Международная конференция ИОИ-10

Графическая вероятностная модель со скрытыми состояниями на основе главных многообразий

Юлин Сергей Сергеевич

Рыбинский государственный авиационный технический университет имени П. А. Соловьева

Научный руководитель: канд. тех. наук, профессор Паламарь И. Н.

Греция, о. Крит
2014 г.

Применение графических вероятностных моделей

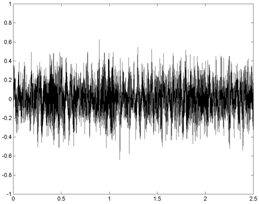
Классификация временных последовательностей используется при решении задач распознавания:

- речи;
- траекторий рукописного текста (online handwriting);
- движений глаз, жестов рук, движений головы.

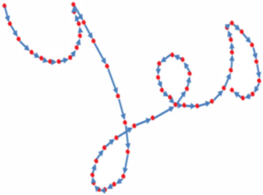
Объект классификации:

$$\bar{x} = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}, x_t \in \mathcal{R}^d,$$

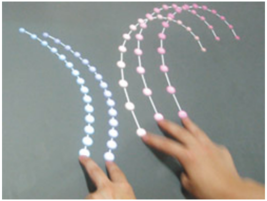
где T – длина последовательности,
d – размерность пространства признаков.



a)



b)



c)

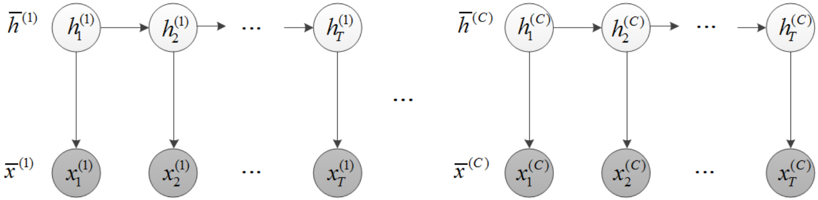
a) http://www.wagner.com/technologies/voice-speech/noisyvoice2.gif, b) https://kaggle2.blob.core.windows.net/competitions/kaggle/3385/media/stroke.png, c) http://www.hci.iis.u-tokyo.ac.jp/assets/images/research/tracking-fingertips1.jpg

Скрытая Марковская Модель

Для моделирования временных последовательностей вводят дополнительную случайную дискретную величину h (скрытое состояние), которая позволяет учитывать в модели изменение анализируемой последовательности во времени.

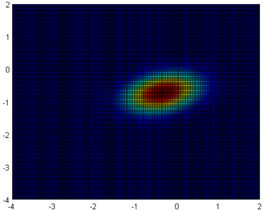
Совместное распределение случайных величин модели:

$$p(\bar{h}, \bar{x}) = \prod_{t=1}^{T} p(h_t \mid h_{t-1}) p(x_t \mid h_t)$$

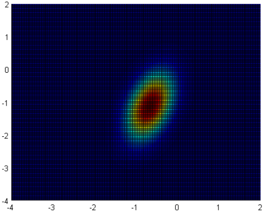


Графическое представление Скрытой Марковской Модели (C – количество классов)

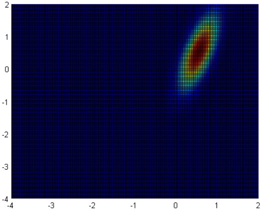
Недостатки Скрытой Марковской Модели



$h = 1$



$h = 2$



$h = 3$

Недостатки Скрытой Марковской Модели (*Hidden Markov Model* - HMM):

- необходимость выбора формы плотности вероятности $f_h(x)$;
- необходимость отсутствия зависимостей в компонентах вектора признаков (в случае гауссовской плотности вероятности **наличие линейной зависимости между компонентами приводит в вырожденности матрицы ковариации** $|\Sigma| = 0$). Плотность гауссовского многомерного распределения:

$$f_h(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \quad x, \mu \in \mathcal{R}^d$$

Устранение недостатков

Недостатки Скрытой Марковской Модели были устранены в дискриминантной модели "Условные случайные поля со скрытыми состояниями"(*Hidden Conditional Random Fields* - HCRF), предложенной *A. Gunawardana et al.* в 2005 году в работе *"Hidden conditional random fields for phone classification"*

Недостатки HCRF:

- целевая функция оптимизации не является выпуклой;
- дискриминантной модели необходимо больше обучающих данных чем порождающей модели для достижения своей асимптотически минимальной ошибки классификации (*Andrew Ng, Michael Jordan "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes", 2002*)

Цель: разработать порождающую графическую модель дающую хорошее качество классификации при малом количестве обучающих данных и не требующую оценки параметров плотности вероятности.

Связь минимизации эмпирического риска и максимизации правдоподобия

Максимизация правдоподобия (MLE):

$$L(X \mid \Sigma,\ \mu) = \sum_{i=1}^{n} \ln f(x_i \mid \Sigma,\ \mu) \to \max_{\Sigma,\ \mu}$$

где n – размер обучающей выборки;
X – обучающая выборка;
$\Sigma,\ \mu$ – параметры модели.

Минимизация эмпирического риска (ERM):

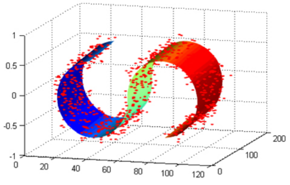
$$Q(X, W) = \sum_{i=1}^{n} E(x_i, W) \to \min_{W}$$

где W – параметры модели.

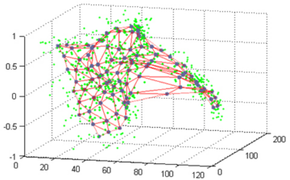
Эквивалентность MLE и ERM

$$-\ln f(x_i \mid \Sigma,\ \mu) = E(x_i, W)$$

Аппроксимация данных нелинейными многообразиями



Аппроксимация нелинейным многообразием[a)]



Аппроксимация нелинейным многообразием в виде сетки узлов

a) http://wissrech.ins.uni-bonn.de/research/pub/feuersaenger/mani.pdf

Для каждого $k = 0,1,\ldots,d-1$ среди всех k-мерных нелинейных многообразий в $\mathcal{R}^d$ найти такое $M_k \subset \mathcal{R}^d$, что сумма квадратов уклонений $x_i, i = 1\ldots n$ от M_k минимальна:

$$Q(X, M_k) = \sum_{i=1}^{n} dist^2(x_i, M_k) \to \min_{M_k}.$$

Задача построения такого многообразия трудоемка, поэтому строят его точечную аппроксимацию такими алгоритмами как **самоорганизующиеся карты Кохонена, упругие карты, алгоритм растущего "нейронного газа"** и другими:

$$Q(X, W) = \sum_{i=1}^{n}\sum_{j=1}^{k} \| x_i - w_j \|_2^2 \to \min_{W},$$

где w_j –- значение j-ого узла сетки в $\mathcal{R}^d$, $w_j \in M_k$;
k – количество узлов в сетке.

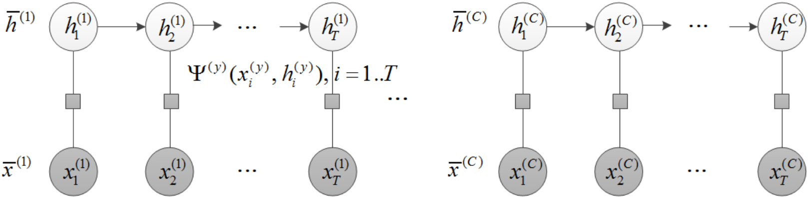
Основываясь на том, что:

- MLE эквивалентно ERM: $-\ln f(x_i \mid \Sigma,\ \mu) = E(x_i, W)$;
- номер узла аппроксимирующий сетки соответствует номеру скрытого состояния, т.е $h = \{1, \ldots, k\}$

получим, что совместное распределение случайных величин модели NPM-PGM:

$$p(\bar{h}, \bar{x}) = \prod_{t=1}^{T} p(h_t \mid h_{t-1}) \Psi(x_t, h_t)$$

$$\Psi(x_t, h_t) = \frac{\exp(-\parallel x_t - w_{h_t} \parallel_2^2)}{\sum_{j=1}^{k} \exp(-\parallel x_t - w_j \parallel_2^2)},\ p(h_t \mid h_{t-1}) = \frac{a_{h_t, h_{t-1}}}{\sum_{j=1}^{k} a_{h_t, j}}$$



Графическое представление модели NPM-PGM (C – количество классов)

NPM-PGM: обучение

Оптимизация функции:

$$L(\bar{h}, \bar{x} \mid W, A) = \sum_{l=1}^{n} \ln \prod_{t=1}^{T} p(h_t \mid h_{t-1}) \Psi(x_t^{(l)}, h_t) \to \max_{W,\ A},$$

где W– множество значений узлов сетки,
A – матрица вероятностей переходов между узлами сетки,

сводится к оптимизации:

$$\sum_{l=1}^{n} \sum_{j=1}^{k} \| x^{(l)} - w_j \|_2^2 \to \min_{W},$$

и вычислению:

$$a_{i,j} = \begin{cases} a_{i,j} + 1, \text{если } i = h_t = \arg\min \| x_t - w_i \|_2^2 \text{ и} \\ \qquad\qquad j = h_{t-1} = \arg\min \| x_{t-1} - w_i \|_2^2 \\ 0, \text{в противном случае} \end{cases}$$

NPM-PGM: вероятностный вывод

Вычисление правдоподобия выполняется **алгоритмом "прямого-обратного" хода** (Р. Л. Стратонович, 1960), также как в моделях HMM и HCRF). Временная сложность алгоритма – $O(k^2T)$.

Так как количество узлов аппроксимирующей сетки k может быть довольно велико, то для ускорения процесса вычисления можно воспользоваться алгоритмами сэмплирования:

– алгоритм Метрополиса-Гастингса;

– сэмплирование по Гиббсу.

Эксперименты: описание данных

UCI: Spoken Arabic Digit Data Set

- *Описание*: слова, произнесенные на арабском языке
- *Количество классов*: 10
- *Размерность пространства признаков*: 13
- *Количество экземпляров каждого класса*: 880
- *Источник*: http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit

UCI: Character Trajectories Data Set

- *Описание*: траектории рукописных букв английского алфавита
- *Количество классов*: 20
- *Размерность пространства признаков*: 3
- *Количество экземпляров каждого класса*: 100
- *Источник*: http://archive.ics.uci.edu/ml/datasets/Character+Trajectories

Эксперименты: результат

Размер обучающей выборки: **90** экземпляров каждого класса

– UCI: Spoken Arabic Digit Data Set

Model	HMM(1 mix)	HMM(16 mix)	HCRF(BFGS)	HCRF(CG)	NPM-PGM
States	5	5	5	5	1024
F-score	0.6018	0.5374	0.6420	0,6274	**0.8043**

– UCI: Character Trajectories Data Set

Model	HMM(1 mix)	HMM(16 mix)	HCRF(BFGS)	HCRF(CG)	NPM-PGM
States	7	7	7	7	256
F-score	0.9329	0.9288	0.9651	0.9516	**0.9809**

Размер обучающей выборки: **800** экземпляров каждого класса

– UCI: Spoken Arabic Digit Data Set

Model	HMM(1 mix)	HMM(16 mix)	HCRF(BFGS)	HCRF(CG)	NPM-PGM
States	5	5	5	5	1024
F-score	0.8778	0.8368	**0.9525**	0.9344	0.9349

Вывод

Достоинства предлагаемой модели NPM-PGM:

- лучший показатель F-score чем у HMM на всех тестах;
- лучший показатель F-score чем у HCRF при малом количестве обучающих данных.

Недостатки предлагаемой модели NPM-PGM:

- сложность выбора оптимального количества узлов аппроксимирующей сетки k, дающего наилучшее качество классификации;
- большое количество скрытых состояний k, что сказывается на времени выполнения классификации (временная сложность алгоритма ”прямого-обратного”хода – $O(k^2T)$).