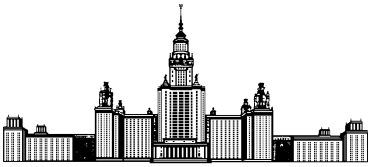
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Отчет по практикуму

Восстановление плотностей распределений

Выполнил:
студент 3 курса 317 группы
Ромов Петр Алексеевич

Москва, 2011

1 Постановка задачи

Задача восстановления плотности распределения формулируется следующим образом. Задано множество точек $X = \{x^{(1)}, \ldots, x^{(m)}\}$ — реализация однородной выборки из неизвестного распределения с плотностью $p(x)$, требуется по выборке X найти некоторое приближение плотности $\hat{p}(x) \approx p(x)$.

Восстановление плотности распределения необходимо для построения Байесовского классификаторы, а также полезно само по себе.

Целью данной практической работы является изучение и сравнение методов восстановления плотности:

1. метод парзеновского окна;
2. параметрическое восстановление плотности;
3. восстановление смеси распределений.

2 Теоретическое введение

2.1 Метод окон Парзена

Оценка плотности Парзена-Розенблатта в одномерном случае имеет вид:

$$\hat{p}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x^{(i)}}{h}\right), \tag{1}$$

где $K(r)$ — ядро (чётная, нормированная функция).

Эту оценку можно обобщить на многомерный случай:

$$\hat{p}_h(x) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{x_j - x_j^{(i)}}{h_j}\right). \tag{2}$$

В [4] приведено обобщение на многомерный случай, использующее функцию расстояния между объектами $\rho(x, x^{(i)})$. Это обобщение удобно в случае, когда представления о природе данных подсказывают верную метрику в пространстве векторов или когда нужно восстановить плотность в пространстве объектов, не представимых числовыми векторами. Трудность подхода с функцией расстояния заключается в необходимости вычисления нормирующего множителя $V(h) = \int K(\frac{\rho(x,x^{(i)})}{h})dx$. В данной работе это обобщение рассмотрено не будет.

Ширина окна h и вид ядра K — структурные параметры метода, от которых так или иначе зависит качество восстановления плотности.

2.2 Параметрические семейства распределений

Параметрическое оценивание опирается на семейства функций плотности, задающиеся при помощи одного или нескольких числовых параметров: $\{p(x;\theta),\ \theta \in \Theta\}$. Один из способов выбрать из этого семейства функцию плотности (наилучшим образом приближающую исходную) — метод максимума правдоподобия:

$$\theta^* = \arg\max_{\theta \in \Theta} \prod_{i=1}^{m} \hat{p}(x^{(i)};\theta), \quad \hat{p}(x) = p(x;\theta^*).$$

Многомерное нормальное распределение. Плотность имеет вид:

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\},$$

Оценки максимального правдоподобия записываются явно:

$$\mu^* = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}, \quad \Sigma^* = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu^*)(x^{(i)}-\mu^*)^\top.$$

2.3 Модель смеси Гауссиан

Модель смеси распределений имеет вид:

$$p(x;\theta) = \sum_{k=1}^{k} w_k p_k(x;\theta_k), \quad \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0, \quad \theta = \{\theta_k, w_k\}_{k=1}^{K}.$$

Здесь $p_k(x;\theta_k)$ — некоторое параметризованное семейство плотностей (например рассмотренных выше). Применение метода максимума правдоподобия "в лоб" приводит к очень сложной задаче, поэтому для настройки параметров смеси используется EM-алгоритм.

Предположим, что элементы смеси имеют Гауссово распределение, т.е. $p_k(x;\theta_k) = \mathcal{N}(x;\mu_k,\Sigma_k)$, $\theta_k = \{\mu_k,\Sigma_k\}$. Тогда шаги EM-алгоритма имеют вид:

Инициализация: Инициализируются значения μ_k и Σ_k.

E-шаг: Используя текущие значения параметров, пересчитать значения γ_{nk}

$$\gamma_{nk} = \frac{w_k \mathcal{N}(x^{(n)};\mu_k,\Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(x^{(n)};\mu_j,\Sigma_j)}$$

M-шаг: Переоценить параметры, используя посчитанные γ_{nk}

$$\mu_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk} x^{(n)}$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(x^{(n)} - \mu_k^{\text{new}})(x^{(n)} - \mu_k^{\text{new}})^\top$$

$$w_k^{\text{new}} = \frac{N_k}{N}$$

На стадии инициализации нужно взять такие значения параметров, чтобы объекты обучающей выборки "подходили" под распределения элементов смеси, иначе значения γ_{nk} принимают очень близкие к нулю значения, что приводит к плохим оценкам матрицы ковариации Σ_k.

В данной работе использовался следующий способ инициализации: значения каждой компоненты берутся случайно из равномерного распределения на отрезке $[\bar{x}_i - \sigma_i, \bar{x}_i + \sigma_i]$, здесь $\bar{x}_i$ — среднее арифметическое i-той компоненты в выборке, σ_i — стандартное отклонение i-той компоненты от среднего значения; начальное значение матрицы $\Sigma_k = I$ — единичная матрица.

Для того чтобы предотвратить вырождение матриц Σ_k (которое происходит из-за плохого начального приближения, а также данных, плохо описываемых смесью гауссиан), после пересчета на M-шаге их диагональ слегка увеличивается (производится регуляризация): $\Sigma_k := \Sigma_k + \varepsilon I$.

2.4 Меры "похожести" распределений

Для оценки качества восстановления плотности по данным, будут использованы два функционала:

$$J(\hat{p},p) = \int (\hat{p}(x) - p(x))^2\, dx,$$

$$\mathrm{KL}(\hat{p}\|p) = \int \hat{p}(x) \ln\left\{\frac{\hat{p}(x)}{p(x)}\right\} dx, \qquad \mathrm{KL}(p\|\hat{p}) = \int p(x) \ln\left\{\frac{p(x)}{\hat{p}(x)}\right\} dx.$$

В отличие от среднего квадратического отклонения $J(\hat{p},p)$, дивергенция Кульбака-Лейблера имеет вероятностное обоснование быть расстоянием между распределениями (подробнее см. [2]).

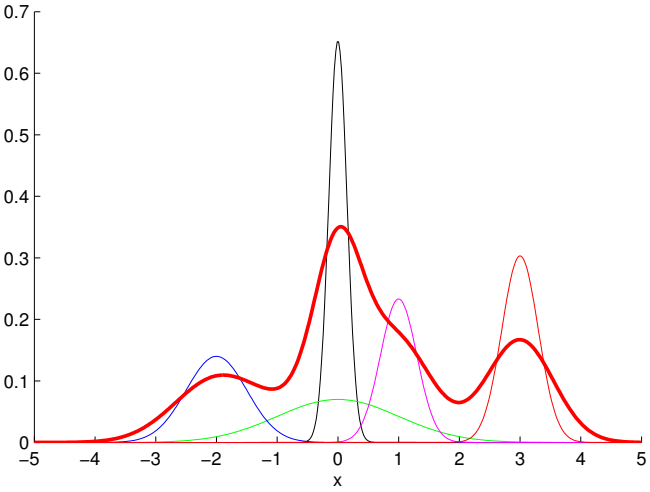


Рис. 1: Модельное распределение, заданное как смесь гауссиан (красная жирная кривая).

3 Эксперименты

3.1 Одномерная модельная задача и метод Парзена

Неизвестное распределение представляет собой смесь гауссиан (Рис. 1). Из этого распределения генерируется случайная выборка по которой строится плотность $\hat{p}(x)$ в форме (1). Во всех экспериментах, кроме эксперимента с различными ядровыми функциями используется ядро Гаусса. Во всех экспериментах, кроме демонстрирующего влияние ширины окна, параметр h настраивается автоматически: методом скользящего контроля оптимизируется правдоподобие выборки.

Ширина окна h. Эксперимент проводился на выборке длины 50. На графиках Рис. 2a–2c показана зависимость качества восстановления плотности от h, на Рис. 2d–2g показаны примеры восстановленной плотности при различных значениях ширины окна.

Оптимальная с точки зрения всех трех функционалов восстановленная плотность получается при ширине окна $h = 0.35$ (Рис. 2e). Из графиков видно, что если взять значение h больше, произойдет “пересглаживание” (Рис. 2f–2g).

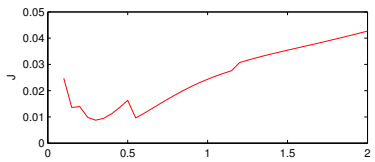
Оптимальная ширина окна зависит от количества элементов выборки (чем меньше элементов выборки, тем сильнее требуется сглаживать восстанавливаемую плотность) и их значений.

Ядровая функция. Эксперимент проводился на выборке размера 200. Для каждого ядра выбиралось оптимальное (по значению правдоподобия на скользящем контроле) значение ширины окна.

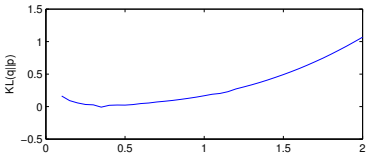
Рассматривались ядровые функции:

1. $E(r) = \frac{3}{4}(1 - r^2)[|r| \leq 1]$ — оптимальное (Епанечникова)
2. $Q(r) = \frac{15}{16}(1 - r^2)^2[|r| \leq 1]$ — квартичное
3. $T(r) = (1 - |r|)[r \leq 1]$ — треугольное
4. $\Pi(r) = \frac{1}{2}[|r| \leq 1]$ — прямоугольное
5. $G(r) = (2\pi)^{-1/2} \exp(-\frac{1}{2}r^2)$ — гауссовское

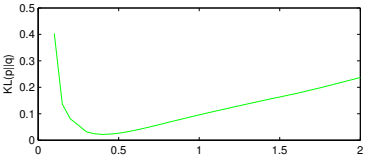
Графики восстановленных плотностей приведены на Рис. 3. В [4] утверждается, что ядровая функция не влияет существенно на восстановленную плотность. Действительно, визуально все графики похожи между собой. Плотности, восстановленные прямоугольным и треугольным ядрами выделяются: в силу их вида, $\hat{p}_\Pi(x)$ является кусочно-постоянной; а $\hat{p}_T(x)$ является кусочно-линейной.



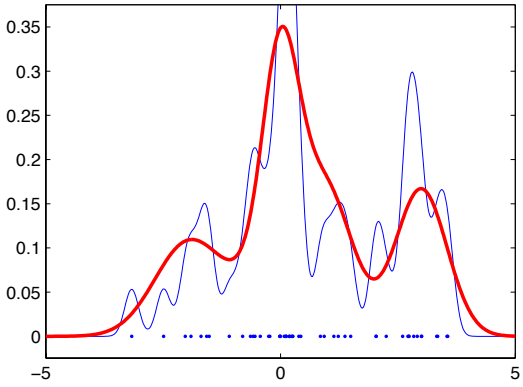
(a) $J(\hat{p}_h, p)$



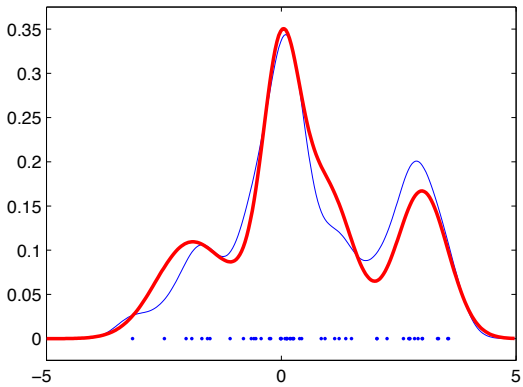
(b) $KL(\hat{p}_h \| p)$



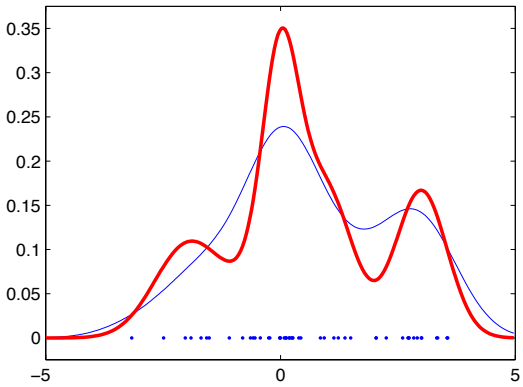
(c) $KL(p \| \hat{p}_h)$



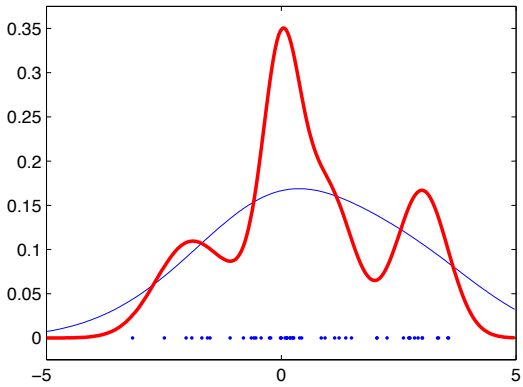
(d) $h = 0.15$



(e) $h = 0.35$

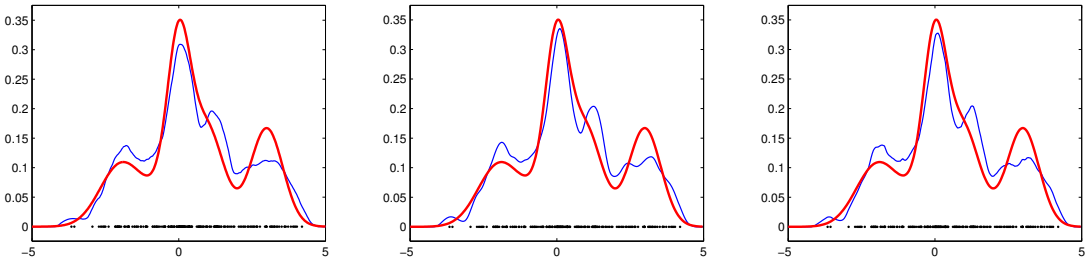


(f) $h = 0.35$

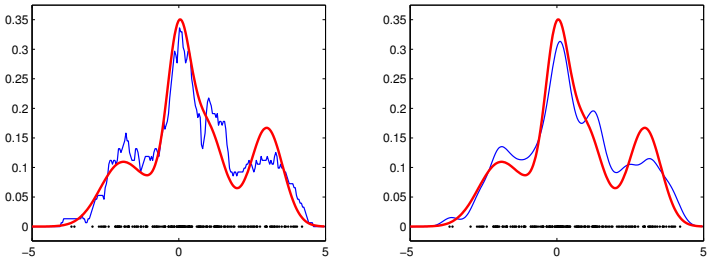


(g) $h = 0.35$

Рис. 2: Влияние ширины окна h в методе Парзена.



(a) оптимальное (b) квартичное (c) треугольное



(d) прямоугольное (e) гауссовское

Рис. 3: Использование различных ядровых функций в методе Парзена.

Число элементов выборки. Следующий эксперимент нужен для ответа на вопос, сколько нужно элементов в выборке для достаточно хорошего восстановления плотности распределения. Так как качество восстановления зависит не только от структурных параметров метода, но и от реализации выборки, для большей объективности результатов в рамках эксперимента каждому размеру выборки генерировались 10 реализаций и оценивалось среднее значение функционала качества и его отклонение от среднего. Графики на Рис. 4 показывают зависимость качества от размера выборки.

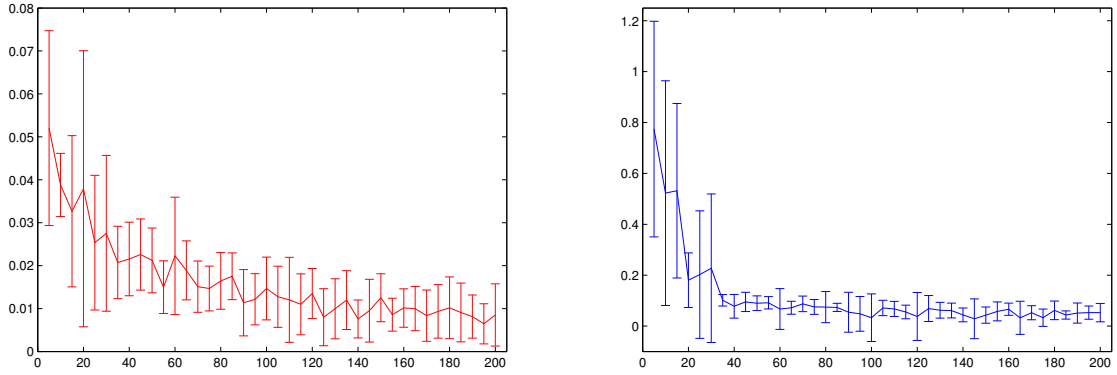
Из приведенных графиков можно сделать вывод, что если важна "вероятностная" точность восстановленной плотности, то достаточно выборки размера 40 — больший размер выборки не улучшает результат с точки зрения KL. Если же важна "физическая" точность плотности — больший размер выборки не помешает; медленное, но верное убывание функционала J при увеличении выборки иллюстрирует результат теоремы Парзена-Розенблатта (см. [4]). Стоит отметить, что для вероятностного моделирования, построения классификаторов важна именно "вероятностная" точность, отклонения плотности как функции (даже существенные) могут не играть роли.

Шум в данных. В этом эксперименте к 50 точкам выборки из исходного распределения добавлялись шумовые точки из равномерного распределения на отрезке $[-5, 5]$. Графики зависимости точности восстановления от количества шумовых точек на Рис. 5 построены как в предыдущем эксперименте.

3.2 Гейзер "Old Faithful"

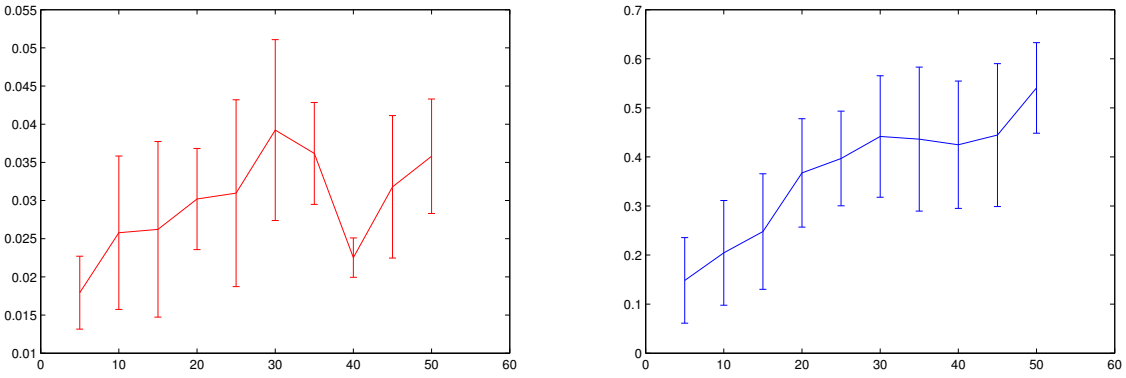
Гейзер, носящий название Old Faithful, находится в Национальном парке Йеллоустоун, США. Гейзер активен, периодически происходят его извержения. Посетители национального парка стремятся стать свидетелями красивейшего события, но гейзер извергается не по расписанию, составленному руководством парка, а по мало известным человеку законам физики. Встает актуальным вопрос о прогнозировании извержений гейзера по некоторым простым наблюдениям.

Имеется 272 наблюдения вида: извержение длилось $x^{(n)}$ минут, после чего прошло $y^{(n)}$ минут до следующего извержения. Наблюдения можно изобразить на плоскости: Рис. 6b. Набор данных взят из [1].



(a) $J(\hat{p}, p)$ (b) $KL(\hat{p}\|p)$

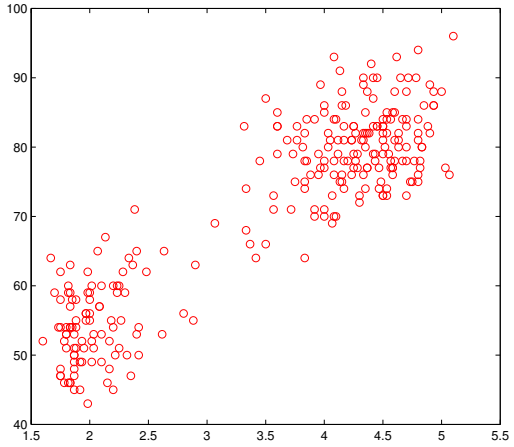
Рис. 4: Качество восстановления плотности методом окон Парзена в зависимости от размера выборки.



(a) $J(\hat{p}, p)$ (b) $KL(\hat{p}\|p)$

Рис. 5: Качество восстановления плотности методом окон Парзена в зависимости от числа шумовых объектов в выборке.

(a) Фото гейзера



(b) Визуализация наблюдений: по горизонтальной оси — длительность извержения, по вертикальной — время до следующего извержения.

Рис. 6: Наблюдение за гейзером Old Faithful.

Если удастся восстановить плотность $p(x, y)$, то можно будет делать прогноз. Пусть, например, прошло извержение гейзера и было засечено время x, мы сможем узнать ожидаемое время следующего извержения μ и точность такого прогноза $\frac{1}{\sigma}$:

$$\mathbb{E}(y|x) = \int y \cdot p(y|x)dy = \frac{1}{p(x)} \int y \cdot p(x, y)dy = \mu$$
$$\text{var}(y|x) = \frac{1}{p(x)} \int y^2 \cdot p(x, y)dy - \mu^2 = \sigma^2$$
$$p(x) = \int p(x, y)dy$$

Интегрировать можно численно, а в случае, если $p(x, y)$ представлено моделью смеси гауссиан (или другой достаточно простой параметрической моделью), интегрировать можно аналитически.

Результат применения методов восстановления плотности для данных наблюдений за гейзером представлены на Рис. 7.

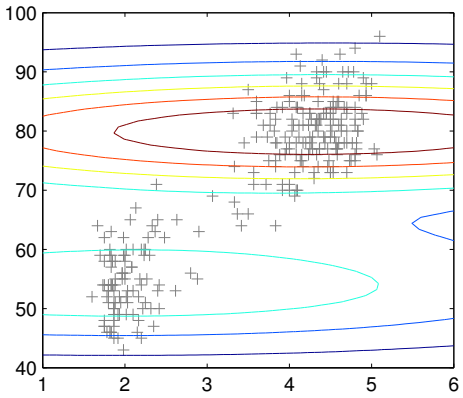
Лобовое применение метода Парзена с ядром Гаусса приводит к достаточно “странным” результатам, так на Рис. 7a видна “горизонтальная вытянутость” плотности: это связано с тем, что диапазон изменения признака x значительно отличается от диапазона изменения y, а ядро сглаживает оба признака одинакого. Для устранания такого эффекта, данные были центрированы и нормированы, в итоге получается: Рис. 7b.

Подгонка двумерного нормального распределения (Рис. 7d) выявляет корреляцию между признаками. Легко видеть, что помимо корреляции признаков, данные разбиты на два четко выраженных кластера. Имеет смысл смоделировать данные смесью из двух гауссиан (Рис. 7e).

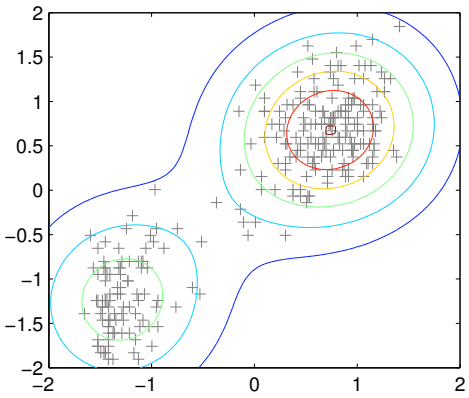
Метод Парзена с автоматической настройкой ширины окна по максимуму правдоподобия построил более плавную плотность (Рис. 7b), если уменьшить ширину, результат будет более близок к плотности смеси гауссиан (Рис. 7c).

3.3 Цвет кожи

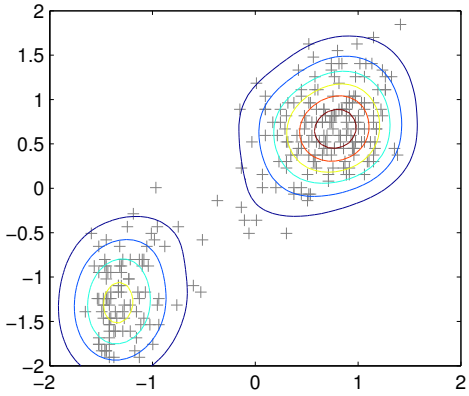
В компьютерном зрении часто возникает задача: по небольшой окрестности пиксела на изображении понять, насколько вероятно пиксел относится к тому или иному объекту; такая информация помогает строить выводы для всего изображения в целом.



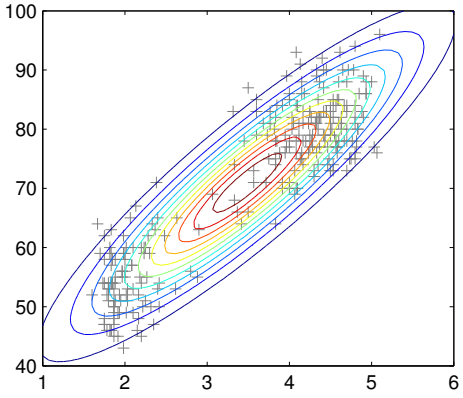
(a) Метод Парзена, исходные данные



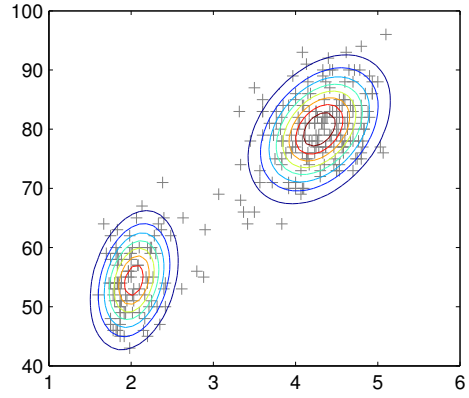
(b) Метод Парзена, нормализованные данные



(c) Метод парзена, с меньшим h

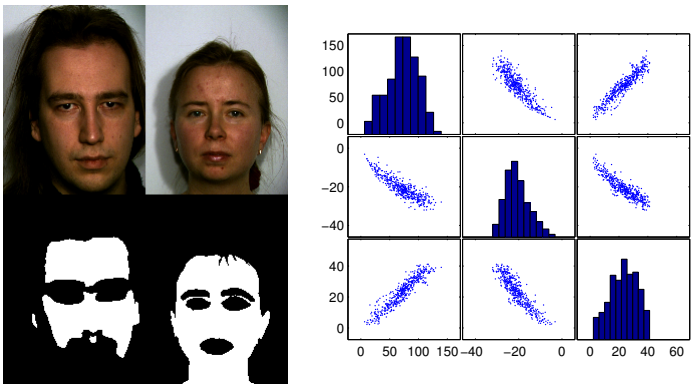


(d) Нормальное распределение



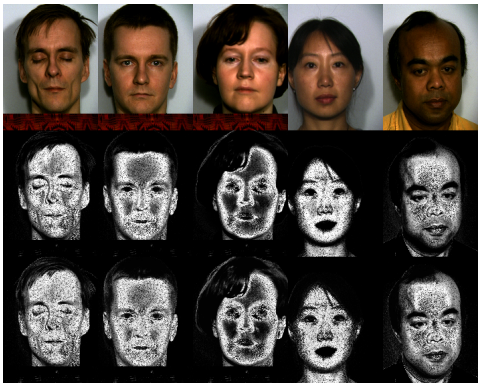
(e) Смесь из двух гауссиан

Рис. 7: Восстановление плотности в задаче о гейзере.



(a) (b)

Рис. 8: Данные для эксперимента с цветом кожи: изображения лиц и маски пикселей, под которыми находится кожа; распределение цветов пикселей в пространстве YCbCr, соответствующих коже.



(a) Изображения лиц. Второй ряд — одна гауссиана, третий ряд — смесь из 5 гауссиан.

(b) Работа обученной модели на произвольном изображении.

Рис. 9: Применение восстановленной плотности для построения “карты кожи”.

Попробуем построить модель для цвета (в хроматическом простнанстве YCbCr), который имеет кожа человека. Такая модель может помочь быстрому нахождению лиц, рук людей на изображениях (подробнее см. [3]). Пример данных, которые были использованы для восстановления плотности, приведен на Рис. 8.

Получать значение плотности нужно быстро. Метод Парзена не подходит по этому критерию, т.к. для получения значения плотности в методе Парзена нужно перебирать все точки тренировочной выборки. В то же время на визуализации расположения точек цвета кожи в хроматическом пространстве (Рис. 8b) видно, что их распределение близко к нормальному.

Имея плотность распределения цвета кожи $p(x), x \in \mathbb{R}^3$, по изображению $I = (x_1, \ldots, x_n)$ в пространстве YCbCr можно построить *карту кожи*: $S = (s_1, \ldots, x_n), s_i = C_0 \log(1 + p(x_i))$.

Для восстановления плотности использовалась модель нормального распределения и модель смеси гауссиан. Предположение о нормальности подтвердилось: использование модели смеси практически не изменило карту кожи (Рис. 9a). Для ускорения настройки параметров выборка случайным образом уменьшалась до 5000 точек.

4 Выводы

Плотности распределения можно восстанавливать. Все методы требуют настройки некоторых структурных параметров. Непараметрический “ленивый” метод Парзена не требует операции обуче-

ния, но вычисление значения плотности в одной точке методом Парзена, как правило, затрачивает гораздо больше времени, чем для этого требуют параметрические методы. Параметрические методы менее гибкие, требуют обучения, но вычисление значения плотности в параметрической модели происходит достаточно эффективно. Параметрические методы позволяют контроллировать гибкость модели. Чем более гибкая параметрическая модель, тем труднее настроить ее параметры.

Для достижения нирваны, автору отчета недостает времени. Если бы была возможность потратить больше времени на развитие отчета по методам восстановления плотности, то автор попытался бы:

- рассмотреть больше методов, тема параметрических методов не раскрыта — непонятно, что делать если гауссовские модели не подходят;
- тема непараметрических методов также не раскрыта — для решения проблем со скоростью их работы существуют методы фильтрации выборки, методы выбора эталонов (см. [5], [4]),
- сделать подробную таблицу сравнения методов;
- рассмотреть больше жизненых ситуаций и проблем, с которыми приходится сталкиваться (например ситуацию, когда распределение плохо приближается гауссовскими моделями);
- поправить трудночитаемые участки текста, а также пунктуацию.

Список литературы

[1] Christopher M. Bishop. *Pattern recognition and machine learning.* Springer, 1st ed. 2006. corr. 2nd printing edition, October 2006.

[2] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003. Available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

[3] Vladimir Vezhnevets, Vassili Sazonov, and Alla Andreeva. A survey on pixel-based skin color detection techniques. In *In ICCGV*, pages 85–92, 2003.

[4] К. В. Воронцов. *Математические методы обучения по прецедентам (теория обучения машин).*

[5] А. Г. Дьяконов. *Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (практикум на ЭВМ кафедры математических методов прогнозирования).* МАКСПресс, 2010.