Оценка плотности совместного распределения*

Мотренко А. П.
anastasia.motrenko@gmail.com
Московский физико-технический институт

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.*

Joint probability density estimation*

A. P. Motrenko
Moscow Institute of Physics and Technology

When solving a classification problem one often has to deal with both discrete and continuous variables. for example, in the logistic regression independent variables are distributed continuously, while a target variable follows Bernoulli distribution. In this paper a method is resented that allows to estimate joint probability distribution which include discrete and continuous variables. A case when no probabilistic assumptions can be made is considered. The methods of nonparametric regression are used. Also a comparison to the classic methods of probability theory is presented. The experiment is conducted on the real and synthetic data. **Keywords**:

joint distribution density, mixed distribution, nonparametric regression, generative classification algorithms..

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

Работа поддержана грантом РФФИ 12-07-31095.

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. 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.

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 смешаной 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

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

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:

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

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, и 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, воспользовавшись следующим свойством плотности совместного распределения:

$$\sum_y p(\mathbf{x}, y) = p(\mathbf{x}),$$

и вторым равенством в (2), получаем

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

Рассмотрим отношение

$$\frac{p(1|\mathbf{x})}{p(0|\mathbf{x})} = \frac{Pp(\mathbf{x}|1)}{(1 - P)p(\mathbf{x}|0)} = \frac{P}{1 - P} \exp\left(c - \mathbf{w}_1^T \mathbf{x} - \mathbf{x}^T \mathbf{w}_2\right), \tag{3}$$

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

$$c = \frac{1}{2}\left(\boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1\right),$$

$$\mathbf{w}_1 = \frac{1}{2}\left(\mu_{\mathbf{1}}{}^T \Sigma^{-1} - \mu_{\mathbf{0}}{}^T \Sigma^{-1}\right),$$

$$\mathbf{w}_2 = \frac{1}{2}\left(\Sigma^{-1}\mu_{\mathbf{1}} - \Sigma^{-1}\mu_{\mathbf{0}}\right).$$

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

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

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

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

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

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

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.

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:

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

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 $C_h(\mathbf{u})$ для вектора непрерывных переменных также определяется как произведение одномерных ядер. Пусть смешанная случайная величина Z включает в себя n непрерывных случайных величин, тогда

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

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

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

Кроме того, оно минимизирует среднеквадратичную ошибку аппроксимации.

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

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

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

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

Здесь использована оценка скользящего контроля

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

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

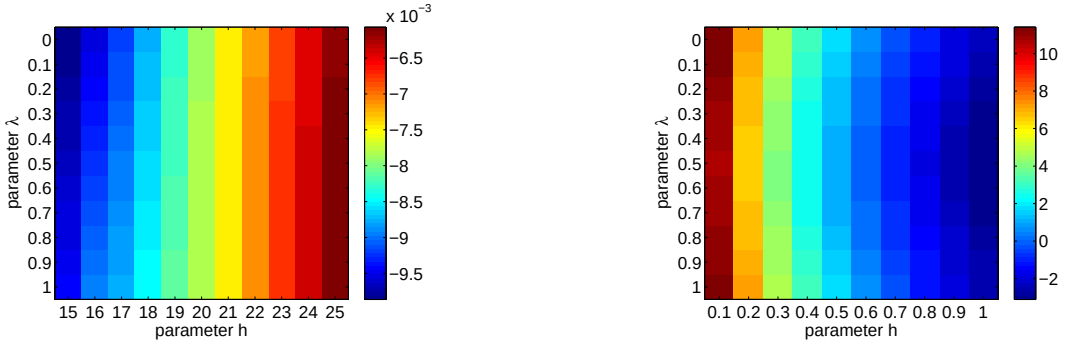
$$\tilde{E}_{\text{MISE}} = \frac{1}{m^2}\sum_{i_1=1}^{m}\sum_{i_2=1}^{m} K^{(2)}_{h,\lambda}(z_{i_1}, z_{i_2}) - \frac{2}{m(m-1)}\sum_{i_1=1}^{m}\sum_{i_2=1,\ i_2\neq i_1}^{m} K_{h,\lambda}(z_{i_1} - z_{i_2}), \tag{7}$$

где $K^{(2)}_{h,\lambda}(z_{i_1}, z_{i_2}) = L^{(2)}_{\lambda}(y_{i_1}, y_{i_2})C^{(2)}_{h}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$, а выражения для $L^{(2)}_{\lambda}(y_{i_1}, y_{i_2})$ и $C^{(2)}_{\lambda}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2})$ определяются как

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

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

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



a. b.

Рис. 1: Значения $\tilde{E}_{\text{MISE}}$ в зависимости от λ и h при больших значениях h (a) и при различных значениях h в логарифмическом масштабе (b).

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 минимизировать 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:

$$\int K(u)du = 1, \tag{8}$$

$$K(u) = K(-u), \tag{9}$$

$$\int u^2 K(u)du = 1. \tag{10}$$

Ограничение (10) снижает вес элементов с $u \approx 0$, препятствуя выбору чересчур узких функций. Учитывая (10), 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 вариация лагранжиана должна равняться нулю. Обозначив через $\Delta K = K - K_0$ малое отклонение от экстремальной функции, получим:

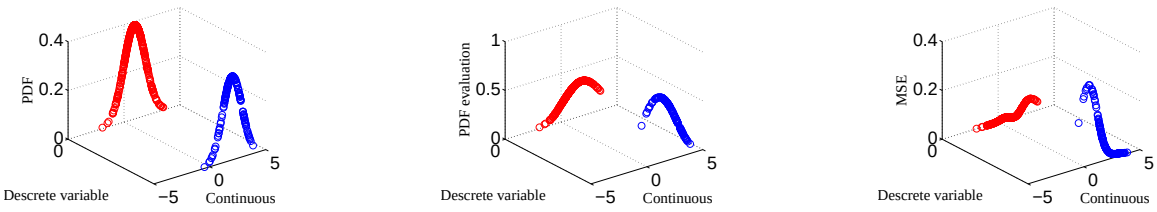
$$2 \int \Delta K(u)(2K(u) + \mu_1 + \mu_2 u^2) du = 0,$$

и, следовательно,

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

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

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



a. b. c.

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

Вычислительный эксперимент

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

$$p(y = 1) = P, \quad p(y = 0) = 1 - P, \tag{11}$$

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

Этот вид смешанной случайной величины описан в разделе 4. Заметим, что если предположения (11), (12) выполняются, то плотность распределения, полученная с помощью (4), есть истинная плотность. Таким образом, для синтетических данных истинная плотность совместного распределения известна. В случае тестовых данных, предположения могут быть слишком грубы, и восстановленная с помощью факторизиции плотность совместного распределения будет отличаться от истинной плотности.

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

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

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

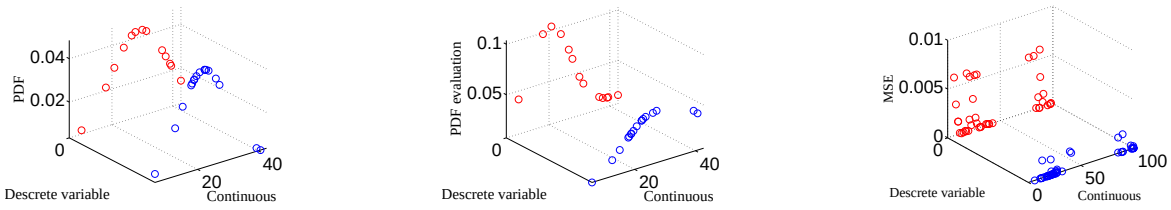
a. b.

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

расстояние Кульбака-Лейблера:

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

Рассмотрим зависимости расстояния между восстановленными плотностями $D_{KL}(\tilde{p}, \hat{p})$ от параметров сглаживания λ, h. Эти зависимости и результаты кросспроверки изображены на рисунке 4. Оба способа приводят к выбору одного значения параметра λ, но дают различные оценки параметра h. Будем использовать расстояние Кульбака-Лейблера, так как его проще вычислять.

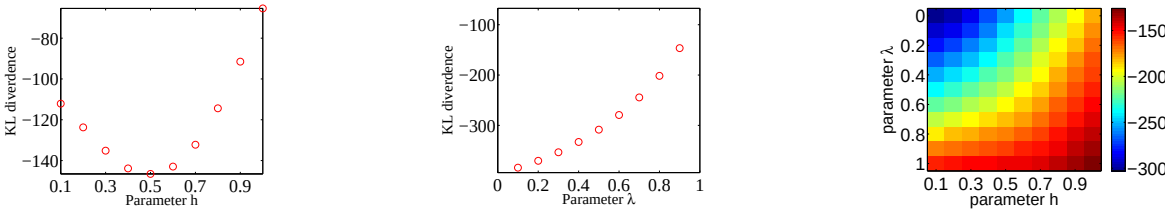


a. b. c.

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

Рассмотрим зависимости расстояния между восстановленными плотностями $D_{KL}(\tilde{p}, \hat{p})$ от параметров сглаживания λ, h. Эти зависимости и результаты кросспроверки изображены на рисунке 4. Оба способа приводят к выбору одного значения параметра λ, но дают

различные оценки параметра h. Заметим, расстояние Кульбака-Лейблера проще вычислять, однако его использование возможно только при известном истинном распределении.
.



a. b. c.

Рис. 5: a. Зависимость расстояния Кульбака-Лейблера $D_{KL}(\tilde{p}|\hat{p})$ от параметра сглаживания h. b. Зависимость расстояния Кульбака-Лейблера $D_{KL}(\tilde{p}|\hat{p})$ от параметра λ при оптимальном значении h. c. Зависимость $\tilde{E}_{\text{MISE}}$ от параметров h и λ.

Рассмотрим также выборку реальных данных. Будем оценивать $p(z)$, предполагая (11) и (12). Функции $\tilde{p}(z)$, $\hat{p}(z)$ и оценка E_{MSE} изображены на рисунке 4. Видно, что предположения оказались слишком сильны, условные распределения $p(\mathbf{x}|1)$ и $p(\mathbf{x}|0)$ не являются нормальными, и функция $\tilde{p}(z)$ далека от истинной функции распределения.

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 удается сделать предположений об условных или маргинальных распределениях этих величин. Предложен способ оценки плотности совместного распределения, основанный на применении методов ядерного сглаживании. 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.

Литература

[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. Annnals 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.