Provably Learning Mixtures of Gaussians and More

Jonathan Huggins

Докладчик: Токмакова Лада

Gaussians Mixture Model (GMM)

Convex combination of k different n-dimensional Gaussians with:

- weights $w_i \in [0,1]$, $\sum_{i=1}^{k} w_i = 1$;
- means $\mu_i \in \mathbb{R}^n$;
- covariance matrices $\Sigma_i \in \mathbb{R}^{n \times n}$

Let $F_i = \mathcal{N}(\mu_i, \Sigma_i) \Longrightarrow$

$$F_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma_i)}} \exp\left(-\frac{1}{2}(\vec{x} - \mu_i)^T \Sigma_i^{-1}(\vec{x} - \mu_i)\right)$$

$F = \sum_{i=1}^{k} w_i F_i$ - the density of the GMM

Defining the Problem - 1

- Пара гауссиан $\mathcal{N}(\mu_1, \Sigma_1)$ и $\mathcal{N}(\mu_2, \Sigma_2)$ являются Δ **- разделенными**, если $\|\mu_1 - \mu_2\| \geq \Delta\ (\sigma_1 + \sigma_2)$.
- Смесь гауссиан является Δ**-разделенной**, если любая пара компонент является Δ-разделенной.

- $F = \sum_i w_i F_i$ являетсяΔ**- статистически обучаемой**, если:
  - $w_{Min} \geq \Delta$
  - $\min\limits_{i \neq j} D\big(F_i, F_j\big) \geq \Delta$, где $D\big(f(x), g(x)\big) = \frac{1}{2}\int_{\mathbb{R}^n} |f(x) - g(x)| dx$

- $\theta = \{(w_i, \mu_i, \Sigma_i), \dots, (w_k, \mu_k, \Sigma_k)\}$
- Пусть p_θ - семейство распределений с параметром $\theta \in \Theta$. Для каждого θ определим **радиус распознавания**:
  - $\mathcal{R}(\theta) = \sup\{r > 0 \big| \forall \theta_1 \neq \theta_2, (\|\theta_1 - \theta\| < r \wedge \|\theta_2 - \theta\| < r) \Longrightarrow \big(p_{\theta_1} \neq p_{\theta_2}\big)\}$
  - Если условие не может быть выполнено, то $\mathcal{R}(\theta) = 0$.

Defining the Problem - 2

- **Радиус распознавания** для θ для GMM:
  - $\mathcal{R}(\theta)^2 = \min\left(\frac{1}{4}\min_{i\neq j}\left\{\left\|\mu_i - \mu_j\right\|^2 + \left\|\Sigma_i - \Sigma_j\right\|_F^2\right\}, w_{Min}^2\right)$
- Набор параметров для смеси из k гауссиан $\hat{\theta} = \left\{\left(\hat{w}_1\hat{\mu}_1, \hat{\Sigma}_1\right), \ldots, \left(\hat{w}_k, \hat{\mu}_k, \hat{\Sigma}_k\right)\right\}$ является ϵ-близкой оценкой для θ, если $\exists\pi \in S_k$: $\forall i \in [k] \Longrightarrow$
  - $\left|w_i - \hat{w}_{\pi(i)}\right| \leq \epsilon$
  - $d\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}\left(\hat{\mu}_{\pi(i)}, \hat{\Sigma}_{\pi(i)}\right)\right) \leq \epsilon$
- Пример расстояния: $D_p\left(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma')\right) = \|\mu - \mu'\| + \|\Sigma - \Sigma'\|_F$.
- Алгоритм A **эффективно изучает** смесь из k гауссиан размерности n, если для $1 > \epsilon > 0$ и $1 > \delta > 0$ A возвращает оценку $\hat{\theta}$, которая находится в ϵ-окрестности реального значения θ с вероятностью $\geq 1 - \delta$, и работает за $poly\left(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{w_{Min}}, \frac{1}{\Delta}\right)$.

Learning GMMs

- Алгоритм содержит три шага:
  - Проекция данных размерности n на одно или несколько подпространств размерности $poly(k), k \ll n$.
  - Определение каждой точки к одной из компонент смеси или использование другого метода оценки параметров.
  - Восстановление компоненты в изначальном пространстве размерности n.

Author	Min. Separation	Mixture Class	Method	Comments
Dasgupta [1999]	$\sqrt{n}$	Gaussian with shared covariance matrix	Random projection	
Dasgupta and Schulman [2000]	$n^{\frac{1}{4}}$	Spherical Gaussian	EM	
Arora and Kennan [2001]	$n^{\frac{1}{4}}$	Gaussian	Distance-based	
Vempala and Wang [2002]	$k^{\frac{1}{4}}$	Spherical Gaussian	Spectral, distance-based	
Kannan et al. [2005]	$k^{\frac{3}{2}}/w_{min}^2$	Log-concave	Spectral, distance-based	need to know w_i
Achlioptas and McSherry [2005]	$k + \sqrt{k \log n}$	Gaussian	Spectral	
Feldman et al. [2006]	> 0	Axis aligned Gaussians	Method of moments (MoM)	no parameter estimation
Belkin and Sinha [2010a]	> 0	Identical spherical Gaussian	Spectral	
Kalai et al. [2010]	≥ 0	Gaussian with two components	Random projections, MoM	
Moitra and Valiant [2010]	≥ 0	Gaussian	Random projections, MoM	
Belkin and Sinha [2010b]	≥ 0	Gaussian	Deterministic projections, MoM	

Spectral Learning - 1

Spectral Algorithm for Learning GMMs [Vempala and Wang, 2002]

$M \leftarrow |S|$

while $S \neq \emptyset$ ***do***

$Compute\ the\ k$

$-\ dimensional\ SVD\ subspace\ W\ of\ S$

$Project\ S\ onto\ W$

$R \leftarrow \max_{x \in S} \min_{y \in S} \|x - y\|$

$S' \leftarrow \left\{ x \in S : \min_{y \in S} \|x - y\| \leq 3\hat{\epsilon} R^2 \right\}$

$G \leftarrow \emptyset$

while $S' \neq \emptyset$ ***do***

$Let\ x, y\ be\ the\ two\ closest\ points\ in\ S'$

$$l \leftarrow \|x - y\|^2 \left(1 + 8 \sqrt{\frac{6 \ln \frac{M}{\delta}}{k}} \right)$$

$$H \leftarrow \{ w \in S' : \|x - w\|^2 \leq l \}$$

$S' \leftarrow S' / H$

$G \leftarrow G \cup \{H\}$

end while

$Report\ each\ H \in G\ with\ variance\ greater\ than \frac{3\epsilon R^2}{k}\ as\ the\ set\ of\ points\ generated\ by\ one\ component\ of\ the\ mixture\ and\ remove\ those\ points\ from\ S$

end while

Spectral Learning – 2

- **Theorem** [Kannan et al., 2005]. Пусть W является k – мерным пространством главных компонент выборки S, где данные порождаются смесью из k компонент. Для каждого $i \in [k]$ пусть $\mu_{i,S}$ - математическое ожидание S_i и $\sigma^2_{i,S}(W)$ - максимальная дисперсия S_i из всех направлений в W. Тогда:

$$\sum_{i=1}^{k} |S_i| d\left(\mu_{i,S}, W\right)^2 \leq k \sum_{i=1}^{k} |S_i| \sigma^2_{i,S}(W)$$

- **Lemma**. Пусть F – логарифмически выпуклое распределение на $\mathbb{R}^n$ с математическим ожиданием μ и вторым моментом

$R^2 = \mathbb{E}_F[(X-\mu)^2]$. Тогда $\exists$ c > 0: $\forall t > 1 \Longrightarrow \Pr(|X-\mu| > tR) < e^{-ct}$

Spectral Learning - 3

Spectral algorithm for learning log-concave mixtures [Kannan et al., 2005]

$m \leftarrow |S|$
while $S \neq \emptyset$ **do**
  $Compute\ the\ k-dimensional\ SVD\ subspace\ W\ using\ a\ subset\ T\ of\ S\ of\ size\ m_0$
  $S \leftarrow S \backslash \mathrm{T}$
  $Project\ S\ onto\ W$
  for all $x \in S$ **do**
  - $Calculate\ the\ set\ S(x)\ consisting\ of\ the\ closest\ \frac{1}{2} w_{Min} m\ point\ to\ x$
  - $Find\ the\ mean\ \mu(x) of\ S(x)$
  - $Form\ the\ matrix\ A(x)\ whose\ rows\ are\ y-\mu(x)\ for\ each\ y \in S(x)$
  - $Calculate\ \sigma(x), the\ largest\ singular\ value\ of\ A(p), i.e. the\ maximum\ standard\ deviation\ of\ S(x) in\ W$

  end for
  $x_0 \leftarrow \arg\max_{x \in S} \sigma(x)$
  $T_0 \leftarrow \left\{ x \in T : \|W(x_0) - W(x)\| \leq \frac{\sqrt{k} log N}{w_{Min}} \sigma(x) \right\},$
  $where\ W(x) denotes\ the\ projection\ of\ x\ to\ W$
  $Report\ T_0\ as\ the\ set\ of\ points\ generated\ by\ one\ component\ of\ the\ mixture\ and\ remove\ those\ points\ from\ T$
end while

Projection Approaches - 1
Random Projection

- Moitra and Valiant [2010] → ϵ-близкая оценка с $D(\cdot,\cdot)$ и расстоянием между гауссианами.

- Kalai et al. [2010] → изучение смеси из двух гауссиан:
  - The 1D Learnability Lemma
  - The Random Projection Lemma
  - The Parameter Recovery Lemma

  ⇓

  Algorithm for learning mixtures of two Gaussians

Projection Approaches - 2
Random Projection

Learning mixtures of two Gaussians [Kalai et al., 2010]

$Pick\ a\ random\ unit\ vector\ u$

$Choose\ n^2 vectors\ u_1, \dots, u_{n^2}\ that\ are\ fairly\ close\ to\ u$

$\boldsymbol{for\ all}\ i \in [n^2]\ do$

$learn\ very\ accurate\ univariate\ parameters\ for\ the\ projection\ of\ the\ mixture\ onto\ u_i$

$\boldsymbol{end\ for}$

$Recover\ the\ true\ n-dimensional\ parameters\ for\ the\ mixture\ with\ high\ probability$

Projection Approaches - 3
Random Projection

- Пусть GMM F из k гауссиан параметризована θ. Тогда GMM $\hat{F}$ из $k' < k$ гауссиан параметризована $\hat{\theta}$ является $\boldsymbol{\epsilon}-$близким **разделением** F если существует сюръективное отображение $\pi: [k] \to [k']$:
  - $\forall j \in [k']: \left|\sum_{i:\pi(i)=j} w_i - \hat{w}_j\right| \leq \epsilon$
  - $\forall i \in [k]: D_p\left(\mathcal{N}(\mu_i, \Sigma_i), \mathcal{N}\left(\hat{\mu}_{\pi(i)}, \hat{\Sigma}_{\pi(i)}\right)\right) \leq \epsilon$

Projection Approaches - 4
Random Projection

Anisotropic algorithm for learning GMMs [Moitra and Valiant, 2010]

$Place\ the\ samples\ in\ isotropic\ position$
$Run\ the\ Hierarchial\ Clustering\ Algorithm\ (HCA)$
$\boldsymbol{if}\ HCA\ returns\ \hat{F}\ \boldsymbol{then}$
    $\boldsymbol{return}\ \hat{F},\quad which\ is\ an\ \epsilon - close\ estimate\ of\ F\ w.h.p.$
$\boldsymbol{else}\ \{HCA\ returns\ partition\ a\ (A, B)\}$
- $Draw\ an\ additional\ sample\ set\ S\ from\ sample\ oracle\ for\ F$
- $Run\ the\ anisotropic\ algorithm\ on\ the\ samples\ S_A \subset S\ that\ are\ in\ A\ to\ get\ \hat{F}_A$
- $Run\ the\ anisotropic\ algorithm\ on\ the\ samples\ S_B \subset S\ that\ are\ in\ B\ to\ get\ \hat{F}_B$

    $\boldsymbol{return}\ \hat{F} = \frac{|S_A|}{|S|}\hat{F}_A + \frac{|S_B|}{|S|}\hat{F}_B$
$\boldsymbol{end\ if}$

Projection Approaches
Deterministic Projection

- Семейство вероятностных плотностей p_θ, параметризованных θ является **полиномиальным семейством**, если каждый l – мерный момент $M_{i_1,\dots,i_l}(\theta) = \int x_1^{i_1} \cdots x_l^{i_l} dp_\theta$ существует и может быть представлен как $poly(\theta^1, \dots, \theta^m)$, и если p_θ однозначно определяется своими моментами.
- **Theorem** [Belkin and Sinha, 2010b]. Пусть p_θ - полиномиальное семейство распределений. Тогда $\exists\, N \in \mathbb{N}: p_{\theta_1} = p_{\theta_2} \Leftrightarrow M_i(\theta_1) = M_i(\theta_2)\ \forall i \in [N]$.
- Пусть $\mathcal{E}(\theta) = \{w \in \Theta: p_w = p_\theta\}$. Тогда ϵ-соседи θ – это объединение окрестностей вокруг каждой точки $\mathcal{E}(\theta)$:
  $\mathcal{N}(\theta, \epsilon)$
  $= \{w \in \Theta: \exists w', \theta' \in \Theta, 0 < \epsilon' < \epsilon \Rightarrow w' \in \mathcal{E}(\theta') \wedge \|w - w'\| < \epsilon' \wedge \|\theta - \theta'\| < \epsilon - \epsilon'\}$.
- **Theorem** [Belkin and Sinha, 2010b]. Для $\theta \in \Theta$ если $\mathcal{E}(\theta) = \{\theta_1, \dots, \theta_k\}$ - ограниченное множество, то $\exists$ такой алгоритм, который для $\epsilon > 0$ возвращает $\hat{\theta}$ с $\epsilon' = \min\left(\epsilon, \min_j \mathcal{R}\left(\theta_j\right)\right)$ для θ_i для некоторого $i \in [k]$ и с вероятностью $1 - \delta$ использует $poly(1/\epsilon', 1/\delta)$ образцов.

Conclusion

$Find\ the\ 2k^2 - dimensional\ coordinate\ subspace\ W\ with\ maximum\ emperical\ \mathcal{R}(\theta)$
$Project\ S\ onto\ W$
$Estimate\ the\ means\ and\ covariance\ entries\ for\ these\ 2k^2\ coordinates\ (by\ Theorem)$
$\boldsymbol{for\ all}\ e_i \notin W\ \boldsymbol{do}$
    $W_i \leftarrow span(W, e_i)$
    $project\ S\ onto\ W_i$
    $estimate\ the\ component\ variance\ and\ means\ along\ e_i (by\ Theorem)$
    $\boldsymbol{for\ all}\ e_j \notin W\ \boldsymbol{do}$
        $W_{ij} \leftarrow span(W, e_i, e_j)$
        $project\ S\ onto\ W_{ij}$
        $estimate\ the\ component\ covariance\ between\ e_i and\ e_j (by\ Theorem)$
    $\boldsymbol{end\ for}$
$\boldsymbol{end\ for}$

Спасибо за внимание!