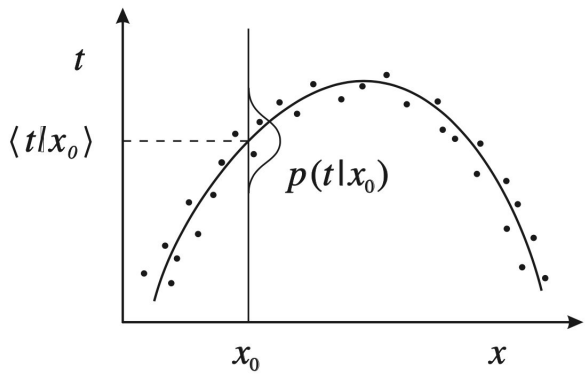
Why learn conditional distribution?

A wide range of problems in machine learning concerns learning the relationship between a variable x and another variable y. In prediction task, the target is to build a deterministic mapping such that $y = f(x)$.



However, such deterministic relationship may not exist. Instead, given $X = x$, it is better to model Y as a random variable with a conditional distribution $\mathcal{P}_{Y|_{X=x}}$.

To learn the conditional distribution is also of great interset to

1. Finance
2. Operation research
3. XXX
4. XXX

Why regularization?

Prediction can overfitting, so does learning the conditional distribution.

Two figures: the overfitting of prediction task, the overfitting of learning the conditional distribution.

In some applications, the dimension of the condition may be quite large, and the number of available samples are limited. This results in the sparsity of the samples' location, which makes overfitting more inclined to happen.

Our approach:

Notation:

The underlying distribution:

$$X \in \mathbb{R}^{n_x}, Y \in \mathbb{R}^{n_y}, (X, Y) \sim \mathcal{P}_{(X,Y)}, X \sim \mathcal{P}_X, Y \sim \mathcal{P}_Y, Y|_{X=x} \sim \mathcal{P}_{Y|_{X=x}} \tag{1}$$

Dataset:

N iid samples of $\{x_i, y_i\}_{i=1}^N$.

We use MDN (Mixture density network) to model the conditional distribution.

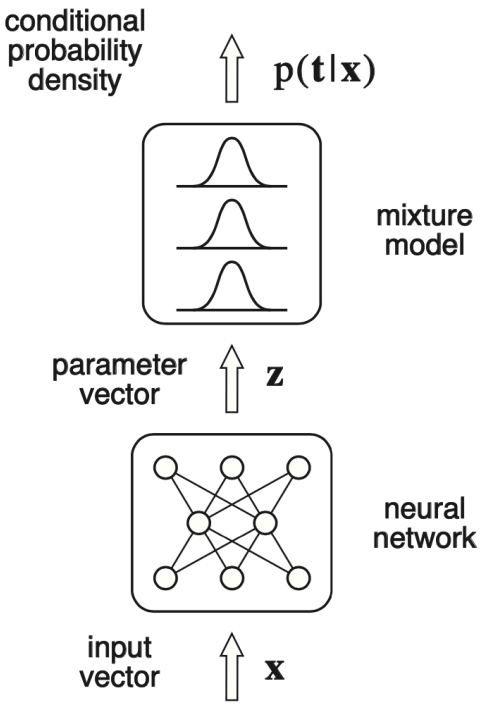


Figure 2: The Mixture Density Network consists of a feed-forward neural network whose outputs determine the parameters in a mixture density model. The mixture model then represents the conditional probability density function of the target variables, conditioned on the input vector to the neural network.

This network models the conditional distribution $\mathcal{P}_{Y|_{X=x}}$ using GMM. (Why GMM? universal approximator)

Input of the network: x

output of the network: parameters of the GMM: weight list $\alpha_i(\mathbf{x})$, mean $\boldsymbol{\mu}_i(\mathbf{x})$, variance $\sigma_i(\mathbf{x})^2$ for $i = 1, 2, \ldots, m$. (m is a hyper-parameter).

The (probability density function) pdf of Gaussian mixture model is

$$p(\mathbf{y} \mid \mathbf{x}) = \sum_{i=1}^{m} \alpha_i(\mathbf{x}) \phi_i(\mathbf{y} \mid \mathbf{x}) \tag{2}$$

with

$$\phi_i(\mathbf{y} \mid \mathbf{x}) = \frac{1}{(2\pi)^{c/2} \sigma_i(\mathbf{x})^c} \exp\left\{ -\frac{\|\mathbf{y} - \boldsymbol{\mu}_i(\mathbf{x})\|^2}{2\sigma_i(\mathbf{x})^2} \right\} \tag{3}$$

We train this network by maximizing the loglikelihood:

$$\text{Log Likelihood} = \sum_{i=1}^{N} \ln\left\{ \sum_{i=1}^{m} \alpha_i\left(\mathbf{x}_i\right) \cdot \phi_i\left(\mathbf{y}_i \mid \mathbf{x}_i\right) \right\} \tag{4}$$

The regularization term:

we hope the conditional distribution at x_1 is close to x_2, when x_i is close to x_2

W-dist: compute distance between high dimensional random variables, well studied and , for two distribution μ_1 and μ_2, $W(\mu_1, \mu_2)$

regularization term

$$\text{Loss} = \sum_{i=1}^{N} \ln\left\{\sum_{i=1}^{m} \alpha_i\left(\mathbf{x}_i\right) \cdot \phi_i\left(\mathbf{y}_i \mid \mathbf{x}_i\right)\right\} + \lambda \cdot \mathbb{E}_{x_0, x_1 \sim \mathcal{P}_X}\left[\max\{W\left(\mu_0, \mu_1\right) - K||x_0 - x_1||, 0\}\right] \quad (5)$$

MDN input: condition x_0, change to $x_1 = x_0 + \Delta x$

Output: parameters of Normal distribution, $m_0 = f_1(x_0)$, $\Sigma_0 = f_2(x_0)$, $m_1 = f_1(x_1)$ and $\Sigma_1 = f_2(x_1)$.

$$W_2(\mu_0, \mu_1)^2 = \|m_0 - m_1\|^2 + \operatorname{trace}\left(\Sigma_0 + \Sigma_1 - 2\Sigma_0^{1/2}\left(\left(\Sigma_0^{1/2}\right)^{\dagger}\Sigma_1\left(\Sigma_0^{1/2}\right)^{\dagger}\right)^{1/2}\Sigma_{0^{\circ}}^{1/2}\right) \quad (6)$$

$W_2(\mu_0, \mu_1)^2 \leq K||x_0 - x_1||$, where $\mu_0 \sim N(m_0, \Sigma_0)$, $\mu_1 \sim N(m_1, \Sigma_1)$, and $m_0 = f_1(x_0)$, $\Sigma_0 = f_2(x_0)$, $m_1 = f_1(x_1)$ and $\Sigma_1 = f_2(x_1)$.

1. add $C \cdot \max\{W_2(\mu_0, \mu_1)^2 - K||x_0 - x_1||, 0\}$ to the training loss, how to sample x_0 or x_1?
2. Define $D_{x_0}(\Delta x) = W_2(\mu_{x_0}, \mu_{x_0+\Delta x})^2$, $\mu_{x_0+\Delta x} \sim N(m_{x_0+\Delta x}, \Sigma_{x_0+\Delta x})$, add $C \cdot ||\frac{\partial D_{x_0}(x)}{\partial x}|_{x=0}||$ to the training loss

---

Output: parameters of GMM, weight list $\alpha_i(\mathbf{x})$, mean $\boldsymbol{\mu}_i(\mathbf{x})$, variance $\sigma_i(\mathbf{x})^2$.

$$W_2(\mu_0, \mu_1)^2 = \|m_0 - m_1\|^2 + \operatorname{trace}\left(\Sigma_0 + \Sigma_1 - 2\Sigma_0^{1/2}\left(\left(\Sigma_0^{1/2}\right)^{\dagger}\Sigma_1\left(\Sigma_0^{1/2}\right)^{\dagger}\right)^{1/2}\Sigma_{0^{\circ}}^{1/2}\right) \quad (7)$$

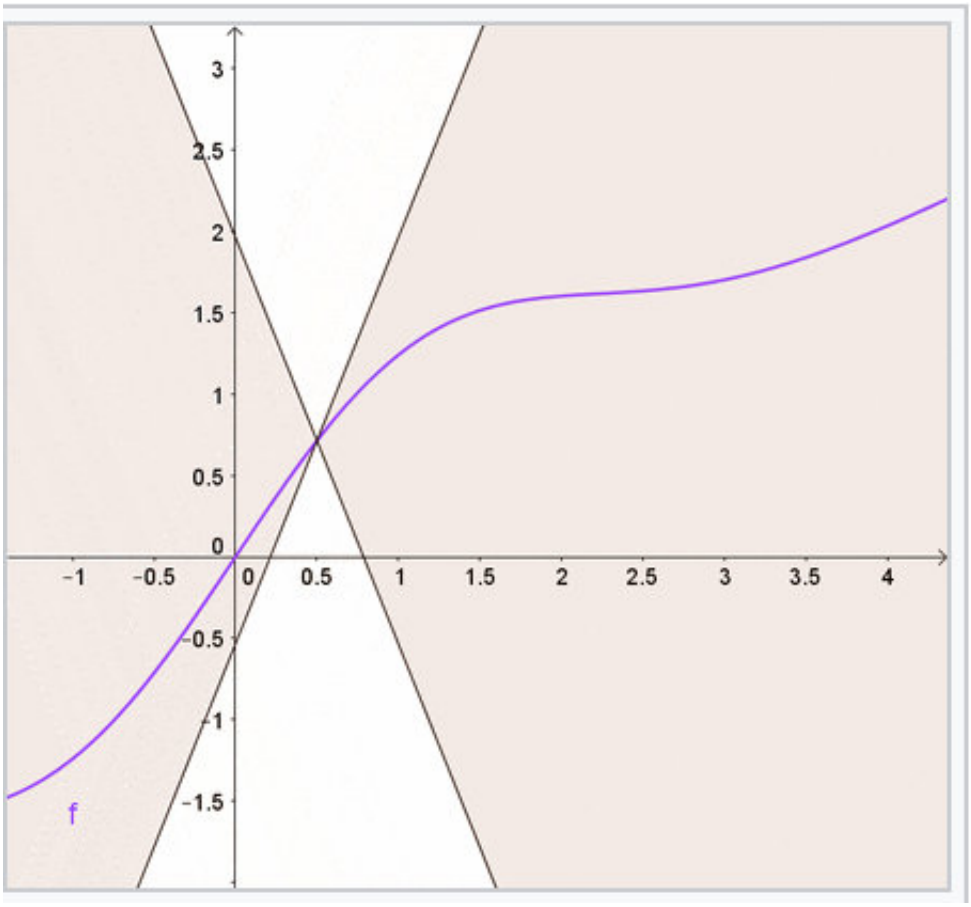
$W_2(\mu_0, \mu_1)^2 \leq K||x_0 - x_1||$, where $\mu_0 \sim N(m_0, \Sigma_0)$, $\mu_1 \sim N(m_1, \Sigma_1)$, and $m_0 = f_1(x_0)$, $\Sigma_0 = f_2(x_0)$, $m_1 = f_1(x_1)$ and $\Sigma_1 = f_2(x_1)$.

1. add $C \cdot \max\{W_2(\mu_0, \mu_1)^2 - K||x_0 - x_1||, 0\}$ to the training loss, how to sample x_0 or x_1?
2. Define $D_{x_0}(\Delta x) = W_2(\mu_{x_0}, \mu_{x_0+\Delta x})^2$, μ_x denotes the GMM given by the network at condition $X = x$. add $C \cdot ||\frac{\partial D_{x_0}(x)}{\partial x}|_{x=0}||$ to the training loss

weight list $\alpha_i(\mathbf{x}_0)$, mean $\boldsymbol{\mu}_i(\mathbf{x}_0)$, variance $\sigma_i(\mathbf{x}_0)^2$.

weight list $\alpha_i(\mathbf{x}_0 + \Delta x)$, mean $\boldsymbol{\mu}_i(\mathbf{x}_0 + \Delta x)$, variance $\sigma_i(\mathbf{x}_0 + \Delta x)^2$.

当 Δx 非常接近于0,
$D_{x_0}(\Delta x) = \sum_{i=1}^{m} \alpha_i(x_0) \cdot W_2(N(\mu_i(x_0), \sigma_i(x_0)), N(\mu_i(x_0 + \Delta x), \sigma_i(x_0 + \Delta x)))^2$



$|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$

take x_1 = 0.5

$|f(0.5) - f(x_2)| \leq K|0.5 - x_2|$

$$< Empty\ Math\ Block > \tag{8}$$

$$X = x, D(Y_{real}|_{X=x}, Y_{network}|_{X=x}), E_{X \sim \mathcal{P}_X}[D(Y_{real}|_X, Y_{network}|_X)] \tag{9}$$