Generative modelling

Fifth & Sixth lectures

E. Scornet

Outline

1. General Motivation

2. No latent variables
   - Parametric density estimation
   - Histograms
   - Kernels
   - Nearest neighbors
   - Fully Visible Belief Nets (FVBN)

3. Latent variables
   - Hidden Markov Model - EM
   - Variational autoencoders
   - Generative Adversarial Networks

4. Applications

Supervised setting

In a supervised setting, we assumed to be given a data set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where X_i is the ith input and Y_i its associated output.

- Discriminative models: estimate the conditional distribution $Y|X$
  1. Linear regression, logistic regression, generalized linear models
  2. Standard Neural Networks, CNN, RNN...
  3. Decision trees, boosting, random forests, kernel methods, k nearest neighbors...
- Generative models: estimate the joint distribution (X, Y)
  1. Naive Bayes
  2. Linear/quadratic discriminant analysis

Generating new data requires to model the joint distribution (X, Y)

→ second class of models.

Unsupervised setting

In a unsupervised setting, we assumed to be given a data set $\mathcal{D}_n = \{X_1 \ldots, X_n\}$ where X_i is the ith observation.

In this setting, we have no output, thus nothing to predict nor discriminate.

Different goals:

- **Descriptive analysis:** detect structure, correlations in the data set using descriptive/graphical tools or using more involved methods (PCA for example)
- **Clustering:** create “homogeneous” groups of observations (usually spending 90% of the allocated time to properly define “homogeneous”)
- **Estimating the distribution of observations:** detect suspect data/behaviour, detect changes in the data set if the data are collected through time
- **Generating new data:** closely related to the previous point.

Outline

1. General Motivation

2. No latent variables
   - Parametric density estimation
   - Histograms
   - Kernels
   - Nearest neighbors
   - Fully Visible Belief Nets (FVBN)

3. Latent variables
   - Hidden Markov Model - EM
   - Variational autoencoders
   - Generative Adversarial Networks

4. Applications

Two different approaches

- **Parametric density estimation**: assume some distribution for the random variables at stake, which depends on some parameters (e.g., mean and variance). Then use the corresponding density functions to compute the likelihood. Choose the parameter values maximizing the likelihood.

Maximum Likelihood Estimation (MLE) is everywhere!

- **Non parametric density estimation** (histograms, kernels, nearest neighbors): do not assume any specific form for the density but learn the density function entirely from data.

Parametric vs. Non parametric
=
Interpretability vs. flexibility

Outline

1. General Motivation

2. No latent variables
   - Parametric density estimation
   - Histograms
   - Kernels
   - Nearest neighbors
   - Fully Visible Belief Nets (FVBN)

3. Latent variables
   - Hidden Markov Model - EM
   - Variational autoencoders
   - Generative Adversarial Networks

4. Applications

Parametric density estimation

Let $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ be a set of densities parametrized by θ.

Assume that you have $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ i.i.d. random variables distributed as $f_{\theta^\star} \in \mathcal{F}$. Our goal is to recover $f_{\theta^\star}$, that is $\theta^\star$ using the data set $\mathcal{D}_n$.

Parametric density estimation

Let $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ be a set of densities parametrized by θ.

Assume that you have $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ i.i.d. random variables distributed as $f_{\theta^\star} \in \mathcal{F}$. Our goal is to recover $f_{\theta^\star}$, that is $\theta^\star$ using the data set $\mathcal{D}_n$.

To do so, assume that $X \sim f_\theta$. Our goal is to find the best θ.

Parametric density estimation

Let $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ be a set of densities parametrized by θ.

Assume that you have $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ i.i.d. random variables distributed as $f_{\theta^\star} \in \mathcal{F}$. Our goal is to recover $f_{\theta^\star}$, that is $\theta^\star$ using the data set $\mathcal{D}_n$.

To do so, assume that $X \sim f_\theta$. Our goal is to find the best θ.

Since the observations are i.i.d. the likelihood is given by

$$\mathcal{L}(\theta, \mathcal{D}_n) = \prod_{i=1}^{n} f_\theta(X_i).$$

Parametric density estimation

Let $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ be a set of densities parametrized by θ.

Assume that you have $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ i.i.d. random variables distributed as $f_{\theta^\star} \in \mathcal{F}$. Our goal is to recover $f_{\theta^\star}$, that is $\theta^\star$ using the data set $\mathcal{D}_n$.

To do so, assume that $X \sim f_\theta$. Our goal is to find the best θ.

Since the observations are i.i.d. the likelihood is given by

$$\mathcal{L}(\theta, \mathcal{D}_n) = \prod_{i=1}^{n} f_\theta(X_i).$$

This leads to the following optimization problem (MLE principle):

$$\hat{\theta} \in \operatorname*{argmax}_{\theta} \mathcal{L}(\theta, \mathcal{D}_n).$$

Parametric density estimation

Let $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ be a set of densities parametrized by θ.

Assume that you have $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ i.i.d. random variables distributed as $f_{\theta^\star} \in \mathcal{F}$. Our goal is to recover $f_{\theta^\star}$, that is $\theta^\star$ using the data set $\mathcal{D}_n$.

To do so, assume that $X \sim f_\theta$. Our goal is to find the best θ.

Since the observations are i.i.d. the likelihood is given by

$$\mathcal{L}(\theta, \mathcal{D}_n) = \prod_{i=1}^{n} f_\theta(X_i).$$

This leads to the following optimization problem (MLE principle):

$$\hat{\theta} \in \operatorname*{argmax}_{\theta} \mathcal{L}(\theta, \mathcal{D}_n).$$

Numerically, it is often simpler to solve

$$\hat{\theta} \in \operatorname*{argmin}_{\theta}(-\log \mathcal{L}(\theta, \mathcal{D}_n)),$$

that is, minimizing the negative log likelihood

$$\hat{\theta} \in \operatorname*{argmin}_{\theta} \Big(- \sum_{i=1}^{n} \log f_\theta(X_i) \Big).$$

Exercises

Exercise 1.

You are given an i.i.d. sample $\mathcal{D}_n = \{X_i, i = 1, \ldots, n\}$ where $X_i \sim \mathcal{N}(\mu^\star, (\sigma^\star)^2)$. Explicit the Maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma^2}$ for this problem.

Exercise 2.

Find the expression of the MLE for the model

$$Y = X\beta + \varepsilon$$

where X is a deterministic matrix of size $n \times d$, $\beta \in \mathbb{R}^d$ are the model parameters and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Solution of Exercise 1

The Gaussian density for a generic pair $\theta = (\mu, \sigma^2)$ is given by

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big). \tag{1}$$

The negative log likelihood $\ell(\theta, \mathcal{D}_n)$ is

$$\ell(\theta, \mathcal{D}_n) = \sum_{i=1}^{n} \left(\frac{1}{2} \log(2\pi\sigma^2) + \frac{(x_i - \mu)^2}{2\sigma^2} \right). \tag{2}$$

To find the minimum of $\ell(\cdot, \mathcal{D}_n)$, we compute its partial derivatives,

$$\frac{\partial \ell}{\partial \mu}(\theta, \mathcal{D}_n) = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2}, \tag{3}$$

which leads to

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{4}$$

Solution of Exercise 1

On the other hand,

$$\frac{\partial \ell}{\partial \sigma}(\theta, \mathcal{D}_n) = \sum_{i=1}^{n} \left(\frac{1}{\sigma} - \frac{(x_i - \mu)^2}{\sigma^3} \right), \tag{5}$$

which leads to

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2. \tag{6}$$

All in all,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \text{and} \quad \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2. \tag{7}$$

Conclusion: Maximum Likelihood Estimates for the mean and the variance are the sampled mean and the sampled variance respectively.

Link with Kullback Leibler

Let us define the Kullback-Leibler divergence between two densities p and q as

$$KL(p\|q) = \int p(\mathbf{z}) \log\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right) d\mathbf{z}. \tag{8}$$

Assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. with density $f_{\theta^\star}$. The empirical distribution $\hat{f}_n$ is defined as

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{X}_i}(\mathbf{x}). \tag{9}$$

Exercise: Prove that Kullback Leibler is related to the likelihood of the model.

Solution

Let us consider the Kullback Leibler divergence between the empirical distribution and the model distribution:

$$KL(\hat{f}_n \| f_\theta) = \int \hat{f}_n(\mathbf{x}) \log \left(\frac{\hat{f}_n(\mathbf{x})}{f_\theta(\mathbf{x})} \right) d\mathbf{x} \tag{10}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\hat{f}_n(\mathbf{X}_i)}{f_\theta(\mathbf{X}_i)} \right) \tag{11}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log(\hat{f}_n(\mathbf{X}_i)) - \frac{1}{n} \sum_{i=1}^{n} \log(f_\theta(\mathbf{X}_i)). \tag{12}$$

Therefore, minimizing the Kullback-Leibler divergence between the empirical distribution and the model distribution is equivalent to

$$\underset{\theta}{\mathrm{argmax}} \sum_{i=1}^{n} \log(f_\theta(\mathbf{X}_i)), \tag{13}$$

which is exactly the log likelihood of the model.

Outline

1 General Motivation

2 No latent variables

- Parametric density estimation
- Histograms
- Kernels
- Nearest neighbors
- Fully Visible Belief Nets (FVBN)

3 Latent variables

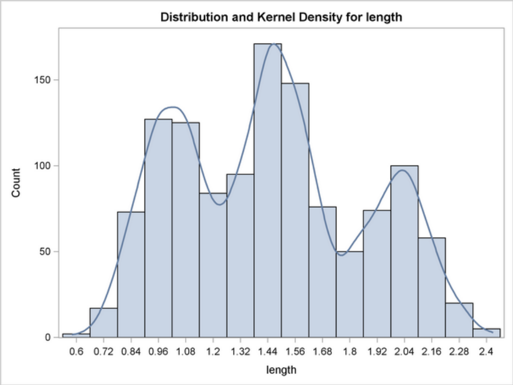
- Hidden Markov Model - EM
- Variational autoencoders
- Generative Adversarial Networks

4 Applications

Histogram

Divide the space into a set of regular intervals of the form

$$I_j = (x_0 + jh, x_0 + (j+1)h], \quad \text{for } j \in \{\dots, -1, 0, 1, \dots\}.$$



In each interval, the density is constant and is proportional to the number of observations falling into this interval.

Histograms

Histogram estimate:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_j \left(\sum_{i=1}^{n} \mathbb{1}_{X_i \in I_j} \right) \mathbb{1}_{x \in I_j}.$$

Several drawbacks:

- Strongly depends on parameters x_0 and h.
- The estimated density is intrinsically discontinuous whereas the true density can be continuous.
- Curse of dimensionality issue: the number of bins should grow exponentially with the number of dimensions. A lot of observations are required in high dimension.

Conclusion: Very useful for visualization but not suited for further analysis in high dimension.

Choice of h

Assume that $X \in [0,1]$ and that, the regular grid is given, for $j \in \{1, \dots, K\}$, by

$$I_j = (((j-1)h, jh],$$

for $K \in \mathbb{N}$ and $h = 1/K$. Consider the estimate

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_j \left(\sum_{i=1}^n \mathbb{1}_{X_i \in I_j} \right) \mathbb{1}_{x \in I_j}.$$

Theorem

Assume that f is twice differentiable and has support included in $[0,1]$. Let h be the bandwidth of the histogram estimator. Then, the Mean Integrated Square Error is given by

$$\mathbb{E}\left[\int_x (\hat{f}_{n,h}(x) - f(x))^2 dx \right] = \frac{h^2}{12} \int_0^1 (f')^2(x) dx + \frac{1}{nh} + O(1/n) + O(h^3). \qquad (14)$$

Exercise: Prove it!

Proof

- Bias-variance decomposition

$$
\begin{aligned}
MISE(\hat{f}_{n,h}) &= \mathbb{E}\left[\int_x (\hat{f}_{n,h}(x) - f(x))^2 dx \right] && (15)\\
&= \int_x \mathbb{E}\left[(\hat{f}_{n,h}(x) - f(x))^2 \right] dx && (16)\\
&= \sum_{j=1}^{K} \int_{I_j} \mathbb{E}\left[(\hat{f}_{n,h}(x) - f(x))^2 \right] dx && (17)\\
&= \sum_{j=1}^{K} \int_{I_j} \left\{ \left[\mathbb{E}(\hat{f}_{n,h}(x)) - f(x) \right]^2 + \mathbb{V}[\hat{f}_{n,h}(x)] \right\} dx. && (18)
\end{aligned}
$$

Since $nh\hat{f}_{n,h}(x)$ is a Binomial random variable of probability $p_j = \int_{I_j} f(x)dx$ for $x \in I_j$, we have

$$\mathbb{E}(\hat{f}_{n,h}(x)) = \frac{np_j}{nh}, \quad \text{and} \quad \mathbb{V}[\hat{f}_{n,h}(x)] = \frac{np_j(1-p_j)}{(nh)^2}. \qquad (19)$$

Proof

Thus,

$$MISE(\hat{f}_{n,h}) = \sum_{j=1}^{K} \int_{I_j} \left\{ \left[\frac{p_j}{h} - f(x) \right]^2 + \frac{p_j(1-p_j)}{nh} \right\} dx \tag{20}$$

$$= \int_0^1 f^2(x)dx + \frac{\sum_{j=1}^{K} p_j}{nh} - \frac{n+1}{nh} \sum_{j=1}^{K} p_j^2 \tag{21}$$

$$= \int_0^1 f^2(x)dx + \frac{1}{nh} - \frac{n+1}{nh} \sum_{j=1}^{K} p_j^2. \tag{22}$$

Asymptotically, we need to have $nh \to \infty$ and $h \to 0$. Note that,

$$\int_{I_j} f^2(x)dx - \frac{p_j^2}{h} = \int_{I_j} \left(f(x) - \frac{1}{h} \int_{I_j} f(u)du \right)^2 dx \tag{23}$$

$$= \frac{1}{h^2} \int_{I_j} \left(\int_{I_j} (f(x) - f(u))du \right)^2 dx. \tag{24}$$

Proof

Since f is twice differentiable, we have

$$f(u) - f(x) = (u-x)f'(a_j) + O(h^2), \tag{25}$$

for $x, u \in I_j$, where a_j is the left border of I_j. Therefore,

$$\begin{aligned}
\int_{I_j} f^2(x)dx - \frac{p_j^2}{h} &= \frac{f'(a_j)^2}{h^2} \int_{I_j} \Big(\int_{I_j} (x-u)du \Big)^2 dx + O(h^4) & (26)\\
&= \frac{f'(a_j)^2}{h^2} \frac{h^5}{12} + O(h^4) & (27)\\
&= \frac{h^3 f'(a_j)^2}{12} + O(h^4) & (28)\\
&= \frac{h^2}{12} \int_{I_j} f'(x)^2 dx + O(h^4). & (29)
\end{aligned}$$

Consequently,

$$\begin{aligned}
MISE(\hat{f}_{n,h}) &= \sum_{j=1}^{K} \Big(\int_{I_j} f^2(x)dx - \frac{p_j^2}{h} \Big) + \frac{1}{nh} - \frac{1}{nh} \sum_{j=1}^{K} p_j^2 & (30)\\
&= \frac{h^2}{12} \int_0^1 f'(x)^2 dx + O(h^3) + \frac{1}{nh} + O(1/n). & (31)
\end{aligned}$$

Proof

The optimum is reached at

$$h_{opt} = \left(\frac{n}{6} \int_0^1 f'(x)^2 dx \right)^{-1/3}, \tag{32}$$

which leads to a rate of convergence of $n^{-2/3}$ for $MISE(\hat{f}_{n,h_{opt}})$.

Outline

1. General Motivation

2. No latent variables
   - Parametric density estimation
   - Histograms
   - Kernels
   - Nearest neighbors
   - Fully Visible Belief Nets (FVBN)

3. Latent variables
   - Hidden Markov Model - EM
   - Variational autoencoders
   - Generative Adversarial Networks

4. Applications

Sliding-window estimate

Simply consider

$$\hat{f}_{n,h}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{X_i \in [x-h, x+h)}. \tag{33}$$

This estimate is also piecewise constant, but we got rid of the origin x_0. Another way to write is

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right), \tag{34}$$

where

$$w(x) = \begin{cases} 1/2 & \text{if } |x| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

General kernel estimate

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{35}$$

where

$$K(x) \geq 0, K(x) = K(-x) \text{ and } \int K = 1. \tag{36}$$

Examples:

- Gaussian kernel: $K(x) = (2\pi)^{-1/2}\exp(-x^2/2)$
- Rectangular kernel: $K(x) = (1/2)\mathbb{1}_{x\in[-1,1]}$
- Triangular kernel: $K(x) = (1-|u|)\mathbb{1}_{x\in[-1,1]}$
- Epanechnikov kernel: $K(x) = (3/4)(1-u^2)\mathbb{1}_{x\in[-1,1]}$

Note that the smoothness of the estimate is the same as that of the kernel. If K is p continuously differentiable then $\hat{f}_{n,h}$ is too.

Theorem

Consider a kernel which satisfy the following assumptions:

1. $\int_{\mathbb{R}} K(u)du = 1$.
2. K is a symmetric function: $K(u) = K(-u)$.
3. $\int_{\mathbb{R}} u^2 |K(u)| du < \infty$.
4. $\int_{\mathbb{R}} K(u)^2 < \infty$.

Assume that f is a twice differentiable function such that $\|f\|_\infty$ and $\|f''\|_\infty$ exist.

Exercise: Find the rate of convergence of a kernel estimate satisfying the previous conditions.

Solution

Bias.
We have

$$\mathbb{E}[\hat{f}_{n,h}(x)] = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right] \tag{37}$$

$$= \frac{1}{nh}\sum_{i=1}^{n}\int_{\mathbb{R}} K\left(\frac{y-x}{h}\right)f(y)dy \tag{38}$$

$$= \int_{\mathbb{R}} K(u)f(x+uh)du. \tag{39}$$

With a Taylor expansion of order 2, we get

$$\mathbb{E}[\hat{f}_{n,h}(x)] = \int_{\mathbb{R}} K(u)\left[f(x) + uhf'(x) + \frac{(uh)^2}{2}f''(\eta_u)\right]du \tag{40}$$

$$= f(x)\int_{\mathbb{R}} K(u)dx + hf'(x)\int_{\mathbb{R}} uK(u)dx + \frac{h^2}{2}\int_{\mathbb{R}} u^2K(u)f''(\eta_u)du. \tag{41}$$

Solution

Therefore,

$$\begin{aligned}
|\mathbb{E}[\hat{f}_{n,h}(x)] - f(x)| &\leq \frac{h^2}{2}\Big| \int_{\mathbb{R}} u^2 K(u) f''(\eta_u) du \Big| && (42)\\
&\leq \frac{h^2}{2} \int_{\mathbb{R}} u^2 |K(u)||f''(\eta_u)| du && (43)\\
&\leq \frac{h^2 \|f''\|_\infty}{2} \int_{\mathbb{R}} u^2 |K(u)| du. && (44)
\end{aligned}$$

Solution

Variance.
We have

$$\begin{aligned}
\mathbb{V}[\hat{f}_{n,h}(x)] &= \frac{1}{(nh)^2}\mathbb{V}\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right] && (45)\\
&= \frac{1}{(nh)^2}\sum_{i=1}^{n}\mathbb{V}\left[K\left(\frac{X_i - x}{h}\right)\right] && (46)\\
&\leq \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)^2\right] && (47)\\
&\leq \frac{1}{nh^2}\int_{\mathbb{R}} K\left(\frac{y - x}{h}\right)^2 f(y)dy && (48)\\
&\leq \frac{1}{nh}\int_{\mathbb{R}} K(u)^2 f(uh + x)du && (49)\\
&\leq \frac{1}{nh}\|f\|_\infty \int_{\mathbb{R}} K^2(u)du. && (50)
\end{aligned}$$

Solution

Finally,

$$
\begin{aligned}
MISE(\hat{f}_{n,h}) &\leq \left(\frac{h^2\|f''\|_\infty}{2}\int_{\mathbb{R}} u^2|K(u)|\right)^2 + \frac{1}{nh}\|f\|_\infty \int_{\mathbb{R}} K^2(u)du. && (51)\\
&\leq C_1^2 h^4 + \frac{C_2}{nh}, && (52)
\end{aligned}
$$

where the optimum is reached at

$$h_{opt} = (C_2/4C_1^2)^{1/5} n^{-1/5}, \tag{53}$$

which leads to

$$MISE(\hat{f}_{n,h_{opt}}(x)) \leq Cn^{-4/5}. \tag{54}$$

The rate of convergence of kernel estimate is faster than that of histograms for twice differentiable density functions.

Outline

1. General Motivation

2. No latent variables
   - Parametric density estimation
   - Histograms
   - Kernels
   - Nearest neighbors
   - Fully Visible Belief Nets (FVBN)

3. Latent variables
   - Hidden Markov Model - EM
   - Variational autoencoders
   - Generative Adversarial Networks

4. Applications

Nearest neighbors density estimate

The k nearest neighbors estimate for density estimation is given by

$$\hat{f}_{n,k}(x) = \frac{k}{nV_k(x)}, \tag{55}$$

where $V_k(x) = Vol(B(x, d_k(x)))$, where $d_k(x)$ is the distance between x and its k nearest neighbor. The volume of the unit ball in $\mathbb{R}^d$ is given by

$$Vol(B(0,1)) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \tag{56}$$

The resulting estimate

- is not a density $\int f \neq 1$
- has a lot of discontinuities
- Even for large regions with no observations, the estimated density is not zero.

Exercise: Explain the intuition behind this estimate and try to find conditions on (k, n) so that the estimate is consistent.

Intuition for Nearest neighbor estimate

For any fixed k,

$$\frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in B(x, d_k(x))} \simeq \mathbb{P}[X \in B(x, d_k(x))], \tag{57}$$

where

$$\mathbb{P}[X \in B(x, d_k(x))] = \int_{B(x,d_k(x))} f(u) du. \tag{58}$$

For a fixed $x \in \mathbb{R}^d$, $d_k(x)$ is small for k/n small enough. In that case, $f(u) \simeq f(x)$ for $u \in B(x, d_k(x))$ that is

$$\int_{B(x,d_k(x))} f(u) du \simeq f(x) Vol(B(x, d_k(x))). \tag{59}$$

Finally, for $k \to \infty$ and $k/n \to 0$, we have

$$\frac{k}{n} \frac{1}{Vol(B(x, d_k(x)))} \to f(x). \tag{60}$$

Outline

1. General Motivation

2. No latent variables
   - Parametric density estimation
   - Histograms
   - Kernels
   - Nearest neighbors
   - Fully Visible Belief Nets (FVBN)

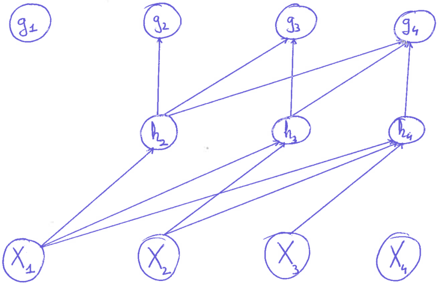
3. Latent variables
   - Hidden Markov Model - EM
   - Variational autoencoders
   - Generative Adversarial Networks

4. Applications

Core idea

Difficult to model a multivariate distribution in high dimensions. The joint distribution P of the vector $\mathbf{X} = (X_1, \ldots, X_d)$ is written as

$$P(\mathbf{X}) = \prod_{j=1}^{d} P(X_j | X_1, \ldots, X_{j-1}).$$



The network is composed of three layers: input layer, hidden layer, output layer. Note that there are no latent variables hence the name "visible".

[“Modeling high-dimensional discrete data with multi-layer neural networks”, Y. Bengio and S. Bengio 2000]

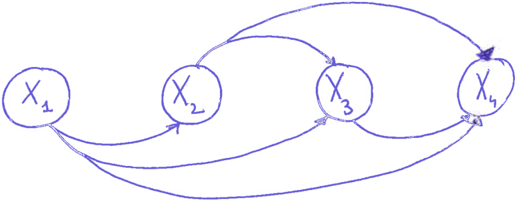
Logistic Autoregressive Bayesian Network (LARC)

Original idea from

[*Graphical models for machine learning and digital communication*, Frey and Frey 1998]

In this paper, the network was only designed for binary discrete random variables: there was no hidden layer and the output function was a logistic. All in all, the corresponding model can be written as

$$\Pr(X_j = 1|X_1, \ldots, X_{j-1}) = \frac{1}{1 + \exp(-w_0 - \sum_{\ell<j} w_j X_j)}$$



Two more ingredients

Groups in input and hidden layer

The input layer is organized in $d-1$ groups corresponding to groups 1 to $d-1$: X_d is not used as an input.
The hidden layer is organized in $d-1$ groups corresponding to groups 2 to d: since the distribution of X_1 is represented unconditionally, it does not require hidden units.

The network is not fully connected

A preprocessing step is done to test the link between X_j and X_ℓ. For each pair (X_j, X_ℓ), we compute the Kolmogorov-Smirnov statistic

$$s_{j,\ell} = \sqrt{n} \sup_{x_j, x_\ell} |\hat{P}_n(X_j \leq x_j, X_\ell \leq x_\ell) - \hat{P}_n(X_j \leq x_j)\hat{P}_n(X_\ell \leq x_\ell)|.$$

The pairs are ranked according their $s_{j,\ell}$ values and only the pairs whose statistic is above a specified threshold (computed via cross-validation) are kept. For these pairs only, the corresponding groups are connected in the network, both between input and hidden layer and between hidden and output layer.

["Distribution free tests of independence based on the sample distribution function", Blum et al. 1961]

Outline

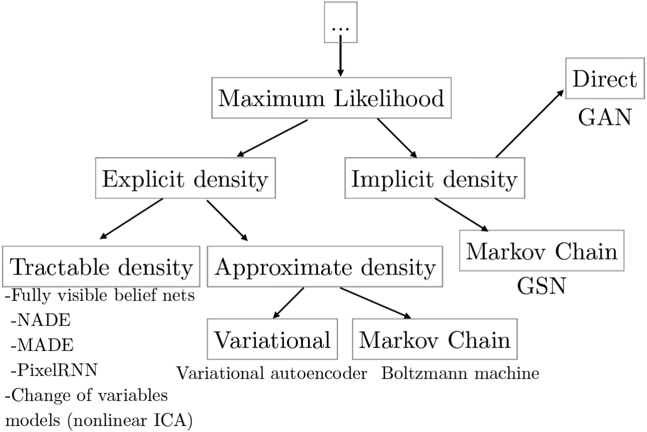
1. General Motivation

2. No latent variables
    - Parametric density estimation
    - Histograms
    - Kernels
    - Nearest neighbors
    - Fully Visible Belief Nets (FVBN)

3. Latent variables
    - Hidden Markov Model - EM
    - Variational autoencoders
    - Generative Adversarial Networks

4. Applications

Summary: Taxonomy


...
Maximum Likelihood
Explicit density
Implicit density
Direct
GAN
Tractable density
-Fully visible belief nets
-NADE
-MADE
-PixelRNN
-Change of variables
models (nonlinear ICA)
Approximate density
Variational
Variational autoencoder
Markov Chain
Boltzmann machine
Markov Chain
GSN