можно прибегнуть и к экспертному выбору набора показателей, а после проведения классификации исследовать однородность каждого показателя во вновь полученных классах. Если распределение некоторого показателя не имеет статистически значимого различия в классах, то можно сделать вывод о нецелесообразности включения этого показателя в список переменных, определяющих попадание объекта в ту или иную группу.

2. Описание методов классификации объектов

Формализуем описанную социологическую задачу и кратко рассмотрим следующие известные методы классификации:

1) дискриминантное правило Фишера;
2) алгоритм, использующий бинарную логистическую регрессию;
3) метод опорных векторов Вапника;
4) «наивное» байесовское правило для упрощенной модели,

а затем построим оптимальный байесовский классификатор для объектов с зависимыми номинальными показателями.

Пусть проведено наблюдение над n объектами, каждый из которых характеризуется m-мерным вектором признаков $X = (X_1, ..., X_m)$. Про каждый исследуемый объект известно, что он относится к классу K_1 или к классу K_2. Набор показателей X представителей классов K_1 и K_2 является обучающей выборкой объема $n = n_1 + n_2$, где n_1 и n_2 – число представителей классов K_1 и K_2 соответственно. Предполагается, что в случае, когда объект принадлежит классу K_i, $i = 1, 2$, распределение случайного вектора признаков X этого объекта принадлежит параметрическому семейству P_{Θ_i}, $\theta_i \in \Theta$, с разными значениями параметров

Журнал НЭА
№2 (14), 2012
С. 27–49.

θ_1 и θ_2. Параметры θ_1, θ_2 неизвестны, но могут быть оценены по обучающей выборке.

Задача состоит в том, чтобы по реализации $x=(x_1,...,x_m)$ вектора признаков $X=(X_1,...,X_m)$ вновь поступившего на исследование объекта определить класс K_i, $i=1,\ 2$, к которому этот объект следует отнести. То есть требуется построить решающее правило $\delta(x)$, которое ставит в соответствие каждому наблюдению $x=(x_1,...,x_m)$ одно из значений множества решений $D=\{d_1,d_2\}$, где решение d_i означает, что объект x определен в класс K_i, $i=1,\ 2$.

Решающее правило определяется дискриминантной функцией $\gamma=\gamma(x)$ следующим образом:

$$\delta(x)=\begin{cases} d_1, & \text{если } \gamma(x)\geq c; \\ d_2, & \text{если } \gamma(x)<c, \end{cases} \qquad (1)$$

где c – некоторое пороговое значение.

Таким образом, всякое дискриминантное правило $\delta(x)$ порождает разбиение выборочного пространства на две непересекающиеся области W_1 и W_2, где $W_i=\{x:\ \delta(x)=d_i\}$.

Перейдем к рассмотрению классифицирующих правил, позволяющих решить сформулированную задачу.

2.1. Прогностическое правило Фишера

Построение этого правила основано на предположении о том, что показатели X объектов класса K_1 имеют многомерное гауссовское распределение $N(m_1,V)$, показатели объектов класса K_2 – многомерное распределение $N(m_2,V)$, ковариационные матрицы V этих распределений одинаковы, а векторы средних m_1 и m_2 различны. Тогда дискриминантная функция оптимального (в смысле минимума байесовского риска) правила (Ивченко, Медведев, 2010, глава 7; Дронов, 2003, глава 9) имеет вид

$$\gamma(x)=\hat{V}^{-1}(\hat{m}_1-\hat{m}_2)^t(x-0,5(\hat{m}_1+\hat{m}_2)),$$

где $\hat{m}_1,\hat{m}_2,\hat{V}$ – оценки неизвестных параметров распределений $N(m_1,V)$ и $N(m_2,V)$, построенные по обучающей выборке. Если ковариационные матрицы в разных классах различаются, то в качестве оценки $\hat{V}$ обычно выбирают усредненную оценку $\hat{V}=0,5\left(\hat{V}_1+\hat{V}_2\right)$, где $\hat{V}_1$ и $\hat{V}_2$ – оценки ковариационных матриц в классах K_1 и K_2. Пороговая константа c в формуле (1) для правила Фишера имеет вид

$$c=\ln(\hat{\pi}_2/\hat{\pi}_1)=\ln(n_2/n_1),$$

а правило Фишера реализовано в пакете программ SPSS.

Журнал НЭА
№2 (14), 2012
С. 27–49.

2.2. Правило, основанное на модели бинарной логистической регрессии

Опишем процедуру классификации, используя модель бинарной логистической регрессии (Hosmer, Lemeshov, 2000). Пусть переменная Y, определяющая принадлежность респондента к классу, принимает значение 1, если респондент относится к классу K_1, и значение 0, если респондент относится к классу K_2.

Рассмотрим регрессионную модель вида

$$P(Y=1 \mid X=x)=\frac{\exp(b_0+b_1x_1+\ldots+b_mx_m)}{1+\exp(b_0+b_1x_1+\ldots+b_mx_m)}, \qquad (2)$$

где $X=(X_1,\ldots,X_m)$, $x=(x_1,\ldots,x_m)$, а $b_0,\ldots,b_m$ – неизвестные параметры регрессионного уравнения. Регрессионные коэффициенты уравнения (2) можно оценить методом максимального правдоподобия (ММП). Имея оценки $\hat{b}_0,\ldots,\hat{b}_m$ неизвестных параметров $b_0,\ldots,b_m$ уравнения (2), решающее правило строят следующим образом:

$$\delta(x)=\begin{cases} d_1, & \text{если } \exp(\hat{b}_0+\sum_{i=1}^{m}\hat{b}_ix_i)\Big/\Big[1+\exp(\hat{b}_0+\sum_{i=1}^{m}\hat{b}_ix_i)\Big]\geq 0{,}5; \\ d_2, & \text{если } \exp(\hat{b}_0+\sum_{i=1}^{m}\hat{b}_ix_i)\Big/\Big[1+\exp(\hat{b}_0+\sum_{i=1}^{m}\hat{b}_ix_i)\Big]< 0{,}5. \end{cases} \qquad (3)$$

Другими словами, респондента с реализацией признаков $x=(x_1,\ldots,x_m)$ следует определить в класс K_1, если оценка условной вероятности попадания в этот класс при таком значении x будет не меньше 0,5.

В пакете программ SPSS построение оценок регрессионного уравнения (2) и классификационное правило (3) реализованы в процедуре Regression Binary Logistic.

2.3. Метод опорных векторов[1]

Рассмотрим метод опорных векторов (Vapnik, 1995), который хорошо зарекомендовал себя при классификации объектов с количественными показателями. Этот метод основан на построении такой разделяющей гиперплоскости H, чтобы объекты из разных классов были удалены от нее на максимально далекое расстояние. В случае линейно разделимой выборки удается найти не только разделяющую гиперплоскость, но и разделяющую полосу, в которую не попадают обучающие объекты, а H проходит по середине этой полосы. Задача построения гиперплоскости H и разделяющей полосы сводится к нахождению константы b и вектора $w=(w_1,\ldots,w_m)$, такого, что норма $\langle w,w\rangle$ минимальна и при этом выполняется n неравенств-ограничений: $y_i\left(\langle w,x_i\rangle-b\right)\geq 1$, $i=1,\ldots,n$, где i – номер объекта; $y_i=1$,

[1] Удачное изложение метода опорных векторов имеется в (Воронцов, 2007).

Журнал НЭА
№2 (14), 2012
С. 27–49.

если объект $x_i = (x_{1i}, ..., x_{mi})$ принадлежит классу K_1; $y_i = -1$, если объект $x_i = (x_{1i}, ..., x_{mi})$ принадлежит классу K_2. Неравенства задают условие непопадания объектов в разделительную полосу, ширина которой равна $2/\langle w, w\rangle$. Тогда оптимальная гиперплоскость H описывается уравнением $\langle w, x\rangle = b$, разделяющая классы полоса представлена множеством $\{x: -1 \le \langle w, x\rangle - b \le 1\}$, а величина $\rho_i = y_i(\langle w, x_i\rangle - b)$ – это отступ объекта x_i от границы классов. Решающее правило имеет вид

$$\delta(x) = \begin{cases} d_1, & \text{если } \langle w, x\rangle - b \ge 1; \\ d_2, & \text{если } \langle w, x\rangle - b \le -1. \end{cases}$$

В рассматриваемой задаче, как и в большинстве практических задач, выборка не является линейно разделимой, и условие непопадания объектов в полосу может быть нарушено. В этой ситуации вводят неотрицательные переменные η_i, $i = 1, ..., n$, определяющие величину ошибки на объекте $x_i = (x_{1i}, ..., x_{mi})$. Затем находят b и вектор w из условия минимума функционала $0,5\langle w, w\rangle + C\sum_{i=1}^{n}\eta_i$ при ограничениях $y_i(\langle w, x_i\rangle - b) \ge 1 - \eta_i$, $\eta_i \ge 0$, $i = 1, \ldots, n$. Здесь $C\sum_{i=1}^{n}\eta_i$ – суммарный штраф за ошибки классификации, а C – некоторый управляющий параметр, который позволяет находить компромисс между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки. Параметр регуляризации C должен задаваться пользователем.

Сформулированная задача является задачей квадратичного программирования и имеет единственное решение. Функция Лагранжа имеет вид

$$L(w, b, \alpha, \beta) = 0,5\langle w, w\rangle - \sum_{i=1}^{n}\alpha_i\left[y_i(\langle w, x_i\rangle - b) - 1\right] - \sum_{i=1}^{n}\eta_i(\alpha_i + \beta_i + C),$$

где $\alpha_1, ..., \alpha_n$ – переменные, двойственные к w, а $\beta_1, ..., \beta_n$ – переменные, двойственные к $\eta_1, ..., \eta_n$.

Учитывая необходимое условие экстремума, получим, что $w = \sum_{i=1}^{n}\alpha_i y_i x_i$, $\sum_{i=1}^{n}\alpha_i y_i = 0$, $\alpha_i + \beta_i = C$. В силу неотрицательности β_i имеем $0 \le \alpha_i \le C$.

Объекты x_i, для которых $\alpha_i > 0$, называют опорными векторами. Таким образом, вектор w определяется только опорными векторами.

Классификационное правило метода опорных векторов будет следующим:

Журнал НЭА
№2 (14), 2012
С. 27–49.

$$\delta(x) = \begin{cases} d_1, & \text{если } \langle w, x \rangle - b \geq 0; \\ d_2, & \text{если } \langle w, x \rangle - b < 0. \end{cases}$$

Важно отметить, что множество опорных векторов можно разделить на две группы. К первой группе относятся объекты x_i, для которых $0 < \alpha_i < C$ и $\rho_i = 1$, ко второй – объекты x_i, для которых $\alpha_i = C$, а $\rho_i < 1$. Объекты первой группы лежат на границе разделительной полосы и классифицируются верно, объекты второй группы попадают внутрь разделительной полосы. Последние называются опорными нарушителями и могут быть классифицированы правильно, если $0 \leq \rho_i < 1$, или неправильно, если $\rho_i < 0$. Если объекты из разных классов сильно «перемешаны», будет возникать большое число опорных нарушителей и качество классификации может ухудшиться. Чтобы избежать этого, вводят преобразование $\varphi(x)$ пространства признаков X в новое пространство Z более высокой размерности. Цель такого преобразования – получение в новом спрямляющем пространстве Z линейно разделимой выборки $\varphi(x_i)$, $i = 1, \ldots, n$. Если обучающая выборка непротиворечива, то найдется пространство размерности не более n, в котором выборка будет линейно разделима. Скалярное произведение в пространстве Z определяется ядерной функцией (ядром) $K(x, y) = \langle \varphi(x), \varphi(y) \rangle$ при условии, что пространство Z наделено скалярным произведением. К сожалению, общие методы построения ядер и спрямляющих пространств до сих пор не разработаны. И подбор ядра надо проводить для каждой конкретной задачи.

2.4. Оптимальное байесовское правило для номинальных дискриминантных признаков

Пусть известно, что каждый признак X_l, $l = 1, ..., m$, измеряется в номинальной шкале и имеет z_l градаций (категорий). Например, пусть признак X_l обозначает образование респондента. Тогда X_l имеет четыре градации: образование ниже среднего; среднее образование; среднее специальное образование; высшее образование. Таким образом, при проведении одного наблюдения (опросе одного респондента) каждый признак X_l, $l = 1, ..., m$, может принимать одно из z_l возможных значений, а реализацией $x_l = (x_{l1}, ..., x_{lz_l})$ признака X_l будет вектор размера z_l, состоящий из $z_l - 1$ нуля и одной единицы. Причем если единица стоит на месте j, то это означает, что в данной реализации признак X_l принял градацию номер j. Случайный вектор X_l, который имеет указанную структуру, можно описать полиномиаль-

Журнал НЭА
№2 (14), 2012
С. 27–49.

ным распределением (Королюк и др., 1978, глава 6.4) с параметрами $(1, p_{l1}, ..., p_{lz_l})$, где $p_{l1}, ..., p_{lz_l}$ – вероятности появления в одном опыте градаций $1, \ldots, z_l$ соответственно. При этом параметры распределения $p_{l1}, ..., p_{lz_l}$ различны для класса K_1 и класса K_2. Обозначим через $p_{lj,(i)}$, $l = 1, \ldots, m$; $j = 1, \ldots, z_l$; $i = 1, 2$, вероятность того, что признак X_l респондента из класса i имеет категорию номер j. Тогда распределение признака X_l, $l = 1, \ldots, m$, для каждого респондента описывается полиномиальным распределением с параметрами $(1, p_{l1,(1)}, ..., p_{lz_l,(1)})$, если респондент принадлежит к классу K_1, и полиномиальным распределением с параметрами $(1, p_{l1,(2)}, ..., p_{lz_l,(2)})$, если респондент принадлежит к классу K_2. Таким образом, распределение каждой компоненты случайного вектора $X = (X_1, ..., X_m)$ – это многомерное полиномиальное распределение с указанными параметрами. Предполагая, что компоненты вектора X являются независимыми случайными величинами, получим, что распределение вектора X описывается произведением m полиномиальных распределений с соответствующими параметрами. Если компоненты вектора $X = (X_1, ..., X_m)$ зависимы, то условную вероятность того, что признак X_l респондента из класса K_i имеет категорию номер j при условии, что произошло событие $\bigcap_{k=1}^{l-1}\{X_k = x_k\}$, обозначим $p_{(lj|1,\ldots,l-1),(i)}$.

Заметим, что вероятности $p_{(lj|1,\ldots,l-1),(i)}$ и $p_{lj,(i)}$, $l = 1, ..., m$; $j = 1, ..., z_l$; $i = 1, 2$, неизвестны, но при достаточно большой выборке могут быть оценены частотами $\hat{p}_{(lj|1,\ldots,l-1),(i)}$ и $\hat{p}_{lj,(i)}$ появления соответствующих градаций по обучающей выборке. Вероятности попадания респондента в классы K_1 и K_2 можно оценить частотами

$$\hat{\pi}_1 = \frac{n_1}{n_1 + n_2}, \ \hat{\pi}_2 = \frac{n_2}{n_1 + n_2} = 1 - \hat{\pi}_1 .$$

2.5. «Наивный» байесовский классификатор

Рассмотрим оптимальное в байесовском смысле правило для классификации объектов, характеризующихся независимыми между собой номинальными признаками. Дискриминантная функция этого правила (Mirkin, 2011) имеет вид

$$\gamma(x) = \sum_{l=1}^{m}\sum_{j=1}^{z_l} x_{lj} \ln \frac{\hat{p}_{lj,(1)}}{\hat{p}_{lj,(2)}}, \qquad (4)$$

в которой

$$x_{lj} = \begin{cases} 1, \text{ если у наблюдаемого респондента признак } X_l \text{ принял градацию номер } j; \\ 0 \text{ – в остальных случаях;} \end{cases} \qquad (5)$$

Журнал НЭА
№2 (14), 2012
С. 27–49.

$\hat{p}_{lj,(i)}$ – оценки соответствующих вероятностей $p_{lj,(i)}$, $l=1,...,m$, $j=1,...,z_l$, $i=1,2$.

Пороговая константа этого правила совпадает с константой правила Фишера. В теории машинного обучения такое правило принято называть «наивным» байесовским классификатором. Следует отметить, что требование независимости показателей $X_1,...,X_m$ является существенным при выводе этого правила.

2.6. Оптимальный байесовский классификатор для объектов с зависимыми показателями

В рассматриваемой задаче классификации респондентов условие независимости показателей $X_1,...,X_m$ не выполняется. Поэтому построим оптимальный байесовский классификатор для объектов, характеризующихся номинальными признаками, которые могут быть зависимыми.

Теорема. *В описанной выше модели байесовское решающее правило* $\delta(x)$ *имеет вид*

$$\delta(x)=\begin{cases} d_1, & \text{если } \gamma(x)\geq \ln(\hat{\pi}_2/\hat{\pi}_1); \\ d_2, & \text{если } \gamma(x)<\ln(\hat{\pi}_2/\hat{\pi}_1), \end{cases} \tag{6}$$

где дискриминантная функция

$$\gamma(x)=\sum_{j=1}^{z_1}x_{1j}\ln\left(\frac{\hat{p}_{1j,(1)}}{\hat{p}_{1j,(2)}}\right)+\sum_{l=2}^{m}\sum_{j=1}^{z_l}x_{lj}\prod_{k=1}^{l-1}\prod_{i=1}^{z_k}x_{ki}\ln\frac{\hat{p}_{(lj|1,...,l-1),(1)}}{\hat{p}_{(lj|1,...,l-1),(2)}}, \tag{7}$$

в которой x_{lj}, $l=1,...,m$; $j=1,...,z_l$, *определены формулой* (5), *а* $\hat{\pi}_2$, $\hat{\pi}_1=1-\hat{\pi}_2$ – *оценки вероятностей попадания в классы* K_2 *и* K_1 *соответственно.*

Следствие 1. Если в обучающей выборке число респондентов в классах K_1 и K_2 одинаково, то $\hat{\pi}_1=\hat{\pi}_2=0,5$ и решающее правило имеет вид

$$\delta(x)=\begin{cases} d_1, & \text{если } \gamma(x)\geq 0; \\ d_2, & \text{если } \gamma(x)<0. \end{cases}$$

Следствие 2. Если компоненты вектора X являются независимыми, то дискриминантная функция (7) совпадает с функцией (4).

Доказательство теоремы. Пусть $L(d,\theta)$ – функция потерь, определяющая потери от неправильной классификации объектов, а значения этой функции $L(d_j,\theta_i)=l(j\,|\,i)$, $j,i=1,\ 2$ – потери, которые имеют место в случае, когда объект из класса K_i отнесен к классу K_j. Естественно считать, что $l(1\,|\,1)=l(2\,|\,2)=0$, т.е. при пра-

Журнал НЭА
№2 (14), 2012
С. 27–49.

вильной классификации потери равны нулю. Определим функцию риска $R(\delta) = (R_1(\delta), R_2(\delta))$, где $R_i(\delta) = M[L(\delta(X), \theta_i)]$.

Вычисляя математическое ожидание, получим $R_i(\delta) = \sum_{j=1}^{2} l(j|i)p(j|i)$, где $p(j|i) = P_{\theta_i}(X \in K_j)$ – вероятность того, что объект из класса K_i будет помещен в класс K_j. Тогда $R(\delta) = (l(2|1)p(2|1), l(1|2)p(1|2))$.

Таким образом, значение риска $R_i(\delta)$ характеризует средние потери, которые имеют место при классификации по правилу δ объекта из класса K_i.

Пусть известно априорное распределение $\pi = (\pi_1, \pi_2)$, где π_i – вероятность того, что произвольный наблюдаемый объект принадлежит классу K_i, $i = 1, 2$. Тогда байесовский риск определяется следующим образом (Ивченко, Медведев, 2010):

$$r(\delta) = \sum_{i=1}^{2} R_i(\delta)\pi_i = \sum_{j=1}^{2}\sum_{i=1}^{2} l(j|i)p(j|i)\pi_i = l(1|2)p(1|2)\pi_2 + l(2|1)p(2|1)\pi_1. \quad (8)$$

Решающее правило δ^*, при котором байесовский риск (8) минимален, т.е. $\delta^* = \arg\min_{\delta} r(\delta)$, называется байесовским решением в задаче классификации. Найдем данное решение, считая, что $l(1|2) = l(2|1) = 1$.

Согласно теореме Байеса, апостериорные вероятности $\pi_i(x)$, $i = 1, 2$, классов при условии $X = x$ равны

$$\pi_i(x) = \pi_i P(x \in K_i) / \sum_{j=1}^{2}\left[\pi_j P(x \in K_j)\right], \quad i = 1, 2.$$

Таким образом, если объект x отнести к классу K_1 (т.е. $\delta(x) = d_1$), условное математическое ожидание функции потерь при условии $X = x$ будет

$$\begin{aligned} M(L(\delta(X), \theta) | X = x) &= \sum_{i=1}^{2} L(d_1, \theta_i)\pi_i(x) = \\ &= l(1|2)\pi_2 P(x \in K_2) / \left\{\sum_{j=1}^{2}\pi_j P(x \in K_j)\right\} = \\ &= \pi_2 P(x \in K_2) / \left\{\sum_{j=1}^{2}\pi_j P(x \in K_j)\right\}, \end{aligned} \quad (9)$$

а если объект x отнести к классу K_2 (т.е. $\delta(x) = d_2$), то

$$M(L(\delta(X), \theta) | X = x) = \pi_1 P(x \in K_1) / \left\{\sum_{j=1}^{2}\pi_j P(x \in K_j)\right\}. \quad (10)$$

Журнал НЭА
№2 (14), 2012
С. 27–49.

Заметим, что знаменатели в выражениях (9) и (10) совпадают, и функция $r(\delta)$ достигнет минимального значения, если решающее правило $\delta^*(x)$ будет построено следующим образом:

$$\delta^*(x) = \begin{cases} d_1, \text{ если } \pi_1 P(x \in K_1) \geq \pi_2 P(x \in K_2); \\ d_2, \text{ если } \pi_2 P(x \in K_2) \geq \pi_1 P(x \in K_1). \end{cases}$$

Таким образом, область W_1 имеет вид

$$W_1 = \left\{ x : \frac{P(x \in K_1)}{P(x \in K_2)} \geq \frac{\pi_2}{\pi_1} \right\} = \left\{ x : \ln \frac{P(x \in K_1)}{P(x \in K_2)} \geq \ln \frac{\pi_2}{\pi_1} \right\}. \quad (11)$$

Вычислим вероятность того, что наблюдение $x = (x_1, ..., x_m)$, где каждая компонента x_l, $l = 1, ..., m$, является вектором размерности z_ℓ, попадает в класс K_i, $i = 1, 2$.

Как было показано выше, каждая компонента случайного вектора $X = (X_1, ..., X_m)$ имеет полиномиальное распределение. Согласно полиномиальному закону распределения (Королюк и др., 1978, глава 6.4), вероятность того, что в одном опыте признак X_l, $l = 1, ..., m$, примет значение x_l, где $x_l = (x_{l1}, ..., x_{lz_l})$, а x_{lj} определено в (5), равна

$$P(X_l = x_l) = \frac{1!}{x_{l1}! \ldots x_{lz_l}!} p_{l1}^{x_{l1}} p_{l2}^{x_{l2}} \ldots p_{lz_l}^{x_{lz_l}} = \prod_{j=1}^{Z_l} p_{lj}^{x_{lj}}.$$

Если компоненты $X_1, ..., X_m$ вектора X независимы, то

$$P(X = x) = \prod_{l=1}^{m} P(X_l = x_l) = \prod_{l=1}^{m} \prod_{j=1}^{z_l} p_{lj}^{x_{lj}},$$

а вероятность принадлежности наблюдения $x = (x_1, ..., x_m)$ классу K_i вычисляется следующим образом:

$$P(x \in K_i) = \prod_{l=1}^{m} \prod_{j=1}^{z_l} p_{lj,(i)}^{x_{lj}}. \quad (12)$$

Если же компоненты $X_1, \ldots, X_m$ вектора X зависимы, то вероятность $P(X = x) = P(X_1 = x_1, ..., X_m = x_m)$ следует вычислять согласно формуле умножения вероятностей

$$P(X = x) = P(X_1 = x_1) \prod_{l=2}^{m} P\left(X_l = x_l \mid \bigcap_{k=1}^{l-1} \{X_k = x_k\} \right).$$

Таким образом, по формуле умножения имеем

$$P(X = x) = \prod_{j=1}^{z_1} p_{1j}^{x_{1j}} \prod_{l=2}^{m} \prod_{j=1}^{z_l} p_{(lj|1,\ldots,l-1)}^{x_{lj} \prod_{k=1}^{l-1} \prod_{i=1}^{z_k} x_{ki}}.$$

Оценивая в последнем выражении неизвестные условные и безусловные вероятности появления градаций признаков в каждом классе соответствующими частотами и подставляя в (11), получим, что

Журнал НЭА
№2 (14), 2012
С. 27–49.

$$W_1 = \left\{ x : \sum_{j=1}^{z_1} x_{1j} \ln\left(\hat{p}_{1j,(1)} / \hat{p}_{1j,(2)} \right) + \sum_{l=2}^{m} \sum_{j=1}^{z_l} x_{lj} \prod_{k=1}^{l-1} \prod_{i=1}^{z_k} x_{ki} \ln \frac{\hat{p}_{(lj|1,\ldots,l-1),(1)}}{\hat{p}_{(lj|1,\ldots,l-1),(2)}} \geq \ln \frac{\hat{\pi}_2}{\hat{\pi}_1} \right\}, \quad (13)$$

а для независимых компонент, согласно (11) и (12) –

$$W_1 = \{ x : \sum_{l=1}^{m} \sum_{j=1}^{z_l} x_{lj} \ln \frac{\hat{p}_{lj,(1)}}{\hat{p}_{lj,(2)}} \geq \ln \frac{\hat{\pi}_2}{\hat{\pi}_1} \}.$$

3. Результаты классификации респондентов

В имеющейся выборке 780 респондентов относятся к группе принимавших участие в благотворительных пожертвованиях, а 820 – к группе не принимавших участие в благотворительных пожертвованиях. Таким образом, оценки вероятностей попадания респондентов в классы K_1 и K_2 соответственно равны 0,4875 и 0,5125.

При проведении классификации в качестве обучающей выборки будем использовать все имеющиеся наблюдения. В классификационных таблицах рассматриваемых решающих правил будут представлены доли верно и неверно классифицированных по классам K_1 и K_2 респондентов.

Классификация, основанная на прогностическом правиле Фишера и учитывающая различие ковариационных матриц в классах, получена при помощи процедуры Discriminant пакета программ SPSS. Соответствующая классификационная таблица представлена в табл. 2. Процент верных классификаций составляет 56,3%.

Таблица 2

Классификационная таблица правила Фишера

Истинная группа	Результат классификации	
	Класс K_1	Класс K_2
Класс K_1	0,585	0,415
Класс K_2	0,459	0,541

Классификация, использующая модель бинарной логистической регрессии (правило (3)), получена с помощью процедуры Regression Binary Logistic пакета программ SPSS. Результаты приведены в классификационной табл. 3. Процент верно классифицированных респондентов составляет 59,3%.

Таблица 3

Классификационная таблица правила, использующего модель бинарной логистической регрессии

Истинная группа	Результат классификации	
	Класс K_1	Класс K_2
Класс K_1	0,573	0,427
Класс K_2	0,389	0,611

Журнал НЭА
№2 (14), 2012
С. 27–49.

К достоинствам этого метода классификации следует отнести то, что имеется возможность проверки значимости как каждого отдельного регрессионного коэффициента, так и уравнения регрессии в целом. Согласно критерию отношения правдоподобия, регрессионная модель (2) с выбранными регрессорами $X_1, \ldots, X_5$ оказалась значимой, значимо отличаются от нуля все коэффициенты $b_1, \ldots, b_5$, а при попытке включения в число регрессоров переменной «материальное положение» оказалось, что соответствующий ей регрессионный коэффициент незначимо отличается от нуля.

Недостаток этого метода заключается в том, что поиск оценок МП параметров $b_0, \ldots, b_m$ осуществляется итерационным методом, так как система уравнений правдоподобия является нелинейной по b.

Перейдем к классификации по методу опорных векторов. В качестве преобразования $\varphi(x)$ естественно выбрать (Вапник, Червоненкис, 1974, глава 3) преобразование, переводящее m-мерный вектор признаков x в бинарный вектор $\tilde{x} = (x_{11}, \ldots, x_{1,z_1}, \ldots, x_{m_1}, \ldots, x_{mz_m})$ размерности $z = z_1 + \ldots + z_m$, где z_l – число категорий у признака X_l, а $x_{lj}, l = 1, \ldots, m; \; j = 1, \ldots, z_l$, определены формулой (5).

К сожалению, рассматриваемая нами задача классификации респондентов осложнена тем, что обучающая выборка является противоречивой, так как 10% респондентов, имеющих одинаковые признаки $x = (x_1, \ldots, x_5)$, относятся к разным классам. Это означает, что нельзя найти такое преобразование $\varphi(x)$, которое перевело бы имеющуюся выборку в линейно разделимую, а указанные 10% объектов заведомо будут опорными нарушителями.

Решая описанную выше задачу квадратичного программирования в среде MATLAB, выбирая указанное преобразование $\varphi(x)$ и параметр регуляризации $C \geq 5$, мы получили 58,13% верно классифицированных объектов. Результаты этой классификации представлены в табл. 4. Если не осуществлять преобразование пространства признаков X и рассматривать функцию ядра, равную скалярному произведению, то процент верных классификаций составляет 56,31% и практически совпадает с тем, который дает дискриминант Фишера. Попытка применения гауссовой ядерной функции в данной задаче не привела к улучшению классификации.

Таблица 4

Классификационная таблица правила, основанного на методе опорных векторов

Истинная группа	Результат классификации	
	Класс K_1	Класс K_2
Класс K_1	0,667	0,333
Класс K_2	0,50	0,50

Журнал НЭА
№2 (14), 2012
С. 27–49.

Проведем классификацию, используя «наивное» байесовское правило, основанное на дискриминантной функции (4). Для того чтобы вычислить функцию (4), необходимо оценить параметры распределения отобранных показателей в каждом классе. Этими параметрами являются вероятности p_{lj} событий, состоящих в том, что признак X_l респондента имеет категорию номер j. Известно, что несмещенными и сильно состоятельными оценками вероятностей событий являются частоты появления этих событий.

Реализации оценок параметров $p_{lj,(i)}$, $l = 1,...,m$; $j = 1,...,z_l$; $i = 1, 2$, распределений дискриминантных переменных для классов K_1 и K_2 приведены в табл. 5–9.

Таблица 5

Оценки параметров распределения переменной X_1 (половая принадлежность респондента)

Класс	$\hat{p}_{11,(i)}$	$\hat{p}_{12,(i)}$
Класс K_1 ($i = 1$)	0,4064	0,5936
Класс K_2 ($i = 2$)	0,4915	0,5085

Таблица 6

Оценки параметров распределения переменной X_2 (тип населенного пункта, в котором проживает респондент)

Класс	$\hat{p}_{21,(i)}$	$\hat{p}_{22,(i)}$	$\hat{p}_{23,(i)}$	$\hat{p}_{24,(i)}$	$\hat{p}_{25,(i)}$
Класс K_1 ($i = 1$)	0,0962	0,2026	0,2115	0,2026	0,2872
Класс K_2 ($i = 2$)	0,1159	0,1659	0,1683	0,178	0,372

Таблица 7

Оценки параметров распределения переменной X_3 (федеральный округ проживания)

Класс	$\hat{p}_{31,(i)}$	$\hat{p}_{32,(i)}$	$\hat{p}_{33,(i)}$	$\hat{p}_{34,(i)}$	$\hat{p}_{35,(i)}$	$\hat{p}_{36,(i)}$	$\hat{p}_{37,(i)}$
Класс K_1 ($i = 1$)	0,2295	0,1038	0,1385	0,2564	0,1205	0,0846	0,0667
Класс K_2 ($i = 2$)	0,2927	0,0878	0,1805	0,172	0,0939	0,0768	0,0963

Таблица 8

Оценки параметров распределения переменной X_4 (возраст респондента)

Класс	$\hat{p}_{41,(i)}$	$\hat{p}_{42,(i)}$	$\hat{p}_{43,(i)}$
Класс K_1 ($i = 1$)	0,3577	0,4038	0,2385
Класс K_2 ($i = 2$)	0,3585	0,3049	0,3366

Журнал НЭА
№2 (14), 2012
С. 27–49.

Таблица 9

Оценки параметров распределения переменной X_5 (образование респондента)

Класс	$\hat{p}_{51,(i)}$	$\hat{p}_{52,(i)}$	$\hat{p}_{53,(i)}$	$\hat{p}_{54,(i)}$
Класс K_1 $(i=1)$	0,0577	0,25	0,4103	0,2821
Класс K_2 $(i=2)$	0,1061	0,311	0,361	0,222

Подставляя найденные оценки в функцию (4) и определив пороговую константу $c = \ln(\hat{\pi}_2 / \hat{\pi}_1) = \ln(n_2 / n_1) = \ln(820/780)$, проводим классификацию и получаем процент верных классификаций, равный 59,625%.

Результаты классификации, основанной на «наивном» байесовском правиле (4), использующем упрощенную модель, представлены в табл. 10.

Таблица 10

Классификационная таблица «наивного» байесовского правила

Истинная группа	Результат классификации	
	Класс K_1	Класс K_2
Класс K_1	0,615	0,385
Класс K_2	0,422	0,578

Напомним, что «наивный» байесовский классификатор является оптимальным в смысле минимума байесовского риска (8) только при условии независимости дискриминантных признаков.

В рассматриваемой задаче имеется зависимость между показателями X_1 (пол респондента) и X_4 (возраст респондента); между X_2 (тип населенного пункта, в котором проживает респондент) и X_3 (федеральный округ проживания); между X_5 (образование респондента) и X_4, X_3, X_2. Поэтому применим разработанное решающее правило классификации (6)–(7), которое оптимально и для зависимых номинальных признаков.

Для вычисления дискриминантной функции (7) требуется оценить в каждом классе безусловные вероятности p_{1j}, $j = 1,...,z_1$, и p_{2j}, $j = 1,...,z_2$ (эти оценки представлены в табл. 5, 6), условные вероятности $p_{(3j|2)}, p_{(4j|1)}$ и $p_{(5j|2,3,4)}$, где $j = 1,...,z_l$.

Оценками условных вероятностей являются «условные» частоты появления соответствующих событий. Оценки условных вероятностей $p_{(3j|2)}$ для классов K_1 и K_2 представлены в матрицах размера $(z_3 \times z_2)$:

Журнал НЭА
№2 (14), 2012
С. 27–49.

$$\hat{p}_{(3\,j|2),(1)} = \begin{pmatrix} 0,0615 & 0,0346 & 0 & 0 & 0 & 0 & 0 \\ 0,0026 & 0 & 0,0205 & 0,1 & 0,0462 & 0,0321 & 0,001 \\ 0,0551 & 0,0282 & 0,0295 & 0,0308 & 0,0321 & 0,0141 & 0,0218 \\ 0,05 & 0,018 & 0,0269 & 0,0539 & 0,0244 & 0,0128 & 0,0167 \\ 0,0603 & 0,0231 & 0,0615 & 0,0718 & 0,0179 & 0,0256 & 0,0269 \end{pmatrix},$$

$$\hat{p}_{(3\,j|2),(2)} = \begin{pmatrix} 0,0854 & 0,0305 & 0 & 0 & 0 & 0 & 0 \\ 0,0378 & 0 & 0,0256 & 0,0402 & 0,0171 & 0,0305 & 0,0146 \\ 0,0463 & 0,0146 & 0,0402 & 0,0293 & 0,0061 & 0,0085 & 0,0232 \\ 0,0463 & 0,0207 & 0,0244 & 0,0171 & 0,0317 & 0,0061 & 0,0317 \\ 0,0768 & 0,0219 & 0,0902 & 0,0854 & 0,0390 & 0,0317 & 0,0269 \end{pmatrix},$$

а оценки условных вероятностей $p_{(4\,j|1)}$ для классов K_1 и K_2 – в матрицах размера $(z_4 \times z_1)$:

$$\hat{p}_{(4\,j|1),(1)} = \begin{pmatrix} 0,1603 & 0,1974 \\ 0,1692 & 0,2346 \\ 0,0769 & 0,1615 \end{pmatrix}, \quad \hat{p}_{(4\,j|1),(2)} = \begin{pmatrix} 0,1939 & 0,1646 \\ 0,1646 & 0,1402 \\ 0,1329 & 0,2037 \end{pmatrix}.$$

Оценка условных вероятностей $p_{(5\,j|2,3,4)}$ является громоздкой и содержит большое количество нулевых значений, поэтому мы ее здесь не приводим.

Результаты классификации, основанной на правиле (6)–(7), приведены в табл. 11. Процент верных классификаций составляет 65,302%.

Таблица 11

Классификационная таблица оптимального байесовского правила

Истинная группа	Результат классификации	
	Класс K_1	Класс K_2
Класс K_1	0,599	0,401
Класс K_2	0,293	0,707

4. Сравнительный анализ классификационных решающих правил

Для сравнения различных дискриминантных правил необходимо выбрать некий инструмент, позволяющий оценивать качество классификации. Удобным средством оценки качества классификационного правила является метод, основанный на анализе так называемой операционной характеристической кривой ROC (Receiver Operating Characteristic – функциональные характеристики приемника). Традиционный ROC-анализ (Паклин, Орешков, 2010; Файнзильберг, Жук, 2009) предусматривает сравнение таких операци-

Журнал НЭА
№2 (14), 2012
С. 27–49.

онных характеристик правила, как чувствительность (доля респондентов, которые верно отнесены правилом к классу K_1) и специфичность (доля респондентов, которые верно отнесены правилом к классу K_2). Интегральной характеристикой оценки качества правила является площадь под ROC-кривой.

При построении кривой нас не интересует то, как построен непосредственно алгоритм классификации, т.е. правило рассматривается как «черный ящик», который на основании информации об объекте принимает одно из двух решений – d_1 или d_2. Для того чтобы построить ROC-кривую, проводится «экзамен» на обучающей выборке и результаты «экзамена» представляются в классификационной таблице (табл. 12).

Таблица 12

Результаты тестирования по обучающей выборке

Истинный результат	Результат тестирования	
	Решение d_1 «относится к классу K_1»	Решение d_2 «относится к классу K_2»
«Относится к классу K_1»	*TP*	*FN*
«Относится к классу K_2»	*FP*	*TN*

Примечание. В таблице приняты следующие обозначения: *TP* (true positive) – число правильных результатов «относится к классу K_1» (истинноположительный результат); *TN* (true negative) – число правильных результатов «относится к классу K_2» (истинноотрицательный результат); *FP* (false positive) – число респондентов, ошибочно отнесенных к классу K_1 (ложноположительный результат); *FN* (false negative) – число респондентов, ошибочно отнесенных к классу K_2 (ложноотрицательный результат).

По данным табл. 12 вычисляются операционные характеристики правила:

а) чувствительность (sensitivity) $S_E = \frac{TP}{TP + FN}$;

б) специфичность (specificity) $S_P = \frac{TN}{TN + FP}$.

Имея эти характеристики, можно представить результаты «экзамена» в двумерном ROC-пространстве, в котором по оси ординат откладываются значения S_E, а по оси абсцисс – значения $1 - S_P$. Таким образом, классифицирующее правило (бинарный классификатор) с фиксированными операционными характеристиками S_E и S_P отображается точкой в ROC-пространстве.

Изменяя пороговое значение c в классификационном правиле (1), будем получать новые последовательности точек (S_E, S_p) в ROC-пространстве. Отсюда вытекает следующий алгоритм построения экспериментальной ROC-кривой:

Журнал НЭА
№2 (14), 2012
С. 27–49.

1) отсортировать наблюдаемые значения дискриминантной функции $\gamma(x)$ в порядке убывания;
2) последовательно уменьшать (с некоторым шагом) порог c, перемещаясь вниз по списку отсортированных значений $\gamma(x)$, и по экзаменационной выборке наблюдений вычислять соответствующие пары значений S_E и S_P;
3) отобразить полученные таким образом последовательности точек (S_E, S_p) в ROC-пространстве.

ROC-пространство дает наглядное графическое представление о диагностической ценности правила и позволяет сравнивать качество различных правил.

Для сравнения качества различных классифицирующих правил удобно использовать интегральную характеристику *AUC* (Area Under Curve) – площадь под ROC-кривой. Понятно, что величина *AUC* определяет среднюю чувствительность теста $\overline{S}_E$ при возможных значениях специфичности $0 \le S_P \le 1$ или среднюю специфичность теста $\overline{S}_P$ при возможных значениях чувствительности $0 \le S_E \le 1$. Ситуация, при которой ROC-кривая совпадает с диагональю (в этом случае $AUC = 0{,}5$), соответствует случайному угадыванию. И правило с такими характеристиками является бесполезным. Для идеального правила величина $AUC = 1$. Поэтому считается, что чем ближе *AUC* к единице, тем качественнее правило. Традиционно принята (Паклин, Орешков, 2010; Файнзильберг, Жук, 2009) следующая экспертная шкала для значений *AUC*, по которой можно судить о качестве модели (табл. 13).

Таблица 13

Характеристики модели по значению *AUC*

Интервал *AUC*	Качество модели
0,9–1,0	Отличное
0,8–0,9	Очень хорошее
0,7–0,8	Хорошее
0,6–0,7	Среднее
0,5–0,6	Неудовлетворительное

Следует отметить, что показатель *AUC* не содержит никакой информации о чувствительности и специфичности модели и предназначен скорее для сравнительного анализа нескольких моделей.

При построении ROC-кривой правила воспользуемся встроенной функцией пакета SPSS. В качестве проверяемой переменной выбирается вектор значений дискриминантной функции, получаемый при помощи разработанной программы; в качестве переменной состояния – переменная, обозначающая истинную группу рассматриваемого

Журнал НЭА
№2 (14), 2012
С. 27–49.

респондента; в качестве значения переменной состояния – значение, которое соответствует положительному значению дискриминантной функции. Графики ROC-кривых для правила, реализованного в процедуре Discriminant пакета SPSS (правило *A*), «наивного» байесовского правила (правило *B*), описанного в настоящей статье оптимального байесовского правила (правило *C*) и правила, основанного на использовании логистической регрессии, реализованного в процедуре Regression Binary Logistic пакета SPSS (правило *D*), представлены на рисунке. Площади под кривыми представлены в табл. 14.

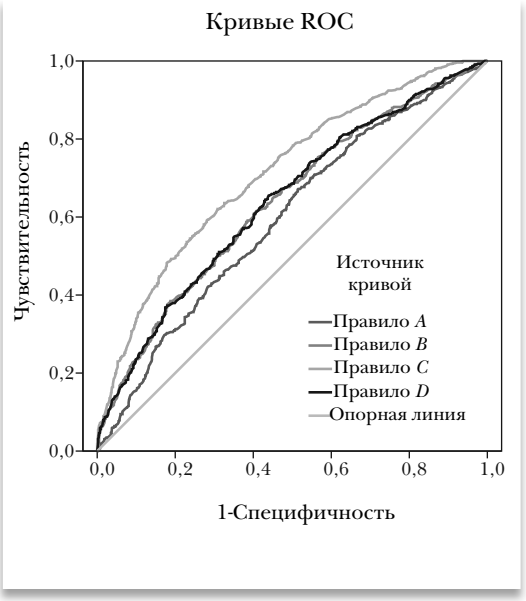

Рисунок.
Графики ROC-кривых

Таблица 14

Значения показателя *AUC*

Правило	*A*	*B*	*C*	*D*
Показатель *AUC*	0,597	0,639	0,713	0,640

Как видно из результатов, представленных в табл. 14, качество модели, основанной на оптимальном байесовском правиле, характеризуется как «хорошее», тогда как качество правил «наивного» байесовского классификатора и логистической регрессии характеризуется как «среднее». Модель, основанная на линейном классификаторе Фишера, является «неудовлетворительной».

Заключение

В работе сформулировано дискриминантное правило бинарной классификации объектов, характеристики которых измеряются в номинальной шкале и могут быть зависимыми. Показана оптимальность (в смысле минимума байесовского риска) предложенного правила. Важно отметить, что оно является достаточно простым в алгоритмическом отношении и не требует оценивания и обращения ковариационной матрицы наблюдений или решения оптимизационных задач. Однако при большом числе связей между признаками может оказаться затруднительным оценивание условных вероятностей.

В реальной задаче разделения респондентов на классы это правило показало более высокое качество классификации по сравнению

Журнал НЭА
№2 (14), 2012
С. 27–49.

с другими правилами. Согласно экспертной шкале ROC-анализа его качество оценивается как «хорошее». Не так хорошо зарекомендовали себя «наивный» байесовский классификатор и бинарная логистическая регрессия, их качество классификации оценивается как «среднее». Важно отметить, что использование модели логистической регрессии позволяет не только классифицировать объекты, но и оценивать вероятности попадания объектов в классы, а также проверять значимость дискриминантных переменных. Оценивая вероятность попадания в класс респондента с заданными характеристиками, исследователь имеет возможность выявить модальную категорию респондентов этого класса. Это означает, что с помощью логистической регрессии можно описать наиболее вероятный «социально-демографический портрет» респондента, участвующего в благотворительных пожертвованиях.

«Наивный» байесовский классификатор оптимален только при условии независимости признаков. Несмотря на то что это условие в рассматриваемой задаче не выполняется, качество «наивного» байесовского классификатора практически совпадает с качеством классификатора, основанного на логистической регрессии. «Наивное» байесовское правило имеет самый простой алгоритм классификации среди всех рассмотренных здесь правил.

Чуть хуже двух предыдущих классификаторов показал себя метод опорных векторов. Причиной этого, на наш взгляд, является наличие большого числа опорных нарушителей, появление которых связано со спецификой выборки. К недостаткам алгоритма опорных векторов следует отнести неоднозначность выбора ядерной функции и параметра регуляризации.

Неудовлетворительное, согласно шкале ROC-анализа, качество имеет линейное правило Фишера. Это объясняется тем, что правило Фишера оптимально для показателей, имеющих многомерное гауссовское распределение, а показатели респондентов имеют полиномиальное распределение.

Таким образом, зная такие объективные социально-демографические характеристики респондента, как пол, возраст, образование, федеральный округ и тип населенного пункта проживания, можно построить качественный прогноз относительно участия (или неучастия) этого респондента в благотворительных пожертвованиях.

Литература

Вапник В.Н., Червоненкис А.Я. (1974). Теория распознавания образов (статистические проблемы обучения). М.: Наука.

Воронцов К.В. (2007). Лекции по методу опорных векторов. [Электронный ресурс] Курс лекций. Режим доступа: http://www.ccas.ru/voron/download/SVM.pdf, свободный. Загл. с экрана. Яз. рус. (дата обращения: 2011 г.).

Журнал НЭА
№2 (14), 2012
С. 27–49.

Дронов С.В. (2003). Многомерный статистический анализ. Барнаул: Изд-во Алт. гос. ун-та.

Ивченко Г.И., Медведев Ю.И. (2010). Введение в математическую статистику. М.: URSS.

Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р. и др. (1989). Факторный, дискриминантный и кластерный анализ /Под ред. И.С. Енюкова. М.: Фин. и стат.

Королюк В.С и др. (1978). Справочник по теории вероятностей и математической статистике. М.: Наука.

Мерсиянова И.В., Якобсон Л.И. (2009). Практики филантропии в России: вовлеченность и отношение к ним населения. М.: Издательский дом ГУ ВШЭ.

Паклин Н.Б., Орешков В.И. (2010). Бизнес-аналитика от данных к знаниям. Спб.: Питер.

Файнзильберг Л.С., Жук Т.Н. (2009). Гарантированная оценка эффективности диагностических тестов на основе усиленного ROC-анализа // *Управляющие системы и машины.* № 5. С. 3–13.

Hosmer D., Lemeshov S. (2000). Applied Logistic Regression. N.Y.: Wiley.

Mirkin B. (2011). Core Concepts in Data Analysis: Summarization, Correlation, Vizualization. L.: Springer.

Schlisterman E.F., Perkins N.J., Liu A. et al. (2005). Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples // *Epidemiology.* Vol. 16. P. 73–81.

Vapnik V.N. (1995). The Nature of Statistical Learning Theory. N.Y.: Springer-Verlag.

Zweig M.H., Campbell G. (1993). Receiver-Operating Characteristic (ROC) Plots: a Fundamental Evaluation Tool in Clinical Medicine // *Clinical Chemistry.* Vol. 39. № 4. P. 561–577.

Поступила в редакцию 01 сентября 2011 года

E.R. Goryainova
National Research University "Higher School of Economics", Moscow
T.I. Slepneva
Moscow Aviation Institute, Moscow

Binary Classification of Objects with Nominal Indicators

In this work a problem is studied of classification of respondents into classes accepting and not participation in a charity actions. An optimal (in Bayes sense) decisive discriminant rule of division of objects on two classes is constructed for the case when all indicators of observable objects are measured in a nominal scale, and there are signs of dependence between them . Using ROC-analysis methods, comparison of the developed rule with a rule implemented in the software package SPSS (Fisher's discriminant rule), «naive» Bayesian classifier, a rule based on support vector machines (SVM) method and implemented in SPSS package binary logistic regression classifier is made. Results of the ROC-analysis have shown that the proposed rule has higher quality than all other mentioned rules of classification of respondents.

Keywords: *discriminant analysis, solving rule, Bayes solution, Fisher's linear rule, binary logistic regression, support vector machines method, ROC-curve, AUC indicator.*

JEL classification: C38, Z13.