Learning probability distributions

Abbas Mehrabian

McGill University

1 August 2018

Co-authors: Hassan Ashtiani, Shai Ben-David,
Luc Devroye, Nick Harvey, Christopher Liaw, Yaniv Plan, and
Tommy Reddad

An example of distribution learning

Generating random faces for computer games

- ✓ Training data consists of actual faces.
- ✓ A probability density function $\mathbf{P} : \mathbb{R}^d \to \mathbb{R}$ is learned from the data.
- ✓ New random faces are generated using the learned distribution.

An example of distribution learning

Generating random faces for computer games

- ✓ Training data consists of actual faces.
- ✓ A probability density function $\mathbf{P} : \mathbb{R}^d \to \mathbb{R}$ is learned from the data.
- ✓ New random faces are generated using the learned distribution.

A popular approach: generative adversarial networks, based on deep neural networks.

Distribution learning in action

Top: generated images using generative adversarial networks
Bottom: training data

Picture from Karras, Aila, Laine, and Lehtinen (NVIDIA and Aalto University), October 2017

Distribution learning task

also known as density estimation

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$, output a distribution $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.

Distribution learning task

also known as density estimation

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$, output a distribution $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.

- ✓ We assume P belongs to some known class $\mathcal{F}$ of distributions.
- ✓ We would like our algorithm to use as a small sample as possible.
- ✓ Closeness is measured by the total variation distance: $\mathrm{TV}(\mathbf{P}, \widehat{\mathbf{P}}) \coloneqq \sup_E |\mathbf{P}(E) - \widehat{\mathbf{P}}(E)| = \frac{1}{2} \int |p(x) - \widehat{p}(x)| \, \mathrm{d}x$

Distribution learning task

Our setup

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$ from a known class $\mathcal{F}$, output some $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$. What is the smallest number of samples needed to guarantee $\mathrm{TV}(\widehat{\mathbf{P}}, \mathbf{P}) \leq \varepsilon$ with probability 99%? $m_{\mathcal{F}}(\varepsilon)$.

Distribution learning task

Our setup

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$ from a known class $\mathcal{F}$, output some $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$. What is the smallest number of samples needed to guarantee $\mathrm{TV}(\widehat{\mathbf{P}}, \mathbf{P}) \leq \varepsilon$ with probability 99%? $m_{\mathcal{F}}(\varepsilon)$.

Main problem

prove bounds for $m_{\mathcal{F}}(\varepsilon)$ for various classes $\mathcal{F}$.

$m_{\mathcal{F}}(\varepsilon)$ is also known as the minimax risk of $\mathcal{F}$.

A heuristic

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

Example

- ✓ $\mathcal{F} =$ Bernoulli distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} =$ Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 2/\varepsilon^2$
- ✓ $\mathcal{F} = d$-dimensional Gaussians: $m_{\mathcal{F}}(\varepsilon) \leq Cd^2/\varepsilon^2$
- ✓ Finite $\mathcal{F}$: $m_{\mathcal{F}}(\varepsilon) \leq C \log |\mathcal{F}|/\varepsilon^2$ Devroye-Lugosi'01

A heuristic

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

Example

- ✓ $\mathcal{F}$ = Bernoulli distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F}$ = Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 2/\varepsilon^2$
- ✓ $\mathcal{F}$ = d-dimensional Gaussians: $m_{\mathcal{F}}(\varepsilon) \leq Cd^2/\varepsilon^2$
- ✓ Finite $\mathcal{F}$: $m_{\mathcal{F}}(\varepsilon) \leq C \log|\mathcal{F}|/\varepsilon^2$ Devroye-Lugosi'01

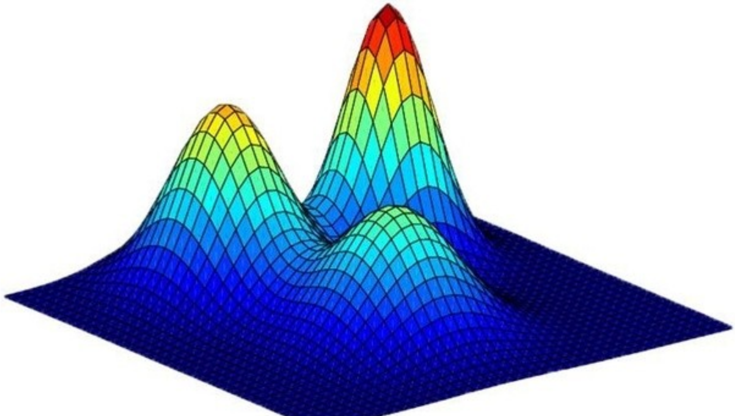
Main result: this heuristic also works for two more complicated classes: mixtures of multidimensional Gaussians, and the Ising model.

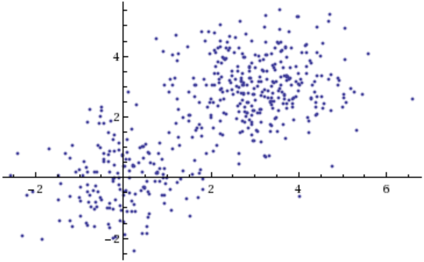
Mixtures of Gaussians

Mixtures of Gaussians

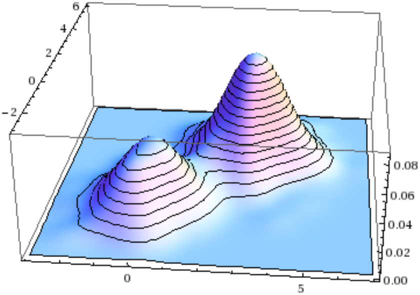
A mixture of k Gaussians in d dimensions has density $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)(x)$, where $w_i \geq 0$ and $\sum w_i = 1$.

$\mathcal{N}(\mu, \Sigma)(x)$ = density of a Gaussian with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$





4
2
-2
-2
2
4
6



6
4
2
0
-2
0.08
0.06
0.04
0.02
0.00
0
5

Main results

mixtures of Gaussians

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = kd^2/\varepsilon^2$ up to polylogarithmic factors.

Main results

mixtures of Gaussians

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = kd^2/\varepsilon^2$ up to polylogarithmic factors.

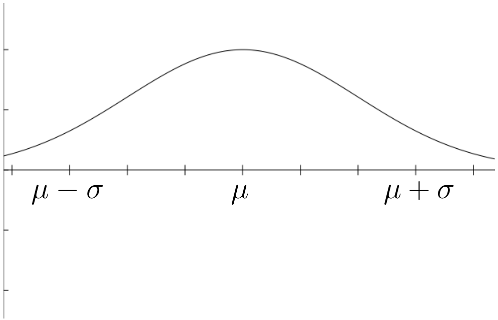
Any density in $\mathcal{F}_{k,d}$ has form $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)(x)$, and Σ_i is $d \times d$, so has $\Theta(kd^2)$ parameters.

Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

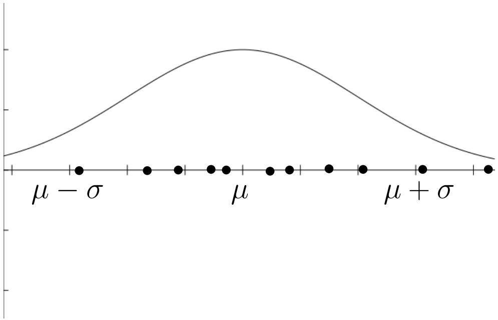


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

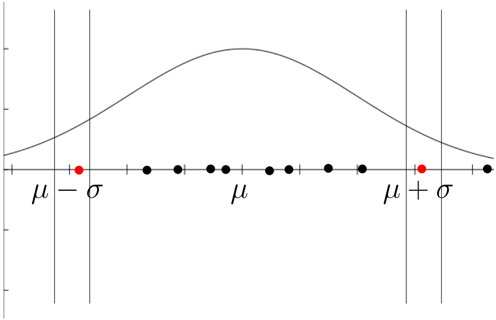


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

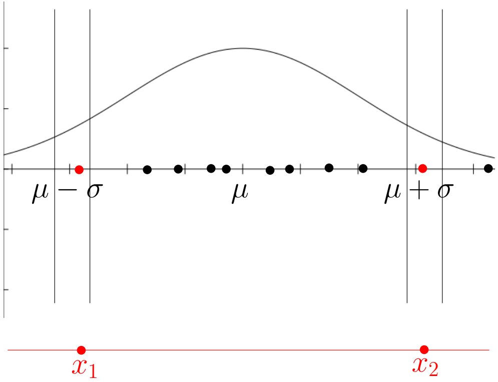


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

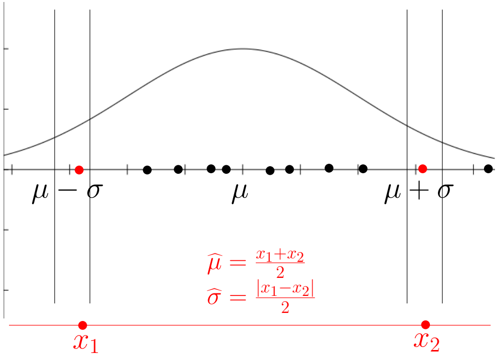


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.



Proof of upper bound: compression

Compressing d-dimensional Gaussians

d-dimensional Gaussians admit $\widetilde{O}(d^2)$-compression.

Compressing mixtures

If $\mathcal{F}$ admits τ-compression, then k-mix$(\mathcal{F})$ admits $(k\tau + k\log k)$-compression.

Compression implies learnability

If $\mathcal{F}$ admits τ-compression, then $m_{\mathcal{F}}(\varepsilon) = \widetilde{O}(\tau/\varepsilon^2)$.

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{O}(kd^2/\varepsilon^2)$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then, $m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

Fano's inequality

Suppose there exist $f_1, \ldots, f_M \in \mathcal{F}$ with

$$\mathrm{KL}(f_i \parallel f_j) = O(\varepsilon^2) \text{ and } \mathrm{TV}(f_i, f_j) = \Omega(\varepsilon) \qquad \forall i \neq j \in [M].$$

Then $m_{\mathcal{F}}(\varepsilon) = \Omega(\log M/\varepsilon^2)$.

To construct this family of $2^{\Omega(d^2)}$ distributions, start with an identity covariance matrix, then choose a random subspace of dimension $d/9$ and slightly increase the eigenvalues corresponding to this eigenspace from 1 to $1 + \varepsilon/\sqrt{d}$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then,
$m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

It is easy to lift this to the class of mixtures, proving
$m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then,
$m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

It is easy to lift this to the class of mixtures, proving $m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.
The logarithmic factors were removed in subsequent work, giving $m_{\mathcal{F}_{k,d}}(\varepsilon) = \Omega(d^2/\varepsilon^2)$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then,
$m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

It is easy to lift this to the class of mixtures, proving
$m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.
The logarithmic factors were removed in subsequent work, giving $m_{\mathcal{F}_{k,d}}(\varepsilon) = \Omega(d^2/\varepsilon^2)$.

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = kd^2/\varepsilon^2$ up to polylogarithmic factors.

The Ising model

Definition

For a graph G on d vertices, and edge weights $\{w_{i,j}\}_{ij \in E(G)}$, the Ising model with parameters $\{w_{i,j}\}_{ij \in E(G)}$ is supported on $\{-1, +1\}^d$ and has probability mass function

$$p_{\mathbf{w}}(x_1, \ldots, x_d) \propto \exp\left(\sum_{ij \in E(G)} w_{i,j} x_i x_j\right)$$

Number of parameters $= |E(G)|$.

The Ising model

Definition

For a graph G on d vertices, and edge weights $\{w_{i,j}\}_{ij \in E(G)}$, the Ising model with parameters $\{w_{i,j}\}_{ij \in E(G)}$ is supported on $\{-1, +1\}^d$ and has probability mass function

$$p_{\mathbf{w}}(x_1, \ldots, x_d) \propto \exp\left(\sum_{ij \in E(G)} w_{i,j} x_i x_j \right)$$

Number of parameters $= |E(G)|$.

Theorem (Devroye, M, Reddad'18)

Let $\mathcal{I}_G =$ Ising models on G. Then, $m_{\mathcal{I}_G}(\varepsilon) \asymp |E(G)|/\varepsilon^2$.

The Ising model

Definition

For a graph G on d vertices, and edge weights $\{w_{i,j}\}_{ij \in E(G)}$, the Ising model with parameters $\{w_{i,j}\}_{ij \in E(G)}$ is supported on $\{-1,+1\}^d$ and has probability mass function

$$p_{\mathbf{w}}(x_1, \ldots, x_d) \propto \exp\left(\sum_{ij \in E(G)} w_{i,j} x_i x_j \right)$$

Number of parameters $= |E(G)|$.

Theorem (Devroye, M, Reddad'18)

Let $\mathcal{I}_G$ = Ising models on G. Then, $m_{\mathcal{I}_G}(\varepsilon) \asymp |E(G)|/\varepsilon^2$.

Lower bound proof uses Fano's inequality again.

Upper bound proof is simpler, uses a technique of Yatracos.

Recap

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

Example

- ✓ $\mathcal{F} =$ Bernoulli distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} =$ Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} = d$-dimensional Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp d^2/\varepsilon^2$
- ✓ Finite $\mathcal{F}$: $m_{\mathcal{F}}(\varepsilon) \leq 9 \log |\mathcal{F}|/\varepsilon^2$ Devroye-Lugosi'01
- ✓ $\mathcal{F}_{k,d} =$ mixture of k Gaussians in d dimensions: $m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Theta}(kd^2/\varepsilon^2)$.
- ✓ $\mathcal{I}_G =$ Ising models on G: $m_{\mathcal{I}_G}(\varepsilon) \asymp |E(G)|/\varepsilon^2$.

Future work

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

1. Does the heuristic works for other classes? For example, graphical models? Distributions generated by neural networks?
2. $\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \leq$ smallest compression size of $\mathcal{F}$. Is the converse true?
3. For binary classification, sample complexity $\asymp$ VC-dimension of the hypothesis class $/\varepsilon^2$. Can we use $\varepsilon^2 m_{\mathcal{F}}(\varepsilon)$ as a natural definition of 'dimension' for class $\mathcal{F}$? Are there connections with other definitions?
4. What about computational complexity?