УДК 519.23

АНАЛИЗ МЕТОДОВ БИНАРНОЙ КЛАССИФИКАЦИИ

© 2014 Ю.С. Донцова

Ульяновский государственный технический университет

Поступила в редакцию 25.06.2014

Рассматривается задача бинарной классификации о принадлежности объекта, характеризующегося заданным вектором признаков, к одному из двух классов. Методами ROC-анализа проведен сравнительный анализ методов бинарной классификации, реализованных на Python(x, y).

Ключевые слова: *дискриминантный анализ, ROC-кривая, показатель AUC, Python(x, y)*

Задача бинарной классификации позволяет решить вопрос о принадлежности объекта к одному из двух классов. Пусть имеется конечное множество объектов $O=\{O_1 \ldots O_n\}$, каждый из которых характеризуется m-мерным вектором признаков ($X_1 \ldots X_m$). Признаки могут быть как числовыми, так и нечисловыми. При этом для некоторых представителей исходного множества известно, к какому классу они относятся. Данные объекты образуют обучающую выборку. Классовая принадлежность остальных объектов неизвестна. Необходимо построить алгоритм, способный определить для произвольного объекта из исходной выборки класс K_j, j=1,2 к которому следует отнести объект. Для решения сформулированной задачи могут быть использованы несколько методов классификации:

- «Случайный лес»
- Градиентный бустинг деревьев решений
- Наивный байесовский классификатор
- Дискриминантный анализ
- Логистическая регрессия

«Случайный лес». Данный метод заключается в использовании множества деревьев принятия решений, каждое из которых представляет собой алгоритм классификации объектов к тому или иному классу на основе независимых переменных (признаков) [2]. Эти деревья могут быть получены различными методами, по разным выборкам наблюдений за одним и тем же объектом, путем привлечения различных характеристик. Такое «многостороннее» рассмотрение задачи, как правило, приводит к лучшему пониманию закономерностей исследуемого объекта. На первом шаге из обучающей выборки генерируется случайная подвыборка с повторением размера n. На втором шаге строится дерево принятия решений, которое классифицирует объекты данной подвыборки. Причем признак, на основе которого происходит разбиение объектов, выбирается не из всех признаков, а только из случайно выбранных признаков (обычно $\sqrt{m}$). Выбор наилучшего признака осуществляется, как правило, с помощью энтропии либо с помощью индекса Джини.

Определение 1. Пусть имеется множество объектов $O=\{O_1 \ldots O_n\}$, r из которых обладают некоторым свойством S, принимающим l различных значений. Тогда энтропия множества O по отношению к свойству S определяется по формуле:

$$H(O,S) = -\sum_{i=1}^{l} \frac{r_i}{n} \log \frac{r_i}{n}.$$

Чем меньше значение энтропии, тем больше информации содержится в варианте разделения и тем лучше разделение объектов.

Определение 2. Для множества объектов $O=\{O_1 \ldots O_n\}$ и свойства S, принимающего l различ-ных значений индекс Джини вычисляется следующим образом:

$$Gini(O,S) = 1 - \sum_{i=1}^{l} \frac{|O_i|}{|O|}.$$

Меньшее значение этого показателя также соответствует лучшему разделению объектов.

Таким образом, дерево строится до полного исчерпания подвыборки. Поскольку в данном методе используется набор деревьев решений, каждое из которых по-своему классифицирует объект, необходимо найти общее решение. Таким образом, прибегают к методу «голосования» или «усреднению». При использовании первого метода объекту приписывается тот класс, которому отдает предпочтение большинство деревьев из набора. В случае задачи регрессионного анализа прогнозируемое значение получается усреднением прогнозов по всем деревьям.

Донцова Юлия Сергеевна, аспирантка. E-mail: ydoncova@yandex.ru

В качестве преимуществ данного метода можно отметить способность к обработке большего числа признаков и классов, нечувствительность к масштабированию, построению деревьев по данным с пропущенными значениями, а также одинаково хорошую обработку как непрерывных, так и дискретных признаков. К недостаткам можно отнести склонность к переобучению, то есть когда получается такой алгоритм, который слишком хорошо работает на примерах, участвовавших в обучении, но достаточно плохо работает на примерах, не участвовавших в обучении и большой размер получающихся моделей.

Градиентный бустинг деревьев решений. Термин «бустинг» от англ. «boosting» означает улучшение и представляет собой процедуру последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. С точки зрения классификации бустинг деревьев решений считается одним из наиболее эффективных методов. Данный алгоритм строит модель в виде суммы деревьев:

$$f(x) = h_0 + v\sum_{j=1}^{M} h_j(x),$$

где h_0 – начальное приближение, $v \in (0,1]$ - параметр, регулирующий скорость обучения и влияние отдельных деревьев на всю модель, $h_j(x)$ – деревья решений.

Новые слагаемые-деревья добавляются в сумму путем минимизации эмпирического риска, заданного некоторой функцией потерь:

$$L(y, y')=L(y, f(x))$$

Функция

$$L(y, y'_1, y'_2) = -\sum_{k=1}^{2}(y = k)\ln(\frac{\exp(y'_k)}{\sum_{i=1}^{2}\exp(y'_k)})$$

предназначена для решения задач классификации на *2* класса.

Байесовский классификатор. В основе байесовского подхода лежит принцип максимального использования имеющейся априорной информации о процессах, ее непрерывного пересмотра и переоценки с учетом получаемых выборочных данных [1]. Применительно к задаче классификации используют так называемые байесовские сети. Строго байесовской сетью называется пара: $N = \langle G, Q \rangle$ где G – ориентированный ацикличный граф, а Q – набор условных распределений. Каждая вершина графа соответствует одной из переменных ($X_1 \ldots X_m$), характеризующих исследуемый объект, для каждой из которых задано условное распределение:

$$Q_{X_i|\prod x_i} = P(X_i \mid \prod x_i)$$

где $\prod x_i$ – множество непосредственных предшественников X_i в графе G.

Совместное распределение при этом определяется по формуле:

$$P(X_1,...,X_m) = \prod_{i=1}^{m} P(X_i \mid \prod x_i)$$

.

Также байесовские сети позволяют оценить условные вероятности $P(Xj|Xi)$ остальных переменных. При классификации объектов на классы граф G условно разделяется на две части: вершина K, соответствующая классу объекта, и все остальные вершины. В случае наивного байесовского классификатора из вершины K проведены стрелки во все входные переменные $X_1 \ldots X_m$, и других ребер у графа G нет (рис.1).

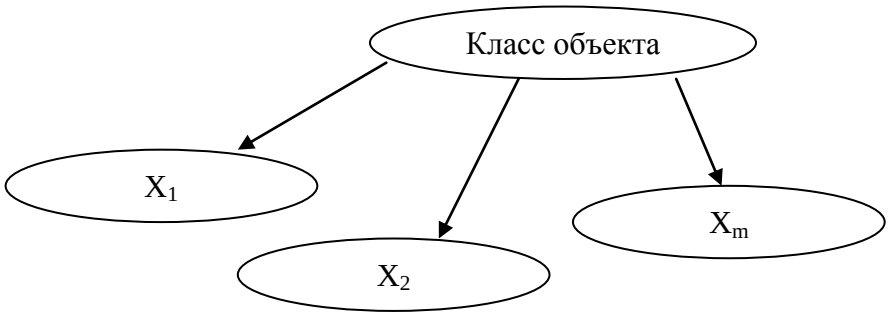


Рис. 1. Структура наивного байесовского классификатора

Обучение классификатора означает оценку условных вероятностей $P(X_i|K)$, а классификация производится с помощью формулы Байеса:

$$P(K=k \mid X=x)=\frac{P(K=k)\prod_{i=1}^{m}P(X_i=x_i \mid K=k)}{\sum_{k'}P(K=k')\prod_{i=1}^{m}P(X_i=x_i \mid K=k')}$$

Использование байесовских сетей позволяет избежать проблемы переучивания (overfitting), то есть избыточного усложнения модели, что является слабой стороной многих методов (например, деревьев решений и нейронных сетей). Также в модели определяются зависимости между всеми переменными, это позволяет легко обрабатывать ситуации, в которых значения некоторых переменных неизвестны. Однако перемножать условные вероятности корректно только тогда, когда все входные переменные действительно статистически независимы. Более того, невозможна непосредственная обработка непрерывных переменных – требуется их преобразование к интервальной шкале, чтобы атрибуты были дискретными

Дискриминантный анализ. Для определения вероятности принадлежности объекта к одному из двух классов используют линейные функции:

$$s_1(x)=q_0^1+q_1^1x_1+\ldots+q_m^1x_m$$
$$s_2(x)=q_0^2+q_1^2x_1+\ldots+q_m^2x_m,$$

где $q_1 \ldots q_m$ – параметры (веса) регрессии, которые находятся, как правило, с помощью метода наименьших квадратов, $s(x)$ – «счет», который содержит достаточное количество информации для того, чтобы различать класс объекта. При этом выбирается тот класс, которому соответствует больший счет.

Логистическая регрессия. Пусть y определяет принадлежность объекта к классу и принимает значение 1, если объект принадлежит классу K_1, и значение 0, если объект принадлежит классу K_2. Тогда делается предположение о том, что вероятность наступления события y=1, равна:

$$P\{y=1 \mid x\}=f(z)$$
$$z=q^Tx=q_0+q_1x_1+\ldots+q_mx_m,$$

где x – вектор-столбец независимых переменных $x_1 \ldots x_m$, q – вектор-столбец параметров (коэффициентов регрессии) $q_1 \ldots q_m$, $f(z)$ – логистическая функция (сигмоида, логит-функция):

$$f(z)=\frac{1}{1+e^{-z}}.$$

Так как y принимает лишь значения 0 и 1, то вероятность второго возможного значения равна:

$$P\{y=0 \mid x\}=1-f(z)=1-f(q^Tx).$$

Таким образом, логистическая регрессия принимает следующий вид:

$$\log\frac{P\{y=1 \mid x\}}{P\{y=0 \mid x\}}=\frac{f(z)}{1-f(z)}=q_0+q_1x_1+\ldots+q_mx_m.$$

Для нахождения параметров $q_1 \ldots q_m$ необходимо составить обучающую выборку, состоящую из наборов значений независимых переменных и соответствующих им значений зависимой переменной y. Формально, это множество пар:

$$(x^{(1)},y^{(1)}),\ldots,(x^{(m)},y^{(m)})$$

где $x^{(i)} \in R^n$ - вектор значений независимых переменных, а $y^{(i)} \in \{0,1\}$ - соответствующее им значение y. Каждая такая пара называется обучающим примером.

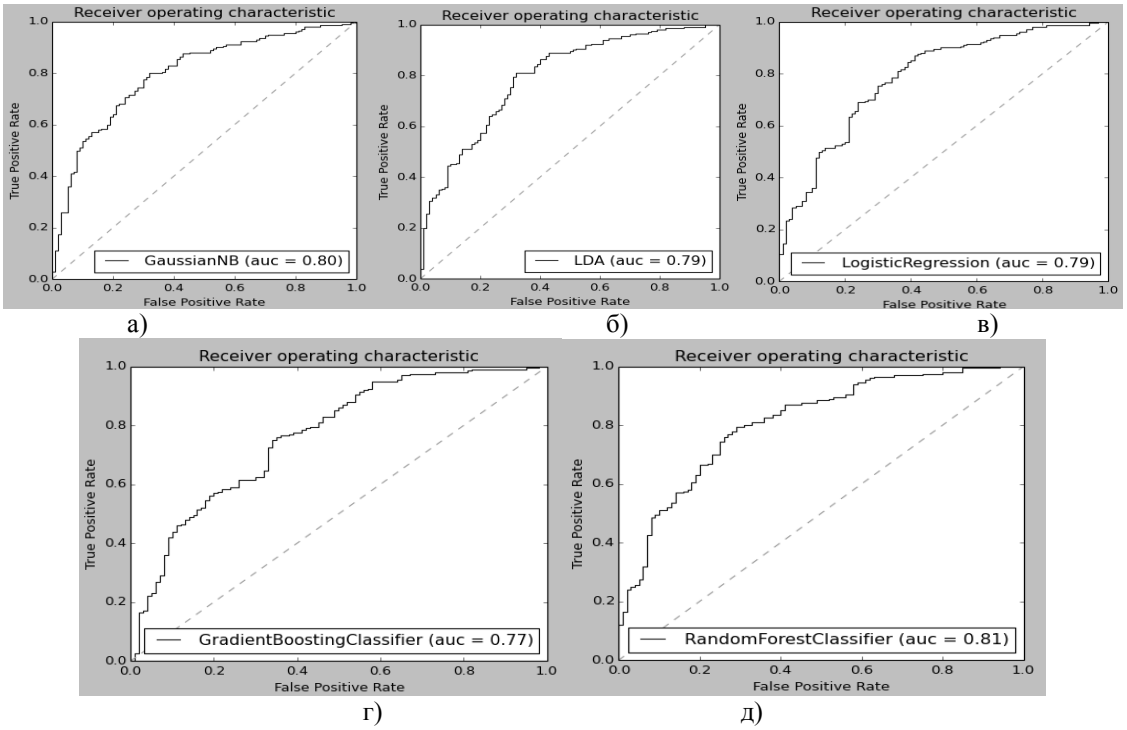
Обычно используется метод максимального правдоподобия, согласно которому выбираются параметры q, максимизирующие значение функции правдоподобия на обучающей выборке. Максимизация функции правдоподобия эквивалентна максимизации её логарифма. Для максимизации этой функции может быть применён метод градиентного спуска или метод Ньютона-Рафсона [3]. Сама задача классификации решается следующим образом: объект x можно отнести к классу y=1, если предсказанная моделью вероятность $P\{y=1 \mid 0{,}5\}>0{,}5$, и к классу y=0 в противном случае. Граничное значение может быть отлично от 0,5. Получающиеся при этом правила классификации являются линейными классификаторами.

Главным преимуществом логистической регрессии является учет ограничений на значения вероятности, которые не могут выходить за рамки 0 и 1, а также работа с входными данными любого рода, благодаря чему данная модель нашла свое применение при нахождении переходных вероятностей состояний системы [4]. К недостаткам следует отнести чувствительность к корреляции между факторами, поэтому в моделях недопустимо наличие сильно коррелированных независимых переменных.

Для оценки качества бинарной классификации прибегают в основном к методам ROC-анализа. ROC-кривая, также известная как кривая ошибок, отображает соотношение между долей верных положительных классификаций от общего числа положительных классификаций

(true positive rate) с долей ошибочных положительных классификаций от общего числа отрицательных классификаций (false positive rate) при варьировании порога решающего правила. Показатель AUC (площадь под ROC-кривой) дает количественную интерпретацию ROC-кривой. Считается, что чем выше показатель AUC, тем качественнее классификатор.

Для решения сформулированной задачи была использована среда программирования Python(x,y), представляющая собой набор библиотек и программного обеспечения для численных расчетов, анализа и визуализации данных на основе Python с готовой реализацией описанных выше моделей. Для сравнительного анализа результатов классификации была сформирована выборка, состоящая из 1000 объектов, каждый из которых характеризуется 20 признаками. Исходная выборка была разделена случайным образом на обучающую выборку (700 объектов) и на тестовую выборку (300 объектов). На рис. 2 изображены ROC-кривые моделей по данным контрольной выборки, а также показатель AUC для каждой модели.



а) б) в)

г) д)

Рис. 2. ROC-кривые: а) байесовский классификатор, б) дискриминантный анализ, в) логистическая регрессия, г) градиентный бустинг, д) «случайный лес»

Результаты ROC-анализа, в том числе и показателя AUC показывают, что, на первый взгляд, расхождения моделей совершенно незначительны и сложно определить качество классификации. Однако при представлении результатов классификации объектов в виде табл. 1, в которой отражено количество верно и ошибочно классифицируемых объектов, преимущество по точности классификации можно отдать методу дискриминантного анализа.

Таблица 1. Результаты классификации объектов различными моделями

Название модели	**Верно классифицированные**		**Ошибочно классифицированные**	
	кол-во (из 300)	%	кол-во (из 300)	%
«случайный лес»	207	69	93	31
градиентный бустинг деревьев решений	220	73,3	80	26,7
наивный байесовский классификатор	225	75	75	25
дискриминантный анализ	227	75,7	73	24,3
логистическая регрессия	222	74	78	26

СПИСОК ЛИТЕРАТУРЫ:

1. *Бидюк, П.И.* Построение и методы обучения байесовских сетей / *П.И. Бидюк, А.Н. Терентьев* // Информатика и кибернетика. 2004. № 2. С. 140-154.
2. *Breiman, W.* "Random Forests"/ Machine Learning. 45(1). 2001. P. 5-32.
3. *Васильев, Н.П.* Опыт расчета параметров логистической регрессии методом Ньютона-Рафсона для оценки зимостойкости растений / *Н.П. Васильев, А.А. Егоров* // Математическая биология и биоинформатика. 2011. Т. 6, № 2. С. 190-199.
4. *Клячкин, В.Н.* Сравнительный анализ точности нелинейных моделей при прогнозировании состояния системы на основе марковской цепи / *В.Н. Клячкин, Ю.С. Донцова* // Известия Самарского научного центра РАН. 2013. Т. 15, № 4(4). С. 924-927.

ANALYSIS OF BINARY CLASSIFICATION METHODS

© 2014 Y.S. Dontsova

Ulyanovsk State Technical University

It is considered the problem of binary classification of objects, described by the attributes vector, to any of two classes. It was performed a comparative analysis of binary classification methods using ROC-analysis method realized by means of Python(x, y) language.

Key words: *discriminant analysis, ROC-curve, AUC, Python(x, y)*

Yuliya Dontsova, Post-graduate Student.
E-mail: ydoncova@yandex.ru