

ОЦЕНКА ПЛОТНОСТИ СОВМЕСТНОГО РАСПРЕДЕЛЕНИЯ

А. П. МОТРЕНКО

Аннотация. В задачах классификации часто возникает ситуация, когда часть переменных распределена непрерывно, а часть — дискретно. Например, в логистической регрессии признаки непрерывны, а переменная отклика подчиняется распределению Бернулли. В работе описан способ оценки плотности совместного неоднородного распределения, включающего дискретные и непрерывные величины. Рассмотрен случай, когда вероятностные предположения о распределении случайных величин сделать не удается. В этом случае применяются методы ядерного сглаживания. В работе также приводится их сравнение с классическими методами теории вероятностей. Эксперимент проводится на реальных и синтетических данных.

Ключевые слова: плотность совместного распределения, смешанное распределение, ядерное сглаживание, порождающие алгоритмы классификации.

1. ВВЕДЕНИЕ

В задачах классификации требуется, по набору наблюдаемых величин, определить метку класса зависящей от них случайной величины. Алгоритмы классификации, включающие оценку плотности совместного распределения зависимых и независимых переменных, называются порождающими, так как с помощью восстановленной плотности совместного распределения можно породить пары зависимых и независимых переменных. Примерами порождающих алгоритмов являются наивный байесовский классификатор [1, 2], скрытые марковские цепи [3, 4].

В случае, когда наблюдаются реализации некоторой непрерывной случайной величины, а зависимая переменная подчиняется дискретному распределению, возникает задача оценки плотности смешанного совместного распределения, включающего дискретные и непрерывные случайные величины. При известных условных и маргинальных [5] плотностях рассматриваемых величин, плотность совместного распределения можно получить аналитически, воспользовавшись определением условной вероятности. Такой способ называется факторизацией, он рассмотрен в работах [6, 7]. В работах [8, 9] рассматривается оценка плотности совместного распределения с помощью копул. В этом случае не делается предположений об условном распределении зависимой переменной при заданном наборе наблюдаемых переменных. Достаточно знать одномерные плотности распределения зависимой и независимых переменных.

В данной работе особое внимание уделяется случаю, когда сделать какие-либо предположения об условной зависимости распределений или о виде копулы не удается.

Date: 2012 г.

Работа поддержана грантом РФФИ 12-07-31095.

В таком случае применяются методы непараметрического [10, 11] восстановления плотности. В данной работе методы ядерного сглаживания применяются для оценки совместного распределения дискретных и непрерывных случайных величин. Вычислительный эксперимент проводится на реальных и синтетических данных.

2. ФАКТОРИЗАЦИЯ

Рассмотрим вначале способ оценки плотности совместного распределения, основанный на разбиении ее на множители. Этот метод будет применяться в вычислительном эксперименте для получения эталонной оценки плотности распределения.

Функцию совместного распределения дискретной случайной величины $Z = (\mathbf{x}, y)$, где $y \in \{0, 1\}$ — дискретная случайная величина, $\mathbf{x} \in \mathbb{R}^n$ — вектор непрерывных независимых случайных величин, определим как

$$(1) \quad P(\mathbf{x}, y) = \sum_{t \leq y} \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} p(\mathbf{s}, t) d\mathbf{s},$$

где $p(\mathbf{x}, y)$ — плотность совместного распределения. Выражение для $p(\mathbf{x}, y)$ можно получить, зная плотность условного распределения одной из величин и маргинальную плотность другой величины:

$$(2) \quad p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|y)p(y).$$

Эта процедура называется факторизацией. Результат факторизации зависит от способа разбиения плотности $p(\mathbf{x}, y)$.

Рассмотрим выборку $D = \{(\mathbf{x}_i, y_i)\}$, $i = 1, \dots, m$, состоящую из m реализаций величины $Z = (\mathbf{x}, y)$.

Предположим, что математическое ожидание величины \mathbf{x} зависит от y , а распределение \mathbf{x} есть смесь гауссовских распределений:

$$p(\mathbf{x}|y) = \exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) - \frac{n}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma|\right), \text{ с вероятностью } p_y, y \in \{0, 1\}.$$

Пусть $p_1 = P$, $p_0 = 1 - P$. Тогда из второго равенства в (2) получаем

$$p(\mathbf{x}) = Pp(\mathbf{x}|1) + (1 - P)p(\mathbf{x}|0).$$

Рассмотрим отношение

$$(3) \quad \frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{Pp(\mathbf{x}|1)}{(1 - P)p(\mathbf{x}|0)} = \frac{P}{1 - P} \exp(c - \mathbf{w}_1^T \mathbf{x} - \mathbf{x}^T \mathbf{w}_0),$$

где параметры c , \mathbf{w}_0 и \mathbf{w}_1 выражаются через параметры нормального распределения следующим образом:

$$\begin{aligned} c &= \frac{1}{2} (\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_1), \\ \mathbf{w}_1 &= \frac{1}{2} (\boldsymbol{\mu}_1^T \Sigma^{-1} - \boldsymbol{\mu}_0^T \Sigma^{-1}), \\ \mathbf{w}_2 &= \frac{1}{2} (\Sigma^{-1} \boldsymbol{\mu}_1 - \Sigma^{-1} \boldsymbol{\mu}_0). \end{aligned}$$

Так матрица Σ симметрична, выражение (3) принимает вид:

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{P}{1-P} \exp(c - \mathbf{w}^T \mathbf{x}), \quad \text{где } \mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2.$$

Учитывая, что $p(1|\mathbf{x}) + p(0|\mathbf{x}) = 1$, получаем

$$p(y|\mathbf{x}) = \left(1 + \left(\frac{1-P}{P} \right)^{2y-1} \exp(-(2y-1)\mathbf{w}^T \mathbf{x} + c) \right)^{-1}, \quad y \in \{0, 1\}.$$

Тогда, воспользовавшись первым равенством в (2), получаем выражение для совместной плотности смешанного распределения:

$$(4) \quad p(\mathbf{x}, y) = p(y|\mathbf{x})(Pp(\mathbf{x}|1) + (1-P)p(\mathbf{x}|0)).$$

Эта плотность выведена из определения условной плотности. Таким образом, если сделанные о характере условных распределений предположения верны, то значение $p(\mathbf{x}, y)$, вычисленное по (4) есть истинное значение плотности в точке (\mathbf{x}, y) . Поэтому в вычислительном эксперименте будем рассматривать оценку (4) как наиболее приближенную к истинному распределению.

3. НЕПАРАМЕТРИЧЕСКАЯ ОЦЕНКА ПЛОТНОСТИ СОВМЕСТНОГО РАСПРЕДЕЛЕНИЯ ДИСКРЕТНОЙ И НЕПРЕРЫВНОЙ СЛУЧАЙНЫХ ВЕЛИЧИН

Может возникнуть ситуация, когда предположения о распределении случайных величин отсутствуют, либо по каким-либо причинам не могут быть использованы. Например, при классификации малых выборок использование плотности совместного распределения зависимых величин и независимых величин вместо плотности условного распределения зависимых переменных может улучшить качество классификации. Однако для этого нужно, чтобы плотность была вычислена без использования предположений об условных распределениях зависимых и независимых величин.

При непараметрическом оценивании плотности случайной величины Z в некоторой точке z , производится усреднение частоты появления в выборке ближайших к ней точек выборки z_i . При этом используются ядерные функции, убывающие с увеличением расстояния между z и Z_i . Обозначим ядерную функцию $K(u)$, тогда оценка плотности величины Z в точке z может быть получена с помощью:

$$(5) \quad \hat{p}(z) = \frac{1}{m} \sum_{i=1}^m K(z - z_i).$$

Определим ядерную функцию смешаной случайной величины $Z = (\mathbf{x}, y)$ в точке $z = (\mathbf{s}, t)$ как произведение дискретного и непрерывного ядер:

$$K_{h,\lambda}(z - z_i) = L_\lambda(t - y_i) C_h \left(\frac{\mathbf{s} - \mathbf{x}_i}{h} \right).$$

Ядерная функция для дискретной переменной имеет вид

$$L_\lambda(u) = \begin{cases} \lambda, & u = 0, \\ 1 - \lambda, & u \neq 0. \end{cases}$$

Ядерная функция для вектора непрерывных переменных также определяется как произведение одномерных ядер

$$C_h(\mathbf{u}) = \frac{1}{h^m} \prod_{j=1}^n c(u_j).$$

В качестве $c(u)$ выберем ядро Епанечникова:

$$c(u) = \frac{3}{4}(1 - u^2)[|u| < 1],$$

где $[|u| < 1]$ – индикаторная функция. Ядро Епанечникова не учитывает влияние точек, отстоящих от \mathbf{s} дальше чем на h и удовлетворяет условию

$$\int_{-\infty}^{\infty} c(u) du = 1.$$

Кроме того, оно минимизирует среднеквадратичную ошибку аппроксимации.

4. ВЫБОР ПАРАМЕТРОВ СГЛАЖИВАНИЯ

Рассмотрим интегральное среднеквадратичное отклонение (MISE) полученной оценки $\hat{p}(z)$ от истинной плотности распределения $p(z)$. Подразумевая под $\int dz$ суммирование по дискретным переменным и интегрирование по непрерывным, запишем выражение для MISE

$$(6) \quad E_{\text{MISE}} = \int (\hat{p}(z) - p(z))^2 dz = \int \hat{p}^2(z) dz - 2 \int \hat{p}(z)p(z) dz + \int p^2(z) dz.$$

Второй член суммы в правой части (6) есть математическое ожидание величины $\hat{p}(Z)$, его можно оценить как

$$\int \hat{p}(z)p(z) dz \approx \frac{1}{m} \sum_{i=1}^m \hat{p}_{\text{LOO}}(z_i) = \frac{1}{n(n-1)} \sum_{i_1=1}^m \sum_{i_2=1, i_2 \neq i_1}^m K_{h,\lambda}(z_{i_1} - z_{i_2}).$$

Здесь использована оценка скользящего контроля

$$\hat{p}_{\text{LOO}}(z_{i_1}) = \frac{1}{n-1} \sum_{i_2=1, i_2 \neq i_1}^m K_{h,\lambda}(z_{i_1} - z_{i_2}).$$

Последний член в (6) можно опустить, так как он не зависит от выбора ширины окна h и параметра сглаживания λ . Получаем:

$$(7) \quad \tilde{E}_{\text{MISE}} = \frac{1}{n^2} \sum_{i_1=1}^m \sum_{i_2=1}^m K_{h,\lambda}^{(2)}(z_{i_1}, z_{i_2}) - \frac{2}{n(n-1)} \sum_{i_1=1}^m \sum_{i_2=1, i_2 \neq i_1}^m K_{h,\lambda}(z_{i_1} - z_{i_2}),$$

где $K_{h,\lambda}^{(2)}(z_{i_1}, z_{i_2}) = L_\lambda^{(2)}(y_{i_1}, y_{i_2})C_h^{(2)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$, а выражения для $L_\lambda^{(2)}(y_{i_1}, y_{i_2})$ и $C_\lambda^{(2)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$ определяются как

$$L_\lambda^{(2)}(y_{i_1}, y_{i_2}) = \sum_{y \in D} L_\lambda(y - y_{i_1})L_\lambda(y - y_{i_2}),$$

$$C_\lambda^{(2)}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sum_{\mathbf{x} \in D} C_\lambda(\mathbf{x} - \mathbf{x}_{i_1})C_\lambda(\mathbf{x} - \mathbf{x}_{i_2}).$$

Метод оценки параметров λ и h , основанный на минимизации выражения (6), называется кросспроверкой.



Рис. 1. Значения \tilde{E}_{MISE} в зависимости от λ и h при больших значениях h (а) и при различных значениях h в логарифмическом масштабе (б).

5. ОБ ОПТИМАЛЬНОСТИ ЯДРА ЕПАНЕЧНИКОВА

Точность оценки функции $p(z)$ зависит не только от выбора ширины окна h и параметра сглаживания λ , но и от выбора ядерной функции $K(u)$. Рассмотрим поведение среднеквадратичной ошибки оценки $\hat{p}(z)$ как функции ядра K . В работе [10] показано, что для минимизации среднеквадратичного отклонения по ядру K необходимо произведение

$$I_{\text{MSE}}(K) = \left(\int K(u)^2 du \right)^2 \int u^2 K(u) du.$$

При этом ядерная функция должна удовлетворять следующим требованиям:

$$(8) \quad \int K(u) du = 1,$$

$$(9) \quad K(u) = K(-u),$$

$$(10) \quad \int u^2 K(u) du = 1.$$

Приходим к задаче минимизации $\int K(u)^2 du$ при ограничениях (8), (9), (10). Запишем лагранжиан этой задачи:

$$L(K, \mu_1, \mu_2) = \int K(u)^2 du + \mu_1 \left(\int K(u) du - 1 \right) + \mu_2 \left(\int u^2 K(u) du - 1 \right).$$

В точке экстремума K_0 его вариация должна равняться нулю. Обозначив через $\Delta K = K - K_0$ малое отклонение от экстремальной функции, получим:

$$2 \int \Delta K(u)(2K(u) + \mu_1 + \mu_2 u^2) du = 0,$$

и, следовательно

$$2K(u) + \mu_1 + \mu_2 u^2 = 0.$$

Заметим, что $K(u)$ обращается в ноль при $u = \pm \left(-\frac{\mu_1}{\mu_2}\right)^{1/2}$. Выбрав ядро с носителем $|u| < \left(-\frac{\mu_1}{\mu_2}\right)^{1/2}$, и учитывая условия (8), (9), (10), получаем ядро перенормированное Епанечникова

$$K(u) = \frac{3}{4 \cdot \sqrt{15}} \left(1 - \frac{u^2}{15}\right) \left[\frac{|u|}{\sqrt{15}} < 1\right].$$

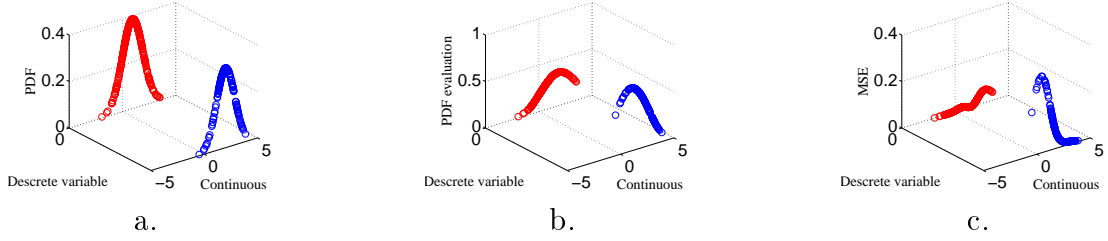


РИС. 2. а. Распределение тестовых данных $p(\mathbf{x}, y)$, восстановленное факторизацией. Так как условные распределения известны, полученная оценка совпадает с истинным совместным распределением. б. Непараметрическая оценка функции $\hat{p}(\mathbf{x}, y)$. с. Зависимость $E_{\text{MSE}}(\mathbf{x}, y)$, среднеквадратичного отклонения \hat{p} от p .

6. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

В вычислительном эксперименте рассмотрена выборка синтетических данных $D = \{z_i\} = \{(\mathbf{x}_i, y_i)\}$, $i = 1, \dots, m$, порожденная таким образом, что:

$$(11) \quad p(y|\mathbf{x}) = \theta^y(\mathbf{x})(1 - \theta(\mathbf{x}))^{1-y},$$

$$(12) \quad p(\mathbf{x}|y) = \mathcal{N}(\boldsymbol{\mu}_y, \sigma_y^2 I).$$

Будем рассматривать функции плотности $\tilde{p}(z)$ и $\hat{p}(z)$, восстановленные с помощью (4) и (5). Графики этих функций изображены на рисунке. Для оценки точности восстановления плотности будем использовать ошибку E_{MSE} :

$$(13) \quad \tilde{E}_{\text{MSE}}(z_i) = (\hat{p}(z_i) - \tilde{p}(z_i))^2.$$

Оценки функции $p(z)$, полученные различными способами, будем сравнивать, ис-



Рис. 3. а. Зависимость расстояния Кульбака-Лейблера $D_{KL}(p|\hat{p})$ от параметра сглаживания h . б. Зависимость расстояния Кульбака-Лейблера $D_{KL}(p|\hat{p})$ от параметра λ при оптимальном значении h .

пользуя расстояние Кульбака-Лейблера:

$$(14) \quad D_{KL}(p, \hat{p}) = \int_z p(z) \log \frac{p(z)}{\hat{p}(z)} dz = \mathbb{E} \left[\frac{p(Z)}{\hat{p}(Z)} \right].$$

На рисунке 6 изображена зависимость $D_{KL}(\tilde{p}|\hat{p})$ от параметров h и λ .

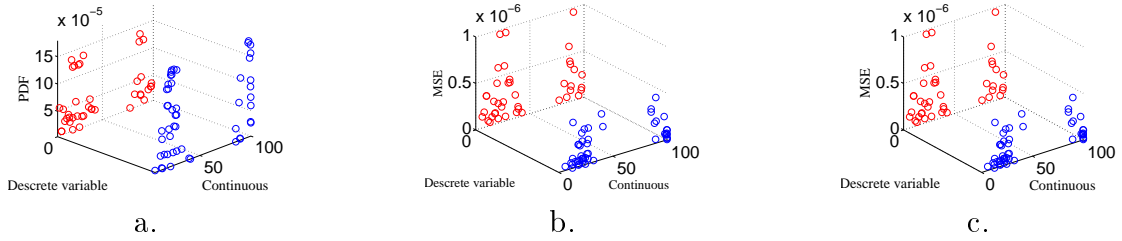


Рис. 4. а. Оценка совместного распределения реальных данных $\tilde{p}(\mathbf{x}, y)$, полученная с помощью факторизации. В отличие от случая с известными условными распределениями, эта оценка может не совпадать с истинной плотностью совместного распределения. Ее точность зависит от точности вероятностных предположений. б. Непараметрическая оценка функции $\hat{p}(\mathbf{x}, y)$. с. Зависимость $E_{MSE}(\mathbf{x}, y)$, среднеквадратичного отклонения \hat{p} от p .

Рассмотрим также выборку реальных данных. Будем оценивать $\tilde{p}(z)$, предполагая (11) и (12). Оценки функций $\tilde{p}(z)$, $\hat{p}(z)$ и E_{MSE} изображены на 6.

Также можно рассмотреть зависимость $D_{KL}(\tilde{p}, \hat{p})$ от параметров λ , h .

7. ЗАКЛЮЧЕНИЕ

В работе рассматривается задача оценки плотности совместного неоднородного распределения. Выборка состоит их дискретных и непрерывных величин случайных, и имеет малый объем, в следствие чего не удастся сделать предположений об условных

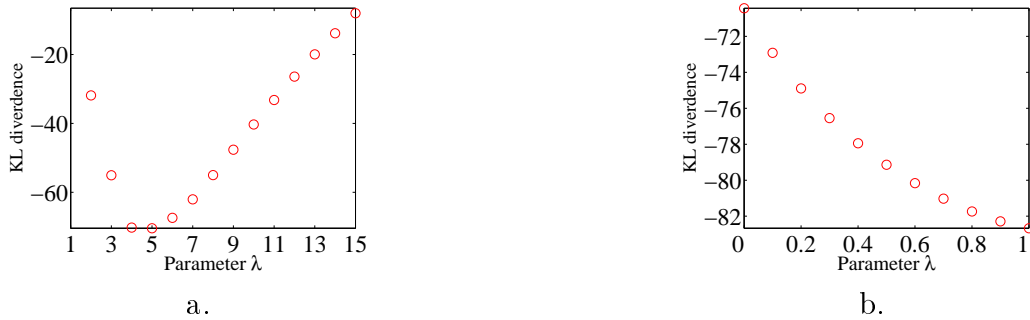


Рис. 5. а. Зависимость расстояния Кульбака-Лейблера $D_{KL}(\tilde{p}|\hat{p})$ от параметра сглаживания h . б. Зависимость расстояния Кульбака-Лейблера $D_{KL}(\tilde{p}|\hat{p})$ от параметра λ при оптимальном значении h .

или маргинальных распределениях этих величин. Предложен способ оценки плотности совместного распределения, основанный на применении методов ядерного сглаживания. В частности, для непрерывных величин используется ядро Епанечникова, а для дискретных — сглаженная индикаторная функция. Для сравнения рассмотрен метод факторизации совместного распределения, то есть разбиение его на условное и маргинальное распределение подмножества случайных величин, образующих многомерную случайную величину, плотность которой необходимо оценить.

СПИСОК ЛИТЕРАТУРЫ

- [1] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2001.
- [2] Rennie J., Shih L., Teevan J., and Karger D. Tackling The Poor Assumptions of Naive Bayes Classifiers. Proceedings of the Twentieth International Conference on Machine Learning (ICML). 2003.
- [3] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77-2(1989), 257–285.
- [4] Stamp, M. A Revealing Introduction to Hidden Markov Models. San Jose State University, 2012.
- [5] Everitt, B. S. The Cambridge Dictionary of Statistics . Cambridge University Press, 2002.
- [6] Olkin, I. and Tate, R. F. Annals of Math. Statistics, 36-1(1965), 343-344.
- [7] McCulloch, C. E. Joint modeling of mixed outcome types using latent variables. Statistical Methods in Medical Research, 17(2008), 53–73.
- [8] Nelsen R. B. An introduction to copulas. Springer, 2006.
- [9] Charpentier A., Fermanian J.-D., Scaillet O. The Estimation of Copulas: Theory and Practice, 2006.
- [10] Хардле, В. Прикладная непараметрическая регрессия. Москва "Мир" 1993.
- [11] Li, Q., Racine, J. Nonparametric Estimation of Distributions with Categorical and Continuous Data, Journal of Multivariate Analysis, 86-2(2003), 266 - 292.

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ, ФУПМ, КАФ. "ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ"