1 Байесовский классификатор

Одним распространённых методов решения задач классификации является так называемый байесовский подход, при котором максимизируется апостериорная вероятность класса. В этом случае решающее правило имеет вид:

$$h(x) = \arg\max_{y \in Y} p(y \mid x).$$

Однако в большинстве случаев мы не можем найти вероятности $p(y \mid x)$, поскольку объект x вообще не встречается в обучающей выборке. По теореме Байеса решающее правило можно переписать в виде

$$h(x) = \arg\max_{y \in Y} p(y \mid x) = \arg\max_{y \in Y} \frac{p(x \mid y)p(y)}{p(x)} = \arg\max_{y \in Y} p(x \mid y)p(y).$$

Найти приближение для условной плотности $p(x \mid y)$ уже значительно проще, однако в случае, если пространство X имеет большую размерность также весьма сложна. Пусть объект x описывается n признаковыми функциями $f_1(x), \ldots, f_n(x)$, значения которых равны $x_1, \ldots, x_n$ (мы отождествляем объект и его признаковое описание), предположим, что значения данных функций являются независимыми случайными величинами. В этом случае мы приходим к решающему правилу

$$h(x) = \arg\max_{y \in Y} \prod_{i=1}^{n} p(f_i(x) = x_i \mid y)p(y).$$

Данный классификатор называется наивным байесовским, для его обучения требуется оценить априорные вероятности классов $p(y)$ и условные плотности $p(x_i \mid y)$. Пусть обучающая выборка X^l состоит из L объектов $x^1, \ldots, x^L$, каждый из которых имеет вид $x^j = [x_1^j, \ldots, x_n^j]$. Обозначим через y^j значение целевой функции на объекте x^j. Поскольку пространство $Y = \{y_1, \ldots, y_K\}$ в задачах классификации является конечным, априорная вероятность класса y_k можно оценить его частотой: $p(y_k) = \frac{|\{j | y^j = y_k\}|}{L}$.

Пусть каждый из признаков f_i принимает значения на конечном множестве X_i. В этом случае условные вероятности $p(f_i(x) = x_i \mid y_k)$ можно оценить на основе частот $c(x_i, y_k) = |\{j \mid y^j = y_k, f_i(x^j) = x_i\}$. Наиболее простой является формула

$$p(x_i \mid y_k) = \frac{c(x_i, y_k) + \alpha}{\sum\limits_{z \in X_i} (c(z, y_k) + \alpha)},$$

при которой условная вероятность $p(f_i(x) = x_i) \mid y_k)$ пропорциональна величине $c(x_i, y_k) + \alpha$, где α — некоторое положительное число, добавляемое, чтобы избежать нулевых значений вероятности. Конкретное значение α выбирается из дополнительных соображений, в дальнейшем мы полагаем $\alpha = 1$.

2 Байесовский классификатор в задачах классификации текстов

Наивный байесовский классификатор традиционно используется в задачах классификации текстов, таких как фильтрация спама, автоматическая рубрикация или определение тональности документа. Наиболее распространены две его разновидности: многомерный (multivariate) и мультиномиальный (multinomial) байесовский классификаторы. Зафиксируем словарь $V = \{v_1, \ldots, v_m\}$. В качестве признакового описания, соответствующего документу d будем брать набор $x = [x_1, \ldots, x_m] = [f_1(x), \ldots, f_m(x)]$, где

$$f_i(x) = \begin{cases} 1, & d \text{ содержит слово } v_i, \\ 0, & \text{иначе.} \end{cases}$$

Более кратко можно записать, что $f_i(x) = Ind(v_i \in d)$. Предположим, что обучающая выборка содержит L_k документов класса y_k, причём L_{ik} из них содержат слово v_i, в этом случае мы получаем следующие формулы для условных вероятностей:

$$\begin{array}{rcll} p(f_i(x) = 1 \mid y_k) & = & p_{ik} & = \frac{L_{ik}+1}{L_k+2}, \\ p(f_i(x) = 0 \mid y_k) & = & 1 - p_{ik} & \end{array}$$

Вероятность $p(f_i(x) = x_i \mid |y_k)$ более компактно переписывается в виде $p(f_i(x) = x_i \mid |y_k) = p_{ik}^{x_i}(1 - p_{ik})^{1-x_i}$, таким образом, классификатор принимает вид

$$\begin{array}{rcl} h(x) & = & y_{\widehat{k}}, \\ \widehat{k} & = & \arg\max\limits_{k=1}^{K} \prod\limits_{i=1}^{n} p_{ik}^{x_i}(1 - p_{ik})^{1-x_i} p(y_k), \end{array}$$

что после логарифмирования приводит к выражению

$$\begin{array}{rcl} h(x) & = & y_{\widehat{k}}, \\ \widehat{k} & = & \arg\max\limits_{k=1}^{K} \sum\limits_{i=1}^{n} ((\log p_{ik} - \log(1 - p_{ik}))x_i) + \sum\limits_{i=1}^{n} \log(1 - p_{ik}) + \log p(y_k). \end{array}$$

Недостатком многомерного байесовского классификатора является то, что он не учитывает количества вхождений слова в документ. Эта проблема решена в мультиномиальном байесовском классификаторе, где в качестве признакового описания для документа d берутся просто входящие в него слова, то есть $f_j(x) = w_j(d)$. где через $w_j(d)$ обозначено j-ое слово,входящее в документ d. Байесовское предположение о независимости признаков означает, что распределение вероятностей $p(w_j(d) = v_i \mid y_k)$ не зависит от позиции j и других слов, входящих в документ. Мы будем также считать, что длина документа не зависит от его класса, в противном случае длину документа необходимо добавить в его признаковое описание. Обозначим через c_{ik} число вхождений слова w_i в документы класса y_k, а через c_k — суммарное число слов во всех документах данного класса, тогда в качестве оценки для вероятности $p(w_j(d) = v_i \mid y_k)$, которую мы в дальнейшем будем обозначать через p_{ik}, разумно взять $p_{ik} = \frac{c_{ik}+1}{c_k+|V|}$.

Фактически, в мультиномиальной байесовской модели документ класса y_k представляет собой последовательность независимых реализаций случайной величины, принимающей конечное множество значений $v_1, \ldots, v_m$ с вероятностями $p_{1k}, \ldots, p_{mk}$. Тогда условная вероятность документа d задаётся формулой

$$p(x \mid y_k) = \prod_{j=1}^{m} p_{jk}^{n_j},$$

где через n_j обозначено число вхождений слова v_j в последовательность x, а решающее правило принимает вид

$$\begin{aligned} h(x) &= y_{\widehat{k}}, \\ \widehat{k} &= \arg\max_{k=1}^{K} \prod_{j=1}^{m} p_{jk}^{n_j} p(y_k). \end{aligned}$$

При логарифмировании получаем

$$\begin{aligned} h(x) &= y_{\widehat{k}}, \\ \widehat{k} &= \arg\max_{k=1}^{K} \sum_{j=1}^{m} n_j \log p_{jk} \log p(y_k). \end{aligned}$$

Таким образом, в качестве признакового описания можно взять величины $n_1, \ldots, n_m$, где n_j равно числу вхождений слова v_i в документ x. Заметим, что в обоих случаях байесовский классификатор можно рассматривать как частный случай линейного классификатора. Мультиномиальный байесовский классфиикатор, по сути, оценивает вероятность документа в соответствие с униграммной языковой моделью. В практических задачах основную проблему представляет правильный подбор словаря V. Чаще всего в словарь не включают наиболее частотные слова (например, служебные), поскольку их частота не зависит от класса. В языках с развитой морфологией в качестве элементов словаря обычно берутся не сами слова, а их леммы, что позволяет существенно уменьшить размер словаря и тем самым упростить вычисления.

Пример 1. Пусть $Y = \{\text{China}, \text{Japan}\}$ (то есть решается задача тематической классификации с двумя классами), а документы выборки содержат следующие слова из словаря и относятся к следующим классам:

Chinese Tokyo Macao	China
Chinese Chinese Beijing	China
Chinese Chinese Shanghai	China
Chinese Tokyo Japan	Japan
Tokyo Japan Tokyo	Japan

Требуется найти, к какому классу отнести документ x, содержащий слова Chinese Chinese Chinese Tokyo Japan.

Доказательство. Вначале рассмотрим многомерную байесовскую модель. Априорные вероятности классов равны $p(China) = 0.6$, $p(Japan) = 0.4$.В этом случае вероятности $p(f_j(x) \mid y)$ задаются таблицей (в ячейке таблицы содержится вероятность, что данное слово содержится в документе данного класса):

	China	Japan
Chinese	0.8	0.5
Macao	0.4	0.25
Beijing	0.4	0.25
Shanghai	0.4	0.25
Tokyo	0.4	0.75
Japan	0.2	0.75

Тогда апостериорные вероятности классов равны с точностью до постоянного множителя:

$$\begin{aligned} p(China \mid x) &= 0.8 \cdot (1-0.4) \cdot (1-0.4) \cdot (1-0.4) \cdot 0.4 \cdot 0.2 \cdot 0.6 = 0.0083, \\ p(Japan \mid x) &= 0.5 \cdot (1-0.25) \cdot (1-0.25) \cdot (1-0.25) \cdot 0.75 \cdot 0.75 \cdot 0.4 = 0.047, \end{aligned}$$

то есть данный документ должен быть отнесён к классу Japan. Если же мы используем мультиномиальную модель, то условные вероятности слов равны

	China	Japan
Chinese	$\frac{6}{15}$	$\frac{2}{12}$
Macao	$\frac{2}{15}$	$\frac{1}{12}$
Beijing	$\frac{2}{15}$	$\frac{1}{12}$
Shanghai	$\frac{2}{15}$	$\frac{1}{12}$
Tokyo	$\frac{2}{15}$	$\frac{4}{12}$
Japan	$\frac{1}{15}$	$\frac{3}{12}$

Тогда апостериорные вероятности классов равны с точностью до постоянного множителя:

$$\begin{aligned} p(China \mid x) &= (\tfrac{6}{15})^3 \cdot \tfrac{2}{15} \cdot \tfrac{1}{15} \cdot 0.6 = 0.00034, \\ p(Japan \mid x) &= (\tfrac{2}{12})^3 \cdot \tfrac{4}{12} \cdot \tfrac{3}{12} \cdot 0.4 = 0.00015. \end{aligned}$$

Таким образом, в данном случае применение различных вариантов байесовского классификатора приводит к различным ответам. □

3 Линейные классификаторы

Одной из основных и наиболее простых разновидностей алгоритмов классификации являются так называемые линейные алгоритмы классификации и задаваемые ими линейные классификаторы. Пусть $X \subset \mathbb{R}^n$ и решается задача классификации с классами, тогда линейным называется классификатор с решающим правилом вида

$$\begin{aligned} h(x) &= y_{\widehat{k}} \\ \widehat{k} &= \arg\max_{1 \leqslant k \leqslant K} (\alpha_k, x) + \beta_k, \end{aligned}$$

где $\alpha_k \in \mathbb{R}^n$, $\beta_k \in \mathbb{R}$, $k = 1, \ldots, K$. Разделяющая поверхность между классами для линейного классификатора является решением линейного уравнения, то есть гиперплоскостью в $\mathbb{R}^n$. В случае двух классов правило принимает вид

$$h(x) = \begin{cases} y_1, & (\alpha_1, x) + \beta_1 \geqslant (\alpha_2, x) + \beta_2, \\ y_2, & \text{иначе.} \end{cases}$$

Полагая без ограничения общности, что $y_1 = 1, y_2 = -1$, а также переобозначив $w = \alpha_1 - \alpha_2$, $w_0 = \beta_2 - \beta_1$, мы приходим к окончательному выражению

$$h(x) = \operatorname{sgn}((w, x) - w_0),$$

где w называется вектором весов, а w_0 — порогом активации. Фактически, линейный классификатор принимает решение в зависимости от того, с какой стороны от разделяющей гиперплоскости $(w, x) = w_0$ находится классифицируемый объект, а величина $(w, x) - w_0$ при этом пропорциональна расстоянию от объекта до этой гиперплоскости. Выборки, для которых можно подобрать подобную плоскость, называются линейно разделимыми.

Определение 1. Выборка $X^L = \{x^1, \ldots, x^l\}$, на которой известны значения целевой функции $Y^L = \{y^1, \ldots, y^L\}$ называется линейно разделимой с отступом $\delta > 0$, если найдётся вектор весов $w = [w_1, \ldots, w_n] \in \mathbb{R}^n$, удовлетворяющий условию $|w| = 1$, и порог активации w_0, что для всех $x^j \in X^L$ выполняются условия:

$$\begin{aligned} y_l = 1 &\quad \leftrightarrow \quad (w, x^l) - w_0) > \delta, \\ y_l = -1 &\quad \leftrightarrow \quad (w, x^l) - w_0 < -\delta. \end{aligned}$$

Условие $|w| = 1$ обусловлено тем, что умножив вектор весов и порог активации на некоторую константу, мы не изменим решающего правила, но поменяем значение отступа. Линейная разделимость означает, что мы не только можем верно классифицировать обучающую выборку некоторым линейным классификатором, но и можем при этом позволить себе ошибаться на δ в определении значения оценивающей функции. В других терминах это означает, что можно подобрать такую прямую, что при проекции на неё точки разных классов будут находиться по разные стороны от некоторой точки w_0, притом в некоторой окрестности этой точки вовсе не будет проекций точек выборки. Разумеется, не для всякой практической задачи выборка является линейно разделимой, однако во многих случаях разделимости можно добиться за счёт правильного подбора признаковых функций.

Зачастую удобно рассматривать линейные классификаторы только вида $h(x) = \operatorname{sgn}(w, x)$. Это не уменьшает общности, поскольку всегда добавить нулевую координату, которая равна w_0 в случае вектора весов и -1 в случае объектов выборки. Подобную операцию будем называть расширением признакового пространства, а полученное пространство — расширенным.

Упражнение 1. Доказать, что в случае конечной выборки условие на δ можно опустить.

Упражнение 2. Доказать, что для линейно разделимой выборки существует бесконечное число разделяющих поверхностей.

Упражнение 3. Доказать, что если конечная выборка линейно разделима, то можно подобрать вектор w' с рациональными координатами и рациональной число w'_0, такие что она будет разделяться и решающим правилом $g(x) = w'^T x - w'_0$.

4 Персептрон Розенблатта

Обучение линейного классификатора по выборке состоит в построении такого вектора весов и порога активации, что полученный линейный классификатор будет наилучшим образом разделять выборку. Одним из первых и простейших алгоритмов является персептрон Розенблатта, который в случае линейно разделимой выборки за конечное число шагов строит разделяющую гиперплоскость. Мы считаем, что объекты выборки заданы в расширенном признаковом пространстве. Отступом объекта x^l будем называть величину $m(x^l) = (w, x^l)y^l$; заметим, что $m(x^l) > 0$ тогда и только тогда, когда объект x^l правильно классифицируется. Персептроном Розенблатта называется следующий итеративный алгоритм:

Алгоритм 1 Алгоритм обучения персептрона Розенблатта.

Вход: Линейно разделимая обучающая выборка $X^L = \{x^1, \ldots, x^L\}$, вектор ответов $Y^L = \{y^1, \ldots, y^L\}$, начальный вектор весов $w_{(0)}$, шаг обучения $\eta > 0$.

Выход: Вектор весов w, такой что для всех l верно условие $(w, x^l)y^l > 0$.

1: $w \leftarrow w_{(0)}$
2: **while** Есть ошибки классификации **do**.
3: Найти вектор x_l, для которого $(w, x^l)y^l > 0$.
4: Положить $w \leftarrow w + \eta x^l y^l$.
5: **end while**

Рассмотрим, как меняется отступ x^l после модификации весов. Пусть $\widetilde{w} = w + \eta x^l y^l$ — обновлённое значение вектора весов, тогда $\widetilde{m}(x^l) = (w + \eta x^l y^l, x^l)y^l = m(x^l) + (\eta x^l y^l, x^l)y^l = m(x^l) + \eta(x^l, x^l) \geqslant m(x^l)$, то есть в случае ненулевого вектора x^l его отступ увеличится, то есть улучшится качество классификации на данном объекте. При этом качество классификации на остальных объектах в конечном счёте не ухудшится, о чём говорит следующая теорема.

Теорема 1. *Для линейно разделимых конечных выборок алгоритм 1 сходится за конечное число шагов независимо от значений шага обучения и начального вектора весов.*

Доказательство. Пусть δ — порог линейной разделимости, $\widehat{w}$ — вектор весов в соответствующем классификаторе, $w_{(t)}$ — вектор весов после t итераций алгоритма, а R — константа, ограничивающаяся длину вектора в выборке X^L, такая константа найдётся в силу конечности выборки.

Запишем скалярное произведение $(\widehat{w}, w_{(t)})$ и выразим его через произведение $(\widehat{w}, w_{(t-1)})$, получим

$$(\widehat{w}, w_{(t)}) = (\widehat{w}, w_{(t-1)} + \eta x^l y^l) = (\widehat{w}, w_{(t-1)}) + \eta y^l (\widehat{w}, x^l) \geqslant (\widehat{w}, w_{(t-1)}) + \eta\delta \geqslant (\widehat{w}, w_{(0)}) + \eta\delta t.$$

Аналогично поступим с произведением $(w_{(t)}, w_{(t)})$, но в данном случае проведём оценку

сверху.

$$(w_{(t)}, w_{(t)}) = (w_{(t-1)} + \eta x^l y^l, w_{(t-1)} + \eta x^l y^l)$$
$$= (w_{(t-1)}, w_{(t-1)}) + \eta^2 (y^l)^2 (x^l, x^l) + 2\eta y^l (w_{(t-1)}, x^l) \leqslant (w_{(t-1)}, w_{(t-1)}) + \eta^2 R^2$$
$$\leqslant t\eta^2 R^2 + (w_{(0)}, w_{(0)}).$$

По неравенству Коши-Буняковского получаем $(\widehat{w}, w_{(t)})^2 \leqslant (|\widehat{w}||w_{(t)}|)^2 = |\widehat{w}|^2 (w_{(t)}, w_{(t)})$. Отсюда по доказанным оценкам получаем

$$((\widehat{w}, w_{(0)}) + \eta\delta t)^2 \leqslant t\eta^2 R^2 + (w_{(0)}, w_{(0)})$$

Поскольку при больших t эта оценка заведомо неверна, получим, что алгоритм может отработать только конечное число итераций, то есть на каком-то шаге мы не сможем найти неправильно классифицируемый вектор. Теорема доказана. □

Пример 2. Пусть элементы выборки разделены по классам следующим образом:

$$+1: \quad [2,1], [1,-1], [0,0]$$
$$-1: \quad [-1,1], [-2,-1].$$

Требуется найти разделяющую плоскость с помощью персептрона Розенблатта.

Доказательство. Перейдём к расширенному признаковому пространству, будем через $\bar{x}^l$ обозначать вектор выборки x^l с добавленной нулевой координатой, равной -1. Возьмём в качестве начального приближения $w_{(0)} = [0,1,2]$, тогда $m_{(0)}(\bar{x}^2) = (w_{(0)}, \bar{x}^2) y^2 = -1 < 0$, поэтому положим $w_{(1)} = w_{(0)} + \eta \bar{x}^2 y^2 = [-1,2,1]$. Тогда $m_{(1)}(\bar{x}^4) = 0$, положим $w_{(2)} = w_{(1)} + \eta \bar{x}^4 y^4 = [0,3,0]$. Теперь $m_{(2)}(\bar{x}^3) = 0$, поэтому возьмём $w_{(3)} = w_{(2)} + \eta \bar{x}^3 y^3 = [-1,3,0]$. Нетрудно поверить, что полученный классификатор $h(x) = \operatorname{sgn}((w_{(3)}, x)) = \operatorname{sgn}(3x_1 + 1)$ верно классифицирует выборку X^L. □

Недостатком полученного алгоритма является требование на линейную разделимость: в случае её отсутствия он может расходиться. При этом заранее проверить линейную разделимость нельзя, так как это можно сделать только построив соответствующий линейный классификатор. Поэтому на практике используют более мощные алгоритмы, такие как машина опорных векторов (support vector machine или SVM).

Список литературы

[1] К. В. Воронцов. Математические методы обучения по прецедентам. Курс лекций. http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

[2] A. McCallum, K. Nigam. A comparison of event models for naive bayes text classification //AAAI-98 workshop on learning for text categorization. – 1998. – Т. 752. – С. 41-48.

[3] К. Мэннинг, П. Рагхаван, Х. Шютце Введение в информационный поиск //М.: Вильямс. – 2011.